

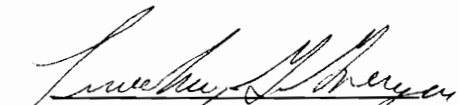
THE ANALYSIS OF LONGITUDINAL ORDINAL DATA

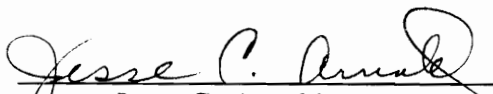
by

Oliver Schabenberger

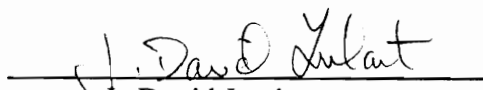
Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Forestry

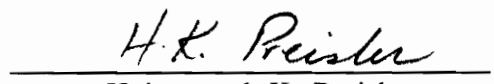
APPROVED:

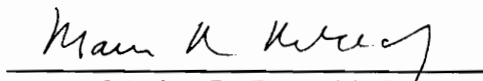

Timothy G. Gregoire, Chairman


Jesse C. Arnold


Eric P. Smith


J. David Lenhart


Haganoush K. Preisler


Marion R. Reynolds Jr.

April, 1995
Blacksburg, Virginia

THE ANALYSIS OF LONGITUDINAL ORDINAL DATA

by

Oliver Schabenberger
Dr. Timothy G. Gregoire, Chairman
Department of Forestry

(ABSTRACT)

Longitudinal data, in which a series of observations is collected on a typically large number of experimental units is one of the most frequent and important sources of quantitative information in forestry. The dependencies among repeated observations for an experimental unit must be accounted for in order to validate statistical estimation and inference in modeling efforts. The recent advances in statistical theory for correlated data created a body of theory which will become of increasing importance as analysts realize the limitations of traditional methods that ignore these dependencies. Longitudinal data fosters research questions that focus on the individual experimental unit rather than the population as in classical cross-sectional data analysis. Mixed model techniques have emerged as powerful tools to address research problems of this kind and are treated extensively in this dissertation.

Over the last years interest in modeling quantal responses that take on only a countable, discrete number of possible values has also increased throughout the discipline. The theory of generalized linear models provides the groundwork to embody quantal response models into the toolbox of applied analysts.

The focus of this dissertation is to combine modern analytical tools for longitudinal data with regression methods for quantal responses. Special emphasis is placed on ordinal and binary data because of their prevalence in ecological, biological, and environmental statistics. The first chapters review the literature and introduce necessary theory. The second part of this dissertation consists of a case study in which binary and ordinal fusiform rust response on loblolly and slash pine is modeled in a longitudinal data base provided by the East Texas Pine Plantation Research Project.

Acknowledgments

This dissertation would not have been possible without the help and dedication of some very special people, whom I want to thank dearly.

In the last three years Dr. Timothy G. Gregoire has not only guided me as major professor through a demanding graduate program and this study but set an example and standard for professional excellence. I am indebted to him for allowing me to draw from his experience and knowledge.

I want to express my lasting gratitude to Dr. J. David Lenhart at Stephen F. Austin University and the sponsors of the East Texas Pine Plantation Research Project for allowing me to work with their data base and helpful criticism.

The Department of Statistics at Virginia Polytechnic Institute and State University provided me with the necessary knowledge and tools to successfully tackle the theoretical problems expounded in this dissertation. I wish to thank the Department of Statistics in its entirety. As members of the doctoral committee Drs. Jesse C. Arnold, Marion R. Reynolds Jr., and Eric P. Smith have provided invaluable comments and criticisms that improved this dissertation substantially.

I owe so very much to the love and support my parents Kurt and Rosemarie Schabenberger provided through the years. Without their encouragement and financial support I would not have come to the U.S and left the confines of my native environment. It is to them, that I dedicate this dissertation with the greatest pleasure.

Table of Contents

- Chapter 1. Introduction..... 1**
- Chapter 2. Literature Review..... 9**
 - 2.1. Ordinal Response Methods..... 9
 - 2.2. Longitudinal Data..... 16
 - 2.2.1 Longitudinal Data in General..... 16
 - 2.2.1.1. Approaches for continuous response..... 17
 - 2.2.1.2. Categorical response..... 21
 - 2.2.1.2.1. Population-Averaged models..... 22
 - 2.2.1.2.2. Subject-Specific models..... 26
 - 2.2.1.2.3. Conditional models..... 33
- Chapter 3. Background..... 39**
 - 3.1. Cumulative Link Models..... 39
 - 3.1.1. Rationale..... 40
 - 3.1.2. The Parallel Lines Assumption..... 42
 - 3.1.3. Model Parsimony and Response Nature..... 46
 - 3.1.4. Choice of the Link Function..... 50
 - 3.1.5. Fitting the POM..... 54
 - 3.1.6. A Convenient Reparameterization..... 59
 - 3.1.7. Predictions..... 60
 - 3.1.8. Judging Model Performance with Classifi-
cation Tables..... 64
 - 3.2. Estimating Function Theory..... 70

3.2.1. Competing Principles.....	70
3.2.2. Optimal Estimating Functions in the Sense of Godambe.....	73
3.2.3. EF and Categorical Data.....	78
3.2.4. A General Estimating Function.....	80
3.3. Issues in Longitudinal Studies.....	91
3.3.1. Longitudinal Data Structures.....	91
3.3.2. Drop-Outs and Censoring.....	97
3.3.3. Research Questions in LD Analysis.....	100
3.3.3.1. Age and cohort effects.....	100
3.3.3.2. PA, SS, and conditional research problems.....	104
3.3.4. How To Model Correlations.....	111
Chapter 4. Statistical Inference for Longi- tudinal Ordinal Data.....	115
4.1. Notation.....	115
4.2. Population-Averaged Models.....	118
4.2.1. Generalized Estimating Equations (GEE).....	118
4.2.1.1. GEE for longitudinal binary data.....	120
4.2.1.2. GEE for longitudinal ordered response.....	129
4.2.1.2.1. GEE0.....	130
4.2.1.2.2. GEE1 and GEE2.....	132
4.3. Subject-Specific, Mixed Models.....	139
4.3.1. Semi-Parametric Methods.....	143
4.3.2. Approximate Parametric Methods.....	149
4.3.3. Fully Parametric Methods.....	153

4.4. Conditional Models.....	157
4.4.1. RC Models for Conditional Events.....	160
4.4.2. RC Models for Joint Occurrences.....	165
4.4.3. CC Models for Ordered Responses.....	169
4.4.3.1. Bonney's CC model for ordered response.....	170
4.4.3.2. Cut-off CC models.....	173
4.4.4. Markov Chains.....	175
4.5. Grouped And Ungrouped Cases.....	178
4.6. Computer Implementation.....	180
4.7. Summary.....	182
Chapter 5. Fusiform Rust.....	185
5.1. Biology.....	185
5.2. Approaches to Estimation and Prediction.....	190
Chapter 6. Material.....	199
6.1. The East Texas Pine Plantation Research Project.....	199
6.1.1. Introduction.....	199
6.1.2. Observed Condition Frequencies And Proportions.....	207
6.1.2.1. Box and whisker plots for time dependent covariates	207
6.1.2.2. Marginal proportions by initial age classes.....	213
6.1.2.3. Transition patterns.....	220
6.1.2.3.1. Slash pine.....	220
6.1.2.3.2. Loblolly pine.....	226

6.1.2.4. One-step transitions by initial age and dbh-classes···	226
Chapter 7. Analysis ·····	233
7.1. Slash Pine·····	233
7.1.1. Marginal Plot Models·····	233
7.1.1.1. Measuring Goodness Of Fit·····	235
7.1.1.2. Biological patterns·····	243
7.1.1.3. Ordinal response models·····	249
7.1.1.4. Binary response models·····	264
7.1.2. Transitional Plot Models·····	274
7.1.2.1. RC models for conditional events·····	275
7.1.2.2. CC model for conditional events·····	291
7.2. Loblolly Pine·····	298
7.2.1. Marginal Plot Models·····	298
7.2.1.1. Biological patterns·····	298
7.2.1.2. Ordinal response models·····	303
7.2.1.3. Binary response models·····	318
7.2.2. Transitional Plot Models·····	324
7.2.2.1. RC models for conditional events·····	324
7.2.2.2. CC model for conditional events·····	334
Chapter 8. Discussion and Summary ·····	341
Chapter 9. Bibliography ·····	352
9.1. References by Subject·····	368
Chapter 10. Appendix ·····	370
Vita ·····	383

List of Figures

Figure 3.1. Depiction of cumulative logit model for an ordinal response with four categories. Adopted from Agresti (1990).	43
Figure 3.2. Relating the latent continuous and ordinal scale through a linear regression model and a series of ordered cut-offs.	44
Figure 3.3. Shape of various link functions for categorical response data. See also Table 3.1 and Figure 3.4.	52
Figure 3.4. Cumulative distribution functions corresponding to the link functions in Table 3.1. and Figure 3.3.	53
Figure 3.5. Basal areas on selected plots by age. a): Cross-sectional; b): Longitudinal data structure. Plot numbers printed in panel b).	101
Figure 4.1. Flow charts of estimation steps in a) GEE0, and b) GEE1.	138
Figure 4.2. Flow charts of estimation steps in semi-parametric (marginal quasi-likelihood) (a) and approximate parametric (pseudo-likelihood) approach (b).	147
Figure 4.3. Possible response profiles for a uni-directional process with 4 categories. Initial state in leftmost column. Three transitions are plotted in the following columns to the right.	159
Figure 5.1. Development cycle of fusiform rust.	186
Figure 6.1. Histograms of time-constant plot level covariates for slash pine development data.	205

Figure 6.2. Histograms of time-constant plot level covariates for loblolly pine development data.	206
Figure 6.3. Box and whisker plots for plot level covariates against fusiform rust condition code for slash pine development data.	208
Figure 6.4. Box and whisker plots for plot level covariates against fusiform rust condition code for loblolly pine development data.	209
Figure 6.5. Box and whisker plots for tree level covariates against fusiform rust condition code for slash pine development data.	211
Figure 6.6. Box and whisker plots for plot level covariates against fusiform rust condition code for loblolly pine development data.	212
Figure 6.7. Observed marginal cumulative proportions in development and evaluation data sets by initial age.	217
Figure 6.8. Observed marginal cumulative proportions in development and evaluation data sets by dbh classes.	219
Figure 6.9. Transition patterns and corresponding sample frequencies for slash pine development data. Frequencies (in parentheses) are set to missing if zero. Columns sum to unity.	221
Figure 6.10. Transition patterns and corresponding sample frequencies for slash pine, development data. Frequencies are rendered conditional on previous response.	222
Figure 6.11. Transition patterns and corresponding sample frequencies for loblolly pine development data. Frequencies (in parentheses) are set to missing if zero. Columns sum to unity.	224
Figure 6.12. Transition patterns and corresponding sample frequencies for loblolly pine development data. Frequencies are rendered conditional on previous response.	225
Figure 6.13. Observed cumulative transition proportions in slash pine data sets by initial age classes. C: 0 = healthy, 1 = branch, 2 = stem infection.	228

Figure 6.14.Observed cumulative transition proportions in loblolly pine data sets by initial age classes. C: 0 = healthy, 1 = branch, 2 = stem infection.	229
Figure 6.15.Observed cumulative transition proportions in slash pine data sets by (2 in) DBH classes. C: 0 = healthy, 1 = branch, 2 = stem infection.	231
Figure 6.16.Observed cumulative transition proportions in loblolly pine data sets by (2 in) DBH classes. C: 0 = healthy, 1 = branch, 2 = stem infection.	232
Figure 7.1. Schematic display of ray arrangement as used in the circular prediction plots for marginal ordinal response models.	241
Figure 7.2. Profile plots for single covariate proportional odds models. Species and data: Slash pine development data.	244
Figure 7.3. Profile plots for single covariate proportional odds models. Species and data: Slash pine development data.	245
Figure 7.4.Observed probability surfaces for Slash pine. Surface interpolated by 5th order polynomials. Covariates: Site index (ft), Quadratic mean diameter (in).	246
Figure 7.5.Predicted probability surfaces for two covariate proportional odds model. Slash pine development data. Covariates: Site index (ft), Quadratic mean diameter (in).	247
Figure 7.6. Observed and predicted proportions in Slash pine model [7.14]. Each segment corresponds to one occasion and a 0-100% interval. Observed values denoted by clockwise arrows, predicted values counter-clockwise. GEE0 with 1-dependence structure.	259
Figure 7.7. (Same as Figure 7.6)	260
Figure 7.8. Observed and predicted proportions in Slash pine model [7.14]. Each segment corresponds to one occasion and a 0-100% interval. Observed values denoted by clockwise arrows, predicted values counter-clockwise. Mixed model fit semi-parametrically.	261
Figure 7.9. (Same as Figure 7.8)	262

Figure 7.10. Receiver-Operating-Characteristic (ROC) for healthy trees. Mixed model: Site index and first occasion random. Sensitivity: Proportion of events, predicted as events. False(+): Proportion of predicted events, that are incorrect. False(-): Prop. of predicted no-events, that are incorrect.	271
Figure 7.11. Receiver-Operating-Characteristic (ROC) for trees free of stem infect. Mixed model: Trees per acre and first occasion random. Sensitivity: Proportion of events, predicted as events. False(+): Proportion of predicted events, that are incorrect. False(-): Prop. of predicted no-events, that are incorrect.	272
Figure 7.12. Receiver-Operating-Characteristic (ROC) for alive trees. Mixed model: Site index random. Sensitivity: Proportion of events, predicted as events. False(+): Proportion of predicted events, that are incorrect. False(-): Prop. of predicted no-events, that are incorrect.	273
Figure 7.13. Observed and forecasted proportions based on RC model for conditional events. Circles delimit initial occasion and remeasure- ments. Initial marginal proportions plotted in innermost segment. Plots with only one measurement are excluded from the analysis.	279
Figure 7.14. (Same as Figure 7.13)	280
Figure 7.15. Observed and forecasted proportions based on separate propor- tional odds models. Circles delimit initial occasion and remeasure- ments. Initial marginal proportions plotted in innermost segment. Plots with only one measurement are excluded from the analysis.	289
Figure 7.16. (Same as Figure 7.15)	290
Figure 7.17. Marginal and transition category probabilities as generated by model [7.20]. Horizontal axis denotes the linear predictor without cut-off points. Prev. 0, 1, 2 denotes previously healthy, branch, and stem infected trees, respectively.	294

Figure 7.18.Observed and forecasted proportions based on covariate conditioned odds model. Circles delimit initial occasion and remeasurements. Initial marginal proportions plotted in innermost segment. Plots with only one measurement are excluded from the analysis.	296
Figure 7.19. (Same as Figure 7.18)	297
Figure 7.20.Profile plots for single covariate proportional odds models. Species and data: Loblolly pine development data.	299
Figure 7.21.Profile plots for single covariate proportional odds models. Species and data: Loblolly pine development data.	300
Figure 7.22.Observed probability surfaces for Loblolly pine. Surface interpolated by 5th order polynomials. Covariates: Quadratic mean diameter (in), Stand height (ft).	300
Figure 7.23.Predicted probability surfaces for two covariate proportional odds model. Loblolly pine development data. Covariates: Quadratic mean diameter (in), Stand height (ft).	302
Figure 7.24.Observed and predicted proportions in Loblolly pine model [7.21]. Each segment corresponds to one occasion and a 0-100% interval. Observed values denoted by clockwise arrows, predicted values counter-clockwise. GEE0 with 1-dependence structure.	310
Figure 7.25. (Same as Figure 7.24)	311
Figure 7.26. (Same as Figure 7.24)	312
Figure 7.27. (Same as Figure 7.24)	313
Figure 7.28.Observed and predicted proportions in Loblolly pine model [7.21]. Each segment corresponds to one occasion and a 0-100% interval. Observed values denoted by clockwise arrows, predicted values counter-clockwise. Mixed model fit semi-parametrically.	314
Figure 7.29. (Same as Figure 7.28)	315
Figure 7.30. (Same as Figure 7.28)	316
Figure 7.31. (Same as Figure 7.28)	317

Figure 7.32. Receiver-Operating-Characteristic (ROC) for healthy trees. Mixed model: Site index and first occasion random. Sensitivity: Proportion of events, predicted as events. False(+): Proportion of predicted events, that are incorrect. False(-): Prop. of predicted no-events, that are incorrect.	321
Figure 7.33. Receiver-Operating-Characteristic (ROC) for trees free of stem infect. Mixed model: Trees per acre and first occasion random. Sensitivity: Proportion of events, predicted as events. False(+): Proportion of predicted events, that are incorrect. False(-): Prop. of predicted no-events, that are incorrect.	322
Figure 7.34. Receiver-Operating-Characteristic (ROC) for alive trees. Mixed model: Site index random. Sensitivity: Proportion of events, predicted as events. False(+): Proportion of predicted events, that are incorrect. False(-): Prop. of predicted no-events, that are incorrect.	323
Figure 7.35. Observed and forecasted proportions based on separate propor- tional odds models. Circles delimit initial occasion and remeasure- ments. Initial marginal proportions plotted in innermost segment. Plots with only one measurement are excluded from the analysis.	330
Figure 7.36. (Same as Figure 7.35)	331
Figure 7.37. (Same as Figure 7.35)	332
Figure 7.38. (Same as Figure 7.35)	333
Figure 7.39. Marginal and transition category probabilities as generated by model [7.26]. Horizontal axis denotes the linear predictor without cut-off points. Prev. 0, 1, 2 denotes previously healthy, branch, and stem infected trees, respectively.	335
Figure 7.40. Observed and forecasted proportions based on covariate condi- tioned odds model. Circles delimit initial occasion and remeasure- ments. Initial marginal proportions plotted in innermost segment. Plots with only one measurement are excluded from the analysis.	337
Figure 7.41. (Same as Figure 7.40)	338
Figure 7.42. (Same as Figure 7.40)	339
Figure 7.43. (Same as Figure 7.40)	340

List of Tables

Table 3.1. Frequently employed link functions and corresponding cumulative distribution functions.	51
Table 3.2. Depiction of cross-classification table for a four category response model.	64
Table 3.3. Elementary longitudinal design, balanced and regularly spaced, no missing data.	92
Table 3.4. Standard LD design (Laird 1988), balanced but irregularly spaced.	93
Table 3.5. Rotating panel design, regularly spaced.	96
Table 3.6. Structures for the within-subject error matrix for 4 repeated measurements.	112
Table 4.1. Overview of Estimation Approaches for Longitudinal Ordinal Data as reviewed in Chapter 4.	183
Table 4.1. (Continued).	184
Table 6.1. Measurement patterns in three wave rotating panel, starting 1982.	201
Table 6.2. Plot and tree level covariates measured at each occasion on all plots. Type = m denotes a measured characteristic, type = d a derived variable. $C(x)$ identifies a categorical variable with x classes.	202

Table 6.3. Number of plots by initial age groups in slash pine development and evaluation data sets along with observed proportions of trees in the four rust categories.	214
Table 6.4. Number of plots by initial age groups in loblolly pine development and evaluation data sets along with observed proportions of trees in the four rust categories.	215
Table 7.1. Sums of squares and related criteria and their utility for multivariate categorical regression models.	242
Table 7.2. Results of fitting model [7.14] to slash pine data. Asymptotic standard errors in parentheses.	252
Table 7.3. Goodness-of-fit measures for the estimation methods in Table 7.2 calculated from evaluation data.	253
Table 7.4. Asymptotic 95 % confidence intervals for predicted proportions in marginal slash pine model [7.14]. Results for mixed model based on fixed effects only, ignoring random effects variability.	256
Table 7.5. Prediction of marginal proportions for model [7.14].	257
Table 7.6. Cross-classification table based on direct probability accumulation in 1-dependent GEE0 fit.	258
Table 7.7. Classification table for a binary response model.	265
Table 7.8. Measures of classification power for binary response models.	266
Table 7.9. Estimates and asymptotic standard errors (in parentheses) for binary response models. Approximate parametric approach.	269
Table 7.10. Parameter estimates and asymptotic standard errors for transition model [7.18]. Fit via Maximum Likelihood. The parameter vector for staying in once category was constrained to zero.	277
Table 7.11. Transition probabilities predicted from RC model [7.18] and marginal frequencies obtained from applying Markov chain to mean initial frequencies for selected settings of Initial Age, Quadratic mean diameter, and Trees per acre. Three forecasts listed.	282

Table 7.11.(Continued).	283
Table 7.12.Parameter estimates and asymptotic standard errors for transition model [7.19]. Fit via Maximum Likelihood. Standard errors italicized.	287
Table 7.13.Asymptotic 95 % confidence intervals for predicted proportions in ordered loblolly pine transition models. Length of intervals in probability units.	288
Table 7.14.Results of fitting model [7.20] to slash pine development data. Standard errors italicized.	293
Table 7.15.Results of fitting model [7.21] to loblolly pine data. Asymptotic standard errors in parentheses.	305
Table 7.16.Results from fitting model [7.22] semi-parametrically.	305
Table 7.17.Asymptotic 95 % confidence intervals for predicted proportions in marginal loblolly pine model [7.22]. Results for mixed model based on fixed effects only, ignoring random effects variability.	309
Table 7.18.Estimates and asymptotic standard errors (in parentheses) for binary response models. Approximate parametric approach.	320
Table 7.19.Parameter estimates and asymptotic standard errors for multinomial loblolly pine transition model. Fit via Maximum Likelihood. The parameter vector for staying in once category was constrained to zero. Standard errors italicized.	326
Table 7.20.Parameter estimates and asymptotic standard errors for ordinal loblolly pine model. Fit via Maximum Likelihood. Standard errors italicized.	327
Table 7.21.Asymptotic 95 % confidence intervals for predicted proportions in ordered loblolly pine transition models. Length of intervals in probability units.	328
Table 7.22.Results of fitting model [7.26] to slash pine development data. Standard errors italicized.	334

Chapter 1

Introduction

As any other scientific discipline, Forest Biometry has its roots, history, and challenges. Among the suites of challenges in this day and age we can identify two broad classes.

On one hand, we are dealing with a rather small set of questions that have pursued the profession for many decades and which seem to offer a timeless appeal and field of scientific endeavor. Among those questions are such apparently simple ones like “*what is the volume of a tree*”, “*how can we depict and model growth*”, “*what is the productive potential of a site*”. Perhaps the currently most important challenge of traditional interest is the proper utilization and appreciation of typical forestry data.

Most forestry data are not gathered through designed experiments. Although designed experiments are not uncommon, many researchers will contend that the

possible influences and extraneous sources of variation are so numerous that designed experiments conducted in the field always will retain an observational character to some degree. Maybe more importantly, it has been common to perform multiple measurements of the units of interest. Trees do not walk away like patients in a clinical experiment. Hence, repeatedly measuring the same trees or plots in an inventory system has become common practice. Observational studies with multiple measurements of units, which are called longitudinal, a term more fully explored in a later chapter, lead to observations that are likely to be correlated.

Correlations may also arise in quite different contexts. Trees may be felled and cut up into 4' sections and the sectionwise volumes are calculated. Like ring widths on an increment bore or basal areas on a plot at different ages, these volumes are not independent.

It is a regrettable fact that with a few exceptions, analysts in forestry have been impervious if not resistant to the presence of such correlations (Gregoire, Schabenberger, and Barrett 1995). Two reasons for this can be identified easily. One, accounting for serial or spatial correlations among measurements in a modeling context does not simplify the analysis. In fact, it can complicate matters to a considerable degree, both conceptually, and numerically. Two, it has often been the impression of researchers that such correlations are a mere nuisance, from which nothing can be gained, and that a classical analysis assuming independent observations is perfectly valid.

The latter conception, though widespread, is troublesome because it is wrong. In the presence of correlated measurements, classical analyses are not valid. Like variance heterogeneity, correlations affect the error structure of a particular model. If the error part in a linear model is misspecified, the coefficient estimates obtained by ordinary least squares (OLS) are still consistent and unbiased. However, it has long been known, that OLS estimates can be very inefficient in such circumstances, and more importantly, that the estimates of their precision are biased and inconsistent.

Incorrectly estimating the precision of model terms affects all inferences under the stipulated model, from partial tests about the significance of regressors to confidence statements and criteria of model performance (prediction errors etc.).

It is also incorrect that correlations are a mere nuisance, their direction and magnitude are important pieces of information. In other terms, a second, correlated observation can be thought of as contributing not an entire degree of freedom, but only a partial one, since its value is not free, but somehow predetermined by knowing a first observation and the strength of their correlation. It thus appears unreasonable from the outset not to utilize such information in the modeling process. Gregoire, Schabenberger, and Barrett (1995) have demonstrated how utilizing information about the correlation pattern, i.e. the error structure can not only validate, but strengthen and improve all aspects of modeling, from the efficiency of coefficient estimates to predictions under the model.

Our increasing ability to handle numerically difficult problems makes it possible to implement efficiently the analytical techniques that are defined by the recent progress in statistical theory for correlated observations. In light of the combined advances in theoretical knowledge and computing power it seems hardly justifiable today to forego proper utilization of a great deal of forestry data on the grounds mentioned above.

The actual time required to analyze data pertaining to the classical questions of Forest Biometry stands in sharp contrast to the time, expense, and effort that goes into the collection and maintenance of longitudinal data bases. Not to match this effort of data collection in the stage of data analysis is counter to the high ethics of the profession.

It appears that the challenges of modern Forest Biometry in its classical suit are linked to the data structures that confront us. This is also true for the more recent questions that begin to pursue the profession and that we are attempting to answer. Many of these were raised by a certain shift of emphasis from a purely volume-production oriented forest management to a more interdisciplinary approach that reaches into the fields of ecology and biology. They are oftentimes linked to experimental outcomes that are not measurements on a continuous scale such as height, diameter, volume, or ring width, but response variables that take on only a countable, usually finite number of possible values. Such outcomes will be called categorical.

Many categorical variables arise naturally, like tree species comprise a set of nominally scaled categories, which are distinct, but unordered. Tree mortality is an event representing a response with only two categories “alive” and “dead”. Such event/no-event outcomes are typically called binary.

Ordinal outcomes are a special type of multi-category scale. The response can take on one of a number of distinct and exclusive values which can be ranked. Grouping salaries into wage classes or diameters into dbh classes generates such an ordinal scale. With ordinal scaling, categories can be meaningfully compared only with respect to their ranking. An example for natural ordering, extending the binary example, is observing the health status of trees with respect to a fungus infection as “healthy”, “weak infection”, “strong infection”, “dead”. Aside from the ranking of the categories no arithmetic is possible on the scale. Even if we assign scores to the categories, e.g.

healthy	≡	0
weak infection	≡	1
strong infection	≡	2
dead	≡	3,

it is not meaningful to refer to differences between states of infection, because such quantities are undefined. One could also assign the scores [-100, -99, 100, 107]. This would still represent the same ordering as [0, 1, 2, 3] and must not lead to any numerical changes.

Modeling categorical responses is a rather new feature in Forest Biometry. The frequent occurrence of categorical data and the increasing interdisciplinary character of our work, however, requires forest biometricians to have some knowledge of the do's and do not's when dealing with such data. Maybe the simplest piece of information one should acquire is that again, classical linear model analysis is not valid. To see this from several perspectives, envision the modeling task as replacing the expectation of an observable random variable by a stipulated model. If for example, the response is binary and can be coded as 1/0 corresponding to event/no-event, we are interested to find a suitable model for $Y \in \{0, 1\}$. But

$$E(Y) = 0 * \Pr(Y = 0) + 1 * \Pr(Y = 1) = \Pr(Y = 1).$$

Thus the expectation of Y , the target of the model is a probability. Any linear model that does not confine the predictions to the $(0,1)$ interval is of limited value.

Like the challenge to classical Forest Biometry to deal with complex error structures the analysis of categorical data places similar demands on analysts. A key feature of categorical data is the functional dependence between means and moments of higher order. For the binary response exemplified above, for example

$$\begin{aligned} \text{var}(Y) &= E(Y^2) - E(Y)^2 = \Pr(Y = 1) - \Pr(Y = 1)^2 \\ &= \Pr(Y = 1)\{1 - \Pr(Y = 1)\} \end{aligned}$$

Since $\Pr(Y = 1)$ depends on the stipulated model, variances are furthermore not homogeneous across observations.

The techniques to model categorical data properly are well established. The theory of Generalized Linear Models (GLM) provides the necessary building blocks. The challenge to Forest Biometry will be to utilize GLM's more fully and in compliance with this body of theory and to advance their usage among biometricians to make them as much of an everyday tool like regular linear regression.

In this dissertation an attempt is made to combine the two challenges, that is to model longitudinal categorical data with special emphasis of ordinal responses. Owing to the very recent nature of the theoretical developments in mathematical statistics, it is important however, to discuss the state of the art, options, problems and alternatives in mathematical terms.

Since significant contributions to both fields, longitudinal data analysis and categorical data analysis, have taken place over the last few years only it is not surprising that by combining the techniques many questions will remain open and unanswered as this is an area of active and vibrant statistical research.

Nevertheless, we are today confronted with longitudinal categorical data and will be so increasingly in the future. It is to be understood that this dissertation comprises a first attempt to successfully and properly model such data structures in a forestry context. It is expected that the attempts and approaches taken here will be surpassed in the future.

The organization is as follows: § 2 reviews the literature on ordinal regression models and techniques for the analysis of longitudinal data in general. Background information for the developing discussion is provided for a number of issues in § 3. Alternative approaches to statistical estimation for longitudinal categorical data with emphasis on ordinal responses are presented in detail in § 4. The remainder of this dissertation is devoted to an application connected to the East Texas Pine Plantation Research Project in which the ordered states of a fungus disease are repeatedly measured over time. § 5 describes the disease process and past attempts of modeling it. § 6 discusses the actual data base in an exploratory fashion. § 7 presents results from analyzing the data using several methods from § 4 pertaining to various research questions. A discussion and summarizing outlook conclude the main body of the text in § 8. An Appendix with results from mathematical statistics that are not discussed in the main text, but presumed necessary for the theoretical development is also added. References to entries of the appendix will be denoted as [A. x], where x stands for a labeled number in the appendix.

Chapter 2

Literature Review

2.1 Ordinal Response Methods

As outlined above, an ordinal response is a characteristic whose distinct values comprise a set of categories among which a ranking exists. The development of generalized linear models (GLM) (Nelder and Wedderburn 1972; McCullagh and Nelder 1989) has spawned effulgent analytical tools to model non-continuous and non-normal responses. Amemiya (1981) reviews some of the qualitative response models that belong to the class of GLMs. The newly introduced ideas by Nelder and Wedderburn were the use of link functions and non-normal error distributions, while at the same time the modeling of continuous and categorical responses was made possible

within the same framework. Link functions are mappings of the expectations of the responses onto the real line, that (i) linearize the expected response in the parameters and (ii) allow predictions that concur with the possible scale of the expected response. For example, a binary response Y , coded 0/1 has an expectation which is confined to $(0, 1)$ since it is a probability. The link function transforms $E(Y)$ into a linear function of covariates and unknown constants to be estimated.

For ordinal response several special issues arise. Since each response belongs to a discrete set of possible outcomes, the response distribution follows the multinomial law, i.e. is multivariate. Secondly, the ordering of the categories needs to be reflected in the model and the predictions made under the model. Several different attempts have been described in the literature to deal with this situation. Greenwood and Farewell (1988) and Agresti (1989, 1990) give a description of the various approaches. One can roughly distinguish between estimation based on adjacent-category links, continuation-ratio links, and cumulative links. A fourth category of ordinal regression models is due to Anderson (1984) and is based on constrained multinomial logistic regression. Finally, mean response models have been proposed to deal with ordered response (Koch *et al.* 1977; Lipsitz 1992). For the case of categorical covariates, some of these models have representations as log-linear models in classification tables, where each cell entry corresponds to the response counts at a particular covariate setting. This offers the analyst the option the use of log-linear model theory as a special case of GLM's (Agresti 1989, 1990, 1992, 1993; Schabenberger 1993). In many studies however, at least some covariates are continuous. Categorizing the possible covariate combinations into sets of distinct covariate settings results in a loss of information. We therefore focus here only on models that accommodate a general covariate structure.

To fix ideas let C be the observed ordinal response with values $j = 1, \dots, J$. If we recode the event $C = j$ with a $(J*1)$ response vector $\mathbf{Y}$ consisting of zeros and a 1 in the j th position, then $\mathbf{Y} \sim MN(1, \pi_1, \dots, \pi_J)$ where MN denotes the multinomial distribution with category probabilities π_j . Let $G^{-1}(\bullet)$ be an invertible link function, $\mathbf{x}$ a $(p*1)$ vector of covariates and β a conformable $(p*1)$ vector of unknown constants. Adjacent-category, continuation-ratio and cumulative link models are cases of composite link functions (Thompson and Baker 1981), since they involve a transform of more than one of the indexing parameters of the response distribution. For sake of exposition it is assumed that the link function is the logit link $G^{-1}(t) = \text{logit}(t) = \log\left(\frac{t}{1-t}\right)$. Adjacent-category logits are then defined as

$$\log\left(\frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})}\right) = \alpha_j + \mathbf{x}'\beta \quad [2.1]$$

continuation-ratio logits give rise to

$$\log\left(\frac{\pi_j(\mathbf{x})}{\sum_{k=j+1}^J \pi_k(\mathbf{x})}\right) = \alpha_j + \mathbf{x}'\beta \quad [2.2]$$

and cumulative logits to

$$\log\left(\frac{\sum_{k=1}^j \pi_k(\mathbf{x})}{\sum_{k=j+1}^J \pi_k(\mathbf{x})}\right) = \alpha_j + \mathbf{x}'\beta \quad [2.3]$$

In all three cases, the linear predictor consists of two components, a category dependent parameter α_j , oftentimes incidental, and a linear function of covariates $\mathbf{x}'\beta$.

As Agresti (1989) points out, [2.1] can be represented as a log-linear model if covariates are categorical. [2.2] is a simple logit of conditional probabilities, i.e.

$$\text{logit}\left(\frac{\Pr(C = j|C \geq j, \mathbf{x})}{1 - \Pr(C = j|C \geq j, \mathbf{x})}\right)$$

and [2.3] can be written as

$$\text{logit}\left(\frac{\Pr(C \leq j|\mathbf{x})}{1 - \Pr(C \leq j|\mathbf{x})}\right).$$

Cumulative link models such as [2.3] have been proposed by McCullagh (1980) and McCullagh and Nelder (1989) and for the logit link are known as proportional odds models (c.f. Fienberg 1987). Continuation ratio models have been used by Armstrong and Sloan (1989), Cox (1988), Berridge and Whitehead (1991), Greenwood and Farewell (1988), and Kuk and Chen (1992). Läärä and Matthews (1985) have established an equivalence between continuation-ratio models and cumulative link models if the link function is the complementary log-log transform that refers to the gradient estimates. Note, that while the cut-off points in a cumulative link model are always ordered, this is not necessarily true for a continuation-ratio link model (Engel 1988, 1989). In [2.2] for example, $\pi_j(\mathbf{x}) = \exp\{\alpha_j + \mathbf{x}'\beta\} \sum_{k=j+1} \pi_k(\mathbf{x})$.

Cumulative link models are invariant under amalgamations of adjacent categories. If two categories are combined, the cumulative link model remains valid under a change of the incidental cut-point parameters, This is not true for the continuation-ratio models (Armstrong and Sloan 1989; Greenwood and Farewell 1988). This is one reason, why the use of cumulative link models has been proposed for the situation

where the ordered categories arise from discretizing a continuous latent variable and continuation-ratio models when the categories are of intrinsic interest. As will be discussed in a separate section, cumulative link models remain valid even if a latent continuous scale does not exist.

Adjacent-category logit models can be expressed as baseline-category models and fit separately using standard software (c.f. Begg and Gray 1984; Jobson 1990).

Anderson (1984) entertained the idea of using multinomial logistic regression to model ordered response through imposing constraints on the estimated response probabilities to ensure stochastic ordering. His model is known as the stereotype model. The models proposed by Greenland (1985) can be viewed as special cases of the stereotype model (Campbell and Donner 1989). Anderson commenced from the multinomial logistic model

$$\Pr(C = j|\mathbf{x}) = \frac{\exp(\alpha_j + \mathbf{x}'\beta_j)}{\sum_{k=1}^J \exp(\alpha_k + \mathbf{x}'\beta_k)}$$

and imposed ordering by letting $1 = \phi_1 > \phi_2 > \dots > \phi_J = 0$, and thus

$$\Pr(C = j|\mathbf{x}) = \frac{\exp(\alpha_j + \phi_j \mathbf{x}'\beta)}{\sum_{k=1}^J \exp(\alpha_k + \phi_k \mathbf{x}'\beta)}. \quad [2.4]$$

The similarity between [2.4] and the multinomial logit model has appeal although the model is more difficult to fit due to the multiplicative involvement of the ordering parameters in the predictor. The stereotype model is appropriate, if one wants to

address the question how adjacent categories behave with respect to covariates (Armstrong and Sloan 1989; Greenwood and Farewell 1988; Holtbrügge and Schumacher 1991) and thus provides a means to check the appropriateness of the ordering. They are less informative in instances where the order of the response is natural rather than assessed (Anderson 1984).

Mean response ordinal regression models include the category scores into the response and usually utilize an identity link function. For example, following Koch *et al.* (1977) one might replace the above logits by $\sum_{j=1}^J u_j \pi_j(\mathbf{x})$ where u_j is the score value of the j th category. Lipsitz (1992) equivalently modeled the expectation of $Z = \mathbf{u}'\mathbf{Y} = \mathbf{x}'\boldsymbol{\beta}$ where $\mathbf{u}$ is the vector of scores and $\mathbf{Y}$ is the indicator response vector defined above. The disadvantage of these mean response models is that additional constraints are needed to ensure that the expected response $\mathbf{u}'\boldsymbol{\pi}(\mathbf{x}) \in (0, 1)$ and that they depend on the scores assigned. Especially the spacing of the scores affects the result. While the score vectors $\mathbf{u}' = [0 \ 1 \ 2 \ 3]$ and $\mathbf{u}' = [1 \ 2 \ 3 \ 4]$ will lead to the same results in Lipsitz (1992), using $\mathbf{u}' = [1 \ 2 \ 3 \ 96]$ will have a considerable effect on model performance, despite the fact that all score vectors represent the same ordinal scale.

Upon reviewing the literature of the past decade it emerges that cumulative link models have found by far the widest applicability. Some examples are Agresti (1989, 1992), Agresti and Lang (1993), Arabatzis and Gregoire (1991), Anderson and Philips (1981), Ashby *et al.* (1986, 1989), Campbell *et al.* (1991), Engel (1988, 1989), Hastie and Tibshirani (1987), Hastie *et al.* (1989), Hedeker and Gibbons (1994), Jansen (1990), Landis *et al.* (1988), McCullagh (1980, 1984), McCullagh and Nelder (1989),

Miller and Landis (1991), Stram *et al.* (1988), Tosteson *et al.* (1989), Ware *et al.* (1988). They offer some specific advantages both in interpretation of results as well as model fitting, that recommend their consideration. A specific section is devoted to a more thorough discussion of cumulative link models with special attention paid to McCullagh's proportional odds model (McCullagh 1980).

2.2 Longitudinal Data

2.2.1. Longitudinal Data in General

Longitudinal data have some specific merits not met by classical cross-sectional data (Diggle *et al.* 1994). Longitudinal data can be viewed as two-level data, where repeated measurements are nested in the actual units of interest. Some authors call the repeated measurements elementary units and the units themselves clusters (Longford 1993). Alternatively, the generic term *subject* is recommended for the unit that is repeatedly measured and will be used here. The appealing feature of this data structure is that it contains implicit information about change, that can not be inferred from cross-sectional data (Uncles 1988). Like time series analysis it allows one to model a subject's behavior over time, but simultaneously allows for comparisons between subjects at any given time point as well as across time points. Not surprisingly, a multitude of research questions can be posed and addressed with longitudinal data which can not be supported by cross-sectional data. This has led to a distinction of different analytical approaches depending on the primary focus of the research question into population-averaged, subject-specific and conditional models (Zeger, Liang and Albert 1988, Schabenberger *et al.* 1995). Other terms used are marginal models (c.f. Davies and Crouchley 1984; Zeger 1988a; Ware *et al.* 1988; Zeger and Liang 1992) as a synonym for population-averaged models. Subject-specific models have also been referred to as random effects models or mixed models (Ware 1985). Conditional

models are sometimes called transitional models (Ware *et al.* 1988; Zeger and Liang 1992). As will be seen subsequently, these terms are not as congruent as their use in the literature might suggest. For the treatment in this chapter, a population-averaged model will be called a model in which the estimated coefficients have an interpretation of a population averaged behavior in the covariates. A subject-specific model is understood as leading to an interpretation in terms of a subject's behavior and the term conditional model is reserved for models in which the transition through states is the focus. Population-averaged models will always be marginal ones in the sense of Ware *et al.* (1988), however, a modeling approach that focuses on the marginal moments of the response distribution may well lead to a subject-specific interpretation of the coefficients (Zeger, Liang and Albert 1988). This distinction has not been clearly made in the literature and may lead to confusion and erroneous conclusions, for example if estimates from different types of models are compared (Rosner and Milton 1988; Rotnitzky and Jewell 1990; Rosner and Tosteson 1990). Before reviewing the literature on longitudinal models for categorical response, a brief discussion of models that apply when the response is continuous seems necessary, since some categorical models build on these.

2.2.1.1. Approaches for continuous response

Next to ignoring the serial correlation among repeated measurements on a subject and performing standard analysis, the approach of elementwise regressions has been proposed in the literature. These methods constitute a two-stage analysis. In the first stage, regression models are fit to the subject-wise or time-wise data separately. The

coefficients obtained in the first stage are then combined in a second stage by various procedures, ranging from computing weighted averages to linear models for the first-stage coefficients. Individualized regressions by time points have been used by Wei and Stram (1988) and Lipsitz *et al.* (1988). Forestry applications of subject-wise models can be found in Ferguson and Leech (1978), Davis and West (1981), and Biging (1985). Korn and Whittemore (1979) introduced the idea for binary response models, Diem and Liukkonen (1988) apply it to continuous response. Individualized two-stage regression procedures are today of historical interest only, because of their limitations. The two level structure of longitudinal data, subjects and repeated measures nested within subjects, is broken up into separate stages of the modeling process. Especially when the number of repeated observations is rather small, the first-stage coefficients are inefficient and unstable. Non-linear models may be difficult to fit to such small individual data sets (cf. Biging 1985). The variance-covariance structure of the second-stage responses is usually built from the variance-covariance estimates of the first stage. They are even more unreliable than the first-stage point estimates. The standard method of analysis at the first stage is usually linear OLS (Ferguson and Leech 1978) or non-linear LS (Biging 1985), which does not properly account for the serial correlations within subjects. The investigation of first-stage model residuals to discern the correlation of the observations is no means, since the residuals exhibit correlations due to the modeling process.

For the case where a sufficient number of observations per subject are available, Davidian and Giltinan (1993a, 1993b, 1995) proposed an improved elementwise regression approach, that borrows strength across subjects by estimating the weight matrix, i.e. the covariance parameters over all subjects. Thus, a typical feature of

efficient methods for analysing longitudinal data, - gaining strength by utilizing the two level nature of the data -, is retained.

Some authors have termed the two-stage approach based on elementwise regressions a random coefficient model since the coefficients of the first stage serve as responses in the second stage. Following the more recent literature on longitudinal data and the classical literature on linear models, the term random coefficients should be reserved for models in which the predictors are functions of random quantities.

Analytical tools to deal with longitudinal data can be viewed as methods to allow for general, complex error structures. Gregoire (1985, 1987) introduced a series of models for longitudinal data into the forestry literature that accounts for serial correlations. In his approach the combined error term of a linear model is broken up into components due to serial correlation, subject heterogeneity and other sources. These models can be expressed as mixed linear models. Laird and Ware (1982) provided the ground work for these mixed linear models that have opened up new approaches to the analysis of responses under general complex error structure. In the notation of Laird and Ware (1982) a linear mixed model is of form

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \quad [2.5]$$

where $\mathbf{Y}_i$ is the (n_i*1) vector of repeated observations on subject i , $\mathbf{X}_i, \mathbf{Z}_i$ are (n_i*p) and (n_i*q) matrices of covariates, $\boldsymbol{\beta}$ is a $(p*1)$ vector of fixed effects and $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$, $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i)$. The marginal distribution of the responses is thus $\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i)$. Estimates for the coefficients are obtained via maximizing the marginal likelihood of the responses. The fixed effects estimates $\hat{\boldsymbol{\beta}}$ are available as generalized least squares estimates (GLS) based on the estimated marginal variance-

covariance matrix $\widehat{\mathbf{V}}_i = \mathbf{Z}_i \widehat{\mathbf{D}} \mathbf{Z}_i' + \widehat{\mathbf{R}}_i$. The solutions for the random effects are empirical Bayes and related to Stein-type shrinkage prediction (Laird and Ware 1982, Robinson 1991). If $\mathbf{R}_i = \sigma^2 \mathbf{I}_{n_i}$ estimates of the covariance parameters can be obtained via the EM algorithm (Dempster, Laird and Rubin 1977, Laird and Ware 1982). Newton-Raphson and Fisher scoring algorithms can be found in Lindstrom and Bates (1988) or Wolfinger *et al.* (1993). Gregoire, Schabenberger and Barrett (1995) applied models of this type successfully in forestry, other applications can be found for example in Diggle (1988), Diggle (1990), Hedeker *et al.* (1994), Jennrich and Schluchter (1986), Jones (1990, 1993), Jones and Ackerson (1990), Jones and Boadi-Boateng (1991), Rutter and Elashoff (1993), Wolfinger (1992), Wolfinger *et al.* (1992).

Model [2.5] is linear in the coefficients and has a continuous response. Both features make it appear uninteresting for categorical response models, where the expectation is non-linear in the parameters. For the continuous response case, most recent work has extended model [2.5] to allow nonlinearity in the parameters. Several approaches exist to fit the non-linear mixed effects model. One way to accomplish this is to take a Taylor series expansion and approximate the marginal distribution of the response for an approximate linear version of the non-linear model. These are either first-order expansions around the expected values or the estimates of the random coefficients $\mathbf{b}_i$ or second-order expansions around the random coefficients and the fixed effects estimates. First order expansions have been used by Vonesh and Carter (1992) and Vonesh (1993), Sheiner and Beal (1980), and Hirst *et al.* (1991). Lindstrom and Bates (1990) and Gregoire and Schabenberger (1995a, 1995b, 1995c) used second-order expansions that lead directly to a linearized model of the Laird and Ware form [2.5]. The model parameters are usually fit by either maximum likelihood or restricted

maximum likelihood, based on the approximated marginal distribution. Schabenberger (1995b) reviews some of the existing problems with the algorithm presented by Lindstrom and Bates and details a more efficient implementation as well as semiparametric estimation. Davidian and Gallant (1993) proposed nonparametric approaches to fit mixed non-linear models for longitudinal data with a smooth random effects density; Racine-Poon (1985) used a Bayesian approach, similar in spirit to elementwise regressions.

As can be seen, both linear and many forms of non-linear mixed models for longitudinal data revolve around the simple model of the Laird and Ware form [2.5]. It allows for serial correlation in two ways: (i) explicitly, by modeling the within-subject error matrix $\mathbf{R}_i$ as a function of covariance parameters, and (ii) through the design of the covariate matrix $\mathbf{Z}_i$. If the elements of at least one column of $\mathbf{Z}_i$ vary for a given subject, the model not only explains subject to subject heterogeneity, but also within-subject variation, which enables one to pick up serial correlations. This is especially useful, if the time-varying columns of $\mathbf{Z}_i$ are functions of the correlation metameter, i.e. the scale to which the size of the correlations is likely to be linked (Ware 1985).

2.2.1.2. Categorical response

As was outlined briefly in the opening of section 2.2.1. a distinction has to be made whether the focus of the research question relates to population-averaged, subject-specific, or conditional (transitional) behavior. The model and estimation principle used has to reflect the research question properly. As can be seen from [2.5]

the random effects in the linear model affect the marginal distribution not through their first moment, but the higher order moments. This is in general not true for non-linear models. Categorical models are always non-linear in the expected response if link functions other than the identity link are used. In the following, models for population-averaged, subject-specific and conditional response are addressed separately. A specific section is devoted later to the features of these model types, their distinction and choice by the subject matter scientist.

2.2.1.2.1. Population-Averaged models

Let Y_{it} denote the categorical response of subject i at time t , assumed univariate. Population averaged models focus on

$$G^{-1}(E(Y_{it})) = \mathbf{x}'_{it}\beta$$

where $\mathbf{x}'_{it}$ is a vector of covariates at time t . Parametric approaches require one to specify the full marginal (joint) distribution of the responses including the dependence pattern over time. Koch *et al.* (1977) and Landis *et al.* (1988) provided solutions to this estimation problem. Their approach represents a categorical analog to multivariate ANOVA. Estimates are obtained using weighted least squares methods described earlier by Grizzle, Starmer and Koch (1969). In a setting with categorical covariates, individuals were stratified into classes of identical covariate and missing value pattern. For a J category response observed at t time points, this leads to J^t possible response profiles. If s covariate settings are possible one estimates the parameters of these s multinomial distributions. The variance-covariance matrix is fully estimable. A limitation of this approach is the requirement to stratify the observations into homogeneous classes. Continuous covariates must be discretized in order to arrive at a

finite number of covariate value combinations. But even under this premise the number s can get quite large as the number of covariates increases. As the number of observations at each setting becomes small, the weight matrix approaches singularity. For ordered responses, Agresti (1989) utilized the weighted least squares approach of Koch *et al.* for modeling ordinal response at two time points with a cumulative link model where the covariates are categorical.

If covariates vary only by subject (e.g. initial conditions), and the responses are binary, the joint distribution can be modeled using a beta-binomial model (Prentice 1988) and parameters estimated via maximum likelihood (ML). If covariates vary within subjects (time-dependent covariates), a likelihood approach is feasible only in the special case of paired observations because of range restrictions for the correlations if cluster size exceeds 2 (Prentice 1988, Rotnitzky and Jewell 1990). A parametric approach for an ordinal response based on the n -way Plackett distribution can be found in Molenberghs and Lesaffre (1992), see also Plackett (1981).

The difficulties with parametric population-averaged approaches have led to the investigation of semi-parametric techniques. The Generalized Estimating Equations (GEE) of Liang and Zeger (Liang and Zeger 1986; Zeger and Liang 1986; Zeger *et al.* 1988, Liang, Zeger and Qaqish 1992; Zeger and Liang 1992) provided the groundwork for a multivariate extension of the quasi-likelihood principle (Wedderburn 1974; McCullagh 1983; McCullagh and Nelder 1989; Nelder and Lee 1992) utilizing results from Estimating Function theory (Godambe 1960, 1976; Godambe and Heyde 1987; Godambe and Thompson 1989; Godambe and Kale 1991, Schabenberger and Gregoire 1995). This approach requires specification of the marginal response distribution only

up to the first two moments. A working structure is assumed for the within-subject variance-covariance matrix. Closeness of the working correlation structure to the true variance-covariance matrix of the marginal response vector increases the efficiency of the GEE estimates.

Liang and Zeger show, that provided the elements of the working matrix can be estimated consistently and the mean model is specified correctly, the estimates of the GEE are asymptotically unbiased and multivariate normal with estimable variance-covariance matrix. Prentice (1988) was the first to introduce the idea of two sets of estimating equations, where the first set models the expected value of the responses, the second set the expected values of association quantities, such as log odds ratios, cross-products or correlations. While with a single set of GEEs the time dependence is treated as a nuisance, although explicitly accounted for, two separate sets of estimating equations allow the analyst to model the time dependence structure explicitly. Depending on whether the two sets of estimating equations are treated as orthogonal or not, the terms GEE1 and GEE2 have been used since.

The GEE methodology provided the first general utility method to deal with correlated categorical data if interest is in a population-averaged model and time-varying as well as time-constant covariates are to be used. Prentice (1988) concluded that in situations where a parametric formulation is possible, the solutions are computationally cumbersome and their relative merit is questionable. GEEs provide a relatively straightforward method of estimation with robustness.

Not surprisingly, they have been entertained frequently, mostly in applications concerned with correlated binary responses. Some references are Fitzmaurice and Laird (1993), Fitzmaurice *et al.* (1993), Liang (1992), Lipsitz *et al.* (1991), Miller and

Landis (1991), Paik (1992), Park (1992), Prentice (1988), Sharples and Breslow (1992), Thall and Vail (1990), Zhao and Prentice (1990), and Zhao and Prentice (1991). Much less work has been done for longitudinal ordered response using the GEE approach (Miller and Landis 1991; Miller *et al.* 1993; Lipsitz *et al.* 1994).

As Liang and Zeger (1986) point out, the independence working assumption normally does not lead to poor gradient point estimates, although the estimates of their precision are distorted and consequently inference about model parameters is invalid. How various methods to model the correlations influence the model properties is still a matter of investigation, as is how the GEE perform with different covariate structures. Goodness of fit criteria are - aside from crude (Thall and Vail 1990) or approximate measures (Wei and Stram 1988) - not available. A contribution to this subject was made by Rotnitzky and Jewell (1990) who proposed adjusted extensions of Wald and Score tests for the estimated coefficients of GEE's.

An approach related to GEE, but in spirit closer to the elementwise regression approach was proposed by Wei and Stram (1988). It is generic in the sense, that it is not restricted to a specific type of population-averaged (marginal) model and can be applied in a parametric as well as a semi-parametric setting. Instead of individualized regressions, separate (univariate) analyses are performed for the different time points in the study. Zeger (1988) points out that this of course requires that the observation times are identical across subjects. Data arising from a continuous time process such as in Gregoire, Schabenberger, and Barrett (1995) can not be used with this method. As Wei and Stram (1988) show, the estimates from the separate time points follow an asymptotic multivariate Gaussian distribution, the variance-covariance matrix of which

is estimable consistently without parametric assumptions (Davis 1991). In the second stage of the process the time specific coefficients are examined for trends over time via hypothesis testing based on the asymptotic distribution. As with the GEE approach the method of Wei and Stram avoids fully specifying the marginal response distribution, only the time specific moments are necessary and the asymptotic distribution of the first-stage estimates provides the basis for the second, inferential stage. Zeger (1988) and Davis (1991) emphasize that the GEE approach yields the same estimates as the method of Wei and Stram if one allows the coefficients in the GEE to vary with occasions and uses an independence working correlation matrix. Stram, Wei and Ware (1988) applied the method of Wei and Stram to an ordinal repeated response based on McCullagh's proportional odds model.

Whilst the GEE method derives one vector of mean parameters using a working correlation matrix (assuming independence is just a special case of dependence), the Wei and Stram method yields as many parameter vectors as there are distinct occasions. In order to retain some reasonable overall significance level in the inferential step, Wei and Stram (1988) advise a sequential test procedure, since depending on the number of time points and parameters, a severe multiplicity problem arises. Schabenberger *et al.* (1995) have drawn additional connections between the Wei and Stram method and GEE's and provide a detailed discussion.

2.2.1.2.2. Subject-Specific models

One of the first subject-specific approaches was proposed by Korn and Whittemore (1979) for longitudinal binary data. Their approach is an individualized regression

model, where in the first stage a univariate logistic regression is fit to the data from each subject, using only within-subject covariates. In the second stage the coefficients from the first stage are the responses of a normal theory regression model on between-subject covariates. (Korn and Whittemore 1979; Stiratelli *et al.* 1984; Ware *et al.* 1988). The approach of Korn and Whittemore was primarily designed for panel studies where the number of repeated observations for a subject is rather large compared to the number of subjects, a condition not met by most longitudinal data. It is well known that logistic regression estimation becomes difficult, when response rates in the data vector are low, indeed becoming impossible when the responses do not vary (Albert and Anderson 1984, Hosmer and Lemeshow 1989). Therefore, individuals with low response rates, or that exhibited the same response at all occasions had to be discarded from the analysis. This introduces considerable selection bias. In the example provided by Korn and Whittemore (1979) almost one third of the data had to be excluded from the analysis. The drawbacks cited earlier in discussing elementwise regression approaches for continuous longitudinal response apply here in the same manner. In addition to the problems with fitting logistic models when the response rates are low, it is questionable, if the required asymptotics of the first-stage estimates hold for the number of repeated measurements usually encountered in longitudinal data. One certainly should not rely on asymptotic normality if the coefficients are based on only six or seven observations per subject.

Subject-specific models account implicitly for subject to subject heterogeneity. Albeit the method of Korn and Whittemore has been termed a random coefficients approach, the model that is commonly referred to as a random coefficients model and incorporates subject to subject heterogeneity uses coefficients that vary across subjects

according to some distributional law. Conditional on the random terms in the model it may be assumed that the observations are independent, a condition termed local independence (Conaway 1990). With longitudinal data, this may or may not hold, since accounting for serial correlation through random terms that vary across subjects requires that the functions that modify the random terms are related to the meter of serial correlation. Therefore, the random terms have been used in different contexts. Preisler (1988) utilizes random effects to account for overdispersion in cross-sectional binary response (see also Liang and McCullagh 1993) and Preisler (1989) analyzes longitudinal bioassay data with nested random effects for conditional probabilities.

Since categorical variables are readily analyzed using generalized linear models a natural extension to the subject-specific case is by applying the mixed linear model idea (see [2.5]) to a generalized linear model (McCullagh and Nelder 1989). As will be discussed in much more detail later, the generalized mixed linear model reckons the expectation of the response conditional on the random terms, $G^{-1}(E(Y_{it}|\mathbf{b}_i)) = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{b}_i$, which is the reason for the subject-specific interpretation of the coefficient estimates. Owing to the nonlinearity of the link functions, the marginal expectation, derived from integrating the conditional expectation over the random effects distribution does not equal the marginal expectation that one would model in a population-averaged model, $G^{-1}(E(Y_{it})) = \mathbf{x}'_{it}\boldsymbol{\beta}$. The consequence is an intransitivity of the coefficient estimates, not encountered with linear mixed models (Zeger, Liang and Albert 1988). Neuhaus *et al.* (1991) investigated this phenomenon for a random effects logit model and described the difficulties in interpretation of the subject-specific effects when covariates do not vary within clusters. Neuhaus *et al.* (1991) and Zeger *et al.* (1988) describe the attenuation in population-averaged

coefficients compared to the subject-specific model. If the same model is analyzed under a population-averaged and under a random effects structure, the population-averaged coefficients are attenuated, they will be closer to zero than the subject-specific estimates.

A great difficulty in estimation in generalized mixed models arises from the nontransitivity of the parameters. While one is able to write out the likelihood conditional on the random effects, parameter estimation proceeds from the marginal likelihood function, much as in mixed models for continuous response (see 2.2.1.). Owing to the nonlinearity of most link functions the integrals involved mostly do not exist in closed form (cf. Preisler 1988, 1989; Stiratelli *et al.* 1984; Zeger, Liang and Albert 1988). Several different ways have been proposed to proceed.

One may use a conditional likelihood approach (McCullagh and Nelder 1989) and render the random effects as nuisance parameters. This would require finding sufficient statistics for the nuisance parameters on which the data can be conditioned. Unfortunately, the sufficient statistics themselves may depend on the parameters of interest and the dimension of the sufficient statistics may be greater than the number of nuisance parameters (see also Laird 1991). Circumventing this difficulty, Cox and Reid (1987) have proposed conditioning on the maximum likelihood estimator of the nuisance parameters. The resulting profile likelihood is then treated as a function of the parameters of interest only. Since profile likelihood methods become more and more inefficient as the number of nuisance parameters increases, Cox and Reid (1987) recommend conditioning on the MLE of orthogonalized parameters. The orthogonalization, however, is generally only possible if the parameter of interest is a

scalar, reducing applicability of the method to small estimation problems. Liang (1987) has drawn a connection between Cox and Reid's approximate conditional likelihood approach and Estimating Function theory (see also Godambe 1976). Exact conditional likelihood is possible, however, if covariates and response are categorical. This is demonstrated in Conaway (1989), Agresti (1993), and Agresti and Lang (1993). The latter authors use a cumulative link model for ordered response and propose a generalization of the Rasch item-response model.

The second way to derive parameter estimates is by means of numerical integration of the marginal likelihood. The numerical technique most frequently employed is some form of Gauss-Hermite quadrature (Abramowitz and Stegun 1965; Davidian and Gallant 1993; Ezzet and Whitehead 1991; Hedeker and Gibbons 1994; Jansen 1990; Preisler 1988, 1989). Laird (1978) discusses nonparametric likelihood methods.

Harville and Mee (1984) used the idea of a threshold regression model, implicitly incorporated in a cumulative link model, and formulated the linear predictor as in [2.5] as consisting of fixed and random quantities. They utilize an approximation to the linear predictor to circumvent numerical integration and call their estimates approximate Bayes. Hedeker and Gibbons (1994) picked up on the idea of the generalized mixed effects threshold model and utilized Gauss-Hermite quadrature instead to evaluate the individual likelihood contributions and the necessary derivatives. Jansen (1990) also uses numerical integration for ordinal response models with random effects, but restricts the model to a single random term. Harville and Mee (1984) and Jansen (1990) use the EM algorithm as a vehicle for estimation, Hedeker and Gibbons

(1994) describe Fisher scoring, which is known to converge much faster than the EM algorithm.

A third way to tackle the problem of evaluating distributions that involve integrals that do not exist in closed form in generalized mixed models is by using the Gibbs sampler, i.e. empirically sampling from the appropriate conditional distributions (Gelfand *et al.* 1990; Gelfand and Smith 1990). The Gibbs sampler is a Monte Carlo method that allows one to estimate the desired posterior probabilities in a Bayesian framework. Note that if one assumes an uninformative prior distribution for the fixed effects, the posterior distribution of the fixed effects and the covariance parameters of the random effects is just a normalized version of the marginal likelihood (Zeger and Karim 1991). Using the notation of section 2.1.1 the Gibbs sampler produces realizations from the joint distributions of β , $\mathbf{b}$, and $\mathbf{D}$ via simulation. The algorithm is easy to implement but computationally very demanding and seems to offer advantages over numerical integration only if the dimension of numerical integration becomes very large. Karim and Zeger (1992) have analyzed the famous salamander data discussed by McCullagh and Nelder (1989). They state that with (only) 360 binary responses and 2000 simulated values from the distributions involved, parameter estimation required 5 hours computer time on a DEC workstation. This renders evaluation of competing models infeasible.

Breslow and Clayton (1993) have discussed Taylor series expansions to approximate the marginal response distribution and generalized the results of Lindstrom and Bates (1990) for non-linear Gaussian mixed models to general non-linear functions in the family of link functions and non-Gaussian errors. The models can be fit based on

a penalized sums of squares criterion described by Lindstrom and Bates (1990) following Green (1987). Breslow and Clayton termed this approach penalized quasi-likelihood. It requires evaluation of the random effects solution at every second step of a two-stage iterative algorithm. If the number of subjects is large, this results in an infeasibly large estimation problem (Schabenberger 1995b). Alternatively, they advise a marginal approach based on Zeger, Liang and Albert's (1988) results of subject-specific generalized models with an attenuation correction for the fixed effects. This estimation problem can be implemented using the GEE approach of Liang and Zeger (1986) after approximating the first two marginal moments from the conditional ones. Since this approach is essentially semiparametric and a Taylor series expansion is used to approximate the marginal moments, the problem of numerical integration does not arise. Solutions for the random effects can be obtained after the fixed effects have been estimated, keeping the estimation problem at a manageable size. Schabenberger (1995b) extends the work of Zeger, Liang and Albert (1988) for categorical response, non-normal errors and general link functions to continuous responses, arbitrary non-linear functions and circumventing distributional assumptions beyond the first two moments.

Wolfinger (1993) and Wolfinger and O'Connell (1993) also use Taylor series in concert with normal theory error assumptions to formulate the generalized mixed linear model in terms of a repeatedly fitted Gaussian linear mixed model, applying the Lindstrom and Bates (1990) algorithm to non-continuous responses.

Zeger, Liang and Albert (1988) elaborate on the point that an important difference between a population-averaged (PA) approach and a subject-specific (SS) approach is the nature of the within-subject variance-covariance matrix. While in SS models the

dependence arises from the shared subject effects $\mathbf{b}_i$, PA models provide a description of the covariance on the within-subject level only. Because of this distinction, the within-subject error matrix can be modeled much more freely in a PA model, allowing the analyst to overcome possible convergence and computational problems encountered with a subject-specific approach. Furthermore, inferences in a SS model depend on the distribution assumed for the random effects.

The approach of Hedeker and Gibbons (1994) as well as the GEE based approach of Zeger, Liang and Albert (1988) are marginal maximum likelihood and marginal quasi-likelihood, respectively. It is important to note, however, that these estimation techniques do not lead to population-averaged parameter estimates. The reason herefore is simply that the marginal error matrix is defined through the random coefficients that vary with subjects and not alone by the within-subject error assumption.

2.2.1.2.3 Conditional models

The rationale behind conditional models is rather simple. The expectation being modeled is the probability of a response conditional on other responses. There are - at least in principal - two ways to achieve this. Like in a regular regression model, the condition is introduced as a predictor variable or a response is constructed whose expectation is the conditional probability of interest. The parameters of a conditional model are interpreted as rates of changes (on the appropriate scale) for all subjects that share a specific condition. To distinguish these approaches a conditional model is called covariate-conditioned (CC) if responses are introduced as covariates and response-conditioned (RC) if a conditional response is constructed.

As mentioned before in the discussion of population-averaged and subject-specific models for categorical response the interpretation of the parameters is crucial. The natural interpretation of parameters in a conditional model relates to the conditional distributions under scrutiny (Rotnitzky and Jewell 1990). It may be possible that a certain covariate has a small effect on the conditional response but is of considerable marginal impact (Laird and Ware 1984). Liang and Zeger (1986), Prentice (1988), and Ware *et al.* (1988) discuss the difficulty to obtain expressions for marginal probabilities from relatively simple conditional models. Because the number of responses upon which an outcome can be conditioned depends on the cluster size (number of repeated measures), the coefficients might vary with cluster size (Prentice 1988). As a consequence, the marginal distribution of the response conditioned on covariates only will depend on all model parameters, including the size of a cluster.

Neuhaus and Jewell (1990) argue in reference to Rosner's model (Rosner 1984, 1989) that the natural interpretation of the parameters arises in a superpopulation where there are no subject (cluster) effects and the responses are uncorrelated. They conclude that the conditional coefficients generally underestimate the population-averaged effects and the subject-specific effects. Rosner and Tosteson (1990) express their strong disagreement with these findings and emphasize the fact that the coefficients of PA, SS and conditional models necessarily differ, because they measure different quantities. Consequently, conditional model constitute a class of models in its own.

For both classes of conditional models the focus in the literature has been on the case where the response is binary, and especially for the RC models, for cases where

the correlated responses on a subject form a cluster, rather than a time-ordered process. For example in ophthalmology, disease propensities are modeled by conditioning one eye's state on the condition of the other eye. An RC model that has attracted considerable attention in this general cluster setting was proposed by Rosner (1984, 1989), and Rosner and Milton (1988).

It is assumed that K primary units of analysis (subjects) have n_i secondary units nested within. The n_i secondary units exhibit a form of association. It is easily seen, how this situation is a special case of longitudinal data. For binary response Rosner assumes that the response probability among the subunits follows a beta distribution and that the response distribution for the t th subunit given the response of all the other units in the cluster (except the t th) is binomial. This leads to a compound beta-binomial model (Prentice 1988). Rosner's contribution was to augment the beta-binomial model to allow for covariates that vary both within and between clusters (subjects). If no covariates are present, Rosner's model simply reduces to a beta-binomial model (Connolly and Liang 1988). Parameter estimation utilizes standard ML principles. Rosner and Milton (1988) discussed a comparison between their RC model and a simple ordinary logistic regression analysis ignoring intracluster-correlation and report that in the incorrect analysis p-values of significance tests were too small and confidence intervals too narrow. In the simple case $n_i = 2, \forall i$, Rosner's model is based on two conditional logistic regressions

$$\log \left[\frac{\Pr(Y_1 = 1 | Y_2 = y_2, \mathbf{x}_1)}{\Pr(Y_1 = 0 | Y_2 = y_2, \mathbf{x}_1)} \right] = \mathbf{x}_1' \boldsymbol{\beta}$$

$$\log \left[\frac{\Pr(Y_2 = 1 | Y_1 = y_1, \mathbf{x}_2)}{\Pr(Y_2 = 0 | Y_1 = y_1, \mathbf{x}_2)} \right] = \mathbf{x}_2' \beta$$

where the subscript i has been dropped. Note, that with increasing cluster size and number of response categories, the number of possible logits of this type increases rapidly. The joint probability $\Pr(Y_1 = y_1, Y_2 = y_2 | \mathbf{x}_1, \mathbf{x}_2)$ can be constructed from the two logistic regressions. It is noteworthy that in the general cluster RC model a response is conditioned on the sum of all other responses in a cluster (Qu *et al.* 1987; Connolly and Liang 1988). While this makes sense in the case where no particular ordering of the cluster elements is present, such as in ophthalmology, where one eye does not have precedence over the other, in longitudinal studies one is interested in transition probabilities. The conditions refer to the history of a subject. In other words, the assumptions of the Rosner model applied to longitudinal data would imply that there is local independence in the responses given we know what happened before the response at time t occurred and given all future responses. This is a rather meaningless construct, since in a practical situation one does not have simultaneous information on past, present and future. It would furthermore imply that a response in category j is of equal weight regardless whether it occurred immediately prior to the current time point or will occur in 20 years.

As an extension of the model proposed in 1984, Rosner (1989) discusses the same RC model in a situation with more than one level of dependence. The success probability on the first level, p_i say, is assumed to follow a beta(a,b) distribution and conditional on p_i the second level response probabilities are assumed to follow a beta($\lambda p_i, 1 - \lambda p_i$) distribution where $\lambda p_i < 1$. Conditional on the second level parameters, local independence is assumed.

Qu *et al.* (1987) extended Rosner's model to allow for negative, positive and zero intracluster-correlation and use the Polya-Eggenberger distribution instead of the beta-binomial. The Qu *et al.* model combines ordinary logistic regression (zero correlation), Rosner's beta-binomial model (positive correlation) and a hypergeometric model (negative correlation). Since Rosner's model is computationally very demanding as cluster size increases, Qu *et al.* use a simplex algorithm during estimation.

Connolly and Liang (1988) finally combined Qu *et al.*'s and Rosner's approach in the framework of a more general class of conditional models. Parameters are estimated using a working likelihood and working score function to ease the burden of complete likelihood evaluation. Liang and Zeger (1989) presented a time-series analog for multivariate time series data where the outcome is binary based on Connolly and Liang's work. As for Rosner's model Connolly and Liang (1988) retain the notion that the conditioning event is the sum of all responses in a cluster except the current one.

A second type of RC models combines conditioning on one response with random effects in the linear predictor (Conaway 1990; Preisler 1989). Since the use of random terms accounts for subject-to-subject heterogeneity, the serial correlation among a subjects repeated measurements may not be completely removed. Conaway therefore advises to condition on the previous response (or responses) and to use random effects simultaneously. The complications using random effects as outlined in section 2.2.1.2.2 apply here in the same fashion. Numerical integration may be required to compute the likelihood, likelihood score, and derivatives. Conaway (1990) interestingly assumes a log-log link function and a log-gamma distribution for the random effects to facilitate computations.

All RC models have in common that they are computationally rather demanding. CC models offer the advantage to work with the same response vector as in a marginal setting and to condition on previous events by augmenting the data matrix. The principal model of interest was proposed by Bonney (1987) and Brooks and Bonney (1987). The joint probability of response in a cluster is expanded into a product of conditional probabilities. For example:

$$\Pr(Y_{i1} = y_{i1}, \dots, Y_{in_i} = y_{in_i}) = \Pr(Y_{i1} = y_{i1}) \Pr(Y_{i2} = y_{i2} | y_{i1}) \Pr(Y_{i3} = y_{i3} | y_{i1}, y_{i2}) \dots \Pr(Y_{in_i} = y_{in_i} | y_{i1}, \dots, y_{in_i-1})$$

By assuming that each of the conditional probabilities follows a generalized linear model where the conditioning responses serve as covariates, the joint probability can be obtained easily. The advantage of models of the Bonney type is, that they can be fit using standard software, if the data matrix is augmented properly (Fahrmeir and Tutz 1994). Rosner (1989) criticizes that the way the conditions are introduced in Bonney's model require that an order exists among the responses in a cluster. As mentioned before, this may not be the case in studies of generally clustered data, but is certainly the case for longitudinal data, where the repeated measurements are strictly ordered in time.

Laird (1991) discusses a connection between Bonney's model and the time series model for continuous response by Rosner *et al.* (1985) and Rosner and Muñoz (1988).

Chapter 3

Background

3.1. Cumulative Link Models

Chapter 2.1. provided a brief introduction into ordinal regression models introducing the three most prevalent model types, cumulative link models, continuation-ratio models and stereotype models.

This section discusses cumulative link models in more detail and gives some guidelines on how to choose among the major model types in a particular context.

3.1.1. Rationale

One way of motivating a cumulative link model is by assuming the existence of an underlying, continuous scale from which an ordinal scale is derived. This construction is always valid if the ordered categories arise from discretizing a continuous scale. For example, yearly wages can be grouped into classes \$0-\$10,000, \$10,001-\$20,000, \$20,001-\$30,000, and so forth. Cumulative link models were first introduced in this context, illustrating why they are sometimes referred to as grouped-continuous models (Anderson 1980; Engel 1988, 1989). However, the existence of the latent continuous scale is not a requirement (McCullagh 1980; Schabenberger 1995a) and cumulative link models apply as well in cases where the categories are not intervals on a continuous scale.

Let Z be the underlying continuous variable and assume that the model

$$Z = -\mathbf{x}'\boldsymbol{\beta} + \epsilon \quad [3.1]$$

holds, where ϵ follows a distribution law with cdf $G(\cdot)$. The subscript i has been suppressed without ambiguity. We assume that the ordered response C takes on value j ($j = 1, \dots, J$) iff

$$\alpha_{j-1} < Z < \alpha_j; \text{ where } \alpha_0 = -\infty, \alpha_J = \infty.$$

Then

$$\begin{aligned} \Pr(C \leq j|\mathbf{x}) &= E(I_{\{C \leq j|\mathbf{x}\}}) = \Pr(Z \leq \alpha_j|\mathbf{x}) \\ &= \Pr(-\mathbf{x}'\boldsymbol{\beta} + \epsilon \leq \alpha_j) \\ &= \Pr(\epsilon \leq \alpha_j + \mathbf{x}'\boldsymbol{\beta}) \\ &= G(\alpha_j + \mathbf{x}'\boldsymbol{\beta}) \end{aligned} \quad [3.2]$$

The parameters α_j ($j = 1, \dots, J - 1$) are called the cut-off points. The negative sign in [3.1] is a convenience so that the predictor in [3.2] appears in standard additive form (Agresti 1990; McCullagh 1980). Since $\alpha_{j-1} < \alpha_j$, $\forall j$ it is implied that

$$\begin{aligned}\Pr(C = j|\mathbf{x}) &= \Pr(C \leq j|\mathbf{x}) - \Pr(C \leq j - 1|\mathbf{x}) \\ &= G(\alpha_j + \mathbf{x}'\beta) - G(\alpha_{j-1} + \mathbf{x}'\beta) > 0\end{aligned}$$

Consequently, the model [3.2] imposes strict stochastic ordering since the response distribution $G(\alpha_j + \mathbf{x}'\beta)$ is stochastically greater than $G(\alpha_{j-1} + \mathbf{x}'\beta)$ (Taylor 1990).

The choice of the link function $G^{-1}()$ is in principal arbitrary. Any invertible mapping from $(0,1)$ onto the real line can be used. However, the motivation through the grouped-continuous case where $\Pr(\epsilon < t) = G(t)$ suggests to choose $G^{-1}()$ as the inverse function of a continuous cumulative distribution function. Traditionally the logit and probit links are used most frequently for binary, nominal, and ordinal responses, corresponding to the quantile functions of the standard logistic and the standard Gaussian distribution. The logit link $G^{-1}(t) = \log\left(\frac{t}{1-t}\right) = \text{logit}(t)$ offers advantages because it is the canonical link of the multinomial distribution (Bickel and Doksum 1977; McCullagh and Nelder 1989). The choice of the link function is discussed more fully in § 3.1.4. Throughout this dissertation it is understood that $\log$ denotes natural logarithms.

If $G^{-1}(t) = \text{logit}(t)$, model [3.2] becomes

$$\Pr(C \leq j|\mathbf{x}) = \mu_j(\mathbf{x}) = \frac{\exp(\alpha_j + \mathbf{x}'\beta)}{1 + \exp(\alpha_j + \mathbf{x}'\beta)} \quad [3.3]$$

and has been termed the proportional odds model (POM) (McCullagh 1980, 1984). The

name POM stems from the fact that the log cumulative odds ratio of response at two different design points $\mathbf{x}_1$ and $\mathbf{x}_2$ is proportional to β , i.e.

$$\log \left\{ \frac{\Pr(C \leq j|\mathbf{x}_1)/\Pr(C > j|\mathbf{x}_1)}{\Pr(C \leq j|\mathbf{x}_2)/\Pr(C > j|\mathbf{x}_2)} \right\} = \text{logit}(\mu_j(\mathbf{x}_1)) - \text{logit}(\mu_j(\mathbf{x}_2)) = (\mathbf{x}'_1 - \mathbf{x}'_2)\beta$$

The log cumulative odds ratios are thus independent of the categories, implying the same effect on this log odds ratio for all categories. It has also been called the parallel lines assumption.

3.1.2. The Parallel Lines Assumption

Figure 3.1 depicts this assumption for an ordered response with 4 categories. The POM fits a series of logistic cdf's shifted across the range of the linear predictor and identical in shape. It is easily seen from Figure 3.1. that the POM reduces to the ordinary logistic model if the number of categories is $J = 2$. The category probabilities $\Pr(C = j|\mathbf{x})$ are obtained by subtracting the appropriate cumulative probabilities and are displayed in Figure 3.1. as vertical intersections. An alternative interpretation of the parallel lines assumption is obtained on the linear log cumulative odds scale, where

$$\log \left\{ \frac{\Pr(C \leq j|\mathbf{x})}{\Pr(C > j|\mathbf{x})} \right\} = \text{logit}(\mu_j(\mathbf{x})) = \alpha_j + \mathbf{x}'\beta .$$

Figure 3.2. depicts the relationship between the latent continuous scale and the ordinal scale for a model with a single regressor and four categories.

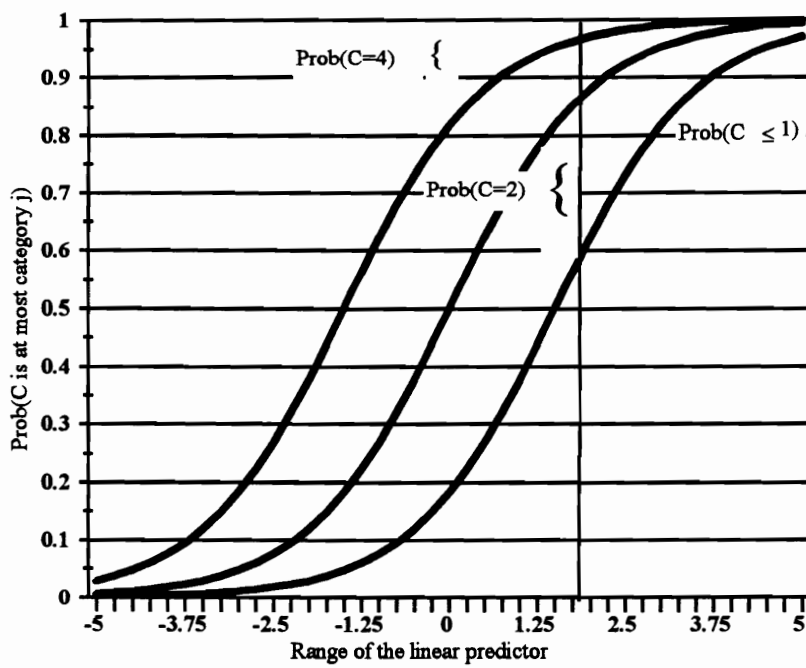


Figure 3.1. Depiction of cumulative logit model for an ordinal response with four categories. Adopted from Agresti (1990).

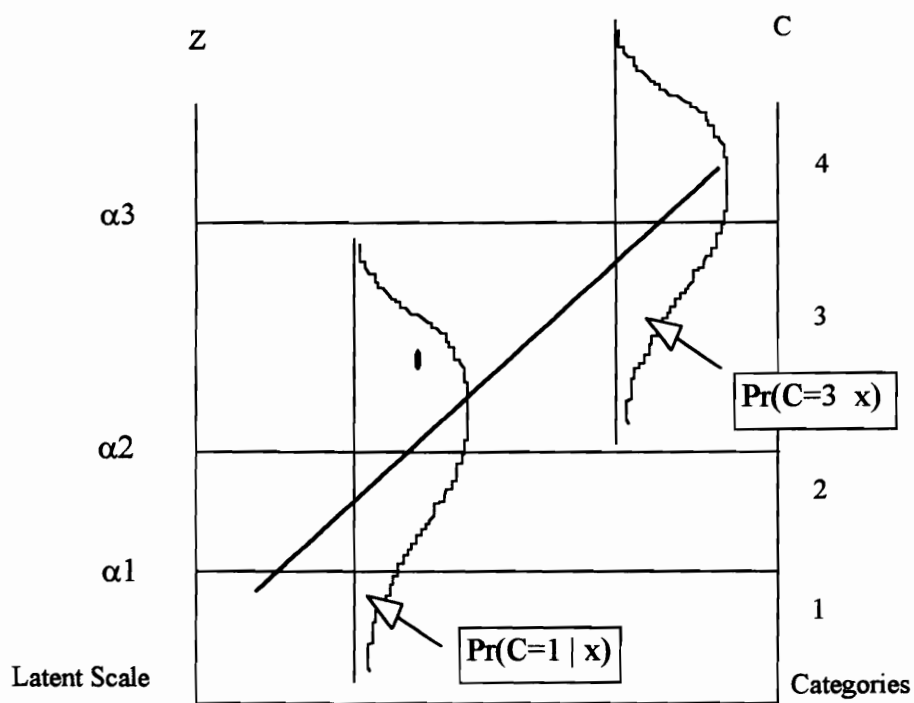


Figure 3.2. Relating the latent continuous and ordinal scale through a linear regression model and a series of ordered cut-offs.

The intercepts on the latent scale are the cut-off points α_j , $j = 1, \dots, J - 1$.

The parallelism assumption of the POM has been questioned repeatedly. Some authors believe, that this assumption imposes too many constraints (cf. Campbell and Donner 1991). Notice, that continuation-ratio logits also impose a parallel lines assumption, but not for the log cumulative odds, but the log conditional odds

$$\log \left\{ \frac{\Pr(C = j|\mathbf{x})}{\Pr(C > j|\mathbf{x})} \right\}.$$

The question of the appropriateness of such an assumption is at the heart of ordinal regression models and addressing it will lead to guidelines for the choice of an ordinal model type. First, the parallel lines assumption is expressed as log odds of cumulative or conditional probabilities. It will be difficult in some applications to develop an intuitive feeling for the meaningfulness of quantities of this particular type. Second, the assumption can - in principle, at least - be checked. If the POM is fit by maximum likelihood, we can compare the two nested alternatives

$$\text{logit}(\mu_j(\mathbf{x})) = \alpha_j + \mathbf{x}'\beta \quad [3.4]$$

$$\text{logit}(\mu_j(\mathbf{x})) = \alpha_j + \mathbf{x}'\beta_j \quad [3.5]$$

Model [3.5] relaxes the parallelity assumption and allows the gradient coefficients to vary with categories. If [3.5] provides a significantly better fit than [3.4] one might be inclined to remove the parallelity constraint. Likelihood ratio theory can be employed to test the hypothesis $H_0 : \beta_j = \beta, \forall j$. However, caution should be exercised for the following reasons. In Figure 3.1 rejecting this null hypothesis would lead to a series of logistic cdf's with varying shape shifted along the range of the linear

predictor. As a consequence, the response distributions will intersect, possibly within the range of covariate values of interest. This happens especially, when the response frequencies are small in some categories, since the cdf's in Figure 3.1 will move closer together. Intersecting curves result in negative predictions for the category probabilities, a quite unsatisfactory result. Even if model [3.5] fits the observed data better than [3.4], the resulting structure may be illogical. This can not be checked in advance, since it is an example of model-data interaction.

3.1.3. Model Parsimony and Response Nature

Considerations of this kind may be viewed as practical in nature. However, they are not entirely motivated by practical issues. From a more theoretical and philosophical angle, they relate to the importance of ordering at all. There has been agreement in the literature that statistical models for ordered response should impose strict stochastic ordering in some fashion. Various ways exist to achieve this goal; the POM with the parallelism constraint is just one of them. Log-linear models for ordinal data utilize the ordered scores (Agresti 1990; Haberman 1978; Hagenaars 1990). Comparing [3.5] and [3.4], where [3.5] jeopardizes strict stochastic ordering in favor of greater flexibility, we see that stochastic ordering is a stringent condition that usually leads to more parsimonious model structures.

Model parsimony in general is a desirable feature, but it should be noted here, that ordered regression models impose parsimony rather than offer the analyst a clear choice. It has been proposed to use multinomial logistic regression for nominal scales even if the outcomes are ordered in the hope that information about the ordering of the

response expresses itself through the estimates (Campbell and Donner 1989; Campbell *et al.* 1991). For example, compare a model for a 5-category response with 6 predictor variables. In the multinomial logistic model

$$\Pr(C = j|\mathbf{x}) = \frac{\exp(\alpha_j + \mathbf{x}'\beta_j)}{\sum_{k=1}^J \exp(\alpha_k + \mathbf{x}'\beta_k)}$$

we estimate $J - 1 = 4$ parameter vectors of length $(7*1)$ each, absorbing the α 's as intercepts. One category, L say, is chosen as a baseline for identifiability with $\alpha_L = 0$, $\beta_L = 0$. (Arabatzis 1990; Arabatzis and Gregoire 1991; Hosmer and Lemeshow 1989). The POM

$$\Pr(C \leq j|\mathbf{x}) = \frac{\exp(\alpha_j + \mathbf{x}'\beta)}{1 + \exp(\alpha_j + \mathbf{x}'\beta)}$$

estimates 1 gradient vector of length $(6*1)$ and $J - 1 = 4$ cut-offs. That is 28 parameters in the nominal multinomial model versus 10 parameters in the POM.

The greater flexibility of the multinomial logistic regression model leads to a less parsimonious model and dubious qualities with regard to the ordering of the response. Nevertheless, such a model may prove superior in terms of prediction (classification) over the strictly ordered alternative (Campbell *et al.* 1991).

It is obvious where this leads: if the response is not ordered or the order is of doubtful quality, using an ordinal regression model may lead to poor fit and poor predictions, since the inflicted stochastic order may not be supported by the data. A nominal multinomial procedure may be superior. If the ordering is correct and informative, however, the ordinal regression model suggests a parsimonious structure which will lead to better predictions (Campbell and Donner 1989). The analyst should

therefore ask how much confidence can be placed on the ordered nature of the response. Two cases can be distinguished, known as *natural* and *assessed* order.

Naturally ordered responses arise, when simple pieces of information lead to the assignment of an event to a response category. The grouped-continuous case is an example. Once the latent scale is divided into adjacent, non-overlapping, and exhaustive intervals the assignment to a category can be made without ambiguity; all categories are naturally distinguishable. Clearly defined classification rules that leave little room for ambiguity and intra-rater variation also lead to natural ordering.

A different situation occurs, for example, when a patient's health condition is assessed and a physician has to assign the patients state with respect to a certain disease to one of a number of ordered categories. Various pieces of information will be combined such as the patient's age, medical history, general constitution, etc.. If the possible categories are healthy, mildly ill, moderately ill, severely ill, and dead, the first and last category might be assessable quite easily. The classification in categories in-between however, depends on the interpreter's perception of the patient's status, the weight he attaches to the different pieces of information, as well as the meaning he derives from the attributes mildly, moderately, severely. Not surprisingly, different interpreters will arrive at different classifications. Such a response has been termed assessed ordered (Anderson 1984). Questions that need to be addressed are of the following kind:

are adjacent categories distinguishable, if not, should they be combined?
is the response in its entirety ordered or should it be treated nominally?

The notion of assessed ordering, category indistinguishability and dimensionality of the process prompted Anderson's investigation of the stereotype model. From this

standpoint it seems intuitive that the stereotype models are nested between strictly ordinal regression models and nominal multinomial logistic models. Put in other words: it is subject to model evaluation and performance if the response is treated ordinally or not.

Greenwood and Farewell (1988) discuss the modeling process based on a stereotype model following Anderson (1984) in the steps:

- Clarify the dimensionality of the model. This issue refers to the specific family of stereotype models introduced in Anderson (1984) and need not be deepened here. Ordinal models are of dimensionality one.
- Reduce the number of parameters in the model by testing distinguishability of categories. This step is inferential.
- Fit the stereotype model. If the estimates of the ϕ_j 's are ordered (see [2.4]), so is the response.

The final step according to Greenwood and Farewell (1988) sheds additional light on the question when a stereotype model should be considered and when a stringent ordinal regression model such as the cumulative link or continuation-ratio model appears appropriate. If the nature of the response is clearly ordered (naturally), categories are of intrinsic interest, and distinguishability is no issue, then a strictly ordinal model such as the POM is in place. If the nature of the response is in doubt, either because assessed or of unknown origin, a model like the stereotype one may be called for, since it is both parsimonious and allows to check the ordering through the $\hat{\phi}_j$.

Category indistinguishability has an interesting interpretation in cumulative link models, not provided by continuation-ratio models. Upon combining adjacent categories, the cutpoints will change, the model itself remains valid (Armstrong and Sloan 1989). Applications and simulations have shown however, that it is desirable not to combine categories, since it results in poorer inference. Holtbrügge and Schumacher (1991) found in a simulation study, that data generated under a stereotype model was fit better by McCullagh's proportional odds model. The ML estimates in the POM exhibited higher precision and smaller finite sample bias.

3.1.4. Choice of the Link Function

Cumulative link models were introduced in Chapter 2 without explicitly specifying the link function. Choosing the canonical link function is desirable for some reasons, stated in § 3.1.5. However, employing the canonical link is not a requirement. In principle any inverse cumulative distribution function will do. For categorical data the most frequently employed link functions are displayed in Table 3.1 together with the corresponding cumulative distributions.

Table 3.1
Frequently employed link functions and corresponding
cumulative distribution functions

Name of link	Link function $G^{-1}(t)$	Cdf $G(t)$	Name of cdf
logit	$\log\left(\frac{t}{1-t}\right)$	$\frac{\exp(t)}{1+\exp(t)} = \frac{1}{1+\exp(-t)}$	standard logistic
probit	$\Phi^{-1}(t)$	$\int_{-\infty}^t \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\}$	standard normal
log-log	$-\log\{-\log(t)\}$	$\exp\{-\exp(-t)\}$	
compl. log-log	$\log\{-\log(1-t)\}$	$1 - \exp\{-\exp(t)\}$	extreme-value (Gompertz)

McCullagh (1980) established, that all models of the form

$$\text{link}\left\{\sum_{k=1}^j \pi_k(\mathbf{x})\right\} = \alpha_j + \mathbf{x}'\beta$$

constitute a family. Common to all members is a proportionality property as discussed in 3.1.2. For the logit link we have seen that the model reduces to the proportional odds model. McCullagh (1980) shows, that for the complementary log-log link $G^{-1}(t) = \log\{-\log(1-t)\}$ one obtains the proportional hazards model, promoted by Cox (1972). Consequently all models of this form are qualitatively similar. Since the shapes of the link functions are usually very similar, the fits will be hardly distinguishable. However, the logit and probit link offer the advantage over the complementary log-log link, that the resulting models are invariant under a reversal of the categories.

Figure 3.3 displays some of the frequently employed link functions and Figure 3.4 the corresponding cumulative distribution functions.

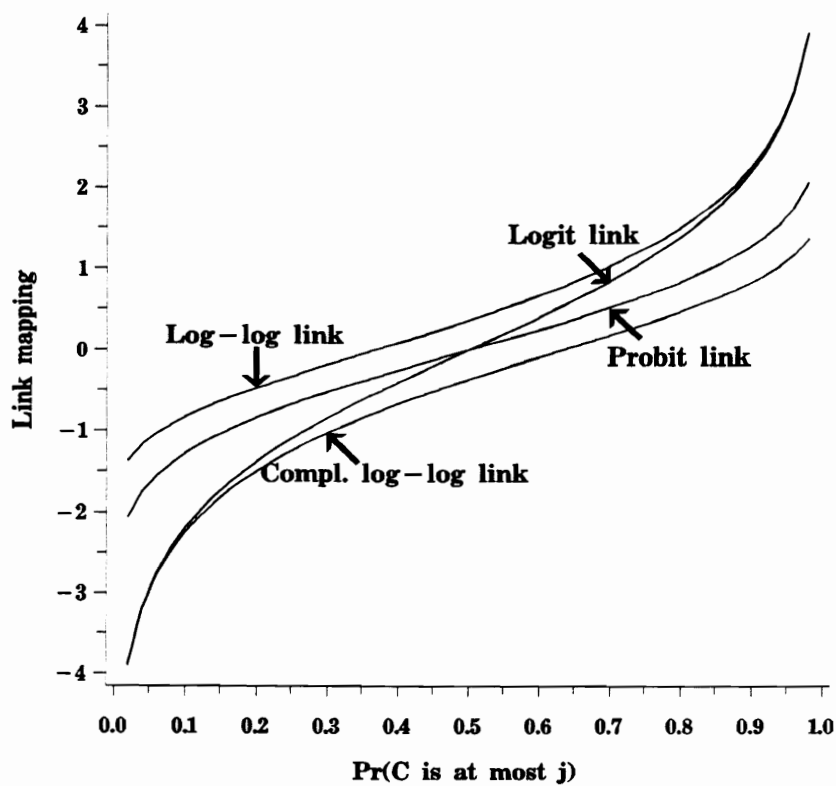


Figure 3.3. Shape of various link functions for categorical response data. See also Table 3.1 and Figure 3.4.

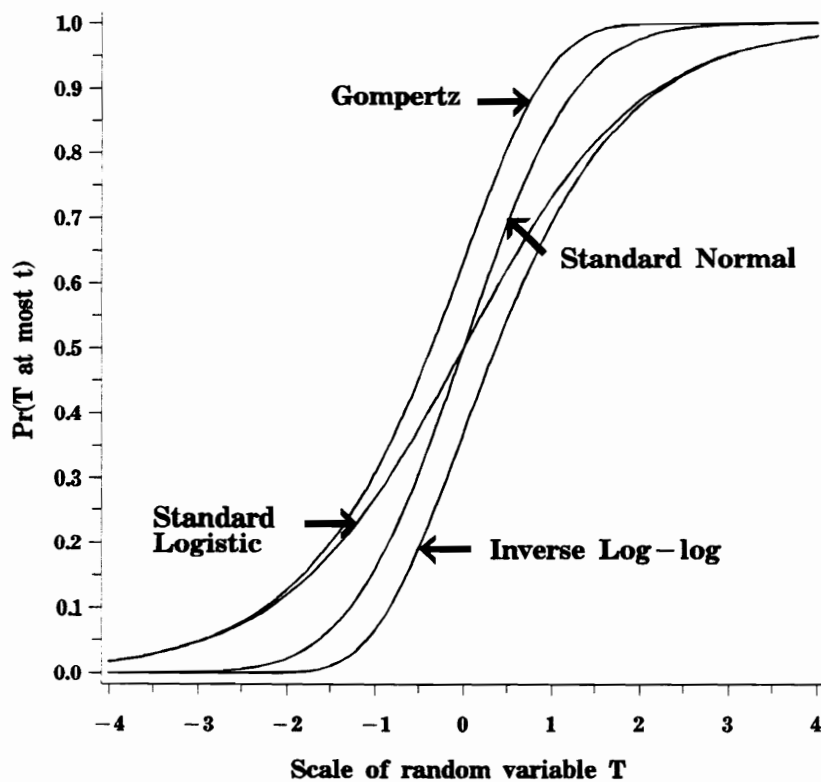


Figure 3.4. Cumulative distribution functions corresponding to the link functions in Table 3.1 and Figure 3.3.

The link functions are not too different in shape. Figure 3.3 sheds some light on the choice of the appropriate link function. The logit and probit link are both symmetric but the logistic distribution has more probability mass in the tails. The logit and complementary log-log link start out with the same mapping (Figure 3.3) but the latter function accumulates probability mass faster. The reverse relation holds for the log-log link and the probit link; while they end up with the same rate of probability accumulation, the probit link accelerates faster for small probabilities.

The asymmetric complementary log-log function assumes that the underlying continuous cumulative distribution function increases faster the higher a particular response on the latent scale but shows only moderate rate of increase for low responses. The log-log link reverses this relationship in that initial acceleration exceeds acceleration for high responses. It is thus comprehensible, why the complementary log-log link was employed in survival analysis, where it is assumed that the underlying distribution is of an exponential type (Agresti 1990). Agresti thus recommends the complementary log-log link when $\Pr(C \leq j|\mathbf{x})$ approaches unity faster than zero.

One advantage of the logit link is the straightforward interpretation of the parameters in terms of odds and odds ratios, not provided by any other link function in Table 3.1.

3.1.5. Fitting the POM

A point in favor of the POM in cross-sectional studies is the availability of fitting routines in widely distributed computer packages. The proportional odds model is the

default of SAS¹/PROC LOGISTIC which provides a maximum likelihood fit. Due to the multivariate nature of the response the POM is not a genuine generalized linear model but a multivariate GLM, requiring a composite link function. This section discusses how to fit the POM and introduces a general purpose algorithm. The relationship of this algorithm to Fisher scoring and Iteratively Reweighted Least Squares (IRLS) is also discussed.

The proportional odds model can be fitted easily with the following algorithm discussed in Schabenberger (1995a) and Schabenberger and Gregoire (1995). A more detailed discussion about the origin of this algorithm is deferred to a subsequent section.

Let $\mathbf{Y}_i$ be a vector of cumulative indicators for the i th response, i.e.

$$Y_{ik} = \begin{cases} 1 & \text{iff } C_i \leq k \\ 0 & \text{otherwise} \end{cases}$$

Note that $E(Y_{ik}) = \Pr(Y_{ik} = 1) = \Pr(C_i \leq k) = \mu_{ik}$. Let μ_i be the vector of expectations of $\mathbf{Y}_i$ corresponding to the first $J-1$ categories. Since $E(Y_{iJ}) = 1$, one category can be omitted, the last one is chosen for convenience. Put $\mathbf{D}_i = \partial \mu_i / \partial \theta$, where $\theta = [\alpha_1, \dots, \alpha_{J-1}, \beta']'$ for the $(J-1)*(J-1+p)$ matrix of first derivatives and $\mathbf{V}_i = \text{var}(\mathbf{Y}_i)$, where $\text{cov}(Y_{ik}, Y_{ik'}) = \mu_{ik}(1 - \mu_{ik'})$, $k \leq k'$. Note, that this is not the variance-covariance matrix of a regular multinomial vector but of a cumulative multinomial vector (McCullagh and Nelder 1989); see also [A.53]. The likelihood score function can be written as

¹SAS is a registered trademark of SAS Institute Inc., Cary, NC. USA.

$$U(\boldsymbol{\theta}, \mathbf{y}) = \sum_{i=1}^K \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \quad [3.6]$$

if it is assumed that the responses from different subjects $i = 1, \dots, K$ are uncorrelated.

A solution to [3.6] is obtained iteratively by updating

$$\hat{\boldsymbol{\theta}}_{u+1} = \hat{\boldsymbol{\theta}}_u + \left(\sum_{i=1}^K \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i \right)^{-1} \left(\sum_{i=1}^K \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i) \right) \quad [3.7]$$

where all matrices and vectors on the rhs of [3.7] are evaluated at the u th solution. The variance-covariance matrix of the estimates can be estimated at convergence as

$$\widehat{\text{var}}(\hat{\boldsymbol{\theta}}) = \left(\sum_{i=1}^K \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i \right)^{-1} \quad [3.8]$$

A *robust* estimator is also available and will be introduced in a subsequent section.

This algorithm has several desirable properties. (i) Only first derivatives of the inverse link function are required, which in general makes it possible to use cumulative distribution functions $G(\bullet)$ that do not exist in closed form, but their inverse can be evaluated numerically. For example SAS provides functions to evaluate the inverse of most continuous members of the exponential family. (ii) The variance-covariance matrix of the estimates is a by-product of the iterative procedure. (iii) The directional correction terms in [3.7] can be shown to be an optimal estimating functions in the sense of Godambe (1960) evaluated at the current solutions (see Chapter 3.2) and thus achieves fast convergence. (iv) The algorithm is multivariate Iteratively Reweighted Least Squares (IRLS) and is thus related to both Fisher-Scoring in the multivariate maximum likelihood framework and Generalized Estimating Equations in quasi-likelihood. (v) The structure of [3.6]-[3.8] is multivariate in nature and can be used as a

principal scheme in longitudinal data analysis after making the necessary adjustments (to be discussed later).

Points (i) - (iii) and (v) are rather practical in nature at this point of the discussion. However, (iv) is important to draw connections between maximum likelihood and quasi-likelihood in the sequel. Therefore, some additional illustrations follow. We will establish the equivalence between the algorithms for the general multivariate case here, section 3.2.4 will discuss the estimating function [3.6] more generally.

IRLS is the algorithm of choice for fitting generalized linear models (McCullagh and Nelder 1989), and with a canonical link in the exponential family it is equivalent to Fisher scoring (this is one of the reasons why we prefer canonical links). IRLS regresses an adjusted dependent variable onto the regressors $\mathbf{x}_1, \dots, \mathbf{x}_p$ linearly using weighted least squares. Both, the dependent variable and the weights are adjusted after each iteration of the mean parameters. To be specific let $G^{-1}(\mu_i) = \eta_i$, where $G^{-1}()$ is the link function of the GLM and expand $G^{-1}(\mathbf{y}_i)$ into a first-order Taylor series around μ_i :

$$\begin{aligned} G^{-1}(\mathbf{y}_i) &\doteq G^{-1}(\mu_i) + \frac{\partial G^{-1}(\mu_i)}{\partial \mu_i}(\mathbf{y}_i - \mu_i) \\ &\doteq \eta_i + \frac{\partial \eta_i}{\partial \mu_i}(\mathbf{y}_i - \mu_i) = \mathbf{z}_i \end{aligned}$$

$$\text{var}(\mathbf{Z}_i) = \left(\frac{\partial \eta_i}{\partial \mu_i} \right) \text{var}(\mathbf{Y}_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right)'$$

Regress $\mathbf{z}_i$ onto $\mathbf{X}_i$ with weight $\text{var}(\mathbf{Z}_i)^{-1}$ in the linear model $\mathbf{Z}_i = \mathbf{X}_i \boldsymbol{\theta} + \epsilon_i$. Since the weights need to be updated when $\boldsymbol{\theta}$ is updated, we use IRLS and can write the

correction term that is added to the current solution at any iterative step according to McCullagh and Nelder (1989) as

$$\delta = \left(\sum_i \mathbf{X}_i' \text{var}(\mathbf{Z}_i)^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_i \mathbf{X}_i' \text{var}(\mathbf{Z}_i)^{-1} (\mathbf{z}_i - \eta_i) \right)$$

which is the weighted partial coefficient vector. But

$$\begin{aligned} \sum_i \mathbf{X}_i' \text{var}(\mathbf{Z}_i)^{-1} \mathbf{X}_i &= \sum_i \mathbf{X}_i' \left(\frac{\partial \mu_i}{\partial \eta_i} \right)' \text{var}(\mathbf{Y}_i)^{-1} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \mathbf{X}_i \\ &= \sum_i \left(\frac{\partial \eta_i}{\partial \theta} \right)' \left(\frac{\partial \mu_i}{\partial \eta_i} \right)' \text{var}(\mathbf{Y}_i)^{-1} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \left(\frac{\partial \eta_i}{\partial \theta} \right) \\ &= \sum_i \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i \end{aligned}$$

Similarly,

$$\begin{aligned} \sum_i \mathbf{X}_i' \text{var}(\mathbf{Z}_i)^{-1} (\mathbf{z}_i - \eta_i) &= \sum_i \mathbf{D}_i' \mathbf{V}_i^{-1} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \left(\left(\frac{\partial \eta_i}{\partial \mu_i} \right) (\mathbf{y}_i - \mu_i) \right) \\ &= \sum_i \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{y}_i - \mu_i) \end{aligned}$$

and combining the last two results we finally get

$$\delta = \left(\sum_{i=1}^K \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1} \left(\sum_{i=1}^K \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{y}_i - \mu_i) \right)$$

which is the form of [3.7] and equivalence between the algorithm [3.6]-[3.8] and multivariate IRLS is established.

[3.7] is an iteratively updated coefficient vector. As with many other non-linear optimization methods one needs starting values to get the estimation process on the way. The structure of the cumulative logit model lends itself to a particularly simple

solution. Since

$$\text{logit}(\Pr(C \leq j | \mathbf{x})) = \alpha_j + \mathbf{x}'\beta$$

it seems logical to start iterations at $\beta_0 = \mathbf{0}$ and to put

$$\text{logit}(\Pr(C \leq j | \beta_0 = \mathbf{0})) = \alpha_j \forall j.$$

Substituting cumulative sample frequencies $\hat{\mu}_j$ for the cumulative probabilities one can obtain the starting values for the cut-off parameters as

$$\hat{\alpha}_{j_0} = \text{logit}(\hat{\mu}_j) \forall j \text{ where } \hat{\mu}_j = \frac{\# \text{ obs. responding } \leq j}{\text{total \# of obs.}}.$$

This procedure ensures stochastic ordering through ordering of the $\hat{\alpha}_{j_0}$ from the outset and has been found to perform well in practice, i.e. leads to satisfactory overall convergence behavior.

3.1.6. A Convenient Reparameterization

It is recommended for different reasons to use a reparameterized version of the POM at least during the fitting process. In non-linear optimization the gradient or local design matrix $\mathbf{D}_i$ typically exhibits a certain degree of collinearity caused by the non-linearity of the mean response. In the POM additional collinearity is caused by the cumulative nature of the cut-offs α_j . Any solution to [3.6] will comply to the constraints $\alpha_j \leq \alpha_{j+1}, \forall j$. However, since the cut-offs $\alpha_2, \dots, \alpha_{J-1}$ can not vary over the entire real line the estimates of the cut-off parameters are typically highly correlated. This may cause instabilities in the process of estimation.

Based on Fahrmeir and Tutz (1993) the cut-offs can be rewritten as

$$\alpha_1 = \psi_1$$

$$\alpha_j = \psi_1 + \sum_{k=2}^j \exp\{\psi_k\}, \quad j > 1.$$

Instead of the actual cut-offs α_j one now estimates the parameters ψ_j that vary over the entire real line without violating the stochastic order, since still $\alpha_j \leq \alpha_{j+1}, \forall j$. Starting values for the ψ_j are obtained easily from the starting values for the regular POM as

$$\hat{\psi}_{1_0} = \hat{\alpha}_{1_0}$$

$$\hat{\psi}_{j_0} = \log\left\{\hat{\alpha}_{j_0} - \hat{\alpha}_{(j-1)_0}\right\} \quad j = 2, \dots, J-1.$$

3.1.7. Predictions

Predictions in categorical models are different from those for models with continuous response, simply because the observed response is discrete, the fitted values however are real numbers on (0,1). Relating the fitted values back to the scale of the response is a classification problem.

To illustrate, consider a binary response, coded as a 0/1 indicator and a model for $\Pr(Y_i = 1)$. Using ordinary logistic regression, we want the predicted probability of response to be close to zero, when a subject responded $Y = 0$ and close to 1 otherwise.

However, there is an indifference zone, where we can not be sure, how to interpret the predicted probability from the regression model. If in the logistic model the predicted probability $\hat{p}(\mathbf{x}_i, \hat{\beta})$ is 0.8 say, one might be inclined to say, that the model predicts a 80% chance of a response from subject i . This is not quite satisfactory, since response is a dichotomous event. Several different approaches have been proposed to overcome this difficulty. The simplest assigns a predicted value 1 iff $\hat{p}(\mathbf{x}_i, \hat{\beta}) > 0.5$ and a response 0 otherwise. The arbitrary cut-off point 0.5 is of course questionable.

As the number of possible response categories becomes greater than 2 an additional difficulty arises, because the response is multivariate. That is for subject i we observe a single discrete response, $C = j$, say. An ordinal regression model for J categories will return J probabilities. To illustrate let $J = 5$ and $C = 2$. Assume the vector of predicted probabilities $\hat{\mathbf{p}}(\mathbf{x}_i, \hat{\theta}) = [\hat{p}(\mathbf{x}_i, \hat{\alpha}_1, \hat{\beta}), \dots, \hat{p}(\mathbf{x}_i, \hat{\alpha}_{j-1}, \hat{\beta})]$ in a proportional odds model is

$$[0.2, 0.4, 0.58, 0.80, 1.0]$$

from which we obtain the category probabilities:

$$[0.2, 0.2, 0.18, 0.22, 0.2].$$

The model thus predicts that an individual whose covariates are $\mathbf{x}_i$ is almost as likely to respond in any category. The slight difference between the fourth and the remaining categories will not alter our judgment. But contrasted with the discrete observed response in category 2 we ask ourselves, what is the actual predicted response category? That is, based on $\hat{\mathbf{p}}(\mathbf{x}_i, \hat{\theta})$ we have to identify one of J categories as the actual predicted one.

Three different methods are used to proceed. If an underlying continuous scale exists and is known, one can use class midpoints of the discretized scale and compute predicted values as weighted averages, where the weights are the elements of the prediction vector $\hat{\mathbf{p}}(\mathbf{x}_i, \hat{\theta})$. Schabenberger (1995a) has used this method to predict tree diameters at arbitrary relative heights where the diameters were obtained in classes forming the ordered categories of the latent diameter scale. The advantage of this procedure is that computing weighted averages over the class midpoints maps the categorical response back onto the real line where they have a natural interpretation.

Anderson and Philips (1981) also related the predicted values to the latent scale, whether it exists or not. The classification rule based on the latent model [3.1] is simply

$$\begin{aligned} -\mathbf{x}'_i\boldsymbol{\beta} &< \alpha_1 \Rightarrow C_i = 1 \\ \alpha_{j-1} &< -\mathbf{x}'_i\boldsymbol{\beta} < \alpha_j \Rightarrow C_i = j \\ -\mathbf{x}'_i\boldsymbol{\beta} &\geq \alpha_J \Rightarrow C_i = J \end{aligned} \tag{3.9}$$

The most natural way to obtain a predicted category without referring to the latent scale is to assign a prediction to that response category, whose predicted probability is a suprema, i.e.

$$\sup_{j=1, \dots, J} \left\{ \hat{p}_j(\mathbf{x}_i, \hat{\theta}) \right\} = \hat{p}_k(\mathbf{x}_i, \hat{\theta}) \Rightarrow C_i = k \tag{3.10}$$

It can be shown, that classification rule [3.10] maximizes the probability of correct assignment (Anderson and Philips 1981; Marshall and Olkin 1968) under the premise that misclassifications are of equal consequence. This is of course hardly ever the case. A misclassification in adjacent categories on an ordered scale will usually be not as severe as an incorrect assignment in a category further away from the true state of

nature. If unequal consequences of misclassifications are to be taken into account a decision theoretic approach is required that specifies consequences according to a loss function (Marshall and Olkin 1968). This not only complicates the analysis but more importantly introduces subjective elements in the process.

In the application Anderson and Philips (1981) consider, their procedure [3.9] was found to classify more satisfactorily than [3.10]. However, Campbell, Donner, and Webster (1991) have found in an extensive simulation study that Anderson and Philips method performed poorer than [3.10] and did not recommend its use. [3.9] seems to be less informative when the latent variable does not exist, since then the incidental parameters α_j are not to be interpreted as actual cut-off points on this scale. In absence of a loss function that specifies the consequences of erroneous decisions, [3.10] seems to be favorable.

The three procedures gear towards matching the observed and predicted measurement scales. If a great number of subjects or aggregates such as trees on a plot are available, one might consider the following, simple prediction rule.

For each observation obtain its vector of predicted multinomial probabilities. If the observed category is j , assign the probabilities to the j th column of a $J \times J$ classification table. This procedure in essence permits fractional predictions, for example one may observe 325 subjects in the first category, but 312.8 subjects are predicted in this category. If the number of subjects is large, this should not be of much concern.

The point of this illustration is, that judging ordinal regression model performance on grounds of predicted values will not only depend on the actual fitted quantities provided by the multivariate model, but also on how these quantities are used to connect them with the discrete response scale. A model can be judged inferior because a poor classification rule has been used, although it might yield a very good fit to the observed data. The most plausible way is to use multiple classification rules and compare their respective results.

3.1.8. Judging Model Performance with Classification Tables

All prediction methods outlined in 3.1.6. lead to a classification table in which observed and predicted values are cross-classified. For a four category response models with categories labeled {0, 1, 2, 3} such a table may look like Table 3.2.

Table 3.2.
Depiction of a cross-classification table for a four category response model

		<i>Observed</i>				<i>Category</i>
		0	1	2	3	Margin
<i>Pre- dicted</i>	0	n_{00}	n_{01}	n_{02}	n_{03}	$n_{0.}$
	1	n_{10}	n_{11}	n_{12}	n_{13}	$n_{1.}$
	2	n_{20}	n_{21}	n_{22}	n_{23}	$n_{2.}$
	3	n_{30}	n_{31}	n_{32}	n_{33}	$n_{3.}$
Margin		$n_{.0}$	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{..}$

In this table n_{ij} denotes the number of instances where we predicted category i , but observed category j . Alternatively, one can depict the table in terms of probabilities, p_{ij} rather than counts. Obviously, in order to judge a model's predictive capabilities as satisfactory, we expect large cell counts on the diagonal and smaller counts in off-diagonal cells. Since misclassifications become more severe as the row and column index are more and more apart, we also would expect a decreasing number of counts as we move away from the diagonal in any given row. For example, in the last row of the table, the largest proportion of the n_3 should be found in cell 33, representing the subjects that responded in category 3 and were predicted to do so. A misclassification in cell 32 will probably be not as severe as one in cell 30. It is here that the ordering of the response can be fully utilized, since it directly leads to a reasonable agreement and association pattern in the cross-classification table. On a nominal scale, rows and columns were interchangeable, leaving no clues about strength and direction of association.

Measures of agreement in square contingency tables have been employed for a long time. As an example we can cite tau measures, gamma measures, Cohen's kappa and many others. The disadvantage of such aggregate measures as described e.g. in Schabenberger (1993), is their attempt to compile a complex interaction structure in a single number. As a result, with increasing number of categories and table size, these measures become more and more dubious and uninteresting. Schabenberger (1993) thus advocates a hierarchical, log-linear modeling approach designed for square contingency tables, where row and column classification variables are identical. This approach is outlined here briefly.

Assume the cell counts of a contingency table and their sum are Poisson random variables (a multinomial sampling scheme leads to the same models). Under the assumption of independence of row and column classification variable, the expected frequency in any cell can be expressed as

$$m_{ij} = n_{..} p_{i.} p_{.j}$$

where $n_{..}$ is the total of all cell counts and $p_{i.}$, $p_{.j}$ are the marginal probabilities for predicting category i and observing category j , respectively.

Putting this relationship on a logarithmic scale yields

$$\ln\{m_{ij}\} = \ln n_{..} + \ln p_{i.} + \ln p_{.j}$$

or after reparametrizing

$$\ln\{m_{ij}\} = \mu + \lambda_i^p + \lambda_j^o \quad [3.11]$$

where μ is a grand mean

λ_i^p is the effect of the i th row (prediction in category i)

λ_j^o is the effect of the j th column (observation in category j)

Under Poisson sampling [3.11] is a generalized linear model with canonical link ([A.34], [A.35]). It can be fit easily to any table by maximum likelihood techniques as long as the marginal totals are non-zero (Agresti 1990).

The underlying relationship of [3.11], $m_{ij} = n_{..} p_{i.} p_{.j}$, says that the probability to observe a count in cell ij is governed by the marginal probabilities. [3.11] is therefore called the independence model.

The rationale of the hierarchical approach is as follows: it is unlikely that the independence model provides an adequate fit of the observed cell counts in most

situations. But it can be viewed as a basic model stripped of any agreement or association beyond chance with $(J - 1)(J - 1)$ degrees of freedom. The key towards understanding the structure of a cross-classification table is to discern something about the association and agreement structure which is part of the interaction terms omitted in [3.11]. Simply adding all interaction terms λ_{ij}^{po} does not help much for two related reasons: it depletes the degrees of freedom to zero and provides a perfect and thus uninteresting fit, and it does not structure the interactions in any kind. What we are looking for are simple interaction parametrizations that can describe both association and/or agreement parsimoniously. A first candidate is the linear-by-linear association model (Agresti 1988, 1990):

$$\ln\{m_{ij}\} = \mu + \lambda_i^p + \lambda_j^o + \beta u_i u_j \quad [3.12]$$

where u_i, u_j are the scores for the row and column classification variable. Since only β is an additional parameter, only one degree of freedom is lost beyond [3.11]. The association term $\beta u_i u_j$ suggests that the association increases/decreases monotonically with the scores, i.e. the distance of categories, but is symmetric with respect to row and column classification. [3.12] is consequently not geared to model agreement specifically, but association, one of the two sources of interactions. Models that focus on beyond chance agreement are the agreement-models that focus on diagonal cells and disagreement models, focusing on off-diagonal cells (Tanner and Young 1985).

The homogeneous agreement model assumes that the amount of beyond chance agreement in the diagonal cells is identical for all categories. A single parameter is used to add agreement to the diagonal cells beyond the independence model:

$$\ln\{m_{ij}\} = \mu + \lambda_i^p + \lambda_j^o + \delta_{ij} \quad [3.13]$$

where $\delta_{ij} = \delta$ if $i = j$, 0 otherwise. If however, one suspects that for example the first and third category are easier to predict than the second, different agreement parameters may be reasonable. This gives rise to a non-homogeneous agreement model with the same structure as the homogeneous agreement model, but now $\delta_{ij} = \delta_i$ if $i = j$, 0 otherwise.

At this point it seems natural to combine agreement and association parameters in a single model. Agresti (1988) calls

$$\ln\{m_{ij}\} = \mu + \lambda_i^p + \lambda_j^o + \beta u_i u_j + \delta_{ij} \quad \delta_{ij} = \delta \quad i = j$$

the agreement-plus-linear-by-linear association model and Agresti (1990) terms

$$\ln\{m_{ij}\} = \mu + \lambda_i^p + \lambda_j^o + \beta u_i u_j + \delta_{ij} \quad \delta_{ij} = \delta_i \quad i = j$$

the quasi-independence model for ordered categories.

Fitting these models in a nested sequence by maximum likelihood allows one to judge goodness of fit and significance of model terms via likelihood ratio theory. If a log-linear model is accepted to provide the appropriate fit, the association structure it implies can be viewed as a likely image of the process that generated the pattern of observed and predicted values. In the context of judging the predictive performance of an ordinal regression model it appears reasonable to start a sequence with the independence model to generate baseline likelihood ratio criteria, then add a homogenous or non-homogenous agreement pattern. Depending on the amount of misclassification, some monotonic association may be left over that could be picked up by the term $\beta u_i u_j$.

Fitting a log-linear model to a classification table that is not too large is not a difficult task and requires only a few seconds of computations. In case this is still viewed as an unnecessary complication of matters, one may resort to the simple agreement measures mentioned above. Some of the simpler ones are Cohen's kappa, the fraction correctly classified and the prediction success index. Cohen's kappa (Cohen 1960) measures the overall level of agreement beyond what is expected by chance. The difference of actual agreement and chance agreement is scaled by its maximum value to yield

$$\kappa = \frac{\sum_{i=1}^J p_{ii} - \sum_{i=1}^J p_{i.} p_{.i}}{1 - \sum_{i=1}^J p_{i.} p_{.i}} .$$

Schabenberger (1993) lists some deficiencies of this otherwise widely used statistic. Perhaps most importantly, while $0 < \kappa < 1$, the same does not hold for the sample version $\hat{\kappa}$. One can observe less agreement than expected by chance, yielding $\hat{\kappa} < 0$ and the upper bound of $\hat{\kappa}$ depends on the marginal totals of the classification table.

A very simple descriptive measure of agreement is the fraction correctly classified, i.e.

$$\text{FCC} = \sum_{i=1}^J \frac{n_{ii}}{n_{..}} .$$

Arabatzis (1990) discusses a similar measure, known as the prediction success index

$$\text{PSI} = \sum_{i=1}^J \left\{ \frac{n_{ii}}{n_{..}} - \frac{n_{i.} n_{.i}}{n_{..}^2} \right\} .$$

3.2. Estimation Function Theory

3.2.1. Competing Principles

The multitude of recently proposed statistical approaches to the analysis of correlated observations, especially for the case of categorical responses, which will be reviewed in § 4, has created a certain confusion among analysts about terminology as well as the theoretical underpinnings of some newly developed methods. Reviewing the literature on the subject, many techniques may appear only remotely connected to each other and unnecessarily distinct. As it turns out, a substantial number of the approaches in § 4 can be brought closer together by incorporating them into the body of Estimating Function (EF) theory that has been around since the early 1960's but not gained much recognition as a unifying approach to statistical estimation. This section provides a brief introduction that should aid to draw lines between alternative approaches as the discussion progresses.

Parameter estimation in statistical models is guided by principles such as the principle of least squares, the principle of maximum likelihood, Bayes principle, etc. These principles have to be separated from the techniques involved in estimation. A technique is the specific analytical tool employed to solve the problem posed by an estimation principle. For example, the maximum likelihood principle asks us to find

those parameter estimates that maximize the likelihood function. This can be achieved by taking analytical derivatives, numerical derivatives, by inspection, etc.

The most important estimation principle from a frequentist's standpoint are least squares (LS) and maximum likelihood (ML). Both appear in a variety of suits, such as non-linear least squares, weighted LS, generalized LS, extended LS, to name a few (see for example Beal and Sheiner (1988) for a comparison and review). ML can be employed as pure ML, conditional ML, restricted ML, profile ML, and so forth.

It seems that the myriad of estimation principles makes it difficult to choose the best in any given situation. It is well known that on one hand different principles are optimal under different conditions and on the other hand that they may lead to the same answer under different premises. Normal theory ML point estimates of mean parameters for example are equivalent to LS solutions.

Examples like this have often served to justify the reluctance to investigate competing principles more thoroughly, especially by applied analysts. As a consequence the standard methods are employed frequently in situations where they do not apply. This attitude is counterproductive: the equivalence of estimation principles under different settings should be recognized as a fortunate coincidence pointing towards a more general, unifying theory of statistical estimation, combining methods that seem remote and distant at the moment.

As pointed out before, any one principle is optimal in a specific situation with respect to some criterion. We thus have to think of optimality of estimation routines

conditional on the information provided by the experiment or conditional on assumptions made by the analyst and with respect to the specific criteria. For example unbiasedness of estimators may be considered of utmost importance and one would strive for estimators with highest precision in the subclass of unbiased estimators, thus defining the criteria. If knowledge of the distribution of the random quantities involved is complete or assumed we can perform UMVUE estimation based on the Rao-Blackwell and Lehman-Scheffé theorems (Bickel and Doksum 1977). If we do not feel confident to make distributional assumptions a semi-parametric estimator with highest precision in the class of unbiased estimators will be sought whose moments of order greater than three are unknown. Both approaches can lead to the same estimator, but may also differ substantially. The difference between them refers to the information provided, not the criteria. With the same information as UMVUE estimation another candidate emerges simply by restating the condition: since ML estimates in the exponential family are usually best asymptotic normal, ML would be the principle of choice if unbiasedness is understood in an asymptotic sense. For any finite sample however, MLE's are oftentimes biased. If the criterion is understood in the unbiasedness for any sample size sense, maximum likelihood will frequently fail to qualify.

To judge one estimator versus the other, i.e. an estimation principle versus another requires an understanding of the important properties the estimator has to possess as well as the assumptions one is willing to make.

3.2.2. Optimal Estimating Functions in the Sense of Godambe

In 1960 Godambe published a fundamental article about the optimality of regular maximum likelihood, from which originated the idea of an estimating function to motivate estimation problems and their solutions.

All our current estimation principles have in common that at some point they utilize functions of the data and the parameters to be estimated, e.g. the normal equations, likelihood equations, moment equations, etc.. Interest however is to find the statistic that satisfies the particular equation and has optimal properties in some sense and serves as an estimate of the parameter of interest. Not surprisingly, sufficient statistics have played a major role in classical parametric statistics and it is well known for example, that maximum likelihood estimates are functions of the minimal sufficient statistics, if they exist (Bickel and Doksum 1977).

Consider for example $X_i \sim iid N(\mu, \sigma^2)$, σ^2 known. The likelihood equation for μ is

$$\frac{1}{\sigma^2} \sum_i (x_i - \mu) = 0 \Rightarrow \sum_i (x_i - \mu) = 0$$

with solution $\hat{\mu} = \bar{x}$. The least squares normal equation is also

$$\sum_i (x_i - \mu) = 0 \Rightarrow \hat{\mu} = \bar{x}.$$

Why do LS and ML coincide in this example? It is not (only) because the MLE is a function of the minimal sufficient statistic for μ , $\sum_i x_i$, but because both principles lead

to the same estimating equation $\sum_i (x_i - \mu) = 0$. Godambe's fundamental idea was to shift emphasis from the statistic to the estimating equation. If optimality of an estimating function is defined, this approach has some important advantages. For example, it is possible that a minimal sufficient statistic does not exist, making it impossible to find the maximum likelihood estimate as above. The likelihood score function however, is a minimal sufficient partitioning function of the sample space (Godambe and Heyde 1987). Sticking with pure ML will prevent us from finding an estimator, whereas focusing on the estimating function itself does not stop us from doing so.

Least squares is known to be inconsistent and biased when the variances of the observations are functions of the mean parameters, for example $y_i = f_i(\alpha) + \epsilon_i$; $\text{var}(\epsilon_i) = \sigma^2 g(\alpha)$. The LS estimates will provide a poor solution, however an optimal estimating function exists which will lead to consistent estimates (Godambe and Kale 1991).

Conditional likelihood analysis requires elimination of nuisance parameters by conditioning on their sufficient statistics. This becomes more and more infeasible as the number of nuisance parameters increases and the number of elements of the sufficient statistic is large. In Chapter 2 the proposal has been discussed to use the profile likelihood after substituting MLE's for the nuisance parameters (Cox and Reid 1987; McCullagh and Nelder 1989). Godambe (1976) and Liang (1987) show the connection of this approach, which is only feasible under parameter orthogonality, with a more general estimation function based solution.

In the following discussion, some regularity conditions are required.

Let $p(\mathbf{x}, \theta)$ be a probability density where $X_i \in \mathfrak{P}, \theta \in \Theta$ is an unknown indexing parameter and $U(\mathbf{x}, \theta) = \partial \log p(\mathbf{x}, \theta) / \partial \theta$ is the score function. In accordance with Godambe (1960) we assume

- Θ is an open interval on the real line.
- first and second derivatives of the log likelihood with respect to θ exist.
- differentiation and integration with respect to $p(\mathbf{x}, \theta)$ and $\log\{p(\mathbf{x}, \theta)\}$ are exchangeable
- the variance of the score function is finite for all $\theta \in \Theta$.

These regularity conditions are known from theorems about the Cramér-Rao bound or the information inequality. (Bickel and Doksum 1977; Rao 1973). Since these conditions are met for densities in the exponential family ([A.1]-[A.11]), estimating function (EF) theory finds most applications there. Thus, as long as the response distribution belongs to this family, these conditions are taken for granted.

An unbiased estimating function (EF) in the class $\mathfrak{S}$ is defined as a function of the data and the parameter $g(\mathbf{x}, \theta)$ such that

$$E(g(\mathbf{x}, \theta)) = 0 \quad \forall \theta \in \Theta.$$

We require furthermore, that $g(\mathbf{x}, \theta)$ is informative with respect to θ , i.e. $\partial g(\mathbf{x}, \theta) / \partial \theta$ exists, $\text{var}(g(\mathbf{x}, \theta)) > 0 \quad \forall \theta \in \Theta$. Godambe (1960) defined a function $g(\mathbf{x}, \theta)$ that has these properties as an optimal estimation function (OEF) provided that

$$\frac{E(g(\mathbf{x}, \theta)^2)}{\left\{E\left(\frac{\partial g(\mathbf{x}, \theta)}{\partial \theta}\right)\right\}^2} \leq \frac{E(g^*(\mathbf{x}, \theta)^2)}{\left\{E\left(\frac{\partial g^*(\mathbf{x}, \theta)}{\partial \theta}\right)\right\}^2} \quad [3.14]$$

where $g^*(\mathbf{x}, \theta)$ is any other unbiased estimating function in $\mathfrak{S}$.

Godambe and Kale (1991) term $g_s(\mathbf{x}, \theta) = g(\mathbf{x}, \theta) / E(\partial g(\mathbf{x}, \theta) / \partial \theta)$ a standardized estimating function (cf. Godambe and Heyde 1987; Godambe and Thompson 1989; Schabenberger and Gregoire 1995). Condition [3.14] is thus equivalent to state that an OEF is the standardized EF with smallest variance among all unbiased EF's. This latter interpretation has intuitive appeal for several reasons.

(i) If $g_s(\mathbf{x}, \theta)$ is linear in the parameter, then $\text{var}(g_s(\mathbf{x}, \theta)) = \text{var}(\hat{\theta})$, where $\hat{\theta}$ is the estimate implied by the EF and obtained by solving $E(g_s(\mathbf{x}, \theta)) = 0$ for θ .

(ii) Assume we have an EF that is nonlinear in the elements of the vector parameter θ and want to solve iteratively by means of Fisher scoring. That is to solve $E(g_s(\mathbf{x}, \theta)) = 0$ one applies directional increments

$$\delta\theta = -E\left\{\frac{\partial g(\mathbf{x}, \theta)}{\partial \theta}\right\}^{-1} g(\mathbf{x}, \theta)$$

to current estimates. The iterative scheme can be written as

$$\hat{\theta}_{u+1} = \hat{\theta}_u + \delta\theta = \hat{\theta}_u - g_s(\mathbf{x}, \theta)|_{\hat{\theta}}.$$

The directional increment is a standardized EF evaluated at the current solution. Selecting the function $g_s(\mathbf{x}, \theta)$ with smallest variance among competing EF's leads to the iterative algorithm with smallest correction term on average and thus to fast

convergence. Additional motivations of the standardization can be found in Godambe and Kale (1991).

We can give a less formal argument how $g_s(\mathbf{x}, \theta)$ is derived than in Godambe (1960) and Godambe and Heyde (1987) but adhere to the latter publication as follows: Assume the likelihood function and thus the score function $U(\mathbf{x}, \theta) = \partial \log p(\mathbf{x}, \theta) / \partial \theta$ unknown. However, we are striving to find an EF that is in some sense close to the score function. Two measures of closeness come to mind:

- (i) Minimize $E(g(\mathbf{x}, \theta) - U(\mathbf{x}, \theta))^2$
- (ii) Maximize $\text{Corr}(g(\mathbf{x}, \theta), U(\mathbf{x}, \theta))$

(i) leads to

$$\min \{ E(g^2(\mathbf{x}, \theta)) + I_{\mathbf{x}}(\theta) - 2E(g(\mathbf{x}, \theta)U(\mathbf{x}, \theta)) \}$$

where $I_{\mathbf{x}}(\theta) = E(\partial \log p(\mathbf{x}, \theta) / \partial \theta)^2$ is Fisher information for θ in $\mathbf{X}$, a constant.

(ii) leads to

$$\max \left\{ \frac{E(g(\mathbf{x}, \theta)U(\mathbf{x}, \theta))}{[E(g^2(\mathbf{x}, \theta))I_{\mathbf{x}}(\theta)]^{1/2}} \right\}$$

Thus the important quantity in (i) and (ii) is the product expectation $E(g(\mathbf{x}, \theta)U(\mathbf{x}, \theta))$. By arguments similar to the proof of the Cramér-Rao bound and

using the regularity conditions mentioned above, we get

$$\begin{aligned}
 E(g(\mathbf{x}, \theta)U(\mathbf{x}, \theta)) &= \int g(\mathbf{x}, \theta) \frac{\partial \log p(\mathbf{x}, \theta)}{\partial \theta} p(\mathbf{x}, \theta) d\mathbf{x} \\
 &= \int g(\mathbf{x}, \theta) \frac{\partial p(\mathbf{x}, \theta)}{\partial \theta} d\mathbf{x} = \int \frac{\partial}{\partial \theta} g(\mathbf{x}, \theta) p(\mathbf{x}, \theta) d\mathbf{x} \\
 &= E\left(\frac{\partial g(\mathbf{x}, \theta)}{\partial \theta}\right)
 \end{aligned}$$

Satisfying (i) thus requires to make $E(g^2(\mathbf{x}, \theta))$ small and $E(\partial g(\mathbf{x}, \theta)/\partial \theta)$ large.

This also satisfies (ii). Combining both conditions leads to

$$\min \left\{ \frac{E(g^2(\mathbf{x}, \theta))}{\left\{ E \frac{\partial g(\mathbf{x}, \theta)}{\partial \theta} \right\}^2} \right\}$$

which by definition is $\text{var}(g_s(\mathbf{x}, \theta))$, since $g(\mathbf{x}, \theta)$ is an unbiased EF.

3.2.3. EF and Categorical Data

Why is the discussion of OEF of importance in the present context?

Categorical data exhibits features that renders some of the most important estimation principles and techniques infeasible. This is especially true for repeated measurements (Schabenberger and Gregoire 1995). Specific features of cross-sectional categorical data are nonlinearity due to the use of link functions and discrete error distributions. In contrast to continuous responses there is usually no question as to which distributional law the observations comply to. If responses are binary, the Bernoulli or Binomial distributions are appropriate, both special cases of the

Multinomial distributions for outcomes with more than two categories. For count data the Poisson framework usually applies.

For most of the important discrete distributions there exists a functional relationship between lower and higher-order moments. Most importantly, variances of the observations are functions of the mean parameters, making ordinary least squares analysis infeasible, since it lacks consistence.

Since the response distributions are known, maximum likelihood is the method of choice for categorical response data (Agresti 1990; McCullagh and Nelder 1989). For longitudinal data, specifying the likelihood requires complete knowledge of the dependence pattern over time. This information is usually not available or only emulated through rather stringent assumptions about higher-order moments (Zhao and Prentice 1990, 1991) which can make the analysis sensitive to misspecifications (Zeger *et al.* 1988). It is thus necessary to identify estimation routines that allow for nonlinearity in the parameters, functional mean-variance relationships but do not require specification of all moments of the response distribution.

The foregoing discussion makes it clear, that such routines will be neither ML nor LS.

Since Godambe (1960) we know that in the presence of knowledge of the joint distribution of the responses the likelihood score function is an optimal estimation function. Principles that relax the assumption about knowledge of all moments thus require evaluation of their relative efficiency relative to ML or in other words, are there EF that relax the parametric assumptions that are optimal in this subclass and if so, how much efficiency is lost by not knowing all features of the response distribution?

It is here that our notion of optimality with respect to specified criteria and conditional on information provided or assumptions made, comes into play. If specification of the full likelihood is impossible, we want to find EF's that are optimal conditional on knowing only the first two moments of a subject's response (the information provided, for example), where optimality is measured as variance of an unbiased estimating function (the criteria).

OEF theory helps us to answer the two important and related questions:

(i) If we do the optimal conditional on what we know or are willing to assume, do we sacrifice much by not knowing (or assuming) more about the estimation problem at hand?

(ii) Being unable to furnish information as in classical cross-sectional data analysis, what is optimal, given our willingness to make certain non-stringent assumptions?

3.2.4. A General Estimating Function

Proceeding with this line of thought we are naturally led to the following question: if we are able to establish optimality of an estimating function, how important is it to relate the implied estimate to a classical estimation principle? If we do not know which principle could have given rise to an estimate derived by estimating function theoretical considerations, there is certainly nothing lost, as long as we can establish optimality. The consequences of this spirit are quite extensive, since we would not attach a label to an estimate as an MLE, an LSE, an MOME and so forth, unless required for

identification. There simply would be only two kinds of estimates: those which are implied by optimal estimating functions and those which are not.

The acceptability of this approach certainly would depend on the existence of a narrow class of estimating functions that covers many of the classical analytical methods. Given the multitude of possible models and estimating principles this seems unlikely. However, for a very broad variety of cases, a generally optimal function exists (Schabenberger and Gregoire 1995).

The purpose of this subsection is to introduce this system of estimating equations, to show how they can be solved and reduced to familiar equations and implied estimates in many situations. Much of the development in Chapter 4 will be based on these equations.

To fix ideas, let a general location model be given as $\mathbf{Y} = f(\boldsymbol{\theta}) + \epsilon$ where $\mathbf{Y}$ is the vector of responses, $f(\boldsymbol{\theta})$ is any function relating the parameter vector $\boldsymbol{\theta}$ to the response and the error term ϵ makes the model observational. The only general assumption made, common to all models that follow, is $E(\epsilon) = \mathbf{0}$, i.e. the model is correct.

Consider the estimating function

$$U(\boldsymbol{\theta}, \mathbf{y}) = \left(\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)' \mathbf{V}^{-1}(\mathbf{y} - f(\boldsymbol{\theta})) = \mathbf{0} \quad [3.15]$$

where $\mathbf{V} = \text{var}(\mathbf{Y})$, which has been used before in section 3.1.4 when fitting the proportional odds model. [3.15] clearly is an unbiased estimating function, since $E(\epsilon) = \mathbf{0} \Rightarrow E(\mathbf{Y}) = f(\boldsymbol{\theta}) \Rightarrow E(U(\boldsymbol{\theta}, \mathbf{y})) = \mathbf{0}$. Since we do not specify $f(\boldsymbol{\theta})$ at the

outset, we also allow for non-transformable non-linearity in the parameters and aim at an iterative solution. Assume we have a best initial guess $\hat{\theta}_0$ of the parameter vector and apply a Newton-Raphson technique with Fisher scoring. This leads to iterative updates:

$$\hat{\theta}_{t+1} = \hat{\theta}_t + (\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})^{-1} \left(\mathbf{D}'\mathbf{V}^{-1}(\mathbf{y} - f(\hat{\theta}_t)) \right), \quad [3.16]$$

where $\mathbf{D} = \partial f(\theta)/\partial \theta$.

At convergence the asymptotic moments of the root of [3.15] are $\hat{\theta} \sim \left(\theta, (\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})^{-1} \right)$. Whenever [3.15] has a non-iterative solutions, asymptotics are not required. [3.16] is found from [3.15] by noticing that

$$-E \left(\frac{\partial U(\theta, \mathbf{y})}{\partial \theta} \right) = \text{var}(U(\theta, \mathbf{y})) = (\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})$$

Do [3.15] and the numeric technique [3.16] imply familiar results? A series of examples is provided:

1) Let $\mathbf{Y} = \mathbf{X}\beta + \epsilon$; $\epsilon \sim (\mathbf{0}, \sigma^2\mathbf{I})$. Then [3.15] becomes

$$f(\beta) = \mathbf{X}\beta, \quad \mathbf{D}'\mathbf{V}^{-1}(\mathbf{y} - f(\beta)) = \mathbf{X}'(\mathbf{y} - \mathbf{X}\beta)$$

the regular least squares normal equations. Similarly, if we assume an initial guess $\hat{\beta}_0$ we obtain from [3.16]

$$\begin{aligned} \hat{\beta} &= \hat{\beta}_0 + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \left(\mathbf{y} - \mathbf{X}\hat{\beta}_0 \right) \\ &= \hat{\beta}_0 + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \hat{\beta}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \end{aligned}$$

the usual least squares estimate under the above assumptions about the errors. Notice, that the initial guess drops out, since the root to [3.15] exists in closed form.

2) Let $\mathbf{Y} = \mathbf{X}\beta + \epsilon$; $\epsilon \sim (\mathbf{0}, \mathbf{V})$. [3.15] becomes

$$U(\beta, \mathbf{y}) = \mathbf{D}'\mathbf{V}^{-1}(\mathbf{y} - f(\beta)) = \mathbf{X}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta)$$

and from the initial guess $\hat{\beta}_0$ we arrive at

$$\begin{aligned}\hat{\beta} &= \hat{\beta}_0 + (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}_0) \\ &= \hat{\beta}_0 + (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} - \hat{\beta}_0 = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\end{aligned}$$

Again, the solution exists in closed form, does not depend on the starting values and is a well-known result, the generalized least squares estimate and the MLE if $\mathbf{Y} \sim N(\mathbf{X}\beta, \mathbf{V})$.

3) Let $\mathbf{Y} = f(\theta) + \epsilon$, $\epsilon \sim (\mathbf{0}, \sigma^2\mathbf{I})$ where now $f(\theta)$ is non-linear in the parameters. [3.15] becomes

$$U(\theta, \mathbf{y}) = \mathbf{D}'(\mathbf{y} - f(\theta))$$

and from the initial guess $\hat{\theta}_0$ we arrive at

$$\hat{\theta}_1 = \hat{\theta}_0 + (\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'(\mathbf{y} - f(\hat{\theta}_0))$$

which is an iterative Gauss-Newton algorithm for non-linear least squares.

The estimating function [3.15] is even more general. So far the response and the parameter function are linearly linked. This is not a requirement.

4) Consider a GLM with canonical link function $G^{-1}(\bullet)$ and response distribution in the exponential family of distributions (Bickel and Doksum 1977, [A.1]). Write the

model for a single observation as

$$Y_i = G(\mathbf{x}_i' \boldsymbol{\theta}) + \epsilon_i, \quad \epsilon_i \sim (0, g(E(Y_i)))$$

In section 3.1.4 we have seen that IRLS for $\boldsymbol{\theta}$ leads to a maximum likelihood solution if the link is canonical. But we have also established that [3.13] is equivalent to IRLS. In fact, we extended the equivalence to multivariate IRLS and a multivariate version of the estimating function [3.15]. Without repeating the algebraic steps outlined in section 3.1.4 we conclude that the solution to [3.15] obtained by [3.16] is a ML solution. Another proof of this fact is given after the next theorem.

It should be obvious that the important part of our algorithm is [3.15], not [3.16]. The Newton-Raphson scheme with Fisher scoring is just a vehicle to solve the estimation problem posted by [3.15]. An interesting question is, why does [3.15] reduce to so many well-known cases? It is not, because we use [3.16] to find a root, but because of the structure of [3.15]. Coincidence of our estimates with maximum likelihood estimates arises if [3.15] is the likelihood score equation in a particular setting. For example $\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$ is the likelihood score equation if $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$ and because of the structure of the Gaussian distribution coincides with the partial derivatives of a sum of squares. Godambe's (1960) result, that the likelihood score equation is optimal tells us, that [3.15] is the optimal estimating function in example 1). Alternatively we could utilize the Lehman-Scheffé theorem and because $\hat{\boldsymbol{\beta}}$ in 1) is a complete sufficient statistic, every linear function of $\hat{\boldsymbol{\beta}}$ is a UMVUE for its expectation. Thus $\hat{\boldsymbol{\beta}}$ is UMVUE for $\boldsymbol{\beta}$. We also could employ the Gauss-Markov Theorem without requiring $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$ but only $\epsilon \sim (0, \sigma^2 \mathbf{I})$ and state that the (unique) closed-form solution $\hat{\boldsymbol{\beta}}$ in 1) is minimally dispersed among all linear unbiased estimators of $\boldsymbol{\beta}$.

All of these results establish optimality of our estimating function and implied optimality of the estimate, but under different assumptions. When we compare examples 1)-4) we find that some estimates exist in closed form, others do not. To establish optimality of [3.15] in general, is thus no simple task, since we have to cover closed-form and iterative solutions simultaneously. One way out is to focus on the asymptotic behavior. This leads to a generalization of the Gauss-Markov theorem. It is weaker, since the result is asymptotic in general, but it reduces to finite sample results, whenever a root of [3.15] can be found in closed form. The theorem and its proof follows arguments in Chapter 9.5 in McCullagh and Nelder (1989).

Theorem: Let a vector of random variables be $\mathbf{Y}$ with $E(\mathbf{Y}) = \mu(\theta)$, $\text{var}(\mathbf{Y}) = \mathbf{V}$ and assume the second moment is finite for each component of $\mathbf{Y}$. In the class of unbiased estimating functions of the form $U(\theta, \mathbf{y}) = \mathbf{H}'(\mathbf{y} - E(\mathbf{Y}))$, linear functions of the root of $U(\theta, \mathbf{y}) = 0$ are minimally dispersed, if $\mathbf{H}' = \mathbf{D}'\mathbf{V}^{-1}$ where $\mathbf{D} = \frac{\partial \mu(\theta)}{\partial \theta}$.

Proof:

Let $U_2(\theta, \mathbf{y}) = \mathbf{H}'(\mathbf{y} - E(\mathbf{Y}))$ with root $\tilde{\theta}$ of $U_2(\theta, \mathbf{y}) = 0$ and $U_1(\theta, \mathbf{y}) = \mathbf{D}'\mathbf{V}^{-1}(\mathbf{y} - E(\mathbf{Y}))$ with root $\hat{\theta}$ of $U_1(\theta, \mathbf{y}) = 0$. According to the theorem $\text{var}(\mathbf{a}'\tilde{\theta}) \geq \text{var}(\mathbf{a}'\hat{\theta})$ which implies $\text{var}(\tilde{\theta}) \succeq \text{var}(\hat{\theta})$, where $\succeq$ is to be read as “matrix ordering” and $\mathbf{A} = \mathbf{P} - \mathbf{Q} \succeq 0 \Rightarrow \mathbf{A}$ is positive semidefinite.

If $\tilde{\theta}$ is the root of $U_2(\theta, \mathbf{y}) = 0$ we can approximate

$$\begin{aligned} U_2(\tilde{\theta}, \mathbf{y}) &\doteq U_2(\theta, \mathbf{y}) + \mathbf{H}'\mathbf{D}(\tilde{\theta} - \theta) \Rightarrow (\theta - \tilde{\theta}) \doteq (\mathbf{H}'\mathbf{D})^{-1}U_2(\theta, \mathbf{y}) \\ &\Rightarrow \text{var}(\tilde{\theta}) \doteq (\mathbf{H}'\mathbf{D})^{-1}(\mathbf{H}'\mathbf{V}\mathbf{H})(\mathbf{D}'\mathbf{H})^{-1}. \end{aligned}$$

If $\mathbf{H}' = \mathbf{D}'\mathbf{V}^{-1}$ we are led to $\text{var}(\hat{\boldsymbol{\theta}}) = (\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})^{-1}$. The theorem follows if

$$\text{var}(\tilde{\boldsymbol{\theta}}) \succeq \text{var}(\hat{\boldsymbol{\theta}}) \Rightarrow \text{var}(\tilde{\boldsymbol{\theta}}) - \text{var}(\hat{\boldsymbol{\theta}}) \succeq \mathbf{0} \Rightarrow \text{var}(\hat{\boldsymbol{\theta}})^{-1} - \text{var}(\tilde{\boldsymbol{\theta}})^{-1} \succeq \mathbf{0}.$$

Substituting we obtain

$$\begin{aligned} \mathbf{D}'\mathbf{V}^{-1}\mathbf{D} - \mathbf{D}'\mathbf{H}(\mathbf{H}'\mathbf{V}\mathbf{H})^{-1}\mathbf{H}'\mathbf{D} &\succeq \mathbf{0} \\ \Rightarrow \mathbf{D}'\left(\mathbf{V}^{-1} - \mathbf{H}'(\mathbf{H}'\mathbf{V}\mathbf{H})^{-1}\mathbf{H}'\right)\mathbf{D} &\succeq \mathbf{0} \end{aligned} \quad [3.17]$$

But the last expression is just a residual covariance matrix, where the linear effect of $\mathbf{H}'\mathbf{y}$ was removed from $\mathbf{D}'\mathbf{V}^{-1}\mathbf{y}$, hence is at least non-negative definite (c.f. McCullagh and Nelder 1989). This completes the proof.

If $\mathbf{D}$ and $\mathbf{H}$ are of full rank, we can establish positive definiteness of [3.17] by putting $\mathbf{X}_1 = \mathbf{T}\mathbf{D}$, $\mathbf{X}_2 = \mathbf{T}'^{-1}\mathbf{H}$, where $\mathbf{V}^{-1} = \mathbf{T}'\mathbf{T}$. Now partition $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2]$ as the design matrix of a linear model and compute the partition of the inverse

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

where $\mathbf{A}_{11} = \left[\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1 \right]^{-1}$ according to the rules of partitioning the inverse of symmetric matrices. Since $\mathbf{X}$ is of full rank, $(\mathbf{X}'\mathbf{X}) \succ \mathbf{0} \Rightarrow \mathbf{A}_{11} \succ \mathbf{0}$, where $\succ$ denotes positive definiteness. Substituting for $\mathbf{X}_1$ and $\mathbf{X}_2$ in $\mathbf{A}_{11}$ we get

$$\mathbf{A}_{11} = \left[\mathbf{D}'\mathbf{V}^{-1}\mathbf{D} - \mathbf{D}'\mathbf{H}(\mathbf{H}'\mathbf{V}\mathbf{H})^{-1}\mathbf{H}'\mathbf{D} \right]^{-1}$$

which establishes positive definiteness of [3.17].

Corollary: Since $U(\theta, \mathbf{y}) = \mathbf{D}'\mathbf{V}^{-1}(\mathbf{y} - E(\mathbf{Y}))$ is the optimal estimating function in the class of unbiased estimating functions that are linear in the observables, it is invariant under linear transforms of $\mathbf{Y}$.

The proof is simple.

Proof: Let $\mathbf{Z} = \mathbf{G}\mathbf{Y}$ where $\mathbf{G}$ is a conformable full rank matrix of real values.

Then $\text{var}(\mathbf{Z}) = \mathbf{G}\mathbf{V}\mathbf{G}'$, $E(\mathbf{Z}) = \mathbf{G}E(\mathbf{Y})$, $\partial E(\mathbf{Z})/\partial\theta = \mathbf{G}\mathbf{D}$. The optimal estimating function for θ is

$$U(\theta, \mathbf{z}) = \mathbf{D}'\mathbf{G}'\mathbf{G}'^{-1}\mathbf{V}^{-1}\mathbf{G}^{-1}\mathbf{G}(\mathbf{y} - E(\mathbf{Y})) = \mathbf{D}'\mathbf{V}^{-1}(\mathbf{y} - E(\mathbf{Y})).$$

Consequently, the basic estimating function is not altered by linear, nonsingular transformations and thus implies the same estimate.

The theorem above states optimality in terms of variance of the root of the estimating equation $U(\theta, \mathbf{y}) = 0$. The theory of estimating function tells us however, that focus should be on the function itself and not on the statistics implied by the function. The reason is (i) conceptual and (ii), that properties of the function not necessarily are inherited by the estimators. A typical example is unbiasedness. Although we are only concerned with unbiased estimating functions, the estimators obtained are oftentimes biased. In order to be complete, we have to check the dispersion of the estimating function. This requires standardization of the function itself. Godambe and Heyde (1987) have worked out necessary conditions for an estimating function of a vector parameter to be optimal in the sense that it maximizes the Hotelling vector correlation between EF and likelihood score function (Hotelling 1936). These are natural extensions of the optimality criteria discussed in 3.2.2. See also Bhapkar (1991).

To be specific, if $U(\theta, \mathbf{y})$ is an unbiased EF linear in the observables $\mathbf{Y}$, maximizing the closeness to the score function requires to maximize

$$E\left(\frac{\partial U(\theta, \mathbf{y})}{\partial \theta}\right)' E(U(\theta, \mathbf{y})U(\theta, \mathbf{y})') E\left(\frac{\partial U(\theta, \mathbf{y})}{\partial \theta}\right)$$

Among two competing EF's $U_1(\theta, \mathbf{y})$ and $U_2(\theta, \mathbf{y})$, $U_1(\theta, \mathbf{y})$ is preferable if

$$E\left(\frac{\partial U_1}{\partial \theta}\right)' E(U_1 U_1') E\left(\frac{\partial U_1}{\partial \theta}\right) - E\left(\frac{\partial U_2}{\partial \theta}\right)' E(U_2 U_2') E\left(\frac{\partial U_2}{\partial \theta}\right) \succeq 0 \quad [3.18]$$

where the arguments have been suppressed for brevity. If [3.18] holds for all other estimating functions in the class $\mathfrak{S}$ of unbiased linear EF's against $U_1(\theta, \mathbf{y})$, then $U_1(\theta, \mathbf{y})$ is the optimal estimating function. Using the notation of the theorem we can generally compare

$$U_1(\theta, \mathbf{y}) = \mathbf{D}'\mathbf{V}^{-1}(\mathbf{y} - E(\mathbf{Y}))$$

$$U_2(\theta, \mathbf{y}) = \mathbf{H}'(\mathbf{y} - E(\mathbf{Y}))$$

But since $\partial U_2(\theta, \mathbf{y})/\partial \theta = -\mathbf{H}'\mathbf{D}$, $E(U_2 U_2') = \mathbf{H}'\mathbf{V}\mathbf{H}$, $\partial U_1(\theta, \mathbf{y})/\partial \theta = -\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}$, $E(U_1 U_1') = \mathbf{D}'\mathbf{V}^{-1}\mathbf{D}$, we can substitute in [3.18] and obtain:

$$\begin{aligned} \mathbf{D}'\mathbf{V}^{-1}\mathbf{D}(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})^{-1} - \mathbf{D}'\mathbf{H}(\mathbf{H}'\mathbf{V}\mathbf{H})^{-1}\mathbf{H}'\mathbf{D} &\succeq 0 \\ \Rightarrow \mathbf{D}'\mathbf{V}^{-1}\mathbf{D} - \mathbf{D}'\mathbf{H}(\mathbf{H}'\mathbf{V}\mathbf{H})^{-1}\mathbf{H}'\mathbf{D} &\succeq 0 \end{aligned}$$

But this is just [3.17]. Interestingly, the optimality of the estimating function in terms of its dispersion carries over immediately to optimality of the estimates.

Reminding ourselves of the general result of Godambe (1960), that the maximum likelihood score equation is an optimal estimating function, when the joint response distribution is completely known, we can give yet another, and final justification for [3.15] in the case of generalized linear models with distributions in the exponential

family. To be specific, assume we observe a single random variable Y and let its density or mass function be

$$p(y, \eta) = \exp\{\eta y + d_0(\eta) + S(y)\}_{I(Y)} \quad [3.19]$$

where the indicator set $\{I(Y)\} = \{1 : \forall y \ni p(y, \eta) \neq 0\}$ does not depend on η . [3.19] is an exponential family distribution in natural form where $\eta = c(\theta)$ is the canonical link. Since $E(e^{tY}) = \exp\{d_0(\eta) - d_0(t + \eta)\}$ is the moment generating function of Y , we find the simple relationships

$$\begin{aligned} E(Y) &= -d'_0(\eta) = \mu \\ \text{var}(Y) &= -d''_0(\eta) = \partial E(Y) / \partial \eta \end{aligned} \quad [3.20]$$

If we model the canonical parameter through a linear model we are led to $\eta = \mathbf{x}'\beta$. By the chain rule we can write out the likelihood score function for β_j

$$\frac{\partial \log p(y, \eta)}{\partial \beta_j} = \frac{\partial \log p(y, \eta)}{\partial \eta} \frac{\partial \eta}{\partial \mu} \frac{\partial \mu}{\partial \beta_j} \quad [3.21]$$

But from [3.20] we see that $\partial \eta / \partial \mu = 1 / \text{var}(Y)$. Using this and the fact that

$$\frac{\partial \log p(y, \eta)}{\partial \eta} = y + d'_0(\eta) = y - E(Y)$$

the score function for β_j is

$$\frac{\partial \log p(y, \eta)}{\partial \beta_j} = \frac{\partial \mu}{\partial \beta_j} (y - E(Y)) \text{var}(Y)^{-1}.$$

This is exactly the form of [3.15] for a single observation. Extending this development for a vector of observations and a vector parameter leads to [3.15].

It should come as no surprise at this point, that [3.15] serves also as a means for estimation in longitudinal settings. The main task will be to accommodate the

multivariate nature of the longitudinal response and the special covariance structure caused by serial dependence. Since fitting the proportional odds model has been shown to fall under the scope of [3.15] it can be expected that enhancing a univariate longitudinal estimating function for ordinal response is feasible. This is the topic of Chapter 4.

3.3. Issues in Longitudinal Studies

The discussion of longitudinal data (LD) in Chapter 2 has only skimmed the surface of some pertinent issues in repeated measurements analysis. Addressing these issues is vital to any longitudinal study and the topic of this section.

3.3.1. Longitudinal Data Structures

Longitudinal data can be thought of as repeated observations nested within subjects. Different longitudinal data structures arise depending on the number of nested units much like in a designed experiment. In addition to the number of units their spacing in time is of essential importance. The distance between measurements in itself is information that must be utilized in the modeling process, since the dependency of repeated observations usually is at least a function of the time elapsed between measurement occasions.

The most elementary longitudinal data (LD) structure is that of balanced and equally spaced observations. Balancing refers to the number of observations per

subject, spacing addresses the distance in time between observations. Table 3.3 schematically depicts this elementary structure for a study with 5 subjects and 4 repeated measurements. The body of the table contains x for every value observed.

Table 3.3
Elementary longitudinal design, balanced and regularly spaced,
no missing data

	Subject				
Time	1	2	3	4	5
1	x	x	x	x	x
2	x	x	x	x	x
3	x	x	x	x	x
4	x	x	x	x	x

This structure is rather artificial, especially in observational studies. Usually some form of aberration occurs. Subjects may not be available or recorded at some time points, creating what is known as missing data. Observation times may be irregularly spaced, although the data are balanced within subjects. Laird (1988) defines the standard LD scheme therefore as a balanced, unequally spaced design as in Table 3.4.

Table 3.4
Standard LD design (Laird 1988), balanced but irregularly spaced

	Subject				
Time	1	2	3	4	5
1	x	x		x	x
2	x			x	x
3		x	x	x	x
4			x	x	x
5		x	x		x
6	x		x	x	
7	x	x			
8	x	x	x		

As can easily be seen, the notions of balancedness, missingness and spacing are entangled (Gregoire, Schabenberger and Barrett 1995). The data in Table 3.4 could be described as balanced irregularly spaced without missing data, or as balanced with 8 repeated measurements and three missing data points per subject. If the time scale is considered continuous, the LD structure automatically becomes irregularly spaced. LD analysis methods that allow for implicit unequal spacing thus assume that measurements arise from a continuous time process (Jones 1993) and automatically take missing observations into account (Gregoire, Schabenberger and Barrett 1995). However, this is only part of the truth, as the reason for the missingness or the irregular spacing needs to be clarified. If the measurements that gave rise to Table 3.4 were made at discrete time points (1,...,8), assuming a continuous time process allows for specific analytical tools, but does not explain why three measurements per subject were missing.

Missingness of observations and its impact on the analysis requires special attention. The literature on the subject is quite extensive and no comprehensive review is intended at this point. Only the basic notions shall be discussed following Laird (1988,1991).

Rubin (1976) introduced a hierarchy of missing data processes (see also Dempster *et al.* 1977; Little and Rubin 1987). These contributors coined the terms missing at random (MAR) and missing completely at random (MCAR). Likelihood theory suggests the description of missing data processes as MCAR, ignorable, and non-ignorable (Laird 1988, 1991; Gibbons *et al.* 1993). The following draws a connection between these ideas.

Assume a balanced design where each subject $i = 1, \dots, K$ is observed at times $t = 1, \dots, T$, the response vector and covariate matrix for each subject are denoted $\mathbf{Y}_i$ and $\mathbf{X}_i$. The distribution of the response depends on some indexing parameter vector θ . Let $\mathbf{R}_i$ be a $(T \times 1)$ indicator vector with typical element $R_{it} = 1$ iff Y_{it} was observed, 0 otherwise. The non-response process may depend on covariates in $\mathbf{Z}$ and parameters in ϕ .

If $\mathbf{Y}_{io}$ are the components of $\mathbf{Y}_i$ for which $R_{it} = 1$ and $\mathbf{Y}_{im}$ the components for which $R_{it} = 0$, partition $\mathbf{Y}_i = \begin{bmatrix} \mathbf{Y}_{io}' & \mathbf{Y}_{im}' \end{bmatrix}'$. The observed data then follows the distribution

$$f(\mathbf{Y}_{io}, \mathbf{R}_i | \mathbf{X}_i, \mathbf{Z}, \theta, \phi) = \int f(\mathbf{Y}_i | \mathbf{X}_i, \theta) f(\mathbf{R}_i | \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}, \phi) d\mathbf{Y}_{im} \quad [3.22]$$

where integration is over the space of the missing responses. That is, the density of the observed data and the response mechanism can be obtained by averaging over the sample space of the unobserved, missing responses.

If $f(\mathbf{R}_i|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}, \phi)$ does not depend on $\mathbf{Y}_i$, the process is said to be missing completely at random in the sense of Rubin (1976) $\Rightarrow f(\mathbf{R}_i|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}, \phi) = f(\mathbf{R}_i|\mathbf{X}_i, \mathbf{Z}, \phi)$. [3.22] then reduces to

$$f(\mathbf{Y}_{io}|\mathbf{X}_i, \theta) = \int f(\mathbf{Y}_i|\mathbf{X}_i, \theta) d\mathbf{Y}_{im} \quad [3.23]$$

that is, the marginal distribution is the one that would be used if we would completely ignore the unobserved responses.

If $f(\mathbf{R}_i|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}, \phi)$ depends on the observed responses but not on what could have been observed were there no missing data, i.e. $\mathbf{Y}_{im}$, the process is said to be MAR in the sense of Rubin (1976) (see also Louis 1988). In this case, the non-response mechanism can be ignored in the estimation process. From [3.22] it follows with $f(\mathbf{R}_i|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}, \phi) = f(\mathbf{R}_i|\mathbf{Y}_{io}, \mathbf{X}_i, \mathbf{Z}, \phi)$ that

$$f(\mathbf{Y}_{io}, \mathbf{R}_i|\mathbf{X}_i, \mathbf{Z}, \theta, \phi) = f(\mathbf{Y}_{io}|\mathbf{X}_i, \theta) f(\mathbf{R}_i|\mathbf{Y}_{io}, \mathbf{X}_i, \mathbf{Z}, \phi) \quad [3.24]$$

If inference about θ is the primary focus, $f(\mathbf{R}_i|\mathbf{Y}_{io}, \mathbf{X}_i, \mathbf{Z}, \phi)$ can be ignored in evaluating the likelihood. Therefore, the MAR condition was termed ignorable with respect to likelihood inference (Laird 1988). As Laird points out, the sampling distribution of the observed data does depend on $\mathbf{R}_i$, consequently MLE's will depend on the non-response model. Only evaluating the likelihood can ignore $f(\mathbf{R}_i|\mathbf{Y}_{io}, \mathbf{X}_i, \mathbf{Z}, \phi)$. Caution should therefore be exercised when using the MLE precision estimates if the response mechanism is MAR.

If finally $f(\mathbf{R}_i|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}, \phi)$ depends on the unobserved responses, the mechanism is said to be non-ignorable. $f(\mathbf{R}_i|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}, \phi)$ must be specified otherwise likelihood methods are invalid (Conaway 1993; Wu and Carroll 1988; Wu and Bailey 1988).

These considerations definitely will have an impact on the analysis. It is naturally desirable that the missing data process is MCAR. However, it might be difficult to distinguish in some applications between MAR and MCAR processes. A case where MCAR can be achieved by design, is when the time points of measurements are predetermined and rotating. Here measurement occasions vary between subjects or aggregates of subjects (e.g. trees on a plot) but in a systematic fashion unrelated to the response of interest. Table 3.5 depicts such a LD structure known as a rotating panel design with 6 subjects, 3 measurement waves and equally spaced observations

Table 3.5
Rotating panel design, regularly spaced.

	Subject					
Time	1	2	3	4	5	6
1	x	x				
2			x	x		
3	x	x			x	x
4			x	x		
5	x	x			x	x
6			x	x		
7	x	x			x	x

The missingness of observations for $i > 2$ at the first occasion is due to the fact that these were selected for measurement at later occasions not depending on their

response but for other reasons (randomization for example). Rotating panel designs are a convenient way to collect LD in studies with many subjects and time consuming measurements per subject. The varying initial states can be accounted for by introducing a covariate that captures distance between measurement date and a baseline date and the initial measurement by keeping track of the absolute time points for each measurement.

3.3.2. Drop-Outs and Censoring

Longitudinal data usually appears as a sequence of events in the life history of a subject. To infer about change or aging processes it is helpful if the start and finishing time of the longitudinal study are in some sense meaningful points within the lifetime. This, however, is unfortunately almost never the case. In the rotating panel design for example, the initial recordings for each subject occur at different time points, hence no equality of initial conditions can be assumed. If the starting date of a longitudinal study is chosen arbitrary, this is a case of left censoring, if the finishing date is arbitrary, one speaks of right censoring. Oftentimes the data are collected in a sequence that is neither started nor finished by a specific meaningful event. Some subjects, however, will respond at some occasion in an absorbing state, such as death. Their sequence of observations has a meaningful end point while the sequences for other subjects are right censored.

It is easily seen that many observational longitudinal studies are both left and right censored. This may cause bias of model parameter estimates (Uncles 1988).

Unfortunately, there is no unified theory in sight that could help the analyst out. Uncles (1988) suggests to do what is

“pragmatic, simple and consistent with prior knowledge”.

One reason for observations being missing are drop-outs in the study. Such drop-outs have to be considered carefully. For example a study in which patients health status is monitored will necessarily lead to drop-outs once a patient is completely healthy and does not return to the examinations, dies, or moves out of the area. Sample attrition due to drop-outs can ruin longitudinal data for unbiased inference if the process that leads to the drop-out is non-ignorable. For example if patients move out of the area, because of their illness, the panel gradually loses its representativeness for the population under scrutiny and results will be biased. If patients die for reasons unrelated to the study, i.e. patients of all health status are equally likely to die from such causes, this leads to sample attrition only. Aside from a decrease in the number of repeated measurements the analysis is not in danger, since the probability of non-response is the same for all individuals.

Ignorability due to death therefore depends on the context. If the response scale includes the state of death in some form, for example as one of a set of ordered categories, the death of a subject creates a valid response in an absorbing state. The series of repeated observations for a subject simply ends with a meaningful event. The process can be interpreted as ignorable, since the probability of non-response at occasion $(t + 1)$ after death has occurred at t is 1 and can fully be explained by knowing the previous response, which was observed. It does not depend on what could

have been observed were there a response after death occurred, since nothing is observable.

In forestry a peculiar situation arises, if for example trees are the subject in the longitudinal study and dead trees are not removed. Then one could repeatedly observe a response in the absorbing state provided the assignment of covariates to such trees is meaningful. If this is appropriate depends on the context, i.e. the population of interest. In managed stands dead trees are likely to be removed. Until they are physically removed from the stand they can be kept in the data base, at removal their sequence of measurements terminates. Especially when the number of repeated measurements is small compared to the number of subjects and the terminating event is a rare event, a pragmatic approach is to terminate the sequence of repeated measures for the tree once it dies and to treat the process of future observations as ignorable.

The important point of this discussion is that there has to be a fine distinction between the techniques employed for the analysis of a particular longitudinal data structure and the techniques necessary to adequately model nature's cause of the structure. Using an estimation technique that allows for continuous time processes but requires missingness completely at random is not adequate when a data structure as in Table 3.3. actually arises from a discrete time process in which data are missing at random. The capability of a certain estimation technique to handle the particular spacing and balancedness of longitudinal data does not imply correctness of inferential procedures.

3.3.3. Research Questions in LD Analysis

One of the main advantages of longitudinal data is to carry implicit information about growth and change. Cross-sectional data are not impervious to change since they can include information about the life history of an individual. For example a cross-sectional study with covariates age of trees and number of years since the last thinning operation locates the data within the lifetime of a tree/stand and includes information about the past. Since cross-sectional data are much simpler to analyze than longitudinal it is important to point out the specific merits of LD. That is, questions we are able to answer with longitudinal data, that can not be addressed with cross-sectional data.

3.3.3.1 Age and cohort effects

A major difference between longitudinal and cross-sectional data lies in their respective representation of change. The information in longitudinal data can be partitioned into changes over time within subjects (ageing effects) and differences in the baseline levels across subjects (Diggle *et al.* 1994). The latter are usually described as cohort effects.

Since cross-sectional data do not possess a nested structure it is usually not possible to separate age from cohort effects in cross-sectional data. Figure 3.5. displays basal area by age measurements on selected plots of Eastern white pine (*Pinus strobus* L.). The complete data was used by Gregoire *et al.* (1995).

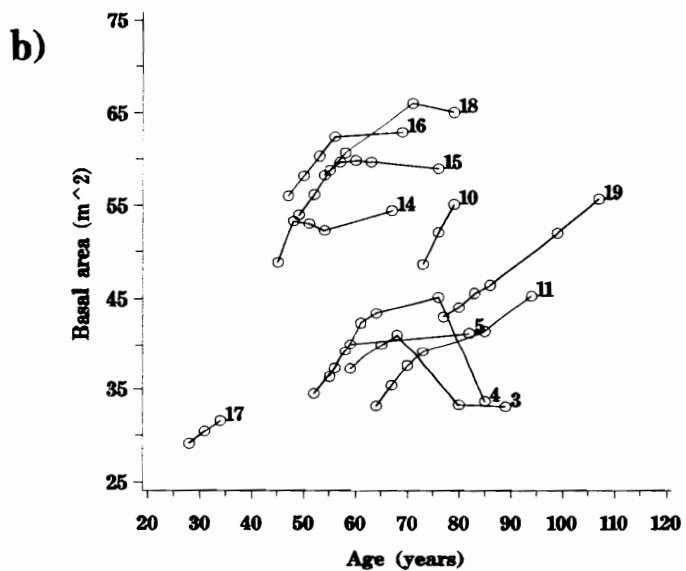
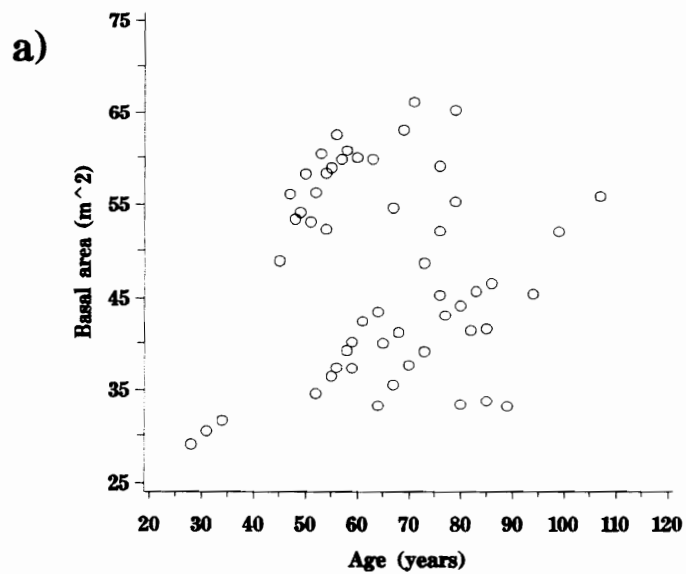


Figure 3.5. Basal areas on selected plots by age.
a): Cross-sectional; b): Longitudinal data structure.
Plot numbers printed in panel b).

The a) panel in Figure 3.5. displays a series of observations as they would appear in a cross-sectional study. It clearly appears that plot basal area increases with age, but also that after age 50 considerable variation exists. The data are actually part of an unbalanced, irregularly spaced longitudinal study, in which the plots are revisited up to ten times (for a complete description of the experiment refer to Gregoire, Schabenberger and Barrett 1995). Panel b) displays the data as longitudinal by connecting the repeated measurements for each plot. Differences in the baseline levels are now clearly distinguishable from the growth effects within each plot. It also becomes clear that the high variability of the measurements in panel a) is caused by two different cohorts. Plots 3, 4, 5, 11, and 19 seem to share a similar baseline, as do plots 14, 15, 16, and 18. If one were to derive a model for Eastern white pine basal area growth from panel a) one would have to assume that cohort effects are absent, i.e. all baselines are the same. Perhaps more importantly, each observation would have to be compared against the overall growth trend in the data. In b) the series of observations allows one to compare the basal area values for each plot against the mean of the particular plot. Each subject serves as its own control (Zeger and Liang 1992). The apparent difference in the baselines of the two cohorts would probably prompt the inclusion of other adjusting covariates or random coefficients.

How dangerous it can be to use cross-sectional data to infer about ageing or growth is apparent when we look at forecasting. Only the average growth trend in the population can be used to predict the basal area at any given age. It is obvious that this will lead to poor predictions for the individual plots. The cross-sectional analysis assumes no cohort effects, equivalent to assuming that plots currently 40 years old will

in 30 years have the basal area of those plots currently 70 years old. Existing differences caused by inhomogeneity of behavior across subjects are ignored. In other words: only under the strong assumption of absence of cohort effects are cross-sectional data really representing growth and not only a snap-shot image of a population at measurement time.

Observational data will never permit us to establish causality. The only way this can be achieved is by assuring that there are no other possible sources for the observed differences than those controlled in the experiment. For example the relation between diameter growth and ‘heaviness’ of thinning regime can only be described as causal when all other possible sources of influence are controlled. From an observational cross-sectional study we thus can not infer that a relationship between thinning regime and diameter growth exists,

- because increasing thinning causes increased diameter growth
- because increased diameter growth results in shorter thinning intervals,
- whether they are related in a cause-feedback relationship,
- or whether some unmeasured factor(s) caused the observed changes.

Cross-sectional relationships can be induced or exaggerated by spurious associations, due to omitted covariates. One may find effects as important in cross-sectional data that were induced because characteristics are omitted that are correlated over time. Longitudinal data allows to account for these effects and are in this sense closer to causality.

3.3.3.2. PA, SS, and conditional research problems

Chapter 2.2.1 introduced the distinction between population-averaged (PA), subject-specific (SS) and conditional models (RC, CC). These models give rise to different parameter interpretations and for categorical responses are usually not transitive. It is therefore important to identify the research problem(s) of interest and to select the adequate model type thereafter. This section discusses the possible research problems and the model types that address them.

Although LD incorporates information about change, the estimation and prediction of individual change may not be the prime interest, rather, one wishes to investigate how covariates relate to the average behavior of subjects in the study. This can focus on subjects individually or on the population as a whole. The latter is the concept of PA models in which the coefficients have an interpretation in terms of average behavior in the population and are in this sense equivalent to models for cross-sectional data. If coefficients vary with time points - provided they are discrete and constant across subjects - one can also focus on occasion-specific population averaged behavior (Ware *et al.* 1988). In the context of a disease process, the following questions typically give rise to a PA model:

- How do covariates such as DBH, tree height, social position etc. affect the disease propensities for a tree species?
- Do disease probabilities change between occasions in the population of interest?
- Does the probability of death induced by the diseases depend on tree age?

As can be seen from the second question, PA models can incorporate questions about change, for example if the coefficients are allowed to vary with measurement occasions. But since they have a PA interpretation like cross-sectional coefficients, confounding of age and cohort effects are possible.

As Zeger and Liang (1988) point out, this can be detected by including for example age at initial occasion and change in age from baseline as covariates in the model. Consequently, a PA model for longitudinal data allows separation of these effects without sacrificing the natural interpretation of the coefficients in terms of ‘mean effects’. In addition, PA models in the sense discussed here are genuine LD tools, accounting for the serial correlation among measurements in order to validate inference about model parameters. Contrary to cross-sectional models, PA models estimate the time-varying effects unbiasedly even if time-independent covariates are omitted, and are thus robust against model misspecification. A PA model furthermore does not require assumptions about heterogeneity of parameters across subjects, a typical feature of SS models and as Zeger *et al.* (1988) put it,

“PA parameters are in this sense one step closer to the data than SS parameters”.

Some authors believe that PA models do not fully utilize the information provided by longitudinal data (see for example Ware *et al.* 1988). This is correct as far as individual changes and individual behavior is concerned. Other model types, however, do not provide a marginal, population-averaged inference which may be of interest in some applications. The relative robustness of PA models and the comparatively weak assumptions are certainly important points in their favor.

Analysts familiar with cross-sectional analytical methods typically appreciate the straightforward interpretation of the model coefficients. The PA model captures changes in the marginal response distribution with changes in the covariates. Ware *et al.* (1988) refer to them as net changes to contrast them with changes in a conditional model.

The above mentioned research questions can easily be rephrased in a subject-specific context. For example

- How do covariates such as DBH, tree height, social position etc. affect the disease propensity of a specific tree?
- Do disease propensities for a tree change between occasions?

SS models are of great interest in studies of growth and development where we are not so much interested how the population as a whole develops, but how the individuals behave. Section 2.2.1.2.2 alluded to the fact that a SS interpretation is achieved by allowing model parameters to vary across subjects according to some distribution. While this offers additional flexibility, it requires the specification of a distribution and makes the model more sensitive to misspecifications. While for a PA model misspecifications refer only to the mean structure, SS models need to specify both the error structure and the mean structure correctly, in order to retain consistency.

To illustrate the differences in PA and SS interpretation of coefficients assume we have a generalized linear mixed model

$$\eta_i = G^{-1}(E(\mathbf{Y}_i|\mathbf{b}_i)) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$$

where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are $(n_i \times p)$ and $(n_i \times q)$ matrices of covariates, $\mathbf{Y}_i$ is the vector of

responses for subject i and $\mathbf{b}_i \sim (0, \mathbf{D})$. The conditional expectations are considered for a fixed value of the random effects $\mathbf{b}_i$ and since these are indexed by subject, η_i has a subject-specific interpretation. If the link function $G^{-1}()$ is the identity function a simplification arises. Let in this case the observational model be given by

$$\begin{aligned}\mathbf{Y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \\ E(\boldsymbol{\epsilon}_i) &= \mathbf{0}; \text{var}(\boldsymbol{\epsilon}_i) = \mathbf{R}_i; \text{cov}(\boldsymbol{\epsilon}_i, \mathbf{b}_i) = \mathbf{0}\end{aligned}$$

It is easy to see that

$$\begin{aligned}E(\mathbf{Y}_i|\mathbf{b}_i) &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i \\ \text{var}(\mathbf{Y}_i|\mathbf{b}_i) &= \mathbf{R}_i\end{aligned}$$

and the marginal moments are

$$\begin{aligned}E(\mathbf{Y}_i) &= \mathbf{X}_i\boldsymbol{\beta} \\ \text{var}(\mathbf{Y}_i) &= \mathbf{R}_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' .\end{aligned}$$

The marginal expectation $\mathbf{X}_i\boldsymbol{\beta}$ is the same that would have been used in a PA model. Thus in the special case of the linear mixed model the fixed effects $\boldsymbol{\beta}$ have a PA interpretation and the subject-specific $\mathbf{b}_i$'s can be interpreted as offsets that modify the population-averaged trend to accommodate a subject's behavior. The random coefficients affect the marginal distribution not in its location but in its dispersion. Note, that normality was not a requirement. The results follow from the linearity of the expectation operator and the uncorrelatedness of $\boldsymbol{\epsilon}_i$ and $\mathbf{b}_i$ only.

With a link other than the identity function the expectation operation does not go through. To illustrate let

$$E(\mathbf{Y}_i|\mathbf{b}_i) = G(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i); \mathbf{b}_i \sim F(0, \mathbf{D})$$

To find the marginal expectation we integrate over the distribution of the random effects:

$$E(\mathbf{Y}_i) = E(E(\mathbf{Y}_i | \mathbf{b}_i)) = \int G(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i) dF\mathbf{b}_i$$

It is not hard to imagine that this expectation will not equal $G(\mathbf{X}_i\boldsymbol{\beta})$ as is the case for the identity link. This will only be achieved if $\mathbf{b}_i = \mathbf{0} \forall i \Leftrightarrow \mathbf{D} = \mathbf{0}$, a trivial case, since this is the PA model. For some link functions the integral will not even exist in closed form. As (Zeger *et al.* 1988) put it:

“the link that transforms $[E(Y_{it} | \mathbf{b}_i)]$ into a linear function of $\mathbf{x}_{it}$ does not also do the same for $[E(Y_{it})]$ ”.

In contrast to the linear mixed model the marginal expectation will depend on the random effects variance-covariance matrix $\mathbf{D}$. Zeger *et al.* (1988) give approximations for the marginal mean for some of the standard link functions (see also Breslow and Clayton 1993). For example, based on the Gaussian approximation to the logistic distribution one finds for the logit link:

$$\text{logit}\{E(Y_{it})\} \doteq |c^2 \mathbf{D} \mathbf{z}_{it} \mathbf{z}_{it}' + \mathbf{I}|^{-q/2} \mathbf{x}_{it}' \boldsymbol{\beta}, \quad c = \frac{13\sqrt{3}}{15\pi}$$

There is attenuation in the PA parameter estimates, they are closer to zero than the subject-specific estimates (Neuhaus and Jewell 1990; Neuhaus *et al.* 1991).

It is important here to reiterate the notion of a marginal model that is used frequently but ambiguously in the literature. All PA models are marginal models since they focus on the marginal response distribution. This does not imply that all estimation methods commencing from marginal distributions lead to PA parameter estimates. If

one proceeds in the above generalized linear mixed model by ML estimation from the marginal distribution, the mean will still depend on the random effects through their dispersion matrix $\mathbf{D}$. The parameters are subject-specific.

Consequently, we will call a model marginal if it is derived from marginal moments, but the parameters population-averaged if the marginal expectation does not depend on subject-wise random terms and subject-specific otherwise. Thus marginal maximum likelihood in a generalized mixed model leads to SS parameter interpretation, both for β and $\mathbf{b}_i$. The corollary is a much more difficult, yet sometimes unclear interpretation of the coefficients. It should be recalled that coefficient interpretation plays a much more important role for categorical responses than for their continuous counterparts.

One further, worthwhile distinction between PA and SS models is that PA models account for serial correlation but treat their estimation as a separate problem (see Chapter 4), SS models incorporate estimation of error components in the algorithm that yields the fixed effects estimates. The same is true for conditional models, the third and last model-type for LD to be discussed here. Serial correlation is also accounted for explicitly in the estimation routine, although using a different vehicle, i.e. information about the subject history. Research questions that give rise to a conditional model are for example:

- How does social position affect the probability of a tree to be infected given that it was healthy at the previous occasion?
- What set of covariates adequately describes the probability of a tree to stay healthy over more than three occasions?

- Does the age effect differ between those trees that make a healthy-healthy and a healthy-infected transition?

Prior information about the response process is essential. Conditional models lead to the narrowest interpretation of regression coefficients, pertaining to groups of subjects that share the same condition, for example health at the previous occasion. Implementing Rosner's model (see Chapter 2.2.1.2.3) Neuhaus and Jewell (1990) demonstrated that the coefficients in the conditional model are further attenuated compared to the population-averaged model.

If covariate effects are not of prime interest in the study these considerations may be regarded of minor importance. But one should have a fairly clear understanding to what quantities the estimates refer.

Traditionally however, the interpretation of coefficients and covariate effects is rendered more important for nonlinear models than for linear ones. This is especially true for categorical models. Link functions like the logit allow interpretation of gradients in terms of changes in odds or log odds, familiar quantities for categorical data analysts. It is thus tempting to interpret coefficients despite problems such as multicollinearity.

It should further be emphasized that what has been described as nontransitivity requires the analyst to have tools handy for the different types of research questions. To fit a marginal, population-averaged model if interest is in the transition probabilities between response states because it appears to be simpler than a conditional model will

not do the job, since conditional coefficients can not be recovered from the PA parameters.

3.3.4. How to Model Correlations

PA and SS models for longitudinal data can be viewed as regression type models with a complex error structure. A location model is augmented by error terms that provide an appropriate description for the marginal variance-covariance pattern. Since the errors across subjects are typically uncorrelated, it is sufficient to focus on an individuals pattern only. While the actual correlations among measurements may differ from subject to subject, the correlation structure will remain the same for all experimental units. In a population averaged longitudinal model there are no random quantities. Thus, a model like

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \quad [3.25]$$

must carry all information about the correlations in $\text{var}(\boldsymbol{\epsilon}_i) = \mathbf{R}_i$. Not only their magnitude, but also the pattern is important in order to make efficient inference. In the process of fitting the model the variance of the error terms is parameterized in terms of usually a small number of parameters, $\mathbf{R}_i(\boldsymbol{\xi})$, say. $\boldsymbol{\xi}$ is estimated across the entire sample, pooling information from all subjects and thus a population parameter vector. Table 3.6. lists some of the more frequently entertained structures for $\mathbf{R}_i(\boldsymbol{\xi})$.

Table 3.6.
Structures for the within-subject error matrix for 4 repeated measurements.

Name of Structure	R_i matrix for a subject with 4 measurements	Parameters in ξ
Simple	$\begin{bmatrix} \xi & 0 & 0 & 0 \\ 0 & \xi & 0 & 0 \\ 0 & 0 & \xi & 0 \\ 0 & 0 & 0 & \xi \end{bmatrix}$	$\xi = [\xi]$
Compound Symmetry	$\begin{bmatrix} \xi + \xi_1 & \xi_1 & \xi_1 & \xi_1 \\ \xi_1 & \xi + \xi_1 & \xi_1 & \xi_1 \\ \xi_1 & \xi_1 & \xi + \xi_1 & \xi_1 \\ \xi_1 & \xi_1 & \xi_1 & \xi + \xi_1 \end{bmatrix}$	$\xi = [\xi, \xi_1]$
Completely Unstructured	$\begin{bmatrix} \xi_1 & \xi_{12} & \xi_{13} & \xi_{14} \\ \xi_{12} & \xi_2 & \xi_{23} & \xi_{24} \\ \xi_{13} & \xi_{23} & \xi_3 & \xi_{34} \\ \xi_{14} & \xi_{24} & \xi_{34} & \xi_4 \end{bmatrix}$	$\xi = [\xi_1, \xi_2, \xi_3, \xi_4, \xi_{12}, \dots, \xi_{34}]$
Autoregressive first order	$\xi \begin{bmatrix} 1 & \xi_1 & \xi_1^2 & \xi_1^3 \\ \xi_1 & 1 & \xi_1 & \xi_1^2 \\ \xi_1^2 & \xi_1 & 1 & \xi_1 \\ \xi_1^3 & \xi_1^2 & \xi_1 & 1 \end{bmatrix}$	$\xi = [\xi, \xi_1], \xi_1 \text{ is a correlation}$
Exponential	$\xi \begin{bmatrix} 1 & \exp\{-\xi_1 d_{12}\} & \exp\{-\xi_1 d_{13}\} & \exp\{-\xi_1 d_{14}\} \\ \exp\{-\xi_1 d_{12}\} & 1 & \exp\{-\xi_1 d_{23}\} & \exp\{-\xi_1 d_{24}\} \\ \exp\{-\xi_1 d_{13}\} & \exp\{-\xi_1 d_{23}\} & 1 & \exp\{-\xi_1 d_{34}\} \\ \exp\{-\xi_1 d_{14}\} & \exp\{-\xi_1 d_{24}\} & \exp\{-\xi_1 d_{34}\} & 1 \end{bmatrix}$	$\xi = [\xi, \xi_1],$ $d_{jj'}$ a distance measure between occasion j and j' .

If there are random effects or coefficients included, the errors are decomposed into within and between subject sources, such as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad [3.26]$$

where $\text{var}(\boldsymbol{\epsilon}_i) = \mathbf{R}_i$ is the within-subject covariance matrix, $\text{var}(\mathbf{b}_i) = \mathbf{D}$. Although the $\mathbf{b}_i$'s are intended to capture subject-to-subject heterogeneity, [3.26] allows one to model *within* subject correlations, since the marginal variance-covariance matrix is a function of $\mathbf{D}$, and $\mathbf{Z}_i$: $\text{var}(\mathbf{Y}_i) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i$. A convenient choice is to assume that, given $\mathbf{b}_i$, there is local independence and hence $\text{var}(\boldsymbol{\epsilon}_i) = \sigma^2\mathbf{I}$. A question of interest is which covariates to select for the fixed effects part, $\mathbf{X}_i$, of the model and which are part of the random effects design matrix $\mathbf{Z}_i$. In theory, only those coefficients can be part of the error structure in [3.26], which can be conceived as outcomes of a random selection from a hypothetical superpopulation of coefficients. This actually requires a sample model for the coefficients (Longford 1993). Typically, this is not how the columns of $\mathbf{Z}_i$ are selected in practice. One researcher's fixed effect will be another one's random coefficient (Louis 1988). Some authors recommend that the $\mathbf{Z}_i$ matrix should be a subset of the $\mathbf{X}_i$ data matrix, otherwise it would be unreasonable to assume $E(\mathbf{b}_i) = 0$ (c.f. Longford 1993). A quite different approach to mixed model analysis presumes that all covariates can be either part of $\mathbf{X}_i$ or $\mathbf{Z}_i$ and the one combination is selected that leads to best fit statistics.

In other applications the correlation structure is of more interest and it will be important that $\text{var}(\mathbf{Y}_i) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i$ provides a reasonable reflection of the actual, unknown correlation pattern. Typically correlations are not assumed equal among the multiple measurements of a subject. It is oftentimes quite reasonable to assume that measurements are higher correlated, the closer they are in some sense. Closeness can

be measured as distance in time, space, or by other means. In order to achieve this type of correlation structure, it is required that $\mathbf{Z}_i$ consists of time-dependent covariates, maybe even polynomials of measurement times since a baseline date. The columns of $\mathbf{Z}_i$ then must provide a meta-measure of correlations. In forestry, age, height growth, diameter growth, height of the bole above ground, etc. can serve as such meta-meters.

It is generally not possible to achieve marginal variance-covariance patterns such as the ones in Table 3.6. when random coefficients are in the model. Although the within-subject error matrix $\mathbf{R}_i$ may be selected according to Table 3.6., the term $\mathbf{Z}_i \mathbf{D} \mathbf{Z}_i'$ in $\text{var}(\mathbf{Y}_i)$ will disrupt this structure. The use of random coefficients for time-varying covariates is limiting ones ability to obtain a specific marginal structure, but are a convenient way to induce correlations of different magnitude within subject's although $\mathbf{b}_i$'s account for subject-to-subject variation.

The motivation for choosing random effects or coefficients over the direct modeling of the variance-covariance pattern in $\mathbf{R}_i$ is often found in a quite different reason. Although the $\mathbf{b}_i$'s are random and not parameters in the typical sense, best linear unbiased predictors (BLUPs), $\hat{\mathbf{b}}_i$, are available to predict their values after the model has been fit. This allows one to individualize the regression model and trace each subject's trajectory, rather than the population average. The term $\mathbf{Z}_i \hat{\mathbf{b}}_i$ serves as a modifier of the average $\mathbf{X}_i \hat{\boldsymbol{\beta}}$ to account for the difference between the i th subject's trend and the population's trend.

Chapter 4

Statistical Inference for Longitudinal Ordinal Data

4.1. Notation

In order to improve readability of this section, some notation will be introduced at this point. In the sequel subscripts are dropped where possible without ambiguity and additional notation introduced if warranted.

Indices and subscripts

$i = 1, \dots, K$	number of subjects (trees)
$t = 1, \dots, n_i$	number of time points at which subject i was measured
$j = 1, \dots, J$	the ordered response categories
$u = 1, \dots$	iteration counter

Responses, covariates and parameters

all vectors are understood as column vectors, row vectors are indicated by transposing them.

C_{it}	ordered response of subject i at time t
Y_{ijt}	indicator for the response of tree i in category j at time t
$Z_{itt'}$	some form of association response between responses at times t and t' ; e. g. $Z_{itt'} = Y_{it}Y_{it'}'$.
p_{ijj}	probability of transition from state j' to state j for subject i
$\mathbf{Y}_{it}$	indicator vector of responses for subject i at time t
$\mathbf{Y}_i$	indicator response vector for subject i at time t obtained by stacking: $\mathbf{Y}_i = [\mathbf{Y}'_{i1}, \dots, \mathbf{Y}'_{in_i}]'$
μ_i	vector of expected values of $\mathbf{Y}_i$
β	a fixed effect parameter vector of length p
$\mathbf{b}_i$	a vector of random coefficients of length q for the i th subject
α	vector of incidental cut-points $\alpha = [\alpha_1, \dots, \alpha_{J-1}]'$
ξ	vector of correlation parameters
θ	general (unspecified) parameter vector
η	linear predictor with typical element η_{ijt}
$\mathbf{x}_{it}$	a $(p \times 1)$ column vector of covariates for tree i at time t
$\mathbf{X}_i$	matrix with t th row given by $\mathbf{x}'_{it}$; usually a covariate matrix for fixed effects

Estimating and score functions

$L(\theta, \mathbf{y})$	likelihood function
$l(\theta, \mathbf{y})$	log-likelihood function
$U^*(\theta, \mathbf{y})$	likelihood score function
$U(\theta, \mathbf{y})$	quasi-likelihood score function
$g(\theta, \mathbf{y})$	general estimating function for θ
$g_s(\theta, \mathbf{y})$	standardized estimating function for θ
$G(t)$	a cumulative distribution function, evaluated at t
$G^{-1}(t)$	inverse of $G(t)$ used as a link function
$\phi(t)$	the standard normal density evaluated at t
$\Phi(t)$	the cumulative standard normal distribution at t
$\psi(t)$	the standard logistic density evaluated at t

$\Psi(t)$ the standard logistic cdf evaluated at t

Matrices

$\mathbf{P} = \text{Diag}(p_i)$ a diagonal matrix $\mathbf{P}$ with typical element p_i

$\mathbf{P} = \text{Diag}(\mathbf{P}_i)$ a block-diagonal matrix with blocks $\mathbf{P}_i$

$[\mathbf{D}]_{ij}$ the typical element of matrix $\mathbf{D}$ in i th row and j th column

$\mathbf{D} = \partial \mathbf{p} / \partial \boldsymbol{\beta}$ $(n \times p)$ Matrix of first derivatives of the elements in $(n \times 1)$ vector $\mathbf{p}$ with respect to the elements in the $(p \times 1)$ vector $\boldsymbol{\beta}$. That is,
 $[\mathbf{D}]_{ij} = \partial p_i / \partial \beta_j$

$\mathbf{P}^{-1}$ regular inverse of $\mathbf{P}$

$\mathbf{P}^-$ a g-inverse of $\mathbf{P}$

$\mathbf{D}_i$ usually a derivative matrix of form $\frac{\partial \mu_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$

$\mathbf{V}_i$ usually $\text{var}(\mathbf{Y}_i)$

Section 4.7. at the end of this chapter contains a tabular summary of the inferential methods discussed in the sequel.

4.2. Population-Averaged Models

4.2.1. Generalized Estimating Equations (GEE)

Generalized estimating equations (GEE) have received much attention since they were first introduced by Liang and Zeger in 1986. To date they are viewed by many primarily as a tool to deal with correlated observations in a semi-parametric framework. The equations themselves are based on the first two marginal moments of the response distributions. This allows the user to estimate model parameters without having to specify a full likelihood. The first moment is determined by the location model for the expectation of the response, i.e. the response function. In the case of categorical data this reduces to the selection of covariates and the link function. The second moments of the response distribution however focus on the association between multiple responses and thus require some knowledge about the correlation pattern in longitudinal data. Liang and Zeger's GEE propose a trick to account for these correlations without placing an undue amount of weight on them, which could make the analysis more sensitive to misspecifications than desired and to ease computational

burden. The trick is to replace the actual correlation pattern by an assumed pattern, called the working correlation assumption. Information about this pattern can be retrieved from the data in the estimation process. The procedure will be described more fully in the succeeding sections.

Currently, GEE are used most frequently as a vehicle to facilitate an estimation problem, i.e. in a technical, generic sense. This does not do justice to its underpinnings. As is often the case, the estimation principle and the technique become confused. To understand how and why GEE works, it is necessary to understand the principles behind estimating function theory, the motivation for inclusion of Chapter 3.2. Some authors view them as an extension of the quasi-likelihood principle. For the reasons outlined in 3.2.4 this comparison is not encouraged here. GEE are rather an extension of the estimating function [3.15]. Since we have shown the optimality of [3.15] in many situations, such as generalized linear models, it is not surprising that this EF can play a role in the analysis of longitudinal categorical data. The examples in 3.2.4 however all assumed complete knowledge of the weight matrix $\mathbf{V} = \text{var}(\mathbf{Y})$. The work of Liang and Zeger showed under which conditions optimality of [3.15] is retained when knowledge about $\mathbf{V}$ is incomplete.

Since their introduction of GEE in 1986 several new forms and varieties have been published. They differ in many regards. Most published work focused on binary response, rather than ordered categorical data. We are discussing the binary case first to lay the groundwork for the extension to the multivariate ordered setting.

4.2.1.1. GEE for longitudinal binary data

Assume the health status of a tree is monitored in the two states, *healthy* and *infected*. Let Y_{it} be an indicator variable for the response of tree i at time t taking on the value 1 if the tree is healthy, 0 otherwise. We are interested in modeling the marginal expectation

$$E(Y_{it}) = \mu_{it} = \Pr(Y_{it} = 1)$$

as a function of covariates $\mathbf{x}_{it}$ and use a link function $G^{-1}(\bullet)$ to map the expectation onto the real line. The marginal model becomes

$$G^{-1}(E(Y_{it})) = G^{-1}(\mu_{it}) = \mathbf{x}_{it}'\boldsymbol{\beta}.$$

While it may seem reasonable to assume that responses from different trees are independent, the repeated measures from a tree are correlated. For the moment, we can proceed however, as if the repeated measures were uncorrelated. Since we will subsequently need to introduce some structure for the correlations we do not attempt to specify the joint marginal distribution of the responses, but seek a semi-parametric approach that depends only on the first two moments of this distribution.

Under the assumption of complete independence the estimation problem reduces to a classical quasi-likelihood problem. That is we know the first two moments

$$\begin{aligned} E(Y_{it}) &= \mu_{it} = G(\mathbf{x}_{it}'\boldsymbol{\beta}) \\ \text{var}(Y_{it}) &= h(\mu_{it}) = \mu_{it}(1 - \mu_{it}). \end{aligned}$$

Under these conditions it has been established ([A.19]) that the integral

$$Q(\mu_{it}, y_{it}) = \int_{y_{it}}^{\mu_{it}} \frac{y_{it} - x}{\text{var}(Y_{it})} dx \quad [4.1]$$

behaves like a log-likelihood function, i.e.

$$\begin{aligned} E(\partial Q(\mu_{it}, y_{it}) / \partial \mu_{it}) &= 0 \\ -E(\partial^2 Q(\mu_{it}, y_{it}) / \partial \mu_{it}^2) &= \frac{1}{\text{var}(Y_{it})}. \end{aligned}$$

Since the distribution remains unspecified, [4.1] is referred to as the log quasi-likelihood function leading to the quasi-likelihood estimating equations

$$\frac{\partial Q(\mu_{it}, y_{it})}{\partial \beta_j} = U(\beta_j, y_{it}) = \frac{\partial \mu_{it}}{\partial \beta_j} \frac{1}{\text{var}(Y_{it})} (y_{it} - \mu_{it}) = 0.$$

This is just the estimating function we encountered in section 3.2.4, where we have identified it as an optimal estimating function in the class of unbiased EF's linear in Y . Stacking the repeated measures for each subject, i.e. $\mathbf{Y}_i = [Y_{i1}, \dots, Y_{in_i}]'$, $\mu_i = E(\mathbf{Y}_i)$, the estimating equation for the complete data can be written in the familiar form

$$U(\beta, \mathbf{y}) = \sum_{i=1}^K \left(\frac{\partial \mu_i}{\partial \beta} \right)' \mathbf{V}_i^{-1} (\mathbf{y}_i - \mu_i) = 0. \quad [4.2]$$

where the variance-covariance matrix $\mathbf{V}_i$ is diagonal with typical element $\mu_{it}(1 - \mu_{it})$. Under the (incorrect) assumption of uncorrelated Y_{ij} 's this is a multivariate form of a quasi-likelihood score equation ([A.20]).

To allow for correlations among the repeated observations, Liang and Zeger (1986) extended [4.2] in the following sense. The matrix $\mathbf{V}_i$ is replaced by a *working* covariance matrix $\tilde{\mathbf{V}}_i$. Specifically we can write $\tilde{\mathbf{V}}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i(\xi) \mathbf{A}_i^{1/2}$ where $\mathbf{A}_i^{1/2} = \text{Diag}(\sqrt{\mu_{it}(1 - \mu_{it})})$ is of dimension $(n_i * n_i)$, $\mathbf{R}_i(\xi)$ is the *working* correlation

matrix for the i th tree and is indexed by a parameter vector ξ , which needs to be estimated.

The main result of the work of Liang and Zeger (1986) is that the estimates obtained by solving [4.2] iteratively are asymptotically unbiased and multivariate normally distributed with estimable variance-covariance matrix, provided the mean model is correctly specified and ξ can be estimated consistently. The closer $\mathbf{R}_i(\xi)$ to the true correlation structure, the more efficient the estimates of the mean parameters. Since [4.2] is of the same form as [3.15] we can apply the same iterative scheme and obtain

$$\hat{\beta}_{u+1} = \hat{\beta}_u + \left(\sum_{i=1}^K \mathbf{D}_i' \tilde{\mathbf{V}}_i^{-1} \mathbf{D}_i \right)^{-1} \left(\sum_{i=1}^K \mathbf{D}_i' \tilde{\mathbf{V}}_i^{-1} (\mathbf{y}_i - \mu_i) \right). \quad [4.3]$$

However, since we have not made a parametric assumption, this iterative scheme is a mere Fisher scoring and coincides with maximum likelihood only if $\tilde{\mathbf{V}}_i$ is the correct covariance matrix, which generally is not the case.

If $\mathbf{R}_i(\xi)$ is the $(n_i \times n_i)$ identity matrix, [4.3] yields the ordinary logistic regression ML estimates under independence.

The fact that only the asymptotic variance is affected by the parametrization of $\mathbf{R}_i(\xi)$ as long as $\hat{\xi}$ is consistent for ξ ([A.27], [A.28]), is a fascinating result of considerable significance. It also sheds light on some of the liabilities for ignoring correlations among observations. Given a correctly specified location model, i.e. a parametrization of the expectation of the random variables under consideration, but ignored serial correlation, the distribution of the estimate $\hat{\beta}$ will not converge to the multivariate Gaussian distribution. It has long been known that in, for example, linear

regression, point estimates of mean parameters are consistent, even if correlations among the observations are not taken into account. However, it is also well-known, that they are inefficient and the estimates of their precision are biased and inconsistent. To see, why this is the case, can be inferred from the asymptotic covariance matrix of the estimates, given by

$$\text{var}(\hat{\beta}) \rightarrow \left(\sum_{i=1}^K \mathbf{D}_i' \tilde{\mathbf{V}}_i^{-1} \mathbf{D}_i \right)^{-1} \left(\sum_{i=1}^K \mathbf{D}_i' \tilde{\mathbf{V}}_i^{-1} \text{var}(\mathbf{Y}_i) \tilde{\mathbf{V}}_i^{-1} \mathbf{D}_i \right) \left(\sum_{i=1}^K \mathbf{D}_i' \tilde{\mathbf{V}}_i^{-1} \mathbf{D}_i \right)^{-1}$$

and estimators thereof. Assuming first that the *working* correlation model $\tilde{\mathbf{V}}_i$ is correct, a model based estimator (Prentice 1988) of $\text{var}(\hat{\beta})$ is

$$\widehat{\text{var}}_M(\hat{\beta}) = \left(\sum_{i=1}^K \hat{\mathbf{D}}_i' \hat{\tilde{\mathbf{V}}}_i^{-1} \hat{\mathbf{D}}_i \right)^{-1} \quad [4.4]$$

since $\text{var}(\hat{\beta}) \rightarrow \left(\sum_{i=1}^K \mathbf{D}_i' \tilde{\mathbf{V}}_i^{-1} \mathbf{D}_i \right)^{-1}$ if $\tilde{\mathbf{V}}_i = \text{var}(\mathbf{Y}_i)$.

An estimator robust against misspecification of $\text{var}(\mathbf{Y}_i)$ is the so-called sandwich estimator (White 1980; Sharples and Breslow 1992; Fitzmaurice *et al.* 1993; Park 1993)

$$\widehat{\text{var}}_R(\hat{\beta}) = \left(\sum_{i=1}^K \hat{\mathbf{D}}_i' \hat{\tilde{\mathbf{V}}}_i^{-1} \hat{\mathbf{D}}_i \right)^{-1} \left(\sum_{i=1}^K \hat{\mathbf{D}}_i' \hat{\tilde{\mathbf{V}}}_i^{-1} (\mathbf{y}_i - \hat{\mu}_i) (\mathbf{y}_i - \hat{\mu}_i)' \hat{\tilde{\mathbf{V}}}_i^{-1} \hat{\mathbf{D}}_i \right) \left(\sum_{i=1}^K \hat{\mathbf{D}}_i' \hat{\tilde{\mathbf{V}}}_i^{-1} \hat{\mathbf{D}}_i \right)^{-1} \quad [4.5]$$

The name stems from the wedging of $\left(\sum_{i=1}^K \hat{\mathbf{D}}_i' \hat{\tilde{\mathbf{V}}}_i^{-1} (\mathbf{y}_i - \hat{\mu}_i) (\mathbf{y}_i - \hat{\mu}_i)' \hat{\tilde{\mathbf{V}}}_i^{-1} \hat{\mathbf{D}}_i \right)$, called the *butter*, between the two model based variance estimators, the *bread*. [4.5] reduces to [4.4] if $\tilde{\mathbf{V}}_i = (\mathbf{y}_i - \hat{\mu}_i) (\mathbf{y}_i - \hat{\mu}_i)'$. The claimed property used to justify *robustness* is consistency under misspecification of $\mathbf{V}_i$. Some authors contend however, that the robustness of [4.5] is questionable. See for example McCullagh's comments in the

discussion of Fitzmaurice *et al.* (1993). Typically, quantities that are functions of the squared errors as the *butter* piece are known to be highly affected by extreme observations.

Both, the *robust* and the model based covariance matrix estimates involve $\hat{\tilde{\mathbf{V}}}_i$. Failure to account for serial correlations, either by ignoring it or by estimating $\hat{\xi}$ inconsistently, will result in lack of convergence of $\widehat{\text{var}}(\hat{\beta})$ to $\text{var}(\hat{\beta})$ and the precision estimates will be poor. An incorrect *working* assumption for $\mathbf{V}_i$ does not corrupt our basis for inference, as long as we can estimate the correlation parameters involved consistently. The closer the *working* structure to the actual correlation pattern, however, the greater the efficiency of the estimating equations and the estimates they imply. Assuming independence is certainly the most extreme choice possible, and one that implies the *working* model $\tilde{\mathbf{V}}_i = \mathbf{A}_i^{1/2} \mathbf{I}_{n_i} \mathbf{A}_i^{1/2} = \mathbf{A}_i$. Since there are no correlation parameters to be estimated, the estimates of precision for $\hat{\beta}$ must be inconsistent, unless trivially, $\xi = \mathbf{0}$. Notice also, that the modes of convergence for $\hat{\beta}$ depend on the mode of convergence of $\hat{\xi}$, see [A.23]-[A.26].

Unfortunately, little guidance can be offered towards the choice of a *working* matrix $\mathbf{R}_i(\xi)$. This involves selection of an appropriate structure as well as consistent estimators for ξ . Table 3.6. listed some pertinent choices especially for continuous responses. For categorical data, the following structures have been proposed correspondingly (c.f. Zeger and Liang 1986; Liang and Zeger 1986; Fitzmaurice *et al.* 1993).

■ Independence structure

$$\mathbf{R}_i(\xi) = \mathbf{I}_{n_i}$$

For a subject with three repeated measures: $\mathbf{R}_i(\xi) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

■ 1-Dependent with common correlation

$\mathbf{R}_i(\xi)$ is tridiagonal, the $t, t+1$ st element is given by α

An estimate can be obtained as

$$\hat{\xi} = \sum_{t=1}^{n-1} \frac{\sum_i^K \hat{r}_{it} \hat{r}_{it+1} / (K-p)}{n-1}$$

where $\hat{r}_{it} = \frac{y_{it} - \hat{\mu}_{it}}{\widehat{\text{var}}(Y_{it})^{1/2}}$ is the Pearson residual and n is the maximum number of time points.

For a subject with three repeated measures: $\mathbf{R}_i(\xi) = \begin{bmatrix} 1 & \xi & 0 \\ \xi & 1 & \xi \\ 0 & \xi & 1 \end{bmatrix}$

■ 1-Dependent with separate correlation estimates

Here different estimates are obtained for the time points - assumed to be

discrete - as $\hat{\xi}_t = \sum_{i=1}^K \hat{r}_{it} \hat{r}_{it+1} / (K-p)$

For a subject with three repeated measures: $\mathbf{R}_i(\xi) = \begin{bmatrix} 1 & \xi_1 & 0 \\ \xi_1 & 1 & \xi_2 \\ 0 & \xi_2 & 1 \end{bmatrix}$

■ Exchangeable correlation structure

This is the structure borrowed from a random effects model also known as compound symmetry and for a subject with three repeated measures

leads to $\mathbf{R}_i(\xi) = \begin{bmatrix} 1 & \xi & \xi \\ \xi & 1 & \xi \\ \xi & \xi & 1 \end{bmatrix}$.

An estimate can be calculated as

$$\hat{\xi} = \sum_{i=1}^K \sum_{t>t'} \hat{r}_{it} \hat{r}_{it'} / \{ (\sum_i \frac{1}{2} n_i (n_i - 1)) - p \}.$$

The above estimates of the correlation parameters are method-of-moment estimates and consistent ([A.13]). Extensions of these patterns for example to n -dependence are straightforward.

The form of GEE discussed so far was first proposed and is workable for both, continuous and categorical responses. A specific feature is that the covariance parameters are treated as nuisance parameters; Prentice (1988) proposed an extension to what has been termed GEE1 and GEE2 subsequently (Zeger, Liang and Qaqish 1992). In light of this terminology, the approach discussed thus far will be called GEE0, a term not yet identified in the pertinent literature.

In GEE1 and GEE2 the correlation structure is not treated as a mere nuisance but subject to modeling. A second set of responses is created that incorporate information about the association of the observations at times t and t' for subject i . Several different association models have been used: empirical residuals $Z_{itt'} = (Y_{it} - \mu_{it})(Y_{it'} - \mu_{it'})$, were employed by Zhao and Prentice (1990) and Zeger and Liang (1992); Lipsitz *et al.* (1991) and Liang *et al.* (1992) used log-odds ratios

$$Z_{itt'} = \frac{\Pr(Y_{it} = 1, Y_{it'} = 1)\Pr(Y_{it} = 0, Y_{it'} = 0)}{\Pr(Y_{it} = 1, Y_{it'} = 0)\Pr(Y_{it} = 0, Y_{it'} = 1)}.$$

Regardless of the choice of $Z_{itt'}$ one proceeds as follows: stack $Z_{itt'}$ as $\mathbf{Z}_i = [Z_{i12}, Z_{i13}, \dots, Z_{i(n_i-1)n_i}]'$, put $E(\mathbf{Z}_i) = \boldsymbol{\nu}_i$, $\text{var}(\mathbf{Z}_i) = \mathbf{W}_i$. Define a second set of estimating equations as

$$U(\boldsymbol{\xi}, \mathbf{z}) = \sum_{i=1}^K \left(\frac{\partial \boldsymbol{\nu}_i}{\partial \boldsymbol{\xi}} \right)' \tilde{\mathbf{W}}_i^{-1} (\mathbf{z}_i - \boldsymbol{\nu}_i) = 0. \quad [4.6]$$

where a model was stipulated for $E(\mathbf{Z}_i) = \boldsymbol{\nu}_i = f(\boldsymbol{\xi})$ which may be either linear or

non-linear and $\tilde{\mathbf{W}}_i$ is a *working* covariance matrix for $\text{var}(\mathbf{Z}_i) = \mathbf{W}_i$. Although we will call the model for $E(\mathbf{Z}_i)$ the association model the structure underlying the second set of estimating equations is a location model $\nu_i = f(\xi)$, where ν_i is somehow linked with the correlation/association among repeated observations. ξ is again a parameter vector to be estimated.

Both systems of GEE, the one for the mean model and [4.6] for the association model can be solved simultaneously using the combined estimating function

$$U(\beta, \xi, \mathbf{z}) = \sum_{i=1}^K \begin{bmatrix} \frac{\partial \mu_i}{\partial \beta} & 0 \\ \frac{\partial \nu_i}{\partial \beta} & \frac{\partial \nu_i}{\partial \xi} \end{bmatrix}' \begin{bmatrix} \mathbf{V}_i & \text{cov}(\mathbf{Y}_i, \mathbf{Z}_i) \\ \text{cov}(\mathbf{Z}_i, \mathbf{Y}_i) & \tilde{\mathbf{W}}_i \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y}_i - \mu_i \\ \mathbf{z}_i - \nu_i \end{bmatrix} = 0 \quad [4.7]$$

If $\text{cov}(\mathbf{Z}_i, \mathbf{Y}_i) = \mathbf{0}$, the system of equations [4.7] has been termed GEE1, GEE2 otherwise. If $\frac{\partial \nu_i}{\partial \beta} = \mathbf{0}$, GEE1 solutions can be obtained by solving the two estimating equations [4.2] and [4.6] separately.

The differences between these three types of GEE for PA models are subtle and important. For GEE0 and GEE1 one arrives at consistent estimates of the mean parameters, given a correctly specified mean model and consistent estimates of ξ , the same model-based assumption made in typical regression. While GEE2 offers greater efficiency over GEE0 and GEE1, both models, the mean and the association model must be specified correctly in order to retain consistency (Liang 1992). This is important, since oftentimes the choice of a proper model for $E(\mathbf{Z}_i) = \nu_i$ will be questionable, as association quantities are not as straightforward in their interpretation as means. While a well established notion may exist which covariates relate to the mean of the responses, $E(\mathbf{Y}_i)$, one usually has less precise knowledge about $E(\mathbf{Z}_i)$.

Another important difference refers to the working assumptions made. In GEE1 and GEE2 this assumption focuses on $\text{var}(\mathbf{Z}_i)$. Since $E(\mathbf{Z}_i)$ is a function of the covariances between repeated measurements and subject to modeling, the predictions from the association models are used to extract the off-diagonal entries for $\text{var}(\mathbf{Y}_i)$. Computational demands increase in the order GEE0, GEE1, GEE2. The second set of estimating equations in GEE1 and GEE2 increases the size of all matrices involved in [4.7] considerably, especially if many observations per subject are available. Putting $\text{cov}(\mathbf{Z}_i, \mathbf{Y}_i) = \mathbf{0}$ in [4.7] is a great simplification of GEE1 over GEE2. In the latter, all matrices must be computed at each iteration and the two sets of estimating equations can not be solved separately, leading to rather large matrix inversions. These considerations become especially important when extending the problem from a binary to a multi-category response.

Zeger and Liang (1986) point out that the choice of $\tilde{\mathbf{W}}_i$ does not affect inference about β very much and recommend an independence assumption for $\mathbf{W}_i$. Zhao and Prentice (1990) discuss other assumptions. A further simplification of [4.7] is possible if β is assumed fixed in $E(\mathbf{Z}_i)$, thus $\frac{\partial \nu_i}{\partial \beta} = 0$. This does not affect the asymptotics (Lipsitz *et al.* 1991). A reasonable manner to carry out the computations in this case is to iterate several times for $\beta \mid \xi$, and after reevaluation of $\hat{\mu}$, to iterate several times for $\xi \mid \beta$ in the association model. This two-step process is continued until convergence (see Fig. 4.1 later in this section).

4.2.1.2. GEE for longitudinal ordered response

The expressions [4.2]-[4.7] hold irrespective whether the response is binary or multinomial, however, some specific changes are required to facilitate the estimation problem. These changes are mainly caused by the complication that the ordinal response for each subject at a single time point is not univariate, but multivariate. The simple logistic model needs to be replaced by an ordinal regression model. Vectors or matrices have to be substituted for scalar quantities in [4.2]-[4.7].

Assume for sake of exposition that the response is J -category ordinal and the categories are labeled $\{1, 2, \dots, J-1\}$. The response at any given point in time t and for any subject can be coded as

$$\mathbf{Y}_{it} = \begin{bmatrix} Y_{it1} \\ Y_{it2} \\ \vdots \\ Y_{it(J-1)} \end{bmatrix} \text{ and across time points } \mathbf{Y}_i = [\mathbf{Y}'_{i1}, \dots, \mathbf{Y}'_{in_i}]'$$

with expectation $E(\mathbf{Y}_i) = \boldsymbol{\mu}_i = [\boldsymbol{\mu}'_{i1}, \dots, \boldsymbol{\mu}'_{in_i}]'$. The last category has again been removed from $\mathbf{Y}_{it}$ because of a linear constraint. Assume that the proportional odds model is a reasonable model to relate covariates to the ordinal response. To set up the problem it is again helpful to code $\mathbf{Y}_{it}$ as a vector of cumulative indicators, i.e.

$$Y_{itj} = \begin{cases} 1 & \text{if tree } i \text{ responds at time } t \text{ in category } j \text{ or less} \\ 0 & \text{otherwise} \end{cases}$$

$$G^{-1}(E(Y_{itj}|\mathbf{x}_{it})) = G^{-1}(\Pr(C_{it} \text{ at most in category } j|\mathbf{x}_{it})) = \alpha_j + \mathbf{x}'_{it}\boldsymbol{\beta}$$

where $G^{-1}(t) = \log(t/(1-t))$. For computational purposes the model may be reparameterized as outlined in § 3.1.7. Since Y_{itj} are cumulative indicators, their variance-covariance matrix is not the usual multinomial covariance matrix, but a $((J-1)*(J-1))$ matrix $\text{cov}(\mathbf{Y}_{it}, \mathbf{Y}_{it})$ with typical elements $\text{cov}(Y_{itj}, Y_{itj'}) = \mu_{itj}(1 - \mu_{itj'})$, $j \leq j'$. Denoting the covariance matrix at any two time points more generally as $\mathbf{V}_{itt'} = \text{cov}(\mathbf{Y}_{it}, \mathbf{Y}_{it'})$ the covariance matrix for any tree can be written as $\text{var}(\mathbf{Y}_i) = \mathbf{V}_i = [\mathbf{V}_{itt'}]$. If $\mathbf{V}_i$ were completely known, the estimating equations for $\theta = [\alpha', \beta']' = [\alpha_1, \dots, \alpha_{J-1}, \beta']'$ would be

$$U(\theta, \mathbf{y}) = \sum_{i=1}^K \left(\frac{\partial \mu_i}{\partial \theta} \right)' \begin{bmatrix} \mathbf{V}_{i11} & \mathbf{V}_{i12} & \dots & \mathbf{V}_{i1n_i} \\ \mathbf{V}'_{i12} & & & \vdots \\ \vdots & & & \vdots \\ \mathbf{V}'_{i1n_i} & \dots & \dots & \mathbf{V}_{in_i n_i} \end{bmatrix}^{-1} (\mathbf{y}_i - \mu_i) = 0. \quad [4.8]$$

[4.8] can be solved as before, given *working* assumptions about the variance-covariance matrix. Comparing [4.2] with [4.8] it is obvious how all scalar quantities in have been replaced by $((J-1)*(J-1))$ matrices.

4.2.1.2.1. GEE0

Under the assumption of uncorrelatedness, $\mathbf{V}_{itt'} = 0, t \neq t'$, [4.8] is a simple generalization of [3.6] as

$$U(\theta, \mathbf{y}) = \sum_{i=1}^K \sum_{t=1}^{n_i} \left(\frac{\partial \mu_{it}}{\partial \theta} \right)' \mathbf{V}_{itt}^{-1} (\mathbf{y}_{it} - \mu_{it}) = 0.$$

The working structure for $\mathbf{V}_i$ required in GEE0 now refers to the relationship between matrices, not scalars. While there is little guidance offered how to select a structure for the longitudinal binary case, even less can be said about the ordered case. The correlation patterns discussed in 4.2.1.1 are likely to carry over directly from the binomial to the multinomial case. The question arises, how to estimate the quantities involved consistently. In the absence of any distributional assumptions and under the premise that only consistent estimates are required, the method-of-moments ([A.13]) suggests itself. A simple estimator for the covariance matrix between observations at times t and t' is

$$\mathbf{C}_{tt'} = \frac{1}{K-p} \sum_{i=1}^K (\mathbf{y}_{it} - \hat{\boldsymbol{\mu}}_{it}) (\mathbf{y}_{it'} - \hat{\boldsymbol{\mu}}_{it'})'$$

As for the case of binary GEE0, $\mathbf{C}_{tt'}$ can be computed for every lag separately, averaged over lags, plotwise etc. The corresponding correlation structures are

- Independence structure

$$\mathbf{C}_{tt'} = \mathbf{0}, \mathbf{V}_i = \text{Diag}(\mathbf{V}_{itt'})$$

- 1-Dependent with common correlation

$$\hat{\mathbf{C}} = \sum_{t=1}^{n-1} \hat{\mathbf{C}}_t / (n-1)$$

$$\hat{\mathbf{C}}_t = \frac{1}{K-p} \sum_{i=1}^K (\mathbf{y}_{it} - \hat{\boldsymbol{\mu}}_{it}) (\mathbf{y}_{i,t+1} - \hat{\boldsymbol{\mu}}_{i,t+1})'$$

For a subject with three repeated measurements: $\tilde{\mathbf{V}}_i = \begin{bmatrix} \mathbf{V}_{i11} & \mathbf{C} & \mathbf{0} \\ \mathbf{C} & \mathbf{V}_{i22} & \mathbf{C} \\ \mathbf{0} & \mathbf{C} & \mathbf{V}_{i33} \end{bmatrix}$

- 1-Dependent with separate correlations by time points

$$\hat{\mathbf{C}}_t = \frac{1}{K-p} \sum_{i=1}^K (\mathbf{y}_{it} - \hat{\boldsymbol{\mu}}_{it}) (\mathbf{y}_{i,t+1} - \hat{\boldsymbol{\mu}}_{i,t+1})'$$

For a subject with three repeated measurements: $\tilde{\mathbf{V}}_i = \begin{bmatrix} \mathbf{V}_{i11} & \mathbf{C}_1 & \mathbf{0} \\ \mathbf{C}_1 & \mathbf{V}_{i22} & \mathbf{C}_2 \\ \mathbf{0} & \mathbf{C}_2 & \mathbf{V}_{i33} \end{bmatrix}$

- Exchangeable correlation structure

$$\hat{\mathbf{C}} = \sum_{i=1}^K \sum_{t>t'} \left(\mathbf{y}_{it} - \hat{\boldsymbol{\mu}}_{it} \right) \left(\mathbf{y}_{it'} - \hat{\boldsymbol{\mu}}_{it'} \right) / \left\{ \left(\sum_i \frac{1}{2} n_i (n_i - 1) \right) - p \right\}$$

For a subject with three repeated measurements: $\tilde{\mathbf{V}}_i = \begin{bmatrix} \mathbf{V}_{i11} & \mathbf{C} & \mathbf{C} \\ \mathbf{C} & \mathbf{V}_{i22} & \mathbf{C} \\ \mathbf{C} & \mathbf{C} & \mathbf{V}_{i33} \end{bmatrix}$.

The estimators above are all consistent ([A.27], [A.28], [A.31], [A.32]). The asymptotics for $\hat{\boldsymbol{\beta}}$ retain. For a similar definition of the correlation patterns see Lipsitz *et al.* (1994).

4.2.1.2.2. GEE1 and GEE2

To accommodate GEE1/GEE2 for ordered response requires coding of the association vectors. Assume we choose simple cross-products of the response indicators as a sort of pseudo-response. Define

$$Z_{itjt'} = Y_{itj} Y_{it'j}$$

$$\mathbf{Z}_{itt'} = [Z_{it0t'0}, Z_{it0t'1}, Z_{it0t'2}, \dots, Z_{it(J-1)t'0}, \dots, Z_{it(J-1)t'(J-1)}]'$$

Using cross-products of the Y_{itj} 's is appealing since then the $Z_{itjt'}$'s are also indicators which makes it relatively easy to find their moments. Specifically

$$\begin{aligned} E(Z_{itjt'}) &= E(Y_{itj} Y_{it'j}) = \text{cov}(Y_{itj}, Y_{it'j}) + E(Y_{itj}) E(Y_{it'j}) \\ &= \rho_{itjt'} \sqrt{\mu_{itj}(1 - \mu_{itj}) \mu_{it'j}(1 - \mu_{it'j})} + \mu_{itj} \mu_{it'j} \end{aligned}$$

where $\rho_{itjt'} = \text{corr}(Y_{itj}, Y_{it'j})$.

To model $E(Z_{itjt'})$ we impose a parametric structure on $\rho_{itjt'}$. To confine the correlation coefficient to $(-1,1)$ a transform is required. Two transforms that have been found to work well for binary response are

(i) the inverse Fisher z-transform

$$\rho_{itjt'} = \frac{\exp(\mathbf{u}'_{itt'}\boldsymbol{\xi}) - 1}{\exp(\mathbf{u}'_{itt'}\boldsymbol{\xi}) + 1}$$

used for example in Lipsitz *et al.* (1991), where $\mathbf{u}'_{itt'}$ is a vector of covariates and $\boldsymbol{\xi}$ a vector of correlation parameters, and

(ii) a continuous AR(1) parametrization

$$\rho_{itjt'} = \exp\left(-\frac{u_{itt'}}{\xi}\right)$$

where $u_{itt'}$ is the positive distance measure between recordings at time t and t' , such as the elapsed time in absolute units or a time substitute such as tree height growth, diameter growth, etc.. While the Fisher z-transform permits positive and negative correlations, the AR(1) model restricts $\rho_{itjt'}$ to be positive.

If $E(\mathbf{Z}_{itt'})$ is estimated, one calculates entries of $\text{var}(\mathbf{Y}_i)$ as

$$\widehat{\text{cov}}(Y_{itj}, Y_{it'j}) = \widehat{E}(Z_{itjt'}) - \widehat{\mu}_{itj}\widehat{\mu}_{it'j}.$$

To set up the estimation process stack $\mathbf{Z}_{itt'}$ to get $\mathbf{Z}_i = [\mathbf{Z}'_{i12}, \mathbf{Z}'_{i13}, \dots, \mathbf{Z}'_{i(n_i-1)n_i}]'$ and let $\boldsymbol{\nu}_i = E(\mathbf{Z}_i)$. For GEE1/GEE2 the working assumption refers to $\text{var}(\mathbf{Z}_i) = \mathbf{W}_i$ which

is structured

$$\mathbf{W}_i = \begin{bmatrix} \text{cov}(\mathbf{Z}_{i12}, \mathbf{Z}_{i12}) & \dots & \dots & \text{cov}(\mathbf{Z}_{i12}, \mathbf{Z}_{i(n_i-1)n_i}) \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ \text{cov}(\mathbf{Z}_{i(n_i-1)n_i}, \mathbf{Z}_{i12}) & \dots & \dots & \text{cov}(\mathbf{Z}_{i(n_i-1)n_i}, \mathbf{Z}_{i(n_i-1)n_i}) \end{bmatrix}.$$

Obviously the structure of $\mathbf{W}_i$ is rather complex and the size of this matrix which has to be built and inverted for every subject at each iteration is considerable. Working assumptions about $\mathbf{W}_i$ are thus chosen with the goal to simplify computations. One such simplification is the so-called structural zero assumption (Zhao and Prentice 1990). Covariances that span more than three measurement occasions are set to zero.

To set all covariance matrices to zero that involve more than two occasions is tantamount to assuming uncorrelatedness of the $\mathbf{Z}_{it\ell}$. $\mathbf{W}_i$ will then be a block-diagonal matrix, $\mathbf{W}_i = \text{Diag}(\mathbf{W}_{it\ell})$ where the elements of $\mathbf{W}_{it\ell}$ are given as

$$\begin{aligned} \text{cov}(Z_{itj'\ell}, Z_{itj''\ell}) &= \nu_{itj'\ell} (1 - \nu_{itj''\ell}) \\ \text{cov}(Z_{itj'\ell}, Z_{itj''\ell'j''}) &= E(Y_{itj} Y_{it'\ell} Y_{itj''} Y_{it'\ell'j''}) - \nu_{itj'\ell} \nu_{itj''\ell'j''}. \end{aligned}$$

But because of the cumulative nature of the Y_{itj} we have

$$\begin{aligned} Y_{itj} Y_{itj''} &= Y_{itj} \text{ if } j < j'' \\ Y_{it'\ell} Y_{it'\ell'j''} &= Y_{it'\ell} \text{ if } \ell' < \ell'' \end{aligned}$$

$$\text{Hence } \text{cov}(Z_{itj'\ell}, Z_{itj''\ell'j''}) = \nu_{it\min(j,j')\ell\min(j'',j'')} - \nu_{itj'\ell} \nu_{itj''\ell'j''}.$$

Needless to say, computing these matrices for each individual is a cumbersome task even under the assumption of uncorrelated $\mathbf{Z}_{it\ell}$'s. But note also, that $\text{var}(\mathbf{Y}_i)$ is not

block-diagonal, but fully specified:

$$\text{var}(\mathbf{Y}_i) = \begin{bmatrix} \mathbf{V}_{i11} & \mathbf{V}_{i12} & \cdots & \mathbf{V}_{i1n_i} \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ \mathbf{V}_{in_i1} & \cdots & \cdots & \mathbf{V}_{in_in_i} \end{bmatrix}$$

where each submatrix is of dimension $(J - 1) \times (J - 1)$. The diagonal submatrices are given by the multinomial nature of the responses and are deterministic, given estimates of the mean parameters.

For GEE2 the covariance matrix $\text{cov}(\mathbf{Y}_i, \mathbf{Z}_i)$ needs to be specified too. It consists of $(J - 1) \times (J - 1)^2$ submatrices with typical elements

$$\text{cov}(Y_{itj}, Z_{itjt'j'}) = \nu_{it\min(j,j')}t'j' - \mu_{itj}\nu_{itjt'j'}.$$

As in the previous section the combined estimating equations are

$$U(\beta, \xi, \mathbf{z}) = \sum_{i=1}^K \begin{bmatrix} \frac{\partial \mu_i}{\partial \beta} & 0 \\ \frac{\partial \nu_i}{\partial \beta} & \frac{\partial \nu_i}{\partial \xi} \end{bmatrix}' \begin{bmatrix} \mathbf{V}_i & \text{cov}(\mathbf{Y}_i, \mathbf{Z}_i) \\ \text{cov}(\mathbf{Z}_i, \mathbf{Y}_i) & \tilde{\mathbf{W}}_i \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y}_i - \mu_i \\ \mathbf{z}_i - \nu_i \end{bmatrix} = 0.$$

It should be obvious that all forms of GEE are based on the same estimating function, which we identified as an OEF in Chapter 3.2.4 provided $\text{var}(\mathbf{Y}_i)$ is known. The differences between the various settings are how the elements of $\mathbf{D}_i$, $\mathbf{V}_i$, and μ_i , are modeled or estimated, respectively.

Owing to the recent development of GEE, many problems need to be addressed in further research. The question of goodness-of-fit, model diagnostics, diagnostics for the choice of the *working* covariance structure, etc. are not sufficiently explored and

discussed in the literature. Given the small number of tools for non-linear models for correlated continuous responses, several years may pass until these problems will be addressed in a workable manner for correlated categorical data.

Based on the asymptotic normality of the estimates, partial significance tests are possible to test hypotheses about the elements of θ , ξ .

Appealing to results in the experimental design literature a simple goodness-of-fit criterion could be $|\widehat{\text{var}}(\widehat{\theta})|$ or $|\widehat{\text{var}}(\widehat{\xi})|$ or a function thereof. Thall and Vail (1990) refer to this measure in the form $H = \log|\widehat{\text{var}}(\widehat{\theta})|^{-1}$ since it is an increasing function of Hotelling's vector correlation between the estimating function and the unknown likelihood score function (Godambe and Heyde 1987; Hotelling 1936). Since the latter is unknown, H can be interpreted only ordinally, i.e. among two competing models the one with larger H statistic is judged to fit the data better. Given a sufficiently large sample, H can be computed both for the model based [4.4] as well as the robust variance estimate [4.5]. Thall and Vail's statistic is obviously related to the concept of design efficiency (Kiefer 1959). Appealing to this concept one could also use statistics such as $\text{trace}\left(\widehat{\text{var}}(\widehat{\theta})\right)$ or $\max(\lambda_j)$, the largest eigenvalue of $\widehat{\text{var}}(\widehat{\theta})$. H should be interpreted carefully if models of different size are compared, since the determinant can increase sharply when adding parameters.

Rotnitzky and Jewell (1990) suggested generalized Wald and Score tests of the hypothesis $H_o : \beta_1 = \beta^*$ where β_1 is a subset of the fixed effects based on the asymptotic normal distribution. One can use either the robust [4.5] or the model based variance estimate [4.4] as weight matrices. To be specific, let β be partitioned as $\beta = [\beta'_1, \beta'_2]'$ and $H_o : \beta_1 = \mathbf{c}$. If $\mathbf{V}$ is a variance estimate based on the principal

elements corresponding to β_1 , then $T_w = (\beta_1 - \mathbf{c})' \mathbf{V}^{-1} (\beta_1 - \mathbf{c}) \stackrel{H_0}{\sim} \chi_s^2$ where s is the rank of the linear hypothesis.

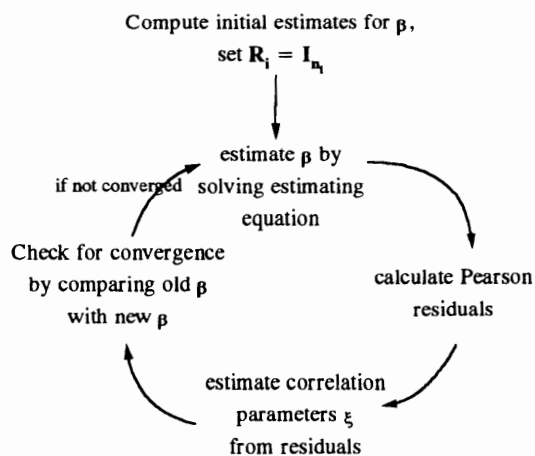
To investigate the appropriateness of the 'working' structure they suggest looking at the eigenvalues of

$$Q = \left(\sum_{i=1}^K \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1} \left(\sum_{i=1}^K \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) (\mathbf{y}_i - \boldsymbol{\mu}_i)' \mathbf{V}_i^{-1} \mathbf{D}_i \right)$$

which is the identity if $\mathbf{V}_i = (\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)'$. Under the null distribution T_w is a weighted sum of Chi-square random variables with 1 degree of freedom each, where the weights are the eigenvalues of Q . Therefore, the average of the first s eigenvalues of Q , e_1 , say, should be close to one. The same is true for the average of the squared eigenvalues, e_2 say, which are the weights in the generalized score test. Thus, Rotnitzky and Jewell (1990) suggest plotting e_1 versus e_2 for different *working* correlation structures. Points close to (1,1) suggest a correct specification. The average eigenvalues can be obtained easily from $se_1 = \text{trace}(Q)$ and $se_2 = \text{trace}(Q^2)$. Unfortunately this investigation depends on the size s of the subset one selects.

Figure 4.1. displays schematic flow charts depicting the computer implementation and estimation steps in GEE0 and GEE1. The double iterative nature of GEE1 that is responsible for some of the additional computational burden, compared to GEE0 becomes obvious in this display.

a) Generalized Estimating Equations - 0



b) Generalized Estimating Equations - 1

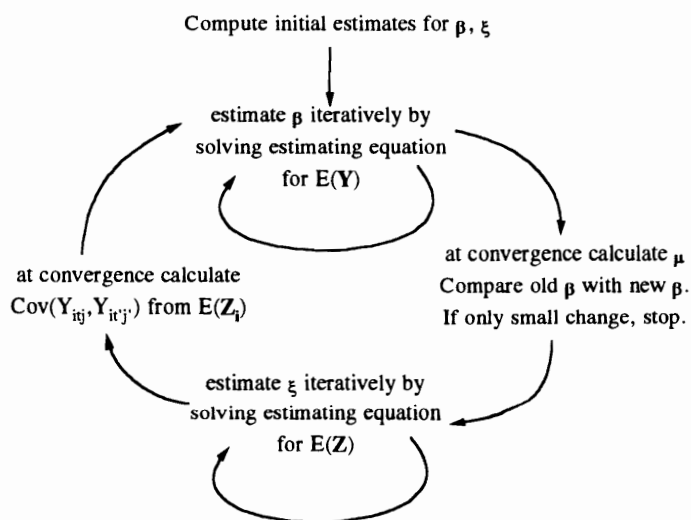


Figure 4.1.: Flow charts of estimation steps in a) GEE0 , and b) GEE1.

4.3. Subject-Specific, Mixed Models

The subject-specific models described in this chapter have in common that the model contains not only unknown constants, but also random variables in place of parameters. We refer to the unknown constants in the location model as the fixed effects, the remainder as the random effects. Whether linear or non-linear models are considered there are many reasons for using random and fixed effects simultaneously. Some were outlined in § 3.3.4. Longford (1993) lists additional motivation.

In designed experiments, random effects broaden the inference space, since one is not only inferring about the factor levels chosen in the study, but about the population of all possible factor levels. In observational studies, such as many longitudinal studies, use of random effects can be motivated by the desire to parsimoniously account for subject-to-subject heterogeneity and to individualize a predictive model. If some parameters of a model are assumed to vary between subjects, one could think about, for example, fitting separate intercept terms, one for each subject. As the number of subjects increases, this leads to an unwieldy estimation/overfitting problem. A parsimonious way of tackling the problem is to ask whether the intercept terms exhibit variation. One assumes that the subject-wise coefficients form a sample from a hypothetical population of coefficients. If there appears to be no variation in this population, the corresponding effect does not appear to vary between subjects. Instead

of estimating as many parameters as there are subjects, one estimates only one quantity, the variation of a random effect.

Some researchers will claim that the variance estimate of the random terms is among the most important results which one can obtain from a mixed model. Because the random effects are associated with subjects, introducing subjects as a source of variation yields a non-diagonal error matrix, i.e. introduces a correlation pattern. In the simplest example let a continuous response $\mathbf{Y}$ be related to explanatory variables by

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \gamma_i + \epsilon_{ij} \quad \text{where}$$

$$\gamma_i \sim N(0, \tau^2), \epsilon_{ij} \sim iid N(0, \sigma^2), \text{cov}(\gamma_i, \epsilon_{ij}) = 0.$$

The error structure $\gamma_i + \epsilon_{ij}$ leads naturally to a model for the correlation among the elements of $\mathbf{Y}_i$, i.e. $\text{var}(\mathbf{Y}_i) = \tau^2 \mathbf{J}_{n_i} + \sigma^2 \mathbf{I}_{n_i}$, with constant correlation $\rho = \tau^2 / (\tau^2 + \sigma^2)$. Note, that γ_i are not parameters. The parameters in the model are $\boldsymbol{\beta}$, τ^2 , σ^2 . But in some applications, the solutions $\hat{\gamma}_i$ that are usually obtained after fitting the model are of considerable interest, since they allow the predictions of the model to be individualized, that is to use the $\hat{\gamma}_i$ as an adjustment to the population averaged prediction $\mathbf{X}_i\hat{\boldsymbol{\beta}}$.

In contrast to the population-averaged version of GEE's discussed in 4.2. where the correlations can be modeled freely, there is no general robustness of mixed models with respect to the specification of the model's random part. The result of PA models for correlated data, that it is only important to estimate the working correlation pattern consistently, does not extend to subject-specific models.

Mixed models for categorical responses form a subclass of the non-linear mixed models, much as GLM's are a special form of non-linear models. We will refer to mixed models for categorical responses as Generalized Mixed Linear Models (GMLM); see [A.34]-[A.36].

The interest in these models has been increasing steadily over the last few years. While after the publication of the papers by Zeger and Liang and Liang and Zeger in 1986 Generalized Estimating Equations seemed to be the ultimate way of analyzing longitudinal categorical data, it soon became apparent, that only certain questions can be addressed satisfactorily through those, i.e. population-averaged research problems. This spawned extensive investigations into subject-specific models for categorical responses, leading to a confusing mix of terminology and nomenclature. Recent approaches have been termed approximate maximum likelihood (Longford 1993), marginal maximum likelihood (Hedeker and Gibbons 1994), marginal quasi-likelihood, penalized quasi-likelihood (Breslow and Clayton 1993), penalized likelihood (Green 1987), and pseudo-likelihood (Wolfinger 1993; Wolfinger and O'Connell 1993). Unfortunately, some of these terms are not quite well defined in particular contexts and are prone to misunderstandings. Rather than following these terms a distinction between (i) semi-parametric, (ii) approximate parametric, and (iii) fully parametric procedures seems to be more helpful and will be followed here.

Approaches in the semi-parametric group are free of distributional assumptions beyond the first few moments. The random effects are specified as having a certain mean and variance. Nothing beyond that is assumed. Since this is in spirit of the estimating function of § 3.2.4, which ultimately gave rise to GEE0-GEE2, the same

type of algorithm can be employed to carry out this procedure. The class (ii) of approximate parametric procedure is based on distributional assumptions, typically the random terms and errors of the location model are assumed to follow a multivariate Gaussian distribution. In order to avoid or simplify integration to arrive at a marginal likelihood function, linearizations are used. They either approximate the integrands by expanding logarithms in series of polynomials (Longford 1993) or approximate the actual mean model to obtain a known marginal distribution (Wolfinger 1993; Wolfinger and O'Connell 1993). Category (iii) of fully parametric procedures replaces analytic integration by numerical techniques and evaluates the actual likelihood function by quadrature methods.

One member from each of these categories will be represented in the following subsections of this chapter. Table 4.1 at the end of this chapter relates the different approaches.

The general formulation of the model in all approaches is as follows: The mean response conditioned on a set of random effects $\mathbf{b}_i$, $E(\mathbf{Y}_i | \mathbf{b}_i)$ is linked to a linear predictor $\eta_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$ through link function $G^{-1}()$:

$$\begin{aligned} G^{-1}(E(\mathbf{Y}_i | \mathbf{b}_i)) &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i \\ E(\mathbf{Y}_i | \mathbf{b}_i) &= G(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i) \end{aligned}$$

The random effects are assumed to vary according to $\mathbf{b}_i \sim (0, \mathbf{D})$ unless stated otherwise.

4.3.1. Semi-Parametric Methods

Zeger, Liang and Albert (1988) approached the problem of fitting this GLMM based on the following idea: the GEE algorithm for estimating the fixed effects β requires only knowledge of the first two marginal moments, $E(\mathbf{Y}_i)$ and $\text{var}(\mathbf{Y}_i)$. The GLMM specifies conditional moments however. If it is possible to approximate $E(\mathbf{Y}_i)$ and $\text{var}(\mathbf{Y}_i)$ from $E(\mathbf{Y}_i | \mathbf{b}_i)$ and $\text{var}(\mathbf{Y}_i | \mathbf{b}_i)$, the algorithm can be used, provided a consistent estimator of $\text{var}(\mathbf{Y}_i)$ can be found. Under this premise the actual distribution of $\mathbf{b}_i$ is not of importance, if the marginal moments are obtained by expectation operations only. Since the approach is subject-specific a *working* correlation matrix does not enter the picture.

One could think of $G(\mathbf{X}_i\beta)$ as an approximation to the subject-specific model when $\text{var}(\mathbf{b}_i) = \mathbf{D} \rightarrow 0$. But from the discussion in 3.3.3 we know that $G(\mathbf{X}_i\beta)$ is the correct marginal mean only if $G(\bullet)$ is the identity function. Alternatively, $G(\mathbf{X}_i\beta)$ would be an approximated marginal mean when expanding the conditional mean $G(\mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i)$ in a first order Taylor series around $E(\mathbf{b}_i) = 0$ and taking expectation.

Zeger *et al.* (1988) discuss several approximations of the marginal means that account for the attenuation effects between marginal and conditional expectations and thus are more efficient than a Taylor series around $E(\mathbf{b}_i)$ alone. They do, however, depend on the link function. If $u_{it} = G(\mathbf{x}'_{it}\beta + \mathbf{z}'_{it}\mathbf{b}_i)$, and μ_{it} denotes the marginal mean, one can write

$$\begin{aligned} \mu_{it} &= \mathbf{x}'_{it}\beta && \text{for } G^{-1}(t) = \text{identity} \\ \mu_{it} &= \exp\left\{\mathbf{x}'_{it}\beta + \mathbf{z}'_{it}\mathbf{D}\mathbf{z}_{it}/2\right\} && \text{for } G^{-1}(t) = \log(t) \\ \mu_{it} &= \Phi\left(|\mathbf{D}\mathbf{z}_{it}\mathbf{z}'_{it} + \mathbf{I}|^{-q/2} \mathbf{x}'_{it}\beta\right) && \text{for } G^{-1}(t) = \Phi^{-1}(t) \end{aligned}$$

$$\text{logit}(\mu_{it}) \doteq |c^2 \mathbf{D} \mathbf{z}_{it} \mathbf{z}_{it}' + \mathbf{I}|^{-q/2} \mathbf{x}_{it}' \boldsymbol{\beta} \quad \text{for } G^{-1}(t) = \log\left(\frac{t}{1-t}\right) \quad [4.9]$$

where $c = (16\sqrt{3})/(15\pi)$. Note that the multiplicative offset in [4.9] is less than one, if $\mathbf{D}$ is positive definite, since the determinant will be greater than one. Thus the front term of [4.9] serves as a shrinkage, forcing the subject-specific coefficients to be larger than the population averaged ones. This is the effect that has been referred to as attenuation in subject-specific models (see 3.3.3).

To obtain a marginal variance approximation we use first principles and write

$$\text{Cov}(\mathbf{Y}_i) = \text{Cov}(\mathbf{E}(\mathbf{Y}_i | \mathbf{b}_i)) + \mathbf{E}(\text{Cov}(\mathbf{Y}_i | \mathbf{b}_i)). \quad [4.10]$$

The approximation comes into play, when linearizing $\mathbf{E}(\mathbf{Y}_i | \mathbf{b}_i)$ through a first-order Taylor series around $\mathbf{E}(\mathbf{b}_i) = \mathbf{0}$.

In what follows, we discuss this for a four-category ordinal response. This requires some additional matrix manipulations due to the multivariate nature of the response, for univariate categorical response see the original paper by Zeger *et al.* (1988), for continuous response see Schabenberger (1995b), and Schabenberger and Gregoire (1995).

Let $Y_{itj} = 1$ if subject i responds in category j or less at time t , 0 otherwise. Define a proportional odds model

$$\mathbf{E}(Y_{itj} | \mathbf{b}_i) = \frac{\exp\{\alpha_j + \mathbf{x}_{it}' \boldsymbol{\beta} + \mathbf{z}_{it}' \mathbf{b}_i\}}{1 + \exp\{\alpha_j + \mathbf{x}_{it}' \boldsymbol{\beta} + \mathbf{z}_{it}' \mathbf{b}_i\}} = u_{itj}$$

for the conditional mean where $\mathbf{b}_i \sim (\mathbf{0}, \mathbf{D})$, $\text{Cov}(\mathbf{b}_i, \mathbf{b}_i) = \mathbf{0}$. Since Y_{itj} is cumulative, $\text{cov}(Y_{itj}, Y_{itj'} | \mathbf{b}_i) = h(u_{itj}, u_{itj'}) = u_{itj}(1 - u_{itj'})$, $j \leq j'$. The approximate marginal

mean follows from [4.9] as

$$\text{logit}(E(Y_{itj})) = \text{logit}(\mu_{itj}) \doteq |c^2 \mathbf{D} \mathbf{Z}_{it} \mathbf{Z}_{it}' + \mathbf{I}|^{-q/2} (\alpha_j + \mathbf{x}_{it}' \boldsymbol{\beta}).$$

To simplify notation, stack the design vectors for the $(J - 1)$ non-redundant categories for each subject to yield $\mathbf{X}_{it}$, and $\mathbf{Z}_{it}$. Furthermore absorb the cut-off design points into $\mathbf{X}_i$ and the α_j 's and $\boldsymbol{\beta}$ into $\boldsymbol{\theta}$. Then, using the identity about marginal and conditional variances above, one obtains:

$$\begin{aligned} \text{Cov}(\mathbf{Y}_{it}) &= \text{Cov}(G(\mathbf{X}_{it}\boldsymbol{\theta} + \mathbf{Z}_{it}\mathbf{b}_i)) + E(\text{Cov}(\mathbf{Y}_{it} | \mathbf{b}_i)) \\ &\doteq \text{Cov} \left(G(\mathbf{X}_{it}\boldsymbol{\theta}) + \frac{\partial G(\mathbf{X}_{it}\boldsymbol{\theta} + \mathbf{Z}_{it}\mathbf{b}_i)}{\partial \mathbf{b}_i} \Big|_{\mathbf{b}_i=0} \mathbf{b}_i \right) \\ &\quad + E \left(h \left(G(\mathbf{X}_{it}\boldsymbol{\theta}) + \frac{\partial G(\mathbf{X}_{it}\boldsymbol{\theta} + \mathbf{Z}_{it}\mathbf{b}_i)}{\partial \mathbf{b}_i} \Big|_{\mathbf{b}_i=0} \mathbf{b}_i \right) \right) \\ &= \frac{\partial G(\mathbf{X}_{it}\boldsymbol{\theta})}{\partial \mathbf{X}_{it}\boldsymbol{\theta}} \mathbf{Z}_{it} \mathbf{D} \mathbf{Z}_{it}' \frac{\partial G(\mathbf{X}_{it}\boldsymbol{\theta})}{\partial (\mathbf{X}_{it}\boldsymbol{\theta})'} + \mathbf{A}_{it} \\ &= \mathbf{L}_{it} \mathbf{Z}_{it} \mathbf{D} \mathbf{Z}_{it}' \mathbf{L}_{it}' + \mathbf{A}_{it} = \tilde{\mathbf{V}}_{it} \end{aligned}$$

since $\frac{\partial G(\mathbf{X}_{it}\boldsymbol{\theta} + \mathbf{Z}_{it}\mathbf{b}_i)}{\partial \mathbf{b}_i} \Big|_{\mathbf{b}_i=0} = \mathbf{Z}_{it} \frac{\partial G(\mathbf{X}_{it}\boldsymbol{\theta})}{\partial \mathbf{X}_{it}\boldsymbol{\theta}}$ and where $\mathbf{A}_{it} = \text{Diag}\{h(\mu_{itj})\}$.

Notice, that (i) $\mathbf{A}_{it}$ is evaluated at the approximated marginal mean, hence computing solutions $\hat{\mathbf{b}}_i$ can be postponed until the estimating equations are solved in case $\hat{\mathbf{b}}_i$ are required, and (ii) that $\mathbf{L}_{it}$ is diagonal. Using the results in 4.2. we can set these estimating equations up as

$$U(\boldsymbol{\theta}, \mathbf{y}) = \sum_{i=1}^K \frac{\partial \mu_i}{\partial \boldsymbol{\theta}'} \tilde{\mathbf{V}}_i^{-1} (\mathbf{y}_i - \mu_i) = \mathbf{0} \quad [4.11]$$

where $\mu_i = [\mu'_{i1}, \dots, \mu'_{in_i}]'$, $\mathbf{y}_i = [\mathbf{y}'_{i1}, \dots, \mathbf{y}'_{in_i}]'$. To build $\tilde{\mathbf{V}}_i$ from $\tilde{\mathbf{V}}_{it}$ let

$\mathbf{L}_i = \text{Diag}\{\mathbf{L}_{it}\}$, $\mathbf{A}_i = \text{Diag}\{\mathbf{A}_{it}\}$, $\mathbf{Z}_i = [\mathbf{Z}'_{i1}, \dots, \mathbf{Z}'_{in_i}]'$, $\mathbf{X}_i = [\mathbf{X}'_{i1}, \dots, \mathbf{X}'_{in_i}]'$. Then

$$\tilde{\mathbf{V}}_i = \mathbf{L}_i \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i' \mathbf{L}_i + \mathbf{A}_i. \quad [4.12]$$

Notice, that although $\mathbf{A}_i$ is block-diagonal, $\tilde{\mathbf{V}}_i$ is not, due to the involvement of $\mathbf{L}_i$ and $\mathbf{Z}_i$.

Estimation is carried out in two steps. In the first step, [4.11] is solved by applying the Fisher scoring algorithm with matrices replaced accordingly, assuming $\mathbf{D}$ and hence $\tilde{\mathbf{V}}_i$ are known. After a new update of all fixed effects (θ) is obtained in the first step, the weight matrix $\tilde{\mathbf{V}}_i$ needs to be updated too. This constitutes the second step. To this end solve [4.12] for $\mathbf{D}$:

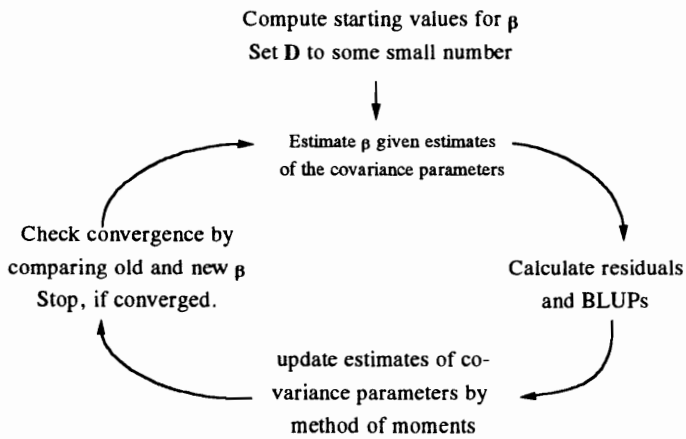
$$\mathbf{D} = (\mathbf{Z}_i' \mathbf{Z}_i)^{-1} \mathbf{Z}_i' \mathbf{L}_i^{-1} (\tilde{\mathbf{V}}_i - \mathbf{A}_i) \mathbf{L}_i^{-1} \mathbf{Z}_i (\mathbf{Z}_i' \mathbf{Z}_i)^{-1}. \quad [4.13]$$

A consistent estimator of $\mathbf{D}$ is the simple moment estimator

$$\hat{\mathbf{D}} = \frac{1}{K} \sum_{i=1}^K (\mathbf{Z}_i' \mathbf{Z}_i)^{-1} \mathbf{Z}_i' \mathbf{L}_i^{-1} \left((y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' - \hat{\mathbf{A}}_i \right) \mathbf{L}_i^{-1} \mathbf{Z}_i (\mathbf{Z}_i' \mathbf{Z}_i)^{-1}. \quad [4.14]$$

[4.14] is a particularly simple estimator and one of the reasons, why MQL is considerably less computer intensive than other approaches to be discussed next. Gregoire and Schabenberger (1995a, 1995b, 1995c) and Schabenberger (1995b) have found $\hat{\mathbf{D}}$ to perform extremely well, even if the number of subjects in the study is as small as 30. Panel a) in Figure 4.2 displays an estimation flow chart for this approach which was termed marginal quasi-likelihood by Breslow and Clayton (1993).

a) Marginal Quasi-Likelihood



b) Pseudo-Likelihood

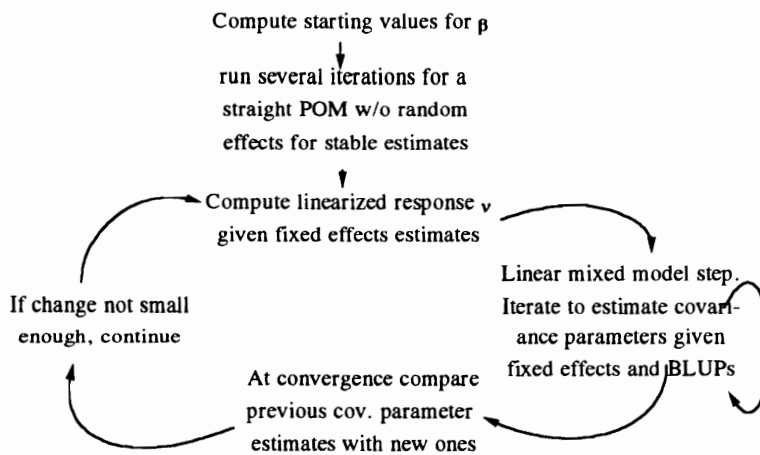


Figure 4.2.: Flow chart of estimation steps in semi-parametric (marginal quasi-likelihood) (a) and approximate parametric (pseudo-likelihood) approach (b).

The cycling between estimating covariance parameters and updating the fixed effects is evident. The semi-parametric approach resembles in its estimation flow very much the GEE0 implementation from which it was derived.

At convergence the best linear unbiased predictors (BLUPs) can be computed for each subject separately based on BLUP theory (Robinson 1991) and the underlying approximations as

$$\hat{\mathbf{b}}_i = \hat{\mathbf{D}}\mathbf{Z}_i'\mathbf{L}_i\mathbf{V}_i^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i). \quad [4.15]$$

A subject's individualized responses are then predicted as $G\left(\mathbf{X}_i\hat{\boldsymbol{\theta}} + \mathbf{Z}_i\hat{\mathbf{b}}_i\right)$.

Since GEE0 are workable for both continuous and categorical responses, it is not surprising that this implementation for GLMM estimation can be adapted for continuous responses and general non-linear responses. This entails a semi-parametric analog of Lindstrom and Bates' approximate Gaussian linear mixed model inference. Details of the procedure can be found in Schabenberger (1995b), applications to forestry in Gregoire and Schabenberger (1995a, 1995b, 1995c) and Schabenberger and Gregoire (1995). A word of caution is in order though. Since the middle piece of [4.14] is obtained by subtraction, $\hat{\mathbf{D}}$ is not necessarily positive definite. The entries on its main diagonal may take on negative values when trying to estimate small variances. By rescaling the columns of the $\mathbf{Z}_i$ matrix, this can partly be avoided. If negative estimates are encountered frequently, this typically is a good indication that the model does not permit subject-to-subject variation. In this case it might be reasonable to reduce the fixed effects part of the model, or resort to a population-averaged approach.

We have seen, that the GEEs can serve as a vehicle for statistical inference in population-averaged as well as subject-specific models. The differences between the two approaches are again important to notice to avoid confusion. In the population-averaged model, a *working* assumption replaces our knowledge about the correct correlation pattern. The correlation matrix can be structured more or less freely. In the subject-specific approach, presented in this section, $\tilde{\mathbf{V}}_i$ arises solely from the specification of the random effects $\mathbf{b}_i$. Once the random part of the model has been specified, $\tilde{\mathbf{V}}_i$ is determined. The efficacy of GEEs in the subject-specific case is affected by (i) the appropriateness of $\mathbf{Z}_i$ as the design matrix for the random terms, and (ii) the quality of the marginal variance approximation. (ii) can hardly be influenced without involving third and fourth order moments which are generally difficult to compute. Hence it is important that the choice of the random terms and the impact it has on the marginal variance $\text{var}(\mathbf{Y}_i)$ is carried out thoughtfully.

4.3.2. Approximate Parametric Methods

The theory of the Gaussian mixed linear models is well established. Lindstrom and Bates (1990) used linearizations to fit a nonlinear mixed model by repeatedly calling a Gaussian mixed linear model routine. The key idea is to apply a linearization so that the marginal model takes on the form of a mixed linear model. The same reasoning can be applied to categorical responses and is described in Wolfinger (1993), Wolfinger and

O'Connell (1993) and Breslow and Clayton (1993). The linearization serves as a vehicle to avoid evaluation of messy, perhaps intractable, integrals.

The derivation starts with an observational GLM

$$\mathbf{Y}_i = G(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i) + \epsilon_i \quad [4.16]$$

where $\epsilon_i \sim (0, \mathbf{A}_i)$ and entertains a series of approximations and assumptions which will be introduced in steps. Let $\mathbf{u}_i = E(\mathbf{Y}_i) = G(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i)$ be the conditional expectation of the response.

Step 1) Expanding $\epsilon_i = \mathbf{y}_i - \mathbf{u}_i$ around $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}_i$ yields

$$\tilde{\epsilon}_i = \mathbf{y}_i - \mathbf{u}_i - \mathbf{G}_i^* \left\{ \mathbf{X}_i(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \mathbf{Z}_i(\mathbf{b}_i - \hat{\mathbf{b}}_i) \right\}. \quad [4.17]$$

Here $\mathbf{G}_i^*$ is a diagonal matrix with entries $\frac{\partial G(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i)}{\partial \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i}$, evaluated at $\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}_i$.

Step 2) The conditional distribution of the approximated error term is assumed Gaussian, in particular $\tilde{\epsilon}_i \mid \hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}_i \sim N(0, \mathbf{A}_i)$.

Step 3) Since it is generally true for GLM's that

$$\frac{\partial G(\eta_{it})}{\partial \eta_{it}} = \left(\frac{\partial G^{-1}(u_{it})}{\partial u_{it}} \right)^{-1},$$

a result of membership in the exponential family ([A.9]), the approximation in step 1 and the distributional assumption in step 2 can be combined. From [4.17] one gets

$$\mathbf{y}_i = \tilde{\epsilon}_i + \mathbf{u}_i + \mathbf{G}_i^* \left\{ \mathbf{X}_i(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \mathbf{Z}_i(\mathbf{b}_i - \hat{\mathbf{b}}_i) \right\}$$

and after rearranging

$$\mathbf{G}_i^{*-1}(\mathbf{y}_i - \mathbf{u}_i) = \left\{ \mathbf{X}_i(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \mathbf{Z}_i(\mathbf{b}_i - \hat{\mathbf{b}}_i) \right\} + \mathbf{G}_i^{*-1}\tilde{\boldsymbol{\epsilon}}_i.$$

But based on the above relationship $\mathbf{G}_i^{*-1}$ is simply $\text{Diag}\left\{\frac{\partial G^{-1}(u_{it})}{\partial u_{it}}\right\}$. This leads to

Step 4) A first order Taylor series of $G^{-1}(\mathbf{y}_i)$ around $\hat{\mathbf{u}}_i$ yields

$$\nu_i = G^{-1}(\hat{\mathbf{u}}_i) + \text{Diag}\left\{\frac{\partial G^{-1}(u_{it})}{\partial u_{it}}\bigg|_{\hat{u}_{it}}\right\}(\mathbf{y}_i - \hat{\mathbf{u}}_i)$$

or

$$\nu_i = G^{-1}(\hat{\mathbf{u}}_i) + \mathbf{G}_i^{*-1}(\mathbf{y}_i - \hat{\mathbf{u}}_i).$$

Step 5) From [4.16] we have $G^{-1}(\hat{\mathbf{u}}_i) = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\hat{\mathbf{b}}_i$. Combining with step 4 finally yields

$$\nu_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{G}_i^{*-1}\tilde{\boldsymbol{\epsilon}}_i, \quad [4.18]$$

a linear mixed model with adjusted within-subject error. As is evident a good number of approximations and assumptions is involved to arrive at [4.18]. The conditional and marginal distributions of the pseudo-response follows from normality of $\tilde{\boldsymbol{\epsilon}}_i$ as:

$$\nu_i | \hat{\mathbf{b}}_i \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \mathbf{G}_i^{*-1}\mathbf{A}_i\mathbf{G}_i^{*-1})$$

and

$$\nu_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{G}_i^{*-1}\mathbf{A}_i\mathbf{G}_i^{*-1}). \quad [4.19]$$

It is informative to compare the moments of the approximate marginal distributions [4.19] with the approximate variance [4.12] in the previous chapter. The matrices $\mathbf{L}_i$ in § 4.3.1. and $\mathbf{G}_i^*$ in this section are of course equivalent, hence $\text{var}(\mathbf{Y}_i)$ and $\text{var}(\nu_i)$ are related by $\text{var}(\mathbf{Y}_i) = \text{var}(\mathbf{L}_i\nu_i)$.

Panel b) in Figure 4.2. outlines the estimation steps to fit the GLMM by this approximate parametric procedure. A complete iteration consists of calculating the pseudo-responses ν_i and iteratively estimating a mixed linear model. The fixed effect estimates and BLUPs can be concentrated out of the likelihood, as described for example in Gregoire *et al.* (1995). The mixed linear model then only scores for the covariance parameters in $\mathbf{D}$. At convergence of this step the covariance parameters are compared to the previous estimates. If changes are sufficiently small, convergence can be assumed. Otherwise ν_i is reevaluated and another cycle entered. Given estimates of the covariance parameters, $\hat{\beta}$ and $\hat{\mathbf{b}}_i$ are given by

$$\hat{\beta} = \left(\sum_{i=1}^K \mathbf{X}_i' \hat{\mathbf{V}}_i^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^K \mathbf{X}_i' \hat{\mathbf{V}}_i^{-1} \nu_i \right).$$

$$\hat{\mathbf{b}}_i = \hat{\mathbf{D}} \mathbf{Z}_i' \hat{\mathbf{V}}_i^{-1} \left(\nu_i - \mathbf{X}_i \hat{\beta} \right)$$

Wolfinger and O'Connell (1993) termed this linearization approach a pseudo-likelihood method since covariance parameters are estimated parametrically, assuming β is fixed at the current estimate $\hat{\beta}$ (Carroll and Ruppert 1988). Breslow and Clayton similarly employed a linearization but used a penalized deviance criterion (Green 1987). They called their approach a penalized quasi-likelihood method. To estimate the covariance parameters the deviance is replaced by the Pearson X^2 statistic, which leads to a normal theory profile likelihood. Wolfinger and O'Connell's approach thus encompasses Breslow and Clayton's method and is a general extension of Lindstrom and Bates' method to categorical responses.

The Taylor series expansions in the work of Wolfinger and O'Connell, Breslow and Clayton, and Lindstrom and Bates are carried out around current solutions of the random effects. Expansions around the fixed effects are implicit in all non-linear optimizations, hence not a point of discussion (Seber and Wild 1989). An alternative to expanding around $\hat{\mathbf{b}}_i$ would be to expand all quantities around the expected value of the random effects, i.e. $E(\mathbf{b}_i) = \mathbf{0}$ (Vonesh and Carter 1992, Vonesh 1993, Sheiner and Beal 1980). Vonesh (1993) refers to expansions around $\mathbf{0}$ as population-averaged, those around $\hat{\mathbf{b}}_i$ as subject-specific. In light of the usage of these terms put forth here this nomenclature will be avoided. There appears to be little right or wrong with either type of expansion. It is simply a question of which provides the more appropriate approximation. One will find that in some instances the numerical results will hardly differ, while in other cases either one approximation may be numerically more stable.

4.3.3. Fully Parametric Methods

The difficulties of evaluating integrals without closed forms are circumvented in the beforementioned methods by linearizations and moment approximations. A brute force method of solving the problem is by numerically computing the integrals involved, i.e. by some form of quadrature routine. Based on work by Harville and Mee (1984), and Jansen (1990), Hedeker and Gibbons (1994) devised a full maximum likelihood approach for clustered ordinal data, that can be used in the longitudinal context.

Harville and Mee (1984) employ what they call the threshold regression concept, now better known as cumulative link models, to allow for the ordering of the responses and approximate estimation via Taylor series expansion. Jansen (1990) and Harville and Mee (1984) implement the EM algorithm.

Hedeker and Gibbons (1994) utilize Fisher scoring which allows multiple random effects in contrast to Jansen (1990) and is known to converge much faster than the EM algorithm. It also avoids linearizations.

To embark on Hedeker and Gibbons' idea let $\eta_{itj} = \alpha_j + \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{b}_i$ be the linear predictor for a cumulative link model with link $G^{-1}(\bullet)$. Then

$$\begin{aligned}\Pr(C_{it} \leq j | \boldsymbol{\beta}, \boldsymbol{\gamma}_i) &= G(\eta_{itj}) \\ \Pr(C_{it} = j | \boldsymbol{\beta}, \boldsymbol{\gamma}_i) &= G(\eta_{itj}) - G(\eta_{it(j-1)})\end{aligned}$$

$G^{-1}(\bullet)$ is typically the logit or probit transform. The origin of η_{itj} is chosen by Hedeker and Gibbons (1994) through $\alpha_0 = -\infty$, $\alpha_1 = 0$. This reduces the size of the estimation problem by one parameter and allows usage of a random intercept without causing rank deficiencies in $\mathbf{X}_{it}$. Maximum likelihood commences as always from the marginal response distribution. For the i th subject the multinomial likelihood ([A.55]) across the n_i time points can be written as

$$L(\boldsymbol{\beta}, \mathbf{b}_i; \mathbf{y}_i) = \prod_{t=1}^{n_i} \prod_{j=1}^J \{G(\eta_{itj}) - G(\eta_{it(j-1)})\}^{y_{itj}}$$

where $Y_{itj} = 1 \Rightarrow C_{it} = j$, 0 otherwise; $\mathbf{Y}_i = [Y_{i11}, \dots, Y_{i1J}, \dots, Y_{in_i1}, \dots, Y_{in_iJ}]$. The marginal likelihood for subject i becomes

$$h(\mathbf{Y}_i) = \int L(\boldsymbol{\beta}, \mathbf{b}_i; \mathbf{y}_i) dF\mathbf{b}_i \quad [4.20]$$

and the complete data likelihood and log-likelihood are

$$L(\beta, \mathbf{b}; \mathbf{y}) = \prod_{i=1}^K h(\mathbf{Y}_i), \text{ and}$$

$$l(\beta, \mathbf{b}; \mathbf{y}) = \sum_{i=1}^K \log\{h(\mathbf{Y}_i)\} \quad [4.21]$$

respectively.

To improve stability during estimation, the random effects part can be reparametrized as $\mathbf{b}_i = \mathbf{T}\theta_i$ where $\mathbf{T}$ is the Cholesky factor of $\text{var}(\mathbf{b}_i) = \mathbf{D}$. Using the Cholesky factor $\mathbf{T}$ instead of $\mathbf{D}$ during iterations ensures that $\hat{\mathbf{D}}$ is positive definite at all times. It also simplifies numerical integration. To implement Fisher scoring, first and second derivatives of the complete data likelihood with respect to α_j 's, β , and the unique elements of $\mathbf{T}$ are required. In general, for arbitrary parameter ψ we have first derivatives

$$\frac{\partial l}{\partial \psi} = \sum_{i=1}^K \frac{1}{h(\mathbf{Y}_i)} \frac{\partial h(\mathbf{Y}_i)}{\partial \psi} \quad [4.22]$$

$$= \sum_{i=1}^K \frac{1}{h(\mathbf{Y}_i)} \int_{\theta} \sum_{t=1}^{n_i} \sum_{j=1}^J Y_{itj} \frac{\partial G(\eta_{itj-1}) - \partial G(\eta_{itj})}{G(\eta_{itj}) - G(\eta_{itj-1})} L(\beta, \theta_i; \mathbf{y}_i) \frac{\partial \eta_{itj}}{\partial \psi} dF\theta_i$$

To evaluate [4.22], some form of quadrature is used, that breaks the integral in a finite sum over a number of specified quadrature nodes with associated weights. If w quadrature nodes are used for each of the q columns in $\mathbf{Z}_i$, the summation is carried out over w^q points. Abramowitz and Stegun (1960) give tables of optimal weights and nodes for Gauss-Hermite quadrature of the standard normal density. As q increases, the number of quadrature nodes can usually be decreased. For a single random term, 5 to 10 nodes are typically necessary to achieve a sufficient accuracy of numerical

integration (Jansen 1990). During computations, the likelihoods $L(\beta, \theta_i; y_i)$ are evaluated by replacing θ_i by the vector of quadrature nodes.

It should be obvious from [4.22] that the computational effort to arrive at estimates is quite formidable, especially when the number of subjects, K , and/or the number of random effects, q , is large. Another disadvantage of this numerical method is that it does not provide solutions for the $\mathbf{b}_i$'s. If they are desired, one has to compute them empirically as modes of a Bayesian posterior distribution (c.f. Davidian and Gallant 1993).

4.4. Conditional Models

This section introduces different ways to model the conditional probability of a subject responding in some state at time t , given it responded in a specified condition previously. To simplify matters and in light of a specific application discussed in following chapters the following constraints are imposed: (i) the ordinal scale consists of 4 categories, to which we assign scores $\{0, 1, 2, 3\}$ for identification. (ii) Once a subject responded in a category j , the subsequent responses are at least j . In other words, the time process is uni-directional. (iii) The last category identifies an absorbing state, from which no transitions are possible. (vi) We are interested in transitions between two consecutive time points. This last constraint is imposed to keep the discussion of statistical methods more comprehensible. It should be obvious from the sequel how adjustments have to be made if any of these constraints are abandoned.

Since at each time point there are 4 response categories, the number of possible response patterns increases rapidly, as the number of events of a subject's life history increases. The uni-directional nature of the response process is of some help here. Let Y_{itj} be a regular indicator variable in this section, not a cumulative one as in Section 4.2. That is,

$$Y_{itj} = \begin{cases} 1 & \text{if subject } i \text{ responds in state } j \text{ at time } t \\ 0 & \text{otherwise} \end{cases}$$

or $Y_{itj} = 1 \Leftrightarrow C_{it} = j$. The assumption of a uni-directional process implies $C_{it} = j \Rightarrow C_{i(t+1)} \geq j$. Consequently if $t > t'$ we have

$$\Pr(C_{it} = j | C_{it'} = j') = 0 \text{ for } j' > j.$$

Figure 4.3. displays the possible transition paths from an initial state across three remeasurements. Conditional on one previous occasion only 9 transitions are possible, rather than 16 if the process were not uni-directional.

It is not difficult to see that the number of transition probabilities still increases rather quickly, while at the same time the number of observed occurrences in any transition class will decrease. This is especially true for the transitions over more than one state, e.g. $0 \rightarrow 3$ or $1 \rightarrow 3$.

Let $p_{itfj} = \Pr(C_{it} = j | C_{i(t-1)} = f) = \Pr(Y_{itj} = 1 | Y_{i(t-1)f} = 1)$ be the conditional probability to observe condition j at time t given that subject i responded f at the last measurement. The p'_{itfj} 's can be arranged in a matrix of transition probabilities in which each row sums to 1:

$$\mathbf{P}_{it} = \begin{bmatrix} p_{it00} & p_{it01} & p_{it02} & p_{it03} \\ 0 & p_{it11} & p_{it12} & p_{it13} \\ 0 & 0 & p_{it22} & p_{it23} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

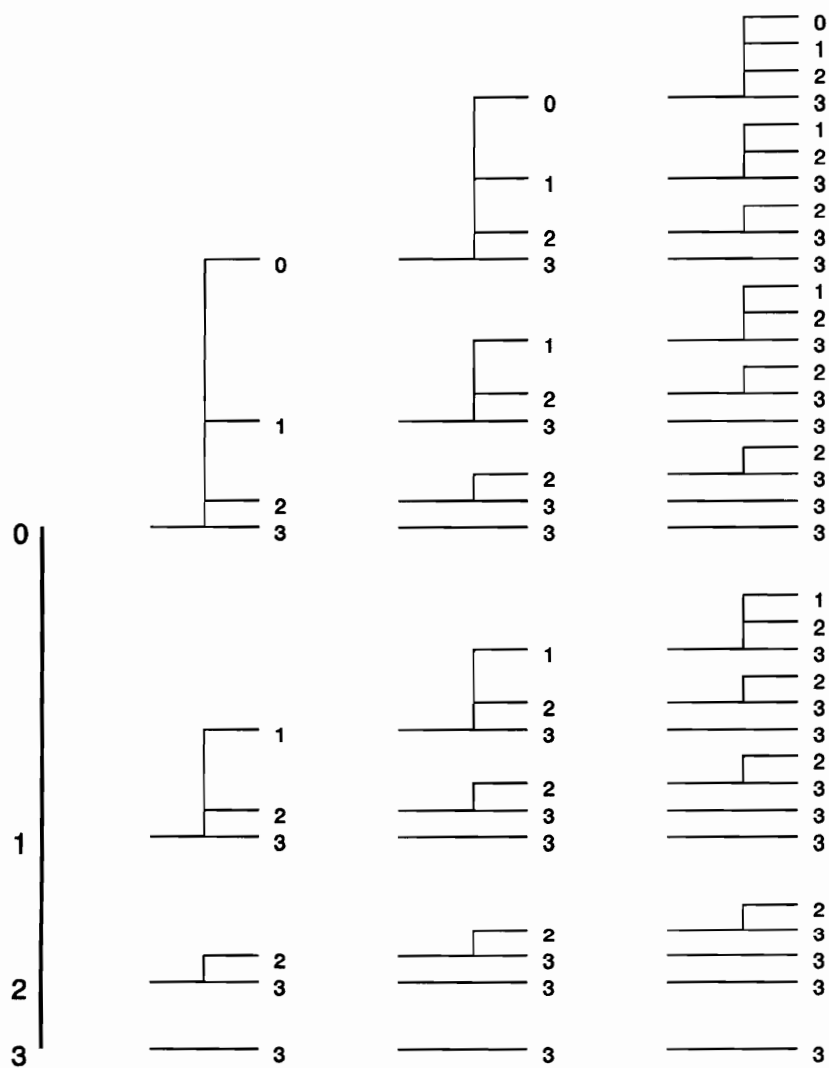


Figure 4.3. Possible response profiles for a uni-directional process with 4 categories. Initial state in leftmost column. Three transitions are plotted in the following columns to the right.

The constraint among the p_{itfj} ,

$$\sum_{j=f}^3 p_{itfj} = 1 \quad \forall \quad j' \quad [4.23]$$

must be reflected in all the models to be developed and predictions must also concur. The modeling task is thus (i) to find a reasonable formulation for p_{itfj} and (ii) to ensure the constraint [4.23]. The latter should be achieved via parametrization of the model instead of using Lagrangian multipliers or other devices for constrained optimization.

4.4.1. RC Models for Conditional Events

The transitions from state j' to j (j'/j for convenience of notation)

$$\begin{array}{cccc} 0/0, & 0/1, & 0/2, & 0/3, \\ & 1/1, & 1/2, & 1/3 \\ & & 2/2, & 2/3 \end{array}$$

can be looked at as sets of multinomial responses with 4, 3, and 2 categories respectively. Therefore it appears natural to investigate some form of multinomial parametrization that achieves [4.23] by construction.

Following Arabatzis *et al.* (1991) one can think of modeling the transition probabilities separately for each previous response. Assume that p_{itfj} does not depend on t and divide the data set into 3 subsets corresponding to the previous responses 0, 1, and 2 respectively. Fit a quadrinomial multinomial regression to the set where

$Y_{i(t-1)0} = 1$, a trinomial multinomial regression to the set where $Y_{i(t-1)1} = 1$, and finally a logistic regression ([A.37]) to the set where $Y_{i(t-1)2} = 1$. Consequently

$$\begin{aligned} p(Y_{it0}, Y_{it1}, Y_{it2}, Y_{it3} | Y_{i(t-1)0} = 1) &\sim MN(1, p_{i00}, p_{i01}, p_{i02}, p_{i03}) \\ p(Y_{it1}, Y_{it2}, Y_{it3} | Y_{i(t-1)1} = 1) &\sim MN(1, p_{i11}, p_{i12}, p_{i13}) \\ p(Y_{it2}, Y_{it3} | Y_{i(t-1)2} = 1) &\sim MN(1, p_{i22}, p_{i23}) \end{aligned} .$$

This is equivalent to assuming that

$$p_{ijj} = \frac{\exp\{\mathbf{x}'_{it} \boldsymbol{\theta}_{jj}\} / \sum_{j'=0/0}^{2/3} \exp\{\mathbf{x}'_{it} \boldsymbol{\theta}_{jj'}\}}{\sum_{l=j'}^3 \frac{\exp\{\mathbf{x}'_{it} \boldsymbol{\theta}_{jl}\}}{\sum_{j'=0/0}^{2/3} \exp\{\mathbf{x}'_{it} \boldsymbol{\theta}_{jj'}\}}} = \frac{\exp\{\mathbf{x}'_{it} \boldsymbol{\theta}_{jj}\}}{\sum_{l=j'}^3 \exp\{\mathbf{x}'_{it} \boldsymbol{\theta}_{jl}\}} \quad [4.24]$$

[4.24] ensures [4.23] since

$$\sum_{l=j'}^3 p_{ijl} = \sum_{l=j'}^3 \frac{\exp\{\mathbf{x}'_{it} \boldsymbol{\theta}_{jl}\}}{\sum_{l=j'}^3 \exp\{\mathbf{x}'_{it} \boldsymbol{\theta}_{jl}\}} = 1 \quad \forall j' .$$

It is common practice to set one parameter vector to zero, corresponding to a baseline (Hosmer and Lemeshow 1989).

Since $p_{ijj} = \Pr(C_{it} = j | C_{i(t-1)} = j')$ the three systems are independent in the sense that $\text{cov}(\hat{\boldsymbol{\theta}}_{jj}, \hat{\boldsymbol{\theta}}_{j'j}) = 0$. The models that refer to different previous responses are unrelated. This is intuitive in the sense that conditioning on the event $C_{i(t-1)} = j'$ narrows the view to observations that share this condition only and ignores any information about other previous responses. If one is not satisfied with this assumption a different approach based on the multinomial parametrization is possible that allows complete interdependence between the $\hat{\boldsymbol{\theta}}_{ijj}$ and will be discussed in Section 4.4.2.

The RC model for conditional events could be fit in separate multinomial regressions as in Arabatzis *et al.* (1991) due to this lack of interrelation. However, this is not necessary. Maximum Likelihood is possible based on the following likelihood function that surfaced in correspondence between Dr. Preisler and Dr. Gregoire.

A subject's individual contribution to the multinomial likelihood in his subset can be written as

$$p_{ij0}^{y_{it0}} p_{ij1}^{y_{it1}} p_{ij2}^{y_{it2}} (1 - p_{ij0} - p_{ij1} - p_{ij2})^{1 - y_{it0} - y_{it1} - y_{it2}}. \quad [4.25]$$

Combining across different previous responses is accommodated by using $y_{i(t-1)j}$ as a multiplier, yielding

$$\phi_{itfj} = \sum_{j=0}^2 y_{i(t-1)j} \left\{ p_{ij0}^{y_{it0}} p_{ij1}^{y_{it1}} p_{ij2}^{y_{it2}} (1 - p_{ij0} - p_{ij1} - p_{ij2})^{1 - y_{it0} - y_{it1} - y_{it2}} \right\} \quad [4.26]$$

and the complete likelihood is

$$L(\theta, \mathbf{y}) = \prod_{i=1}^K \prod_{t=2}^4 \delta_{it} \phi_{itfj} \quad [4.27]$$

where $\delta_{it} = 1$ if tree i was measured at occasions t and $(t-1)$, 0 otherwise. This indicator is proper if the missing data process is at least MAR (see 3.3.1). The purpose of $y_{i(t-1)j}$ as a multiplier in [4.26] is to select the correct set of transition probabilities depending on the value of the previous response. Due to the role of $y_{i(t-1)j}$ the log-

likelihood can be written as

$$l(\theta, \mathbf{y}) = \sum_{i=1}^K \sum_{t=2}^4 \delta_{it} \sum_{j=0}^2 y_{i(t-1)j} \left\{ y_{it0} \log(p_{ij0}) + y_{it1} \log(p_{ij1}) + y_{it2} \log(p_{ij2}) \right. \\ \left. + (1 - y_{it0} - y_{it1} - y_{it2}) \log(1 - p_{ij0} - p_{ij1} - p_{ij2}) \right\} \quad [4.28]$$

In formulas [4.25]-[4.28] it is understood that $p_{ijj}^{y_{ijj}} = 1$ due to the unidirectionality. Thus [4.26] actually expands to

$$\phi_{itj} = y_{i(t-1)0} \left\{ p_{i00}^{y_{i00}} p_{i01}^{y_{i01}} p_{i02}^{y_{i02}} (1 - p_{i00} - p_{i01} - p_{i02})^{1-y_{i00}-y_{i01}-y_{i02}} \right\} \\ + y_{i(t-1)1} \left\{ p_{i11}^{y_{i11}} p_{i12}^{y_{i12}} (1 - p_{i11} - p_{i12})^{1-y_{i11}-y_{i12}} \right\} \\ + y_{i(t-1)2} \left\{ p_{i22}^{y_{i22}} (1 - p_{i22})^{1-y_{i22}} \right\} .$$

Many choices for parametrizing the conditional probabilities are possible. The multinomial parametrization is preferred here because it assures the sum-to-one constraint. Even if other formulations can be found that relate the parameters across different previous responses the multiplier $y_{i(t-1)j}$ would eliminate the contributions for different responses in the derivatives of the log-likelihood function.

As is the case for all conditional models a local independence assumption is implicit, that is after conditioning the responses from a single subject are assumed uncorrelated. This surfaces in [4.27] as the product $\prod_{t=2}^4$ over occasions.

Once the log-likelihood is obtained, estimation of the parameters proceeds as in regular maximum likelihood. To decrease the size of the estimation problem it is reasonable to constrain one parameter vector to zero for each previous response. In

many applications of multinomial logistic regression in biostatistics, it is appealing to choose the *healthy* condition as the baseline. In the context of a uni-directional process, where every subject is either a *mover* or a *stayer* it seems reasonable to choose the baseline as the event to remain in one's category. Thus we put $\theta_{00} = \theta_{11} = \theta_{22} = \theta_{33} = 0$. Let $\theta = [\theta'_{01}, \theta'_{02}, \theta'_{03}, \theta'_{12}, \theta'_{13}, \theta'_{23}]'$ and apply Fisher scoring to [4.28]. The iterative update of θ is obtained as

$$\hat{\theta}_{u+1} = \hat{\theta}_u + \left(\frac{\partial^2 l(\theta, \mathbf{y})}{\partial \theta \partial \theta'} \right)^{-1} \left(\frac{\partial l(\theta, \mathbf{y})}{\partial \theta} \right). \quad [4.29]$$

As discussed in 3.2. and [A.46], $l(\theta, \mathbf{y})$ is an optimal EF in the sense of Godambe (1960) and the correction term in [4.29] is an optimal standardized EF evaluated at $\hat{\theta}_u$ thus assuring fast convergence. The variance of the estimates can be obtained at convergence from the inverse of the Hessian matrix $\left(\frac{\partial^2 l(\theta, \mathbf{y})}{\partial \theta \partial \theta'} \right)$.

The issue of 'independence' of the three sets of multinomial regression models fit simultaneously in [4.28] and [4.29] is apparent when observing that

$$\frac{\partial^2 l(\theta, \mathbf{y})}{\partial \theta_{jj} \partial \theta'_{j''j}} = 0 \quad j'' \neq j$$

which makes it easy to invert the Hessian as a block-diagonal matrix. Notice that [4.28] and [4.29] are equivalent to the OEF in Section 3.2.4. Since the multinomial distribution is in the exponential family ([A.1]), and a canonical link ([A.35]) is employed, the EF of 3.2.4 is the likelihood score function and the general purpose algorithm introduced there is the same as [4.29] (see also [A.43]-[A.46]).

4.4.2. RC Models for Joint Occurrences

The missing interrelation of conditional probabilities that refer to different previous responses in the above model may be considered unrealistic, although unavoidable if one models conditional probabilities directly in the fashion discussed. The RC model for conditional events ensures for example that the probability of a response 2 given a subject responded 0 previously depends on the probability of a 1 response after observing 0, which is natural.

However, if one believes that for example the probability to be in state 2 after being in category 0 initially is related to the probability of a response in 2 after being in state 1, different previous states are involved and [4.28] is not appropriate.

Given the discussion in 4.4.1. derivation of a model that allows for this kind of interrelation across previous states is straightforward. The key idea is to focus first on the joint probabilities

$$\begin{aligned}\pi_{itjj} &= \Pr(C_{it} = j, C_{i(t-1)} = j') \text{ rather than} \\ p_{itjj} &= \Pr(C_{it} = j | C_{i(t-1)} = j').\end{aligned}$$

Taking only one previous response into account, one again arrives at a set of 9 possible events, but an underlying 9-element multinomial probability vector

$$\pi_{it} = [\pi_{it00}, \pi_{it01}, \pi_{it02}, \pi_{it03}, \pi_{it11}, \pi_{it12}, \pi_{it13}, \pi_{it22}, \pi_{it23}]'.$$

Assuming again that the probabilities do not depend on t and using a multinomial parametrization we have

$$\pi_{ijj} = \frac{\exp\{\mathbf{x}'_{it} \boldsymbol{\theta}_{jj}\}}{\sum_{l'/l=0/0}^{2/3} \exp\{\mathbf{x}'_{it} \boldsymbol{\theta}_{l'l}\}} \text{ and } \sum_{j/j=0/0}^{2/3} \pi_{ijj} = 1.$$

Since the conditional probabilities of interest are not modeled directly, they have to be obtained by first principles from

$$p_{ijj} = \frac{\pi_{ijj}}{\sum_{l=j}^3 \pi_{ijl}}$$

which guarantees [4.23].

Fitting such a model may require to include a previous response as a covariate since the repeated observations are not quite locally independent by looking at joint events only. This approach can then be described more aptly as a mixture of a RC and a CC model. However, the responses $C_{it} = j, C_{i(t-1)} = j'$ do incorporate information about the past of a subject implicitly.

The log-likelihood can be written under the assumption of local independence as

$$l(\boldsymbol{\theta}, \mathbf{y}) = \sum_{i=1}^K \sum_{t=2}^4 \delta_{it} \left(\pi_{i00}^{y_{i00}} \pi_{i01}^{y_{i01}} \dots \pi_{i23}^{y_{i23}} \right).$$

Solving the score equation can proceed along the Fisher scoring algorithm of section 4.4.1 where now the Hessian matrix is not block-diagonal. Alternatively one can use a GEE approach based on the EF of 3.2.4, which in principle allows to account for serial correlation if the local independence assumption appears violated. To set up

the GEE (here GEE0) put

$$\mathbf{Y}_i = \begin{bmatrix} \mathbf{Y}_{i2} \\ \vdots \\ \mathbf{Y}_{i n_i} \end{bmatrix}, \mathbf{Y}_{it} = [Y_{itj}]$$

$Y_{itj} = 1$ if subject i responds in state j at time t and in j' at time $(t - 1)$

$$\pi_i = E(\mathbf{Y}_i)$$

$$\boldsymbol{\theta} = [\theta'_{01}, \theta'_{02}, \theta'_{03}, \theta'_{11}, \theta'_{12}, \theta'_{13}, \theta'_{22}, \theta'_{23}]'$$

$$\mathbf{D}_i = \partial \pi_i / \partial \boldsymbol{\theta}$$

and the score equation is again

$$U(\boldsymbol{\theta}, \mathbf{y}) = \sum_{i=1}^K \mathbf{D}'_i \text{var}(\mathbf{Y}_i)^{-1} (\mathbf{y}_i - \pi_i) = 0. \quad [4.30]$$

Solutions are obtained as discussed in Chapter 4.2. The typical elements of the weight matrix $\text{var}(\mathbf{Y}_i)$ are

$$\begin{aligned} \text{cov}(Y_{itj}, Y_{itj''}) &= \pi_{itj}(1 - \pi_{itj}) \quad j' = j'', j = j''' \\ &\quad - \pi_{itj}\pi_{itj''} \quad j' \neq j'' \text{ or } j \neq j''' \end{aligned}$$

Under the local independence assumption $\text{var}(\mathbf{Y}_i)$ is block-diagonal and the solution to the score equation is regular maximum likelihood in multinomial regression.

Any one of the GEE0 structures introduced in 4.2. can be used to account for serial correlation on the off-diagonal blocks of $\text{var}(\mathbf{Y}_i)$.

The complete interrelation of parameters that refer to different previous response states does not come at no cost. A disadvantage of the RC model for joint occurrences

is that the coefficient estimates have an interpretation in terms of joint events, rather than conditional events. Thus, if one wants to interpret the magnitude of the elements of $\hat{\theta}_{jj}$'s, caution should be exercised because they refer not to a probability in a subset of the population that shares the condition (same response j') but to a probability with respect to the entire population.

If the coefficient estimates are not of primary concern, but the probabilities p_{ijj} , then the RC model for joint occurrences is appropriate, since these quantities can be recovered after the model is fit.

Interpretation of single coefficients will also become less informative as the number of response patterns increases, due to more response categories or a higher number of included previous events.

For example with 6 covariates and 9 response patterns one estimates 48 parameters in the RC model for joint occurrences. This large number of estimates seems prohibitive with respect to comparisons between coefficients either by inspection or via hypothesis testing.

The RC model for conditional events is somewhat more parsimonious, however, the number of estimates is still large. Due to a constraint in each of the subsets one would estimate 36 parameters. Compared to a population-averaged model with 6 parameters a large number.

If the estimates are furthermore allowed to vary by time points, the estimation problem becomes unwieldy. For example, with 4 time points, 6 covariates, and 2 previous responses, the RC model for joint occurrences would estimate $(4-1)*6*15 = 270$ parameters, the model for conditional events $(4-1)*6*10 = 180$ parameters.

4.4.3. CC Models for Ordered Responses

All transition models assume that correlations are induced by the influence past responses have on current outcomes. CC models take this concept further and think of past responses as predictors, assuming this influence can be used in an explanatory fashion. Fahrmeir and Tutz (1993) call CC models for this reason *data-driven*.

The joint probability of the n_i responses of subject i , $\Pr(C_{i1}, C_{i2}, \dots, C_{in_i} | \mathbf{x}_{it})$, is expanded into a product of an initial response probability and a series of conditional probabilities:

$$\begin{aligned} \Pr(C_{i1}, C_{i2}, \dots, C_{in_i} | \mathbf{x}_{it}) = & \quad [4.31] \\ \Pr(C_{i1} | \mathbf{x}_{it}) \Pr(C_{i2} | C_{i1}, \mathbf{x}_{it}) \dots \Pr(C_{in_i} | C_{i1}, \dots, C_{i(n_i-1)}, \mathbf{x}_{it}). \end{aligned}$$

The initial occasion provides a natural starting point and all subsequent observations can be ordered relative to it. Such a structure has been termed *asymmetric* by Fahrmeir and Tutz (1993).

The different CC methods vary in how these probabilities are parameterized and how many past events are included. If for example, only the immediately previous response is presumed influential for the presence, one may write

$$\begin{aligned} \Pr(C_{i1}, C_{i2}, \dots, C_{in_i} | \mathbf{x}_{it}) = \\ \Pr(C_{i1} | \mathbf{x}_{it}) \Pr(C_{i2} | C_{i1}, \mathbf{x}_{it}) \dots \Pr(C_{in_i} | C_{i(n_i-1)}, \mathbf{x}_{it}), \end{aligned}$$

a Markov-type transition model, assuming future states are determined uniquely through the present condition. CC models offer a great deal of flexibility in this regard. Two types of CC models are discussed in this section. Bonney's model (Bonney 1987)

is by now a classical representative in this class. Second, a reparameterization based on Bonney's model is introduced which is parsimonious and allows uni-directionality.

4.4.3.1. Bonney's CC model for ordered response

Bonney (1987) introduced a CC model he called a *regressive logistic model* for binary response data (Fahrmeir and Tutz 1993). The model is directly extended in the following to accommodate an ordinal response using a cumulative link model.

Given repeated responses of subject i at times $t = 1, \dots, n_i$, C_{it} and covariates in $\mathbf{x}_{it}$ let $\Pr(C \leq j | \mathbf{x}) = G(\alpha_j + \mathbf{x}'_{it}\boldsymbol{\beta})$ be a cumulative link model for the set of $j = 1, \dots, J$ ordered categories. With $G(t) = \exp(t)/(1 + \exp(t))$ a proportional odds model is obtained.

For the conditional probabilities define

$$\Pr(C_{it} \leq j | C_{i1}, \dots, C_{i(t-1)}, \mathbf{x}_{it}) = G(\eta_{ij(t-1)})$$

where $\eta_{ij(t-1)} = \alpha_j + \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{i1}\boldsymbol{\gamma}_1 + \mathbf{z}'_{i2}\boldsymbol{\gamma}_2 + \mathbf{z}'_{i(t-1)}\boldsymbol{\gamma}_{t-1}$. In this linear predictor $\mathbf{z}_{it}$ is a $(J - 1)$ row vector of zeros with a 1 in position j if response at time t occurred in category j . $\boldsymbol{\gamma}_t$ is a conformable $(J - 1)$ column vector of fixed, unknown constants to be estimated, called the dependence parameters. Note that the conditioning events cover the entire history of a subject up to the current occasion ($C_{i1} - C_{i(t-1)}$). By selecting only a subset of the $\mathbf{z}'_{it}$ s one can easily condition on any number of events of a subject's history.

This is a simple proportional odds model where the set of covariates depends on the previous responses of a subject. As discussed in Chapter 2.2.1.2.3. in contrast to Rosner's model, the $\mathbf{z}_{it}$ are ordered by time points. The conditioning is in terms of a subject's history rather than all occurrences recorded for a subject as implied by models of Rosner's type. The likelihood can be written easily. Let

$$\phi_{itj} = \Pr(C_{it} = j | C_{i1}, \dots, C_{i(t-1)}, \mathbf{x}_{it}) = G(\eta_{ij(t-1)}) - G(\eta_{i(j-1)(t-1)})$$

and obtain

$$L_i(\boldsymbol{\theta}, \mathbf{y}) = \prod_{t=1}^{n_i} \prod_{j=0}^J \phi_{itj}^{y_{itj}}$$

where y_{itj} is a 0/1 indicator for response in state j at time t . The parameter vector $\boldsymbol{\theta}$ contains the cut-off points $\alpha_1, \dots, \alpha_{J-1}$, β , and the γ_t 's. The complete data likelihood is then

$$L(\boldsymbol{\theta}, \mathbf{y}) = \prod_{i=1}^K L_i(\boldsymbol{\theta}, \mathbf{y}) \quad [4.32]$$

The Bonney model has some very important advantages. The likelihood is a simple likelihood for a multinomial response and standard software can be used to fit the model. For example, SAS PROC LOGISTIC can be used after the model matrix has been augmented with the $\mathbf{z}_{it}$'s. The regression parameters and the dependence parameters γ_t are modeled jointly.

Given that $G(\eta_{ij(t-1)})$ is a proper model for $\Pr(C_{it} \leq j | C_{i1}, \dots, C_{i(t-1)}, \mathbf{x}_{it})$, the likelihood principle applies and no additional accounting for serial correlation is necessary.

Most importantly, the Bonney model allows one to investigate the order and the structure of the dependence. One can use different functions $\mathbf{z}_{it}$ and delete some of the dependence terms. For example one might want to include only the previous responses in one model and two previous responses in another model. Since the structures are nested, likelihood ratio tests are possible to check the quality of such competing models.

If more than one previous response is accounted for, the RC models of 4.4.1 and 4.4.2 are of high dimension, limiting their utility in practice to conditioning only on the directly preceding occasion. The Bonney model easily allows conditioning on the complete subject history or only on a subset thereof.

The parts $\mathbf{z}'_{it}\gamma_t$ of the linear predictor can be viewed as offsets to $\alpha_j + \mathbf{x}'_{it}\beta$. Since this offset is applied to cumulative probabilities, the predictions from the model remain strictly stochastically ordered. To demonstrate, assume we condition on the previous response only and model $\Pr(C_{it} \leq j | C_{i(t-1)}, \mathbf{x}_{it}) = G(\alpha_j + \mathbf{x}'_{it}\beta + \mathbf{z}'_{i(t-1)}\gamma)$. Then

$$\begin{aligned}\Pr(C_{it} \leq j | C_{i(t-1)} = j', \mathbf{x}_{it}) &= G(\alpha_j + \mathbf{x}'_{it}\beta + \omega_{j'}) \\ \Pr(C_{it} \leq j+1 | C_{i(t-1)} = j', \mathbf{x}_{it}) &= G(\alpha_{j+1} + \mathbf{x}'_{it}\beta + \omega_{j'}) \\ \Pr(C_{it} = j | C_{i(t-1)} = j', \mathbf{x}_{it}) &= G(\alpha_{j+1} + \mathbf{x}'_{it}\beta + \omega_{j'}) - G(\alpha_j + \mathbf{x}'_{it}\beta + \omega_{j'}).\end{aligned}\tag{4.33}$$

The last term in [4.33] is always positive since the cut-off points are ordered. All conditional probabilities can be generated accordingly and ensure [4.23]. Unfortunately, [4.33] points toward a deficiency of the Bonney model in the present context of a uni-directional process. Since [4.33] is always greater than zero this is also true for $\Pr(C_{it} = j | C_{i(t-1)} = j'', \mathbf{x}_{it})$, $j' < j''$. In other words, the Bonney model does not allow uni-directionality and estimates all transition probabilities with positive quantities.

Two ways are offered to circumvent this shortcoming of the Bonney model. One can set the predicted probabilities for *impossible* transitions to zero and rescale the remaining probabilities such that they sum to one. One can also proceed as with the RC model for conditional events and fit three separate proportional odds models for subsets of the data. A POM with four categories for the subjects that previously responded in the 0 state, a POM with three categories for the subjects previously responding 1 and a logistic regression for the ones that responded in state 2. This requires to set up the problem similarly as in Section 4.4.1. with three simultaneously iterated, independent likelihood estimating equations. In contrast to the conditional model discussed in 4.3.1, this approach would account for the ordering of the responses in each subset.

4.4.3.2. Cut-off CC models

The augmentation of the design matrix in the Bonney model by $\mathbf{z}_{it}$'s adds a vector or matrix of zeros and ones to the data. An alternative way to use past responses as covariates is to adjust the cut-offs in the proportional odds model. The initial proportions are modeled by a regular POM, all subsequent models use cut-offs that are functions of the previous responses. A model that conditions on only one previous response would be set up as follows:

$$\begin{aligned} G^{-1}(E(C_{i1} \leq j)) &= \alpha_j + \mathbf{x}'_{i1}\beta \\ G^{-1}(E(C_{it} \leq j | C_{i(t-1)} \leq j')) &= \psi_j + \alpha_j + \mathbf{x}'_{it}\beta. \end{aligned} \tag{4.34}$$

Construction [4.34] combines the idea of separate cumulative link models with Bonney's CC model through the thresholds only. Response probabilities for a subject

who responded in state 1 previously will be generated with the cut-offs $[\psi_1 + \alpha_1, \psi_2 + \alpha_1]$ corresponding to a cumulative link model with 3 categories. This accommodates uni-directionality. A previously 0 responder will be assigned cut-offs $[\psi_0 + \alpha_0, \psi_1 + \alpha_0, \psi_2 + \alpha_0]$ as necessary to generate a 4 category multinomial expectation and a previously 2 responder $[\psi_2 + \alpha_2]$ resulting in a logistic regression model.

This concept can be developed further. In [4.34], only one set of cut-offs is obtained initially and all others are generated according to the previous responses. One can also separate the cut-offs and estimate

$$G^{-1}(E(C_{i1} \leq j)) = \alpha_j + \mathbf{x}'_{i1}\beta$$

$$G^{-1}(E(C_{it} \leq j | C_{i(t-1)} \leq j')) = \psi_{jj'} + \alpha_j + \mathbf{x}'_{it}\beta.$$

The α_j 's are thresholds for the initial occasion, $\psi_{jj'}$ are offsets that now depend also on the past response. For a subject responding 0 previously the three cut-offs used to calculate the expectation are $[\psi_{00} + \alpha_0, \psi_{10} + \alpha_0, \psi_{20} + \alpha_0]$. For a previously 1 responder $[\psi_{11} + \alpha_1, \psi_{21} + \alpha_1]$.

Models of this type can be fit by standard maximum likelihood techniques.

4.4.4. Markov Chains

The matrix of one step transition probabilities encountered in the previous section,

$$\mathbf{P}_{it} = \begin{bmatrix} p_{it00} & p_{it01} & p_{it02} & p_{it03} \\ 0 & p_{it11} & p_{it12} & p_{it13} \\ 0 & 0 & p_{it22} & p_{it23} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

is a stochastic matrix (Feller 1957), that is, an array of non-negative elements with rows summing to 1. Together with an initial frequency distribution the sequence $\{\mathbf{P}_{it}\}$ defines a Markov chain if the following assumptions hold.

(i) The future behavior of a subject, given a set of covariates, is uniquely defined by the knowledge of the present state, i.e.

$$\begin{aligned} \Pr(C_{it} = j \mid C_{i0}, C_{i1}, \dots, C_{it-1}, \mathbf{x}_{it}) = \\ \Pr(C_{it} = j \mid C_{i(t-1)} = j', \mathbf{x}_{it}) = p_{itj'j} \end{aligned}$$

(ii) $p_{itj'j} = \Pr(C_{it} = j \mid C_{i(t-1)} = j', \mathbf{x}_{it})$ does not depend on t .

If (i) and (ii) hold, the Markov chain is of first order and homogeneous, the simplest case possible (Bailey 1967; Feller 1957). It then follows that $\{\mathbf{P}_{it}\} = \mathbf{P}_i = [p_{kl}]$.

In the sequel the subscript i will be dropped, since the chain can be applied to a single subject, or aggregates (see also chapter 4.6). The n -step transition probability $p_{kl}^{(n)}$ is understood as

$$p_{kl}^{(n)} = \Pr(C_{(t+n)} = k \mid C_t = l, \mathbf{x}_{(t+n)}) \quad [4.35]$$

and by (ii) we have

$$\begin{aligned} p_{kl}^{(1)} &= p_{kl} \\ p_{kl}^{(n)} &= [\mathbf{P}^n]_{kl} \end{aligned} \quad [4.36]$$

The marginal probability to be in any given state after n transitions can be written as

$$\mathbf{p}^{(n)} = \mathbf{P}^n \mathbf{p}^{(0)}$$

where $\mathbf{p}^{(0)}$ is the initial frequency vector. Under the premise of a uni-directional process the chain is irreducible, since not every state can be reached from any other state. This is easily seen from the triangular structure of $\mathbf{P}$. The state $C = 3$ is an absorbing state, or in the jargon of Markov processes, a closed set.

The attractive feature of Markov chains is that under its assumptions a series of interesting probabilities can be calculated that are of considerable interest after fitting a conditional model. These calculations refer to recurrence time and waiting time distributions.

The recurrence time distribution is the set of probabilities that the first return to state j occurs at time n . A recurrence time distribution is associated with each state j . Let $f_j^{(1)} = p_{jj}$, the probability to remain in state j in a one-step transition. The recursive definition of the recurrence time distribution is

$$f_j^{(n)} = p_{jj}^{(n)} - \sum_{t=1}^{n-1} f_j^{(t)} p_{jj}^{(n-t)} \quad [4.37]$$

The probability that after commencing in state j a return to state j ever occurs is

$$f_j = \sum_{n=1}^{\infty} f_j^{(n)}. \quad [4.38]$$

A moment's thought reveals that with triangular transition matrix $\mathbf{P}$, $p_{jj}^{(n)} = p_{jj}^{(1)n} = p_{jj}^n$, thus the recurrent time distribution simplifies to

$$\begin{aligned} f_j^{(n)} &= p_{jj} \quad n = 1 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

and $f_j = f_j^{(1)}$ is the probability that the system stays in state j . A more informative construct in the case of triangular $\mathbf{P}$ is the distribution of waiting times for a passage through state j , after the system started in state i . These quantities, denoted $f_{ij}^{(n)}$ are also defined recursively as

$$\begin{aligned} f_{ij}^{(1)} &= p_{ij} \\ f_{ij}^{(n)} &= p_{ij}^{(n)} - \sum_{t=1}^{n-1} f_{ij}^{(n-t)} p_{ij}^{(t)}. \end{aligned} \quad [4.39]$$

The sum

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)} \quad [4.40]$$

denotes the probability, that starting in state i , the system ever reaches j . In contrast to the recurrence time distribution, the waiting time distribution is not trivial for triangular $\mathbf{P}$.

4.5. Grouped And Ungrouped Cases

Longitudinal data can be thought of as two-level nested data; repeated measurements are nested within subjects. The estimation techniques discussed in § 4 operate on the subject level, i.e. during any part of the algorithm, one sums across subjects, building the necessary vector and matrix pieces subject by subject. This is a very general and usually not inefficient way of setting up the estimation problem and facilitated by the assumption that subjects are independent.

With categorical responses a somewhat different scenario may arise. Since the number of possible responses is countable and usually finite, we can collect all observations that share a common response and covariate pattern. If for example, one were interested in modeling an ordinal response with plot-level covariates but only tree data were available, all responses on a plot in either category can be collected (counted). Assuming four response categories and 100 trees on a specific plot, we can use the treewise data building 100 cumulative response indicators of the form $[1, 1, 1]$ for trees that responded in state 0, $[0, 1, 1]$ for trees in state 1, $[0, 0, 1]$ for responses in state 2, and $[0, 0, 0]$ for the last response category. Alternatively, we can use the response counts in a single vector, e.g. $[40, 60, 70]$ would correspond to 40 state 0 trees, 20 state 1 trees, 10 state 2 trees, and 30 state 3 trees on the plot. These 100 observations share the same plot-level covariates.

The advantage of the second procedure, known as grouped case analysis is that a program only cycles through aggregates of the subjects, increasing the speed of estimation by possibly several orders of magnitudes. The changes that have to be made in the algorithms of chapters 4.2. - 4.4. are rather straightforward. Instead of probabilities, the model now generates counts along the following lines:

Let m^+ be the total number of observations that share covariate pattern $\mathbf{X}^+$. If $\mathbf{Y}$ is the vector of counts and $\boldsymbol{\mu}^+ = [\Pr(C = j | \mathbf{x}^+)]$, then $E(\mathbf{Y}) = m^+ \boldsymbol{\mu}^+$ and

$$\begin{aligned}\text{var}(Y_j) &= m^+ \mu_j^+ (1 - \mu_j^+) \\ \text{cov}(Y_j, Y_{j'}) &= -m^+ \mu_j^+ \mu_{j'}^+ \quad j \neq j'.\end{aligned}$$

This follows directly from the nature of the multinomial distribution, see also [A.49]-[A.54]. If cumulative indicators or counts are used instead, the moments change in an obvious manner. One can also fit the proportions Y_{ij}/m^+ . The results are identical.

4.6. Computer Implementation

It is an unfortunate fact, that as of this writing, almost no software is available to deal with the data structure and estimating techniques outlined in this dissertation.

Although some of these techniques have gained considerable attention and popularity in recent years, their wide spread application has been hampered by the fact that none of the commercially available statistics packages has implemented tools for successfully dealing with longitudinal categorical data.

Even for continuous response data, which in many regards are much easier to deal with, there is still a considerable lack of easy-to-use and efficient software. Much of the software is still custom coded for a specific application. Gregoire, Schabenberger and Barrett (1995) document the commercially available tools and freeware with respect to linear and non-linear mixed models.

For univariate categorical responses, a collection of SAS macros is available from SAS Institute via anonymous ftp from ftp.sas.com in directory /techsup/download/samples/stat_and_iml called GLIMMIX.SAS. It is an implementation of the work by Wolfinger (1993) and Wolfinger and O'Connell (1993), encompassing Breslow and Clayton (1993). For ordinal responses, where the link is composite, they can not be applied. A stand-alone DOS program, written by Donald Hedeker, is available via anonymous ftp from ftp.stat.ucla.edu in directory /pub/apps/msdog/hedeker called

MIXOR.ZIP (as of this writing, the directory actually was called msdog, not msdos). It performs marginal maximum likelihood via quadrature for ordinal response based on cumulative logit or probit models (see chapter 4.3.2.).

The lack of otherwise available software and the desire to put all tools, necessary for the analysis of the data described in the subsequent chapters of this dissertation, in a single program, has prompted the author to custom code most of the estimation techniques discussed in chapter 4 for both grouped and ungrouped categorical data. The high-level matrix programming language GAUSS¹ has been used for this purpose².

Whenever possible, all algorithms have been double-checked with existing software. The authors implementation of marginal maximum likelihood with quadrature has been compared to results obtained by Hedeker's program. The GEE0's under an independence assumption must produce the same results as SAS PROC LOGISTIC. The approximate parametric (pseudo-likelihood) implementation for GMLM must be identical to GLIMMIX.SAS. This has been verified.

¹ GAUSS is a trademark of Aptech Systems, Inc., 23804 S.E. Kent-Kangley Road, Maple Valley, Washington, USA

² Because of the size of the program no printout of the source code is provided with this dissertation. The compiled and source code are made available upon request to the advisory committee along with the necessary data files.

4.7. Summary

A considerable number of estimation methods were discussed in this chapter. They obviously can be quite confusing to the unfamiliar reader. In order to expedite the connections and differences between the approaches Table 4.1. lists the most pertinent methods that can be applied to longitudinal ordinal data.

Table 4.1.
Overview of Estimation Approaches for Longitudinal Ordinal Data as reviewed in Chapter 4.

Population averaged			
Covariances across time are modeled freely, based on working assumptions and possibly association models			
Method	Basis for Inference	Characteristics	Chapter/Section
Generalized Estim. Equations (GEE0)	Estimating Functions	Covariance matrix specified in terms of nuisance parameters, estimated by method of moment techniques. Semi-parametric.	4.2.1.2.1
Generalized Estim. Equations (GEE1)	Estimating Functions	Separate models for mean and associations. Covariances for mean model derived from predicted values of the association model. The two models can be fit separately. Semi-parametric.	4.2.1.2.2
Generalized Estim. Equations (GEE2)	Estimating Functions	Like GEE1, but cross-correlations between mean and association response are permitted. Models fit jointly. Semi-parametric.	4.2.1.2.2

Subject specific			
Covariances are derived from shared subject effects, assumed to vary across subjects according to a probability law			
Method	Basis for Inference	Characteristics	Chapter/Section
Pseudo-Quasi-Likelihood (PQL)	Approximate Max. Likelihood	Taylor series linearization and normal theory error assumptions lead to an approximate mixed linear model. Deviance criterion employed.	4.3.1.1
Pseudo-Likelihood (PL)	Approximate Max. Likelihood	Taylor series linearization and normal theory error assumptions lead to an approximate mixed linear model. Sums of squares criterion.	4.3.1.2
Marginal Quasi-Likelihood (MQL)	Estimating Functions	GEE0 is adapted by approximating marginal moments from conditional ones. Covariance parameters estimated by method of moments. Attenuation corrections improve approximation over simple linearizations in PQL and PL.	4.3.1.3
Marginal Max. Likelihood (MML)	Max. Likelihood	Fully parametric approach. Evaluation of marginal log-likelihood and its derivatives by quadrature methods.	4.3.2

Table 4.1.
(Continued)

Response Conditioned (RC) Models			
Response constructed is an indicator for present and past conditions			
Method	Basis for Inference	Characteristics	Chapter/Section
Models of Rosner's Type	Max. Likelihood	Unordered sequence of observations for each subject. Observations within a subject/cluster are interchangeable.	2.2.1.2.3
Models for conditional events	Max. Likelihood	Based on separate multinomial regression models for past outcomes. Coefficient estimates from different systems are uncorrelated by construction. Hessian block-diagonal.	4.4.1
Models for joint events	Estimating Functions	All possible transitions are fit simultaneously based on multinomial parameterizations. All coefficients are correlated unless aliased Hessian not block-diagonal.	4.4.2

Covariate Conditioned (CC) Models			
Past responses used as regressors. Marginal probabilities expanded in an initial and a series of conditional probabilities			
Method	Basis for Inference	Characteristics	Chapter/Section
Bonney's regressive model	Max. Likelihood	Previous responses coded as indicators that augment the design data matrix. Standard maximum likelihood for generalized linear models is employed.	4.4.3.1
Cut-off conditioned models	Max. Likelihood	Parameters responsible for stochastic ordering are constructed conditionally on previous outcomes to ensure each set of conditional probabilities is strictly ordered.	4.4.3.2

Chapter 5

Fusiform Rust

5.1. Biology

The biology of fusiform rust (*Cronartium quercuum* [Berk.] Miyabe ex Shirai f. sp. *fusiforme*) is well understood and documented (cf. Powers *et al.* 1981). Fusiform rust is a heteroecious fungus that alternates between pine and oak species. On pines, the fungus causes galls on branches and stems. On these galls aecia develop in early spring (February - March) and discharge yellow-orange aecidio spores. After wind dissemination (Kelley and Runion 1991) the spores infect succulent, newly developing oak leaves.

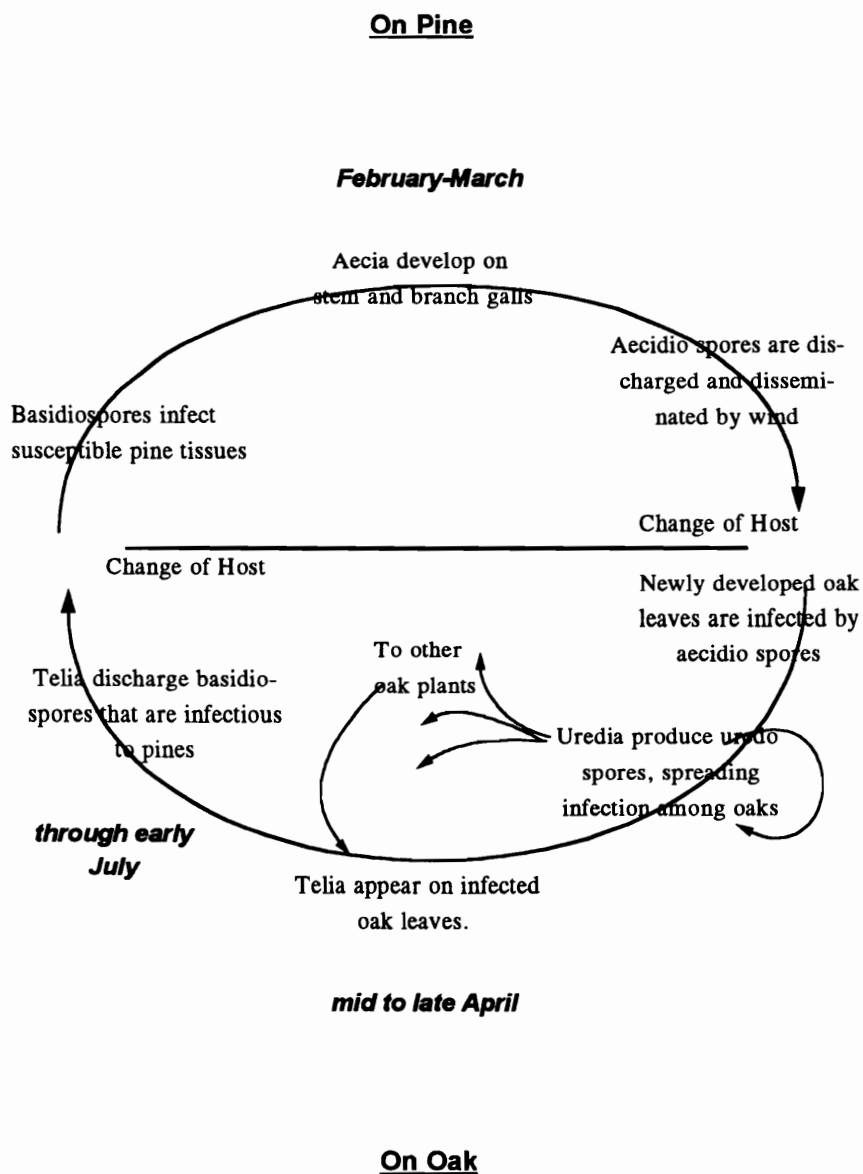


Figure 5.1. Development cycle of fusiform rust.

Within only a few days uredia appear and produce spores that help spread the fungus among oak leaves and oak trees. By mid- to late April telial horns appear on the oak leaves and discharge airborne basidiospores. According to Kelley and Runion (1991) the production of viable basidiospores can continue through early July.

These spores infect susceptible pine tissues and complete the life cycle of fusiform rust. Powers *et al.* report that only young shoots and needles of the pines can be infected. Figure 5.1. depicts the development cycle schematically.

Infection depends greatly on weather and micro-climatic conditions in the vicinity of the susceptible tissues upon infection, a visible gall appears on the branch or stem after some swelling within 6 to 12 months. As with other *Cronartium* species, the infection on the uredia-host causes only minor damage to the leaves and is economically inconsequential. The actual economical damage is always associated with the aecidio-host, the pine species in the case of *Cronartium quercuum*.

Since fusiform rust is a pine-oak rust, its abundance must be considered in relation to the distribution of oak species that can serve as alternate hosts (Snow *et al.* 1986). Powers *et al.* (1993) cite water oak (*Quercus nigra* L.) and laurel oak (*Quercus laurifolia* Michx.) as highly susceptible species. In the course of inoculation experiments, Powers *et al.* (1991) found that common red oak (*Quercus rubra* L.) can serve as an almost universal uredial host for rusts in the pine-oak complex.

Although fusiform rust has been reported to infect many pine species, loblolly pine (*Pinus taeda* L.) and slash pine (*Pinus elliottii* Engelm.) are known to be most

susceptible. Since these two species are the most important pines in the southern U.S. it is imperative that the ramifications of fusiform rust infections be understood.

The damage caused by this disease in southern pine plantations has been estimated by various authors. Due to differences in their data bases and assumptions, the actual numbers vary considerably. Nevertheless they are stunning. Powers *et al.* (1974) estimated economic loss in slash and loblolly pine plantations due to fusiform rust in the southern US as \$30 million per year. Geron and Hafley (1988) cite annual losses in excess of 560 million board feet of sawtimber and 200 million cubic feet of total volume. Based on Anderson and Mistretta (1982), Powers *et al.* (1993) and Arabatzis *et al.* (1991) cite losses excluding mortality to amount to \$128 million annually. Despite the differences in the absolute numbers there is agreement among authors that fusiform rust is the most important forest tree disease in the southern pine region (Burton *et al.* 1985, Anderson *et al.* 1986, Hunt and Lenhart 1986).

The damaging effect on pines are related to the appearance of stem galls, which can easily develop into cankers since they are perennial and under continuing attack by other fungi, bacteria, and insects. If a stem gall girdles the trunk of a stem, the tree dies. In case the tree can fight off complete girdling, the stem is prone to breakage at the location of the gall or canker. Additional causes of rust-associated mortality (RAM) are incremental mortality from all influences that inflict additional stress on the weakened infected tree, such as insects, other fungi, etc. that would not have killed the tree were it not infected (Devine and Clutter 1985).

Infected trees that survive to the rotation period are unlikely to be suitable for saw-logs. As Geron and Hafley (1988) report, the utilization as pulpwood is not affected by the presence of stem galls. If the proportion of stem infected trees is large, a plantation established to achieve high-quality products may end up entirely as pulpwood material. The degrading effect of stem galls is especially apparent because most infections take place in the early years after establishment, thus are in the lower portion of the stem, degrading the most valuable segment of the tree bole (Geron and Hafley 1988). Rust-associated mortality reduces the stocking level and can deplete a plantation to a degree that warrants liquidation and re-establishment.

Fusiform rust incidence forces management decisions of considerable magnitude and liability and success of many pine plantations in the southern U.S. that were established with non-resistant stock depends on our ability to make the correct decisions at the right time during a rotation period. Should a plantation be abandoned and re-established or is a sanitation thinning possible? Should the rotation period be decreased in light of the current state of rust infection in a plantation? Making the correct decision depends on the amount and quality information available about the state of the disease and its likely development over time. For example, whether a sanitation thinning should concentrate on the larger and more vigorous trees or on the smaller trees depends on the relationship between rust infection and tree size (cf. Dell and Driver (1963). Whether the rotation period should be altered or not depends on the development of the disease over time, i.e. the transition probabilities among fusiform rust disease states. Consequently, it is imperative to provide reliable estimates of the current and future situation of either a tree, a group of trees (plot) or an entire plantation.

5.2. Approaches to Estimation and Prediction

Although our understanding of fusiform rust biology is clear, little information is available to guide a management decision. While there have been many attempts to predict fusiform rust disease probabilities and transition frequencies, most of them seem to be of limited value. A good number of studies were of an exploratory nature, describing the prevalence and development of the disease in a certain region. For East Texas for example, Hunt and Lenhart (1986), Lenhart *et al.* (1988), Lenhart *et al.* (1993) and Lenhart *et al.* (1994) depicted the trends in slash pine and loblolly pine plantations and presented a comparison of the two species.

An attempt to relate fusiform rust hazard to commonly measured forest inventory characteristics was presented by Anderson *et al.* (1986). A coarse classification scheme into areas of high, medium, and low rust hazard was associated with inventory data to arrive at *guesstimates* of fusiform rust probabilities. Anderson *et al.* (1988) developed further rust hazard maps for Virginia, the Carolinas, Georgia, and Florida based on forest inventory data along the same lines.

Nance *et al.* (1981) and Shoulders and Nance (1987) used simple frequency estimators to obtain estimates for the rust transition probabilities by counting the number of transitions among disease states and dividing them by appropriate totals.

There exists considerable variation in rust hazards among plantations in any given region, let alone across regions, owing to differences in resistance of the growing stock, site preparation method, irrigation, control of mixing oaks, thinning regiment, etc. A hazards map of the resolution such as in Anderson *et al.* (1988) thus can provide only a crude indication of the possible hazards and not specific enough information about the proper management of existing plantations.

Frequency estimators of the type used in Nance *et al.* (1981) and Shoulders and Nance (1988) require a rather large sample to provide efficient estimators, especially to estimate some of the rare transitions. More importantly, such procedures make it rather difficult to relate disease probabilities to other tree and site specific characteristics such as stand density, site quality, etc. Nance *et al.* utilized a grouping procedure in which the frequency estimates were computed by site index classes, for example. Although not very efficient, this study has shown some interesting relationships which are to be discussed subsequently and which have been used in other studies as well.

In light of these problems it is not surprising that a modeling approach has received some attention early on. Probably the first approach was presented by Dell and Driver (1963) who tried to answer the question whether rust infection rates are related to tree size by graphically comparing these rates across tree diameters between infected and uninfected trees. The first numeric prediction approach is probably Wells and Dinus (1978) who modeled rust-associated mortality at age 10 as a linear function of the number of infected stems at age 5.

Schmidt *et al.* (1979) modeled the percentage of infected trees between age 6 and 10 in loblolly and slash pine as a linear regression on the infection percentages in the

immediately preceding year. Assuming the disease progress curves for the two species are alike, they used one model for both species.

Borders and Bailey (1986) raised the important point that no models exist that predict infection levels from plot-level covariates such as stand density measures, site quality, etc. They presented a logistic model where the logit of rust infection probability was linearly related to the plantation age and site index. Borders and Bailey should be credited with being the first authors to acknowledge the fact that modeling proportions or counts can not be accomplished by linearly relating covariates to the percentages of infection or mortality, such as in Schmidt *et al.* (1979) and Wells and Dinus (1978). Borders and Bailey report on the moderate quality of fit their models achieved, an important point to which we will return shortly.

Geron and Hafley (1988) modeled RAM as a function of surviving trees per acre at stand height 17 feet and the proportion of infected stems at age 5, drawing on the results by Wells and Dinus (1978) that infections occurring prior to age 5 are of special importance in predicting future mortality. Their model is a Bertalanffy type growth function (Seber and Wild 1989).

Arabatzis (1990) and Arabatzis *et al.* (1991) presented generalized linear models for plot data to predict rust incidence in 4 disease states: *healthy*, *branch infected*, *stem infected*, and *dead* trees. This extended the scope of several preceding studies considerably by (i) utilizing a broader response pattern as just {infected, not infected} or {alive, dead} that allows to simultaneously look at mortality and infection

probabilities; (ii) incorporating more covariates such as slope of the plot area, exposition, geographical region, etc., (iii) using external model validation.

These two references also present the first works that deal with the prediction of one-step transition probabilities from plot-level covariates. The approach taken was a separate multinomial regression approach after splitting the data based on the previous responses.

The important shortcomings of the Geron and Hafley (1988) and the Borders and Bailey (1986) approach, as well as of the Arabatzis (1990) and Arabatzis *et al.* (1991) works is that the data bases that served for model development were longitudinal in nature. The serial correlation present in this data was ignored by all authors. This issue can not be taken lightly, especially if the models should clarify how and which covariates are actually related to the disease probabilities. It is an unresolved question whether fusiform rust is related to tree size, individual tree characteristics, site index etc. Studies such as the ones by Borders and Bailey (1986) aimed at clarifying these points. If serial correlations are not taken into account however, estimates of the precision of the model coefficients are biased and inconsistent. The basis for a sound statistical inference and model discrimination is corrupted, invalidated and most inferential results worthless.

To date there are also no tree-level models available that could be used to predict a propensity to infection by fusiform rust and its transitional behavior based on plot- and tree-level characteristics.

Attempts to relate frequency estimators to tree- and plot-level covariates such as site index, height growth, etc. by grouping procedures were, not surprisingly, inconclusive. While some authors found increasing rust infections with site quality (Nance *et al.* 1981, Borders and Bailey 1986), a recent progress report from North Carolina State University cites increasing fusiform rust infection rates with decreasing soil expectation values¹.

Burton *et al.* (1985) presented an analysis of variance approach in an investigation of rust infections of slash pine seedlings up to age 6. The authors found a decrease in height growth after a rust infection occurred and an increase of rust infection rates with tree sizes as did Nance *et al.* (1981). However, plantation growth has also been found unaffected by rust infection rates (Nance *et al.* 1981). This stems from the fact that the studies followed plantations only through their early stages where trees that develop stem galls usually die from girdling before they can impact the growth development of their neighbors.

Borders and Bailey (1986) provide a good review of the sometimes contradictory findings. Without reiterating their development, it should be emphasized that a good deal of such inconsistencies may be caused by improper statistical analyses applied to count or rate data, the ignoring of serial correlations and the considerable variability of rust development across species, sites, silvicultural treatments, etc.

None of the modeling attempts yet has accounted for spatial information about the tree and rust distribution in a stand. Are rust infections clustered in certain areas? Are

¹ Project NC04118, Agency: CSRS NC., Investigator: Roise, J.P.,
Title: Improving Management Decision Making in Forest Resources
Progress report 88/10 - 91/09; available as document 9136041 on USDA world wide web server

these clusters associated with the presence of oak species nearby? How do infected trees that survive past age 10 affect the disease development of their neighbors?

Since the life cycle of *Cronartium quercuum* alternates between two distinct species (see Figure 5.1.), it could be assumed as a working hypothesis that infection of a particular pine tree does not increase the infection likelihood of its neighbors. However, this effect could be confounded by the presence of alternate hosts and the neighboring trees could be more susceptible to infections because a nearby group of oaks periodically discharges basidiospores in their direction. Not much work has been done to incorporate information about the alternate host in rust prediction systems. Snow *et al.* (1986) found a gradient of rust infections related to the proximity of nearby oak species. A project progress report from the University of Florida² found the contrary. Citing from this report:

"... indicated that removing oak in plantings of slash and loblolly pine and in a 500 ft border zone around the plantings did not reduce rust incidence in the oak-free (OF) areas compared with the oak-present (OP) areas. Average rust incidence was significantly greater on the OF area at three sites."

This is certainly a surprising result and another indication of the inconclusiveness of much of the findings in the literature about fusiform rust.

The effect of wheather on the incidence of a rust infection has been completely ignored in all previous modeling attempts. Powers *et al.* (1981) report that infections of both oaks and pines is highly variable from year to year, and they cite so-called *rust*

² Project: FLA-FOR-03064, Agenct: CSRS, FLA, Investigator: Schmidt, R.A.
Title: Epidemiology and Management of Fusiform Rust on Southern Pines
Progress report 91/10 - 92/09; available as document 9152967 on USDA world wide web server

years which occur in 3-5 year intervals and manifest themselves in stem galls at the same height across trees in a stand. It is not difficult to imagine that rust years are associated with environmental conditions under which the production of viable basidiospores flourishes and pines are especially susceptible. Conditions that favor infections on pines are warm, wet weather, humidity above 95 % and temperatures of 15° C to 28° C (Powers *et al.* 1993). To cite Powers *et al.* (1981) referring to the situation in the southern US,

"it is difficult to see any climate factor as limiting"

fusiform rust development. Snow *et al.* (1986) point out, that the viable basidiospores are discharged preferably when oak leaves are wet from dew, fog, or rain. Consequently, a major portion of the basidiospores will be distributed during the night hours, when air movement is commonly weaker than during the day. This leads the author to conclude that the infections spreading from oaks are mostly occurring over short range. Notice, however, that this is in contradiction with the project progress report cited above.

The foregoing discussion leads to a definite challenge:

Forest managers need tools that allow them to predict the rust disease probabilities at any point in time as well as over time to make justifiable management decisions. These decisions are warranted not only in the seedling state of a plantation, but have to be considered and reconsidered throughout the rotation period. The magnitude of initial rust infection and the age at which it occurs, is important for the future development of a stand. This is governed by climatic and environmental factors in the early years of a

plantation, the size, number, and direction of alternate host plants that can serve as pools for viable *Cronartium quercuum* basidiospores.

Information of this kind is usually not available. At later stages of the stand development, after the initial rust infections have taken place, what governs the disease process, is unclear. What we are interested in is to find out if the rust probabilities and transitions can be predicted from information that is usually accessible by forest managers, i.e. plot- and tree-level covariates such as site index, stand density, stand growth, tree size, social position of trees etc. If we succeed, simple predictive models could be implemented which might be of actual use in a given situation.

It should be emphasized that we are not interested to explain what causes rust infection and whether a change in tree size, for example, *causes* a change in rust propensity. The data envisaged are observational in nature and thus do not allow to establish causality. We are interested to find out which variables can serve as reasonable predictors. In doing so, longitudinality of the data - if present - needs to be accounted for to arrive at a sensible model.

Covariates as site index have been found to serve *well* as explanatory variables in some instances. But in light of the biology of fusiform rust it is not very surprising if only moderate amounts of variation can be explained by typical inventory data. As a matter of fact, one could view it as quite an ambitious task to try and relate rust incidence to plot- and tree-level covariates excluding information about weather conditions, alternate hosts, and spatial distributions. One may end up trying to explain pure white noise through a highly structured, complicated, and uninteresting model.

It is easy to imagine that in the complex relationship of host conditions, pine prevalence, viability of spores, wind direction, temperature, humidity etc. the occurrence of a fusiform rust infection can be a quite unpredictable event. One could call it nature's coin flip, whether any nearby pine will get infected from the basidiospores discharged from an oak tree. Attempts to model such processes with classical inventory data may be futile. The inconsistent results among authors could be credited to their failure or success to detect something that appears to be structure in the noise.

Chapter 6.

Material

6.1. The East Texas Pine Plantation Research Project

6.1.1. Introduction

The material used in this and the following chapters was gathered as part of the East Texas Pine Plantation Research Project (ETPPRP) conducted within the Center of Applied Studies at the School of Forestry at Stephen F. Austin State University.

The ETPPRP is a long-term research program investigating factors that affect management of loblolly and slash pine plantations in East Texas. Participating forest industries are represented by

Champion International Corporation
International Paper Company
Louisiana-Pacific Corporation
Temple-Inland Forest Products Corporation.

Data for the ETPPRP are collected on 270 permanent field plots in separate plantations, 84 of which are located in slash pine (*Pinus elliottii* Engelm.) plantations, 186 in loblolly pine (*Pinus taeda* L.) plantations. Each plot consists of two subplots of 10,000 ft^2 (0.23 *acres*), delimited by a 33-foot wide buffer zone from their surroundings.

The data gathered on one of the two subplots is used for development, the other for evaluation purposes in various aspects of modeling. These separate data sets will be referred to as the development (D) and evaluation (E) data, respectively. The plots are revisited and measured in three year intervals. Because of the large number of plots and their distribution throughout East Texas, not all plots can be measured in a given year. Hence a rotating panel design with three measurement waves was established.

Starting in 1982, about one third of the plots were initially measured. These plots were revisited in 1985, 1988, and 1991 and comprise the first measurement wave. The plots of the second wave were initially measured in 1983, the third wave in 1984. As of this writing, four measurement cycles have been completed.

Due to natural causes of destruction and harvesting, not all plots have contributed four measurements over time. Table 6.1 depicts the measurement pattern.

Table 6.1.
Measurement patterns in three wave rotating panel, starting 1982

Wave	Initial Measurement	Cycle	Years
1	1982	1	1982-84
2	1983	2	1985-87
3	1984	3	1987-90
		4	1991-93

A suite of plot and tree level covariates were measured at each point in time or derived from measured characteristics. Table 6.2 lists the most important variables together with descriptions. Other variables were created as model development continued, but all are descendants of the characteristics listed in Table 6.2.

The basic characteristics are typical forest inventory data, aside from the fusiform rust health condition of a tree. The plot level variables are either time-independent, such as slope and aspect, or time-dependent, such as stand height. The tree variables are all time dependent. Although information about the geographical location of the individual plots is available, data about the spatial location of trees within a plot are unavailable.

The fusiform rust condition was assessed for each tree according to the following classification scheme. Trees clear of any rust infection were assigned rust condition code 0. This is understood as *healthy* with respect to fusiform rust.

Table 6.2.

Plot and tree level covariates measured at each occasion on all plots.
 Type = m denotes a measured characteristic, type = d a derived variable.
 $C(x)$ in the Units field identifies a categorical variable with x classes.

Variable	Level	Type	Time dependence	Units	Description
Landform	Subplot	m	No	$C(7)$	Flood Plain (FP), Stream Terrace Upland Flat (UF), Lower Slope (LS), Side Slope (SS), Upper Slope (US), Ridge(R)
Slope	Subplot	m	No	%	Slope of the plot terrain
Aspect	Subplot	m	No	degrees	Aspect of the plot terrain
Site index	Subplot	d	No	feet	Base age = 25 years
Stand height	Subplot	d	Yes	feet	Height of dominant trees
QMD	Subplot	d	Yes	inches	Quadratic mean diameter
Trees per acre	Subplot	d	Yes	count	Only alive trees are counted
Basal area per acre	Subplot	d	Yes	feet ²	Calculated from alive trees only
Relative spacing	Subplot	d	Yes	unitless	$\frac{\sqrt{43560/Trees\ per\ acre}}{Height\ of\ dominant\ trees}$
Site preparation	Plot	m	No	$C(5)$	KG bladed/sheared, not windr. KG bladed/sheared, windrowed KG bladed/sheared, bedded KG bladed/sheared, burned any other site preparation
Age	Plot/Tree	m	Yes	years	Age is advanced one unit in July
DBH	Tree	m	Yes	inches	Diameter at breast height
Height	Tree	m	Yes	feet	
Live crown height	Tree	m	Yes	feet	Height to live crown from ground
Crown class	Tree	m	Yes	$C(4)$	dominant, codominant, intermediate, suppressed
Fusiform rust condition	Tree	m	Yes	$C(4)$	0 = healthy, no infection 1 = canker on branch only 2 = stem canker 3 = death due to fusiform rust

In what follows a tree that does not exhibit a rust infection may well be of poor health for other reasons. Fusiform rust cankers on a branch more than 12 inches from the main stem were assigned to condition class 1, henceforth labeled *branch infection*. These trees were free of stem cankers. Cankers on live branches closer than 12 inches to the main stem or cankers on the stem were assigned to category 2, labeled *stem infection*. If trees were found dead because of their fusiform rust infection, they were assigned to condition class 3, *dead*. If trees were found dead for unknown or not rust related reasons, no condition class was associated to avoid over-representation of fusiform rust as cause of death. Dead trees were removed from the data and did not contribute responses at subsequent measurements.

The response of interest to be modeled is fusiform rust condition. This outcome variable is categorical with four categories. The labels *healthy*, *branch*, *stem*, *dead* suggest an ordinal response. The assignment of a tree to a category at any point in time follows the above mentioned classification scheme, however. Such assignments are always subject to misclassifications. At a subsequent measurement a tree is approached from a different direction and a previously noted branch canker may go undetected. As trees grow older the visibility of upper portions of the crown decreases. Field interpreters will exhibit variability in their decisions whether a tree clearly died of fusiform rust.

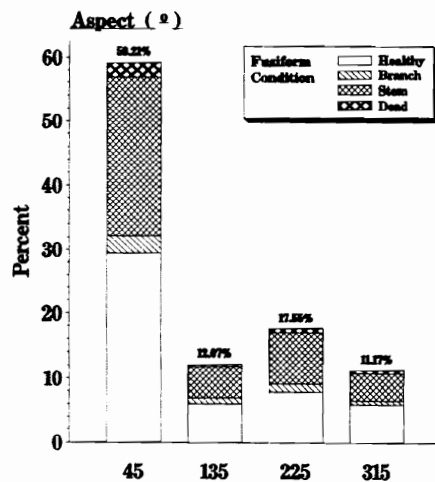
There exists no possibility to assess the correctness of classification over time. The only check was to ensure that a tree's response pattern over time complies to the unidirectionality of the process. A stem infected tree can not reappear as a healthy one at a later occasion, since recovery from a stem infection is impossible. Branch cankers

could be removed by natural pruning. Given the young ages of the plantations (see below) this is rather unlikely.

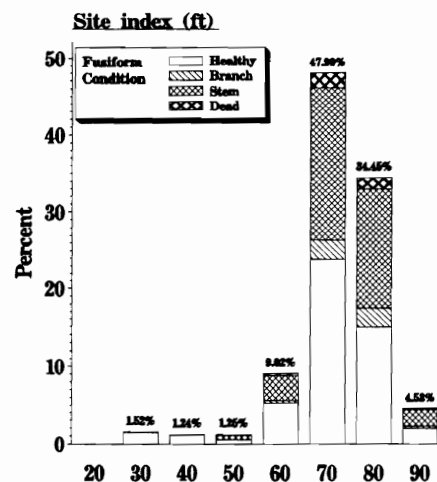
The classifications were assumed to be more reliable at early measurement occasions since trees are smaller and more easily examined, the subsequent responses were mapped to equal the immediately preceding response to avoid illogical response patterns over time. Whether such an adjustment will introduce bias depends on the correctness of the category assignment as well as the appropriateness of the unidirectionality of the disease process.

Figure 6.1 shows histograms of the time-constant plot level covariates for the slash pine development data. The continuous covariates Aspect, Site index, and Slope were grouped for the purpose of plotting only. Figure 6.2 displays histograms for loblolly pine. Most of the slash pine plots have an aspect between 0 and 90 degrees, a slope less than 2% and a site index between 70 and 80 feet. The distribution of responses in the four fusiform rust categories are plotted within each bar. There does not appear that these distributions differ much across the bars of a panel. The hypothesis, that trees of better sites are more likely to obtain fusiform rust cankers and die can not be confirmed from Figure 6.1b) or Figure 6.2b). For loblolly pine the plots are more evenly distributed across slopes, but plots rarely have slopes greater than 10%. While Upland Flat is the dominating land form for slash pine plots, loblolly pine is most frequently found on Lower and Side Slopes. The sloped land forms account for more than 2/3 of all loblolly pine plots.

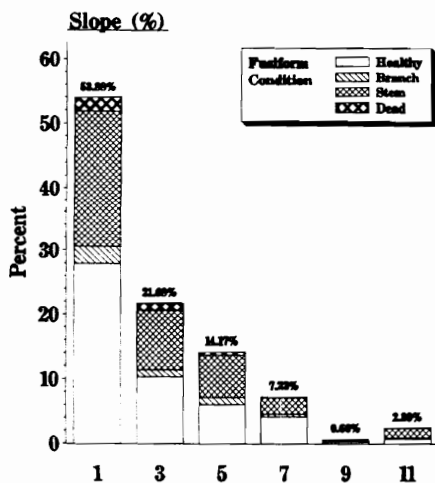
a)



b)



c)



d)

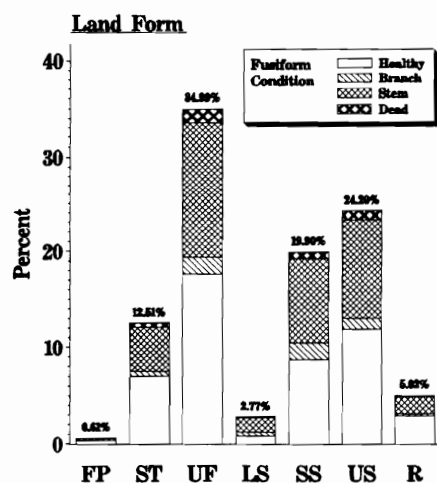
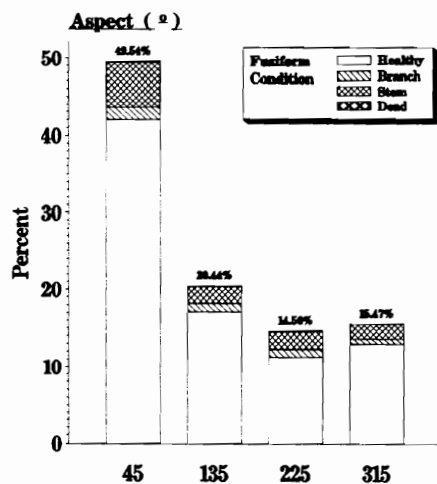
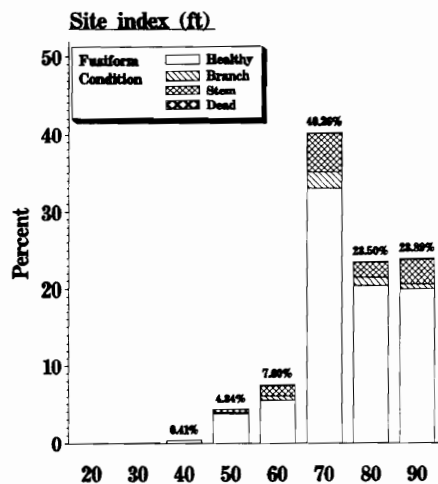


Figure 6.1 : Histograms of time–constant plot level covariates for slash pine development data.

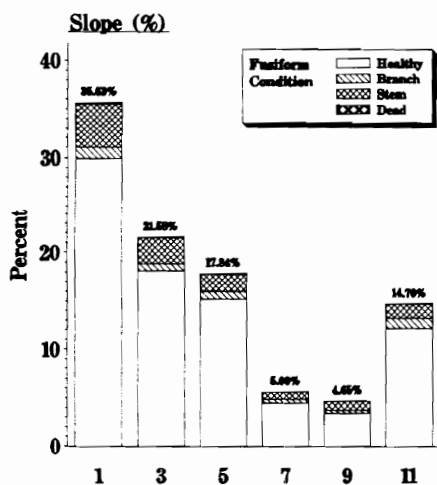
a)



b)



c)



d)

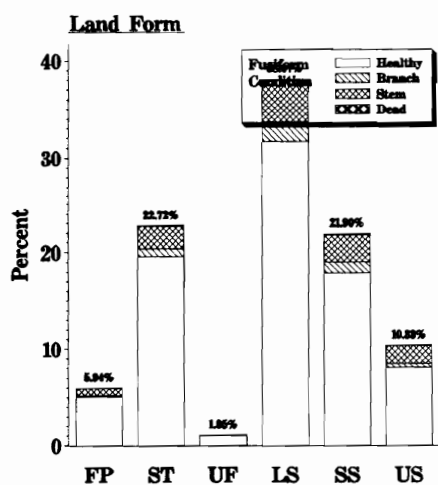


Figure 6.2 : Histograms of time—constant plot level covariates for loblolly pine development data.

6.1.2. Observed Condition Frequencies And Proportions

6.1.2.1. Box and whisker plots for time dependent covariates

Box and whisker plots of the time dependent plot characteristics are presented in Figures 6.3 and 6.4 for slash and loblolly pine, respectively. Tree characteristics are displayed in Figures 6.5 and 6.6. The boxes are drawn from the 25th (Q_1) to the 75th (Q_3) percentile of the empirical distribution of each covariate. The whiskers extend from the end points of the box to either the maximum and minimum observation or 1.5 times the inter-quartile-range ($Q_3 - Q_1$). The median is denoted with a box midline, the arithmetic mean by a filled triangle. Outliers, i.e. observations that would fall below or above the whiskers, are usually plotted as symbols. Owing to the large number of observations in the data sets only a few outliers would clutter the plots considerably. As an alternative the vertical axis is labeled to encompass the smallest and largest observations (the ticks are rounded upwards to simplify the display).

From Figures 6.3 and 6.4 one sees a general trend of the medians across condition code: for age, stand height, basal area pre acre, and quadratic mean diameter the medians increase as the health deteriorates. That is, on average healthy trees are younger, found in stands of lesser height, lower basal area and quadratic mean diameter. For relative spacing and trees per acre the direction is reversed.

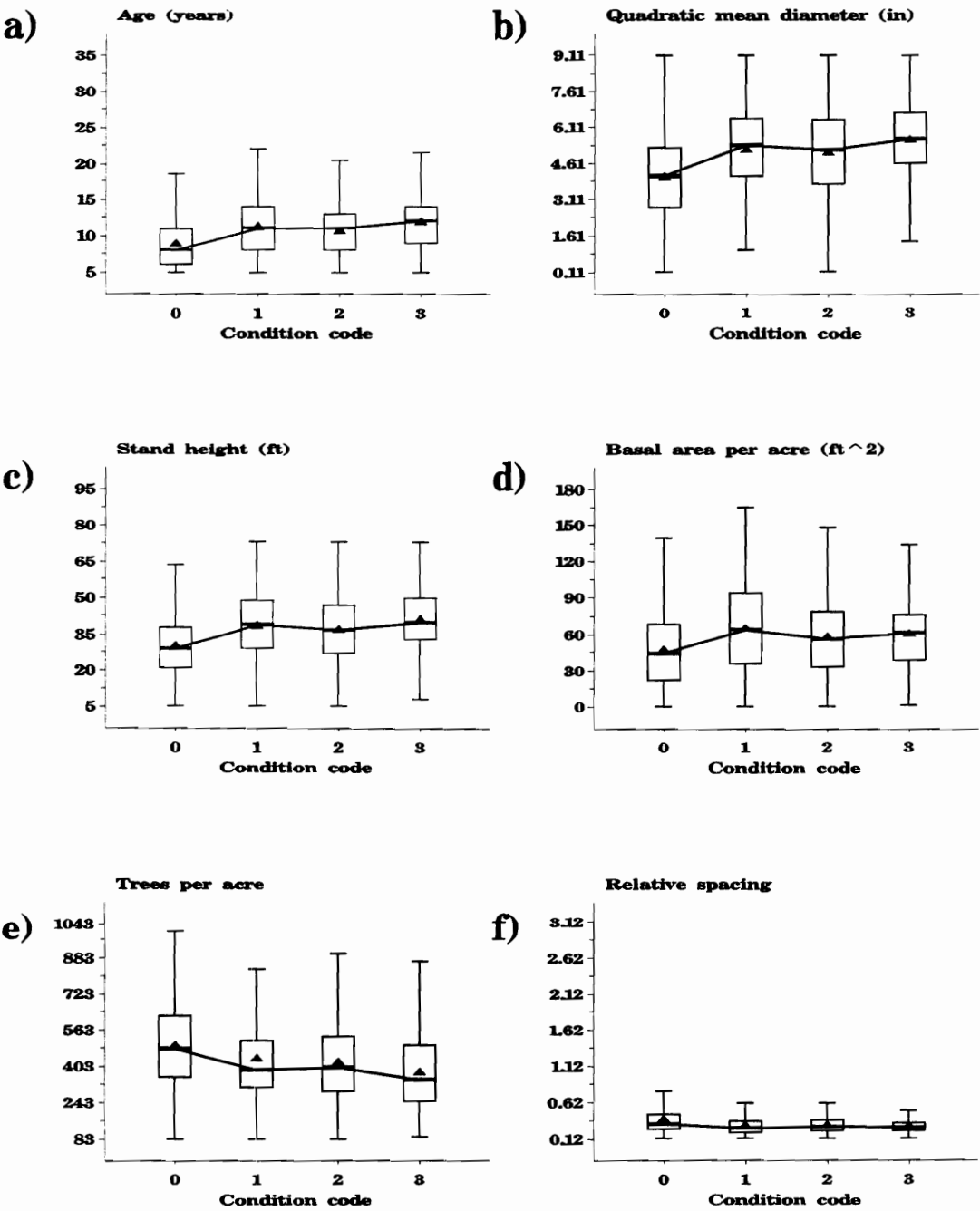


Figure 6.3 Box and whisker plots for plot level covariates against fusiform rust condition code for slash pine development data.

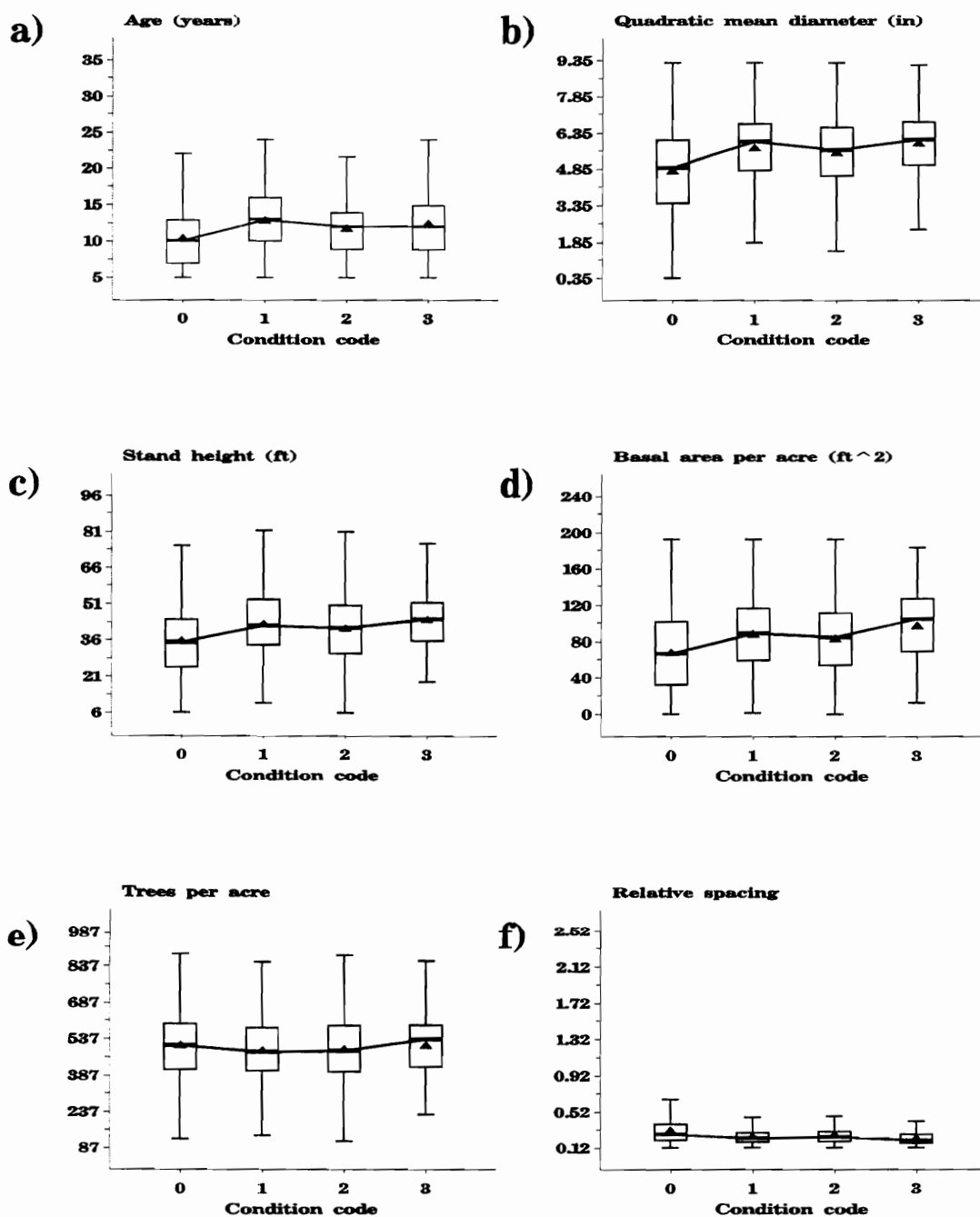


Figure 6.4 Box and whisker plots for plot level covariates against fusiform rust condition code for loblolly pine development data.

For both species, the branch infected state appears to be a little off the trend. One would expect the median age to increase continuously with condition code. Higher averages for condition 1 than for condition 2 are reasonable however, if branch infected trees remain branch infected and have a low transition probability to a stem infection and stem infections are arrived at with high likelihood directly from the healthy state. Addressing questions like these requires to focus on the transitional behavior rather than marginal frequencies (see 6.1.2.3.). For the tree level covariates in Figures 6.5 and 6.6 one notices an increase of the mean/median diameter at breast height and tree height to the branch infected state, then a decrease through states 2 and 3. Height to live crown displays an increasing trend of the medians over condition codes. For dead trees, height to live crown was assumed to equal tree height. Hence as the disease progresses trees are smaller and thinner on average with a smaller crown. The reason herefore may be that infected trees are retarding in growth compared to the healthier competitors. It could also be, that trees of lesser social status are more likely to become infected.

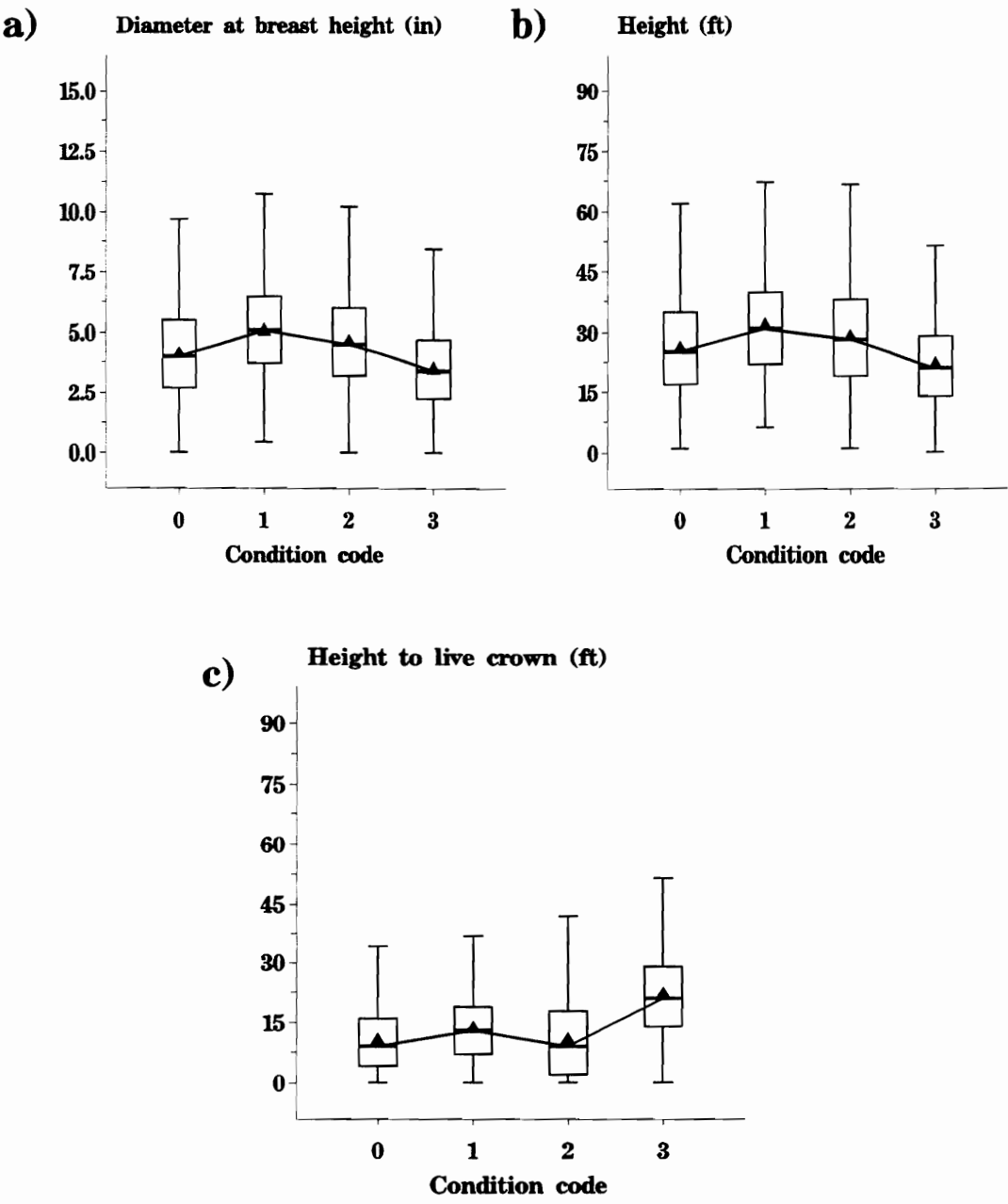


Figure 6.5 Box and whisker plots for tree level covariates against fusiform rust condition code for slash pine development data.

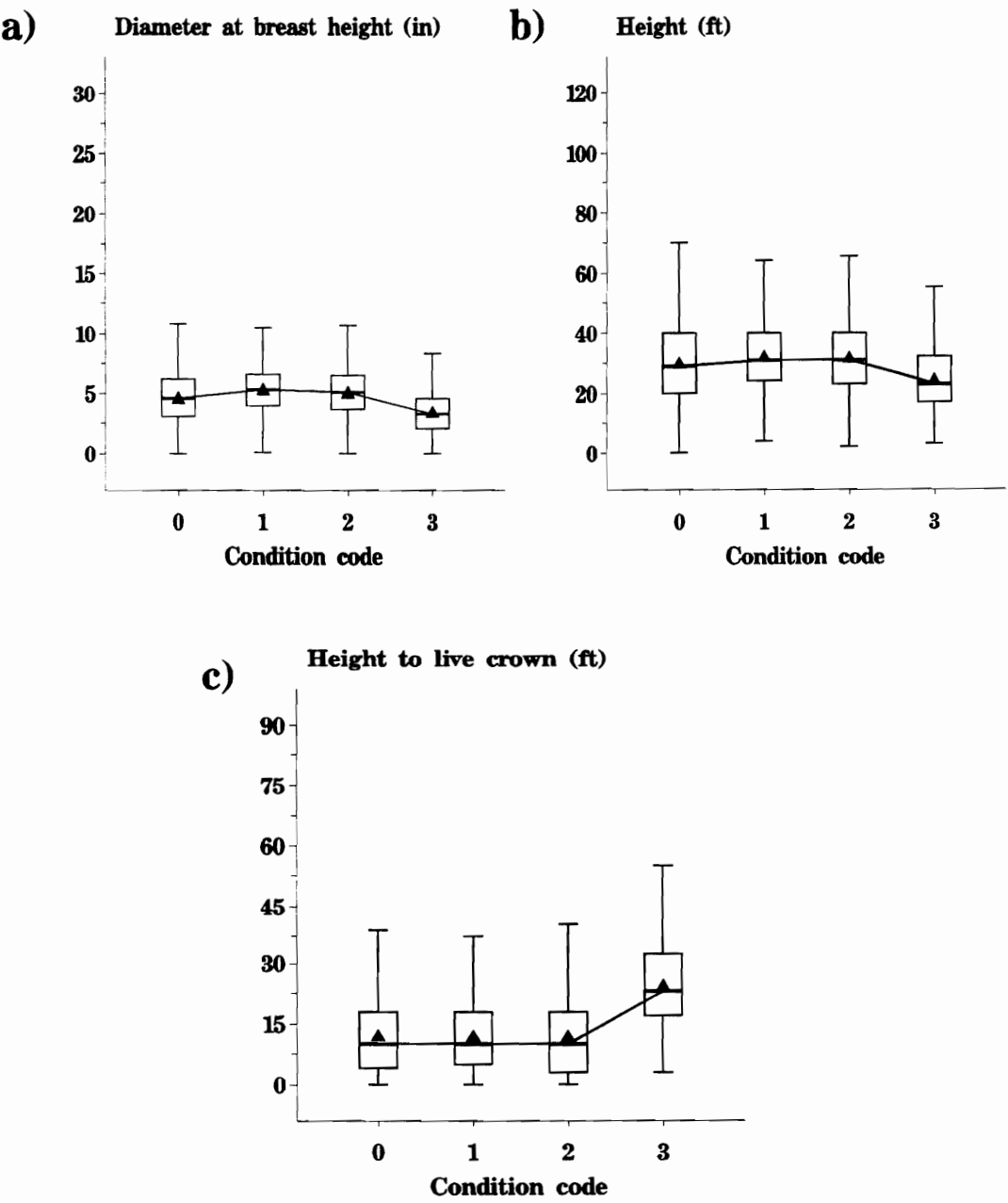


Figure 6.6 Box and whisker plots for tree level covariates against fusiform rust condition code for loblolly pine development data.

6.1.2.2. Marginal proportions by initial age classes

Table 6.3 lists the number of plots and the proportions of trees responding in the four categories by initial age groups for slash pine, Table 6.4 for loblolly pine.

At the initial measurement most of the plots were rather young. 51% of slash pine and 42% of loblolly pine plots were younger than 6 years at their first visit. Arabatzis (1990), who analyzed the same ETPPRP data, but with fewer remeasurements based analysis on plots that were initially at least 5 years, reducing the total number of observations considerably. The common argument is that stem infections at age 5 have been found to be especially significant for the further development of a tree (Wells and Dinus 1978). Since we are among other things concerned with behavior over time and five year old plots will eventually appear above this (questionable) threshold at later measurements, it is believed that the deletion of younger plots merely on the basis of this criterion is not justified. Hence, all plots have been retained in this analysis regardless of the initial age of the trees. The total numbers of observations in the four data sets are

Slash pine development data:	23,653
Slash pine evaluation data:	24,271
Loblolly pine development data:	59,125
Loblolly pine evaluation data:	58,584

Table 6.3.

Number of plots by initial age groups in slash pine development and evaluation data sets along with observed proportions of trees in the four rust categories.

Initial Age	Number of Plots	Observed proportion of responses in category			
		0	1	2	3
Development Data					
5	18	.601	.024	.334	.041
6	25	.556	.042	.371	.031
7	17	.388	.061	.487	.064
8	6	.456	.072	.432	.041
9	3	.283	.153	.545	.019
10	1	.073	.031	.756	.140
11	5	.313	.115	.509	.063
12	1	.481	.120	.380	.019
13	1	.270	.035	.605	.090
14	4	.218	.082	.641	.059
16	1	.254	.215	.523	.008
20	1	.154	.077	.769	.000
21	1	.273	.143	.584	.000
Evaluation Data					
5	18	.624	.028	.318	.030
6	25	.580	.038	.352	.030
7	17	.356	.056	.520	.068
8	6	.457	.067	.430	.046
9	3	.312	.176	.481	.030
10	1	.199	.088	.593	.119
11	5	.400	.129	.422	.049
12	1	.354	.194	.448	.004
13	1	.195	.160	.599	.046
14	4	.258	.149	.535	.058
16	1	.281	.267	.438	.014
20	1	.320	.020	.660	.000
21	1	.368	.039	.592	.000

Table 6.4.

Number of plots by initial age groups in loblolly pine development and evaluation data sets along with observed proportions of trees in the four rust categories.

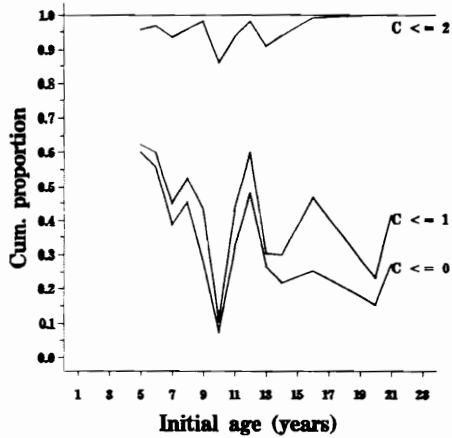
Initial Age	Number of Plots	Observed proportion of responses in category			
		0	1	2	3
Development Data					
5	41	.872	.021	.105	.003
6	38	.866	.031	.099	.004
7	42	.845	.032	.119	.004
8	15	.739	.087	.171	.004
9	5	.830	.055	.111	.003
10	10	.819	.070	.106	.005
11	10	.800	.058	.136	.006
12	11	.726	.087	.179	.008
13	3	.790	.055	.154	.002
14	1	.827	.055	.105	.014
15	5	.852	.060	.086	.002
16	2	.837	.078	.078	.007
17	1	.589	.168	.238	.005
18	1	.673	.038	.288	.000
Evaluation Data					
5	41	.864	.025	.108	.003
6	38	.863	.033	.100	.003
7	42	.827	.032	.137	.005
8	15	.782	.074	.139	.005
9	5	.812	.051	.133	.003
10	10	.808	.061	.127	.005
11	10	.782	.071	.139	.008
12	11	.793	.055	.142	.009
13	3	.844	.036	.119	.001
14	1	.844	.085	.067	.004
15	5	.879	.044	.074	.003
16	2	.793	.146	.061	.000
17	1	.712	.087	.168	.034
18	1	.625	.109	.266	.000

The observed proportions in Tables 6.3 and 6.4 are marginal frequency estimates for the plots in an initial age group and do not reveal much information about behavior over time. However, an interesting pattern emerges from both the slash pine and loblolly pine proportions by age groups. In Table 6.3 the healthy and stem infected trees appear with highest frequency. At young ages the proportion of healthy trees is greatest but decreases with age. At the same time the proportion of stem infections increases considerably and to a lesser degree the branch infections. The rust associated mortality seems to depend little on the initial age of the trees. While in principle the same pattern holds for the loblolly pine frequencies in Table 6.4, fewer trees are infected for this species across all initial age classes and the process appears less dynamic. The infection rates in slash pine are quite dramatic. Proportions of healthy trees as low as 20-30% like the ones in Table 6.3 for older plots may require total removal or shortened rotation periods and subsequent replanting with more resistant stock. It is clearly important to enable projections of the likely future infection rates at an early stage of the rotation period to keep economic damage at a minimum.

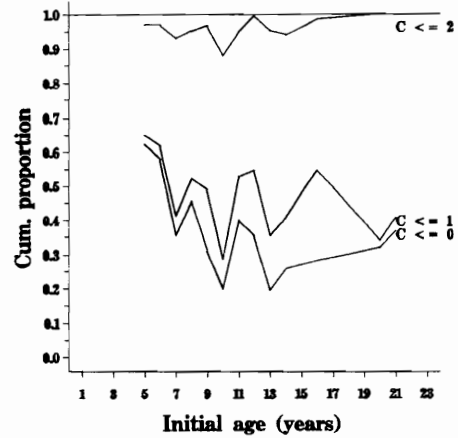
Figure 6.7 depicts the cumulative marginal proportions for the four data sets by initial age groups. The rapid decline of healthy trees in slash pine and the more stable, less dynamic pattern in loblolly pine are articulate. The cumulative proportions also show, that the likelihood to observe branch infected trees follows closely the frequency of healthy trees ($C \leq 0 \Rightarrow C = 0$). The peak around initial age 16 for slash pine is caused by the sparseness of observations in these groups.

Marginal Proportions in Slash Pine

a) Development data

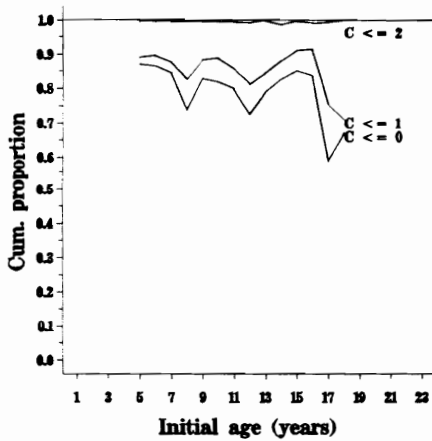


b) Evaluation data



Marginal Proportions in Loblolly Pine

c) Development data



d) Evaluation data

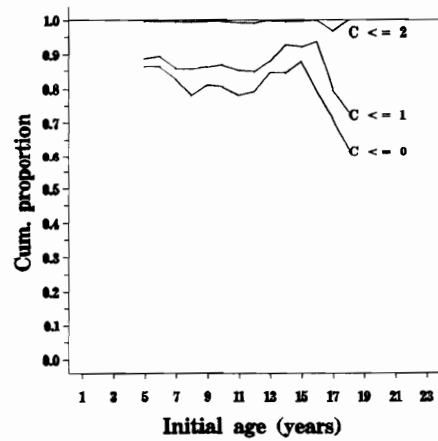


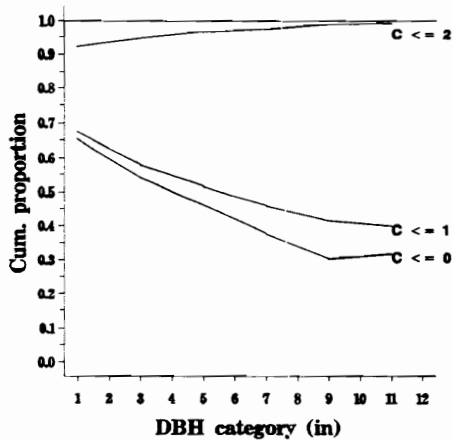
Figure 6.7. Observed marginal cumulative proportions in development and evaluation data sets by initial age.
C: 0 = healthy, 1 = branch, 2 = stem infection.

Figure 6.7 also shows the similarities between the development and evaluation data set. Based on the limited information presented so far, the evaluation data appears to be similar, but with challenges, two criteria for model evaluations to be useful.

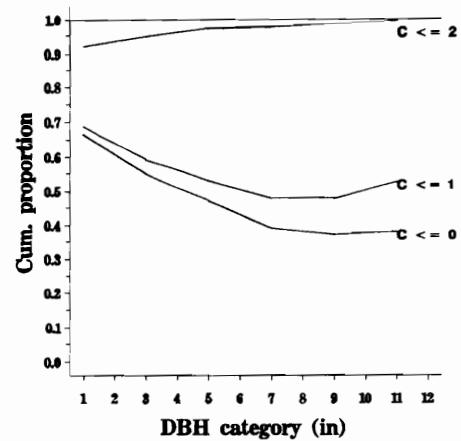
Age at the time of measurement in plantations is as much a plot characteristic as well as a tree characteristic. Dell and Driver (1963) investigated how tree diameter at breast height (DBH) is related to fusiform rust incidence. It has since been a matter of debate if tree characteristics are useful descriptors of rust status and development. With this argument in mind, the exploratory grouping procedure has been applied to tree DBH. Figure 6.8 shows the marginal proportions across all plots by 1 inch DBH classes. Since tree size is an age substitute to some degree, it is not surprising that the same type of decline of healthy trees surfaces as in Figure 6.7. The pattern does not appear as irregular as in Figure 6.7. however, and the branch infected trees do not just follow the trajectory of healthy trees, but increase their proportion as DBH increases. Dell and Driver (1963) also conducted a graphical analysis of marginal proportions by DBH for slash pine and found no difference between the diameters of healthy and diseased trees. From Figure 6.8 it appears however, that larger trees are associated with a higher disease rate than smaller trees and that this effect is less pronounced in loblolly pine. This was also found in Figures 6.5 and 6.6 where the mean/median DBH for branch and stem infected trees is larger than for the healthy trees. It is noted however, that these graphical analyses are of an exploratory nature and should aid in the formulation of reasonable model structures in the sequel. Concentrating on one characteristic alone presumes holding all other characteristics constant. For tree size related variables, this is of course unreasonable, as DBH, for example changes with age, stand density, etc.

Marginal Proportions in Slash Pine

a) Development data

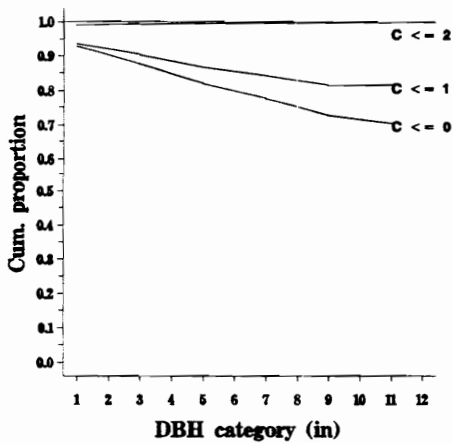


b) Evaluation data



Marginal Proportions in Loblolly Pine

c) Development data



d) Evaluation data

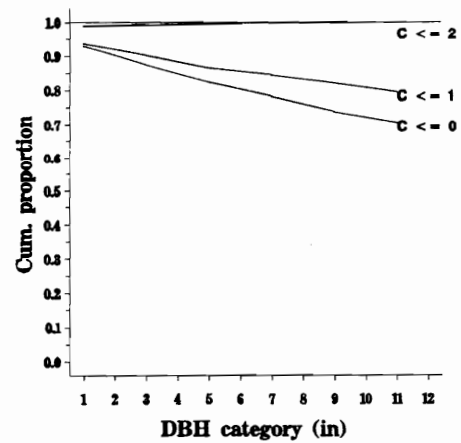
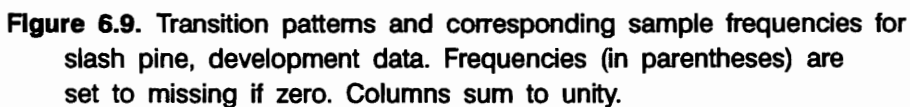


Figure 6.8. Observed marginal cumulative proportions in development and evaluation data sets by dbh classes.
C: 0 = healthy, 1 = branch, 2 = stem infection.

6.1.2.3. Transition patterns

6.1.2.3.1. Slash pine

The dynamic of the fusiform rust disease process over time can only be discussed through usage of transition probabilities. Given the unidirectional nature of the process and the four response categories, only a limited number of transitions is possible and not all 4^k patterns are realized for a transition over k occasions. Nevertheless, the number of transitions grows rapidly, as more occasions are taken into consideration. For the four measurement cycle data a maximum of three transitions is possible. Observed transition frequencies for slash pine (D data) are displayed in Figure 6.9. The numbers 0 through 3 denote the rust condition class. There are 10, 20, and 35 possible patterns over 1, 2, and 3 transitions, respectively, depicted as the vertical panels in Figure 6.9. The observed frequencies are arranged to sum to one for each transition panel. Thus over all plots and occasions, healthy trees (0) are most likely in the slash pine development data, followed by stem infections (2). Among the one-step transitions, the 0-0 and 2-2 pattern are most likely, accounting for more than $\frac{3}{4}$ of all one-step transitions. Among the two-step transitions, 0-0-0, and 2-2-2 again emerge as most frequent but 0-2-2 also has an 8% share. This suggests that once a tree is stem infected there is a high probability of remaining in this category for at least two more occasions, i.e. 6 years. Three-step occasions are most frequently of the 0-0-0-0 and 2-2-2-2 type. More than 50% of all three-step transitions are trees that were either healthy or stem infected throughout the entire period.



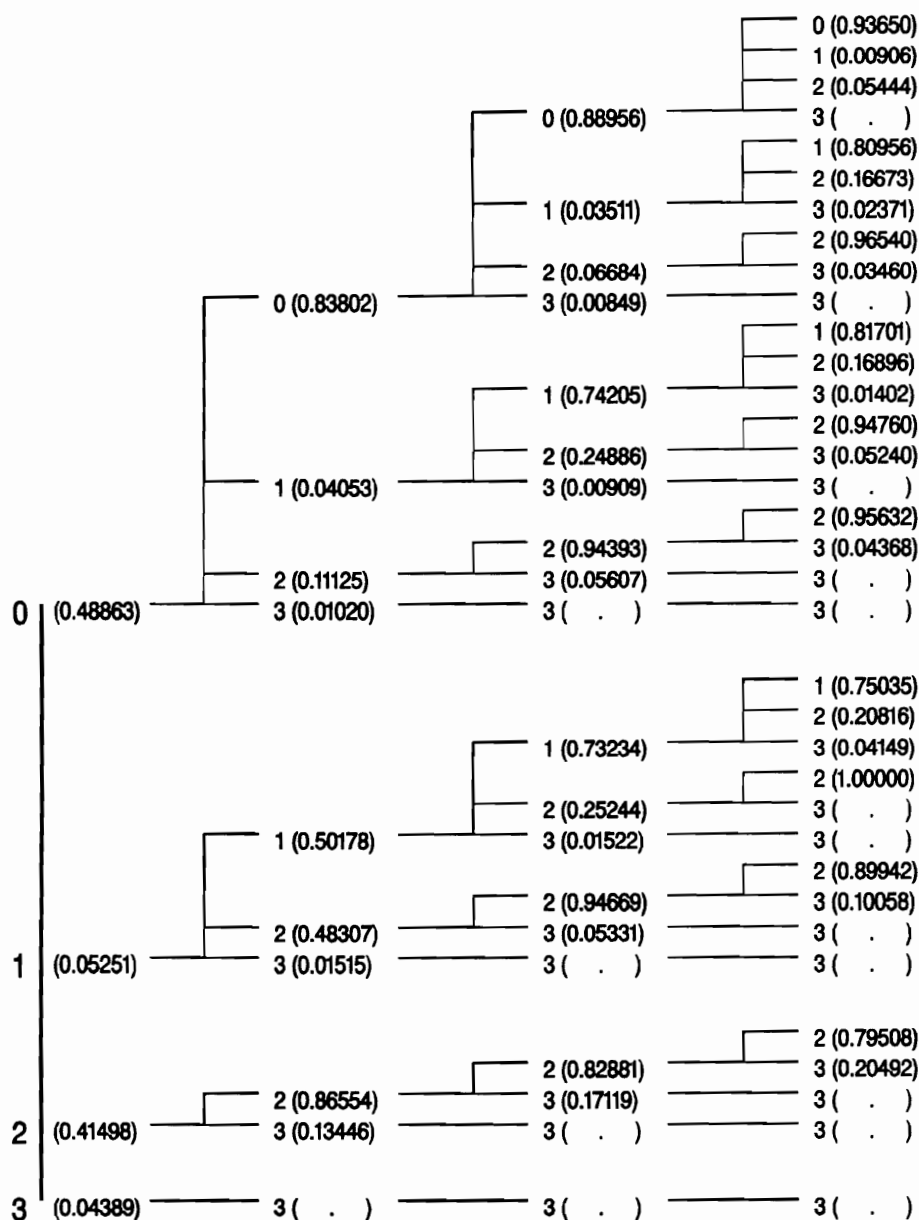


Figure 6.10. Transition patterns and corresponding sample frequencies for slash pine, development data. Frequencies are rendered conditional on the previous response.

Figure 6.10 depicts the same transition patterns for the slash pine D-data, but the observed frequencies are conditional on the immediately preceding outcome. That is, given a slash pine tree was initially healthy, there is a 83% chance that it remains healthy over 1 transition, and a 93% chance that it remains healthy through three occasion. Figure 6.10 also allows one to make comparisons of the following kind: whether a tree is initially clear of infections, branch or stem infected, once it arrives at condition 2, 2-2 transition probabilities are similar regardless of a different state in the past. This is not true for 1-2 transitions, however. It does not seem to make a difference whether a tree was healthy once or twice, a branch infected tree has approximately a 25% chance to be stem infected at the next occasion. For the initially branch infected trees this transition probability is almost twice as large (48%). The rust induced mortality is also of interest. 2-3 transitions are more likely if a tree is initially of poor condition. For example slash pines that were recorded as 0-0-2 have a 3% chance of rust induced death while trees that responded 2-2-2 die with 20% chance of fusiform rust. This phenomenon can be caused by multiple effects. One, the physiology of older trees may enable them to contain infected areas more efficiently; two, stem girdling by cankers and subsequent death is less likely if the stem has larger dimension at the time of infection and three, late infections allow less time for the canker to develop.

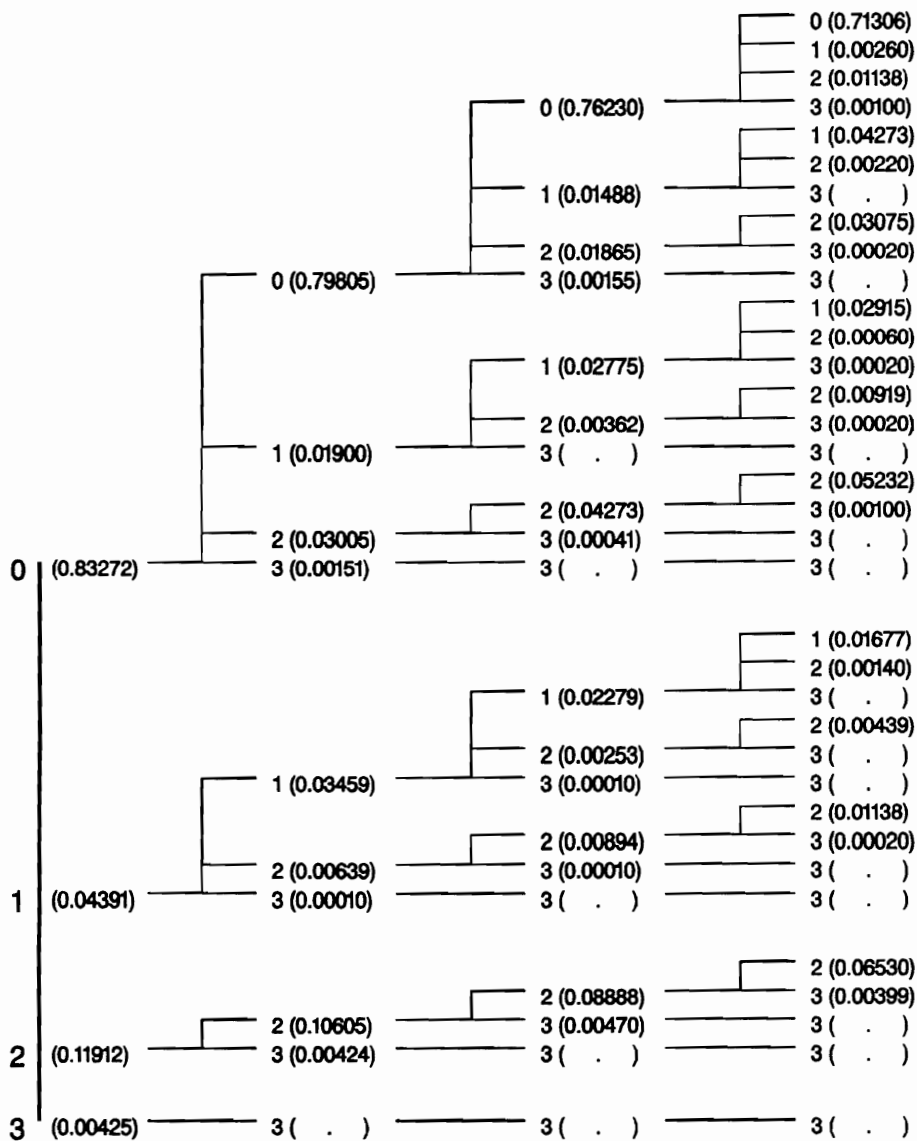


Figure 6.11. Transition patterns and corresponding sample frequencies for loblolly pine, development data. Frequencies (in parentheses) are set to missing if zero. Columns sum to unity.

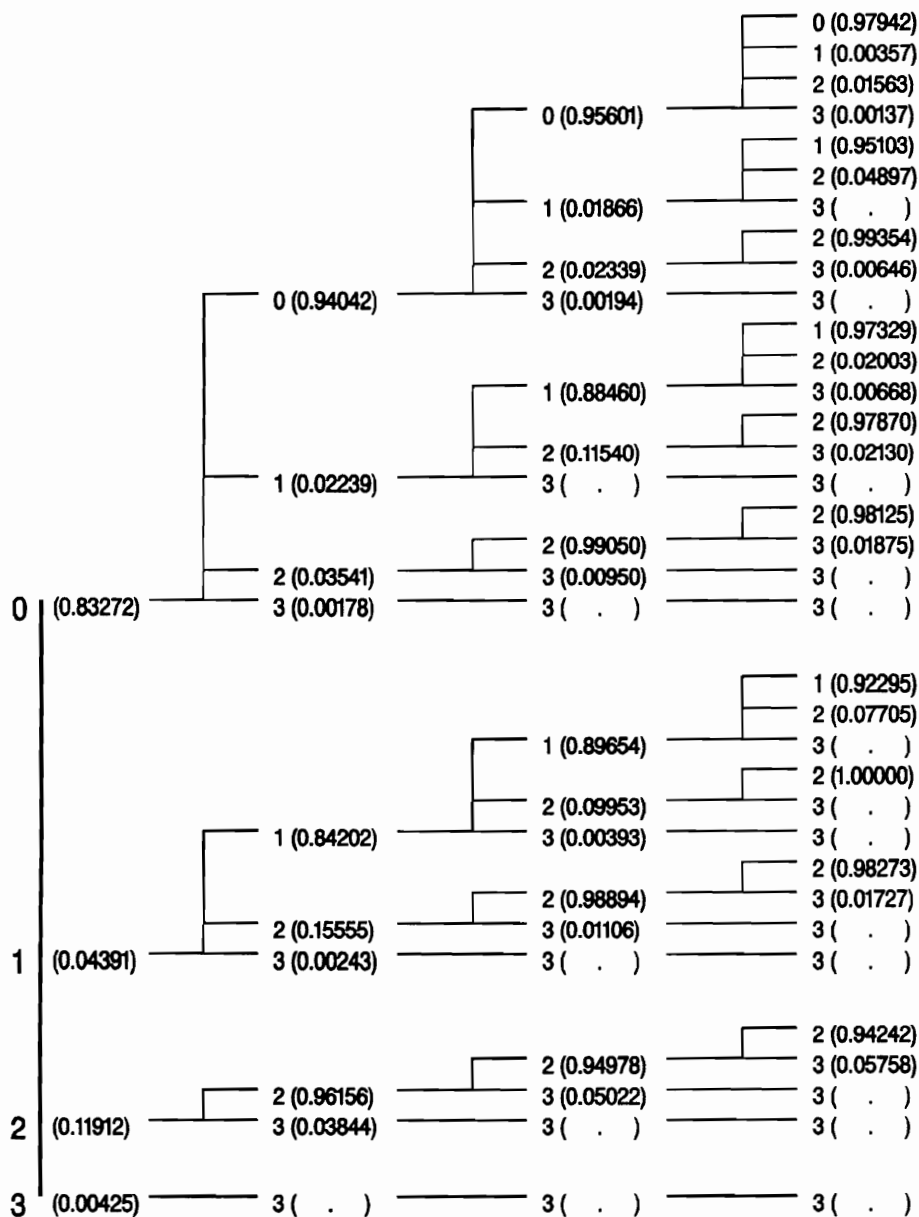


Figure 6.12. Transition patterns and corresponding sample frequencies for loblolly pine, development data. Frequencies are rendered conditional on the previous response.

6.1.2.3.2. Loblolly pine

Corresponding to Figures 6.9 and 6.10 for slash pine, 6.11 and 6.12 display the transitional behavior of loblolly pine based on the development data set. The dominating states of the system are 0, 0-0, 0-0-0, and 0-0-0-0 which make up 83% of the cross-sectional observations, 79% of the one-step transitions, 76% of the two-step transitions, and 71% of the three-step transitions, respectively. Transitions into stem infected states come in second place but are much less frequent. While slash pine overall appears to tend to stem infections, loblolly pine tends to stay healthy. The conditional frequencies in Figure 6.12 show how much less dynamic loblolly pine stands in the ETPPRP are with respect to fusiform rust. The probabilities to remain in the previous state (stayers) are at least 94% for healthy trees, 84% for branch infected trees and 96% for stem infected trees. Only a small portion of trees advance in a more severe disease category between occasions. But even for these eventual movers are the subsequent transition probabilities hardly different from those arising for trees that stay in their category throughout. For example, only 2% of the initially healthy loblolly pines advance to a branch infection in one step. But of those that do, 88% remain branch infected at the next step, comparable to the trees that make a 1-1 transition at the first or the second occasion.

6.1.2.4. One-step transitions by age and dbh-classes

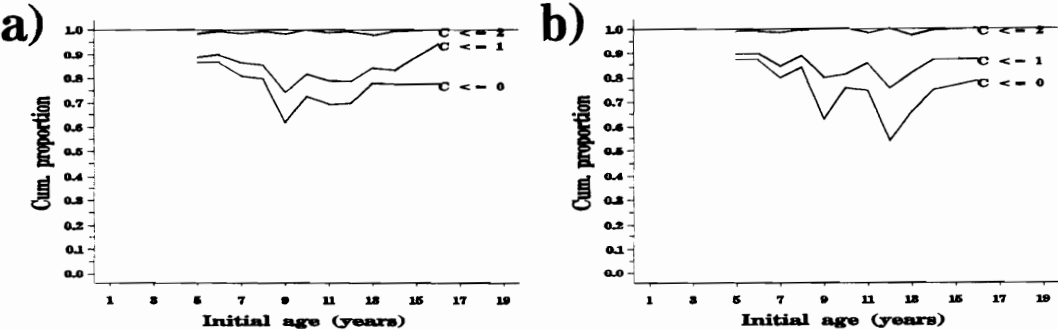
The more transition patterns included in an analysis of the dynamic behavior the more informative the analysis in principle. However, the number of possible patterns

makes modeling efforts of more than one-step transitions cumbersome. Even for only ten possible transitions, the number of model parameters becomes quite large. Figures 6.9 through 6.12 also indicate that the probabilities for staying in a certain category are quite stable and suggest an analysis of transitional behavior based on one-step transitions, i.e. conditioning on the directly preceding outcome only. In this section some exploratory graphs of one-step transition probabilities are presented. In Figure 6.13. cumulative transitions are plotted for the slash pine data sets by age classes. With the exception of previously branch infected trees the probability to stay in the previous category ($C \leq 0$ in 6.13a) and b), $C \leq 1$ in 6.13c) and d), $C \leq 2$ in 6.13e) and f)) do not vary much over different initial ages. The probability to remain healthy stays around 75%. Initial age of 5 years does not appear to be of much significance, although it appears for slash pine that trees on plots older than age 5 have a declining probability to remain healthy. This trend is weak however. Panels 6.13e) and f) show that almost regardless of the initial age a constant proportion of about 10% of the stem infected trees die due to their fusiform rust condition between measurement occasions. The most dynamic panels are 6.13c) and d), displaying transitions from the branch infected state. The probability to remain in this state declines sharply as initial age of the trees increases. After age 8, it drops sharply and an increasing portion of the branch infected trees become stem infected. Overall only a few trees advance more than one category between occasions, independent of the initial age. Consequently, before a tree dies it typically will be found during at least one measurement cycle in the stem infected state. This underlines stem girdling through growing cankers and stem breakage at cankers as important causes of rust induced death.

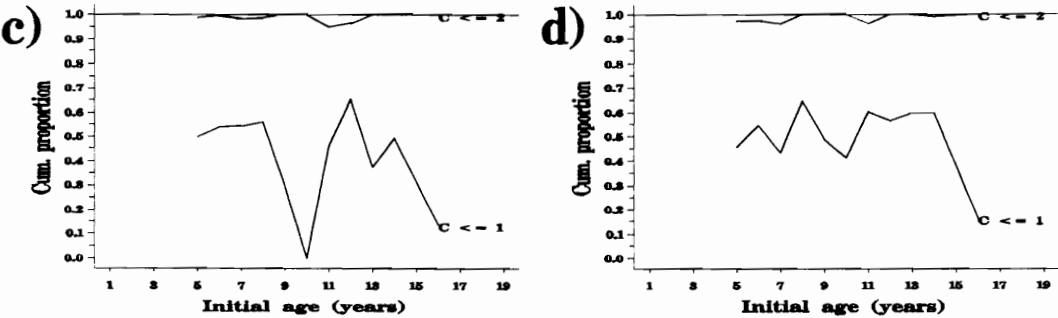
Development data

Evaluation data

Previous state: healthy (0)



Previous state: branch infected (1)



Previous state: stem infected (2)

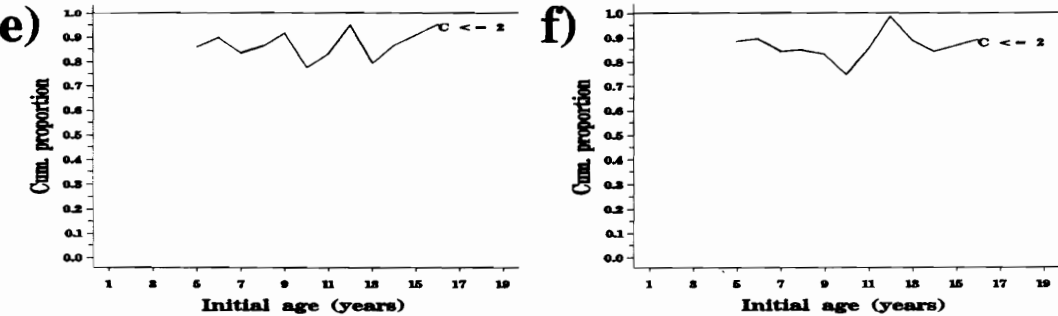
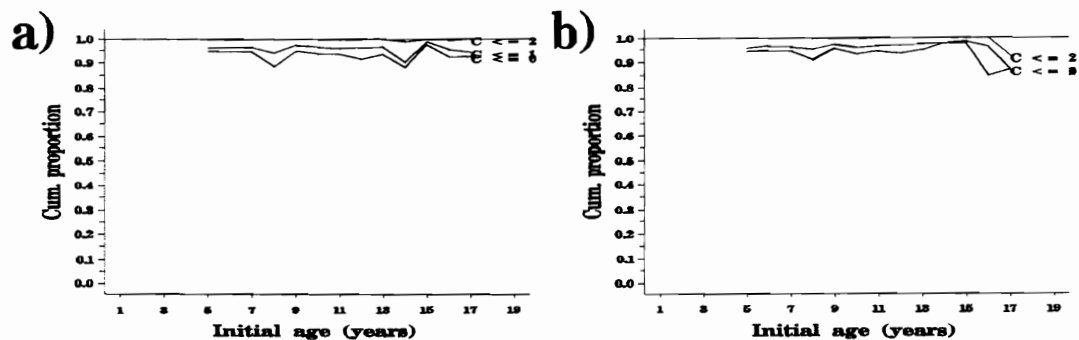


Figure 6.13. Observed cumulative transition proportions in slash pine data sets by initial age classes.
C: 0 = healthy, 1 = branch, 2 = stem infection.

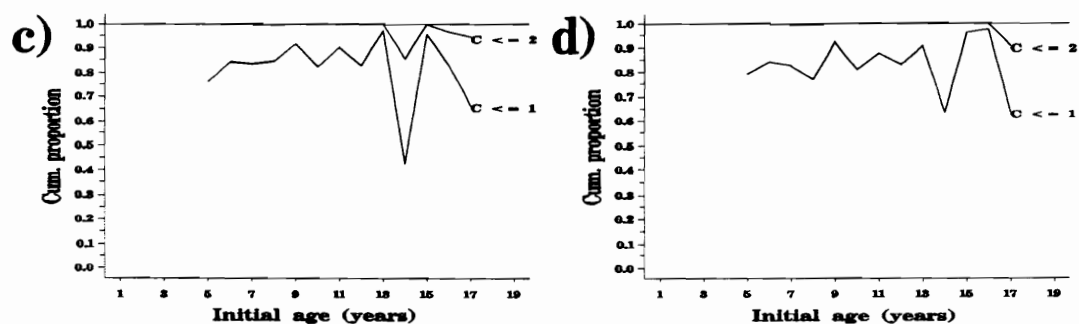
Development data

Evaluation data

Previous state: healthy (0)



Previous state: branch infected (1)



Previous state: stem infected (2)

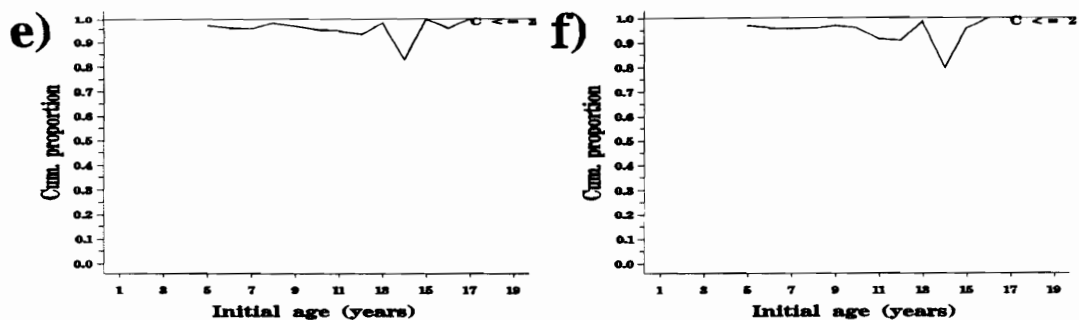


Figure 6.14. Observed cumulative transition proportions in loblolly pine data sets by initial age classes.
C: 0 = healthy, 1 = branch, 2 = stem infection.

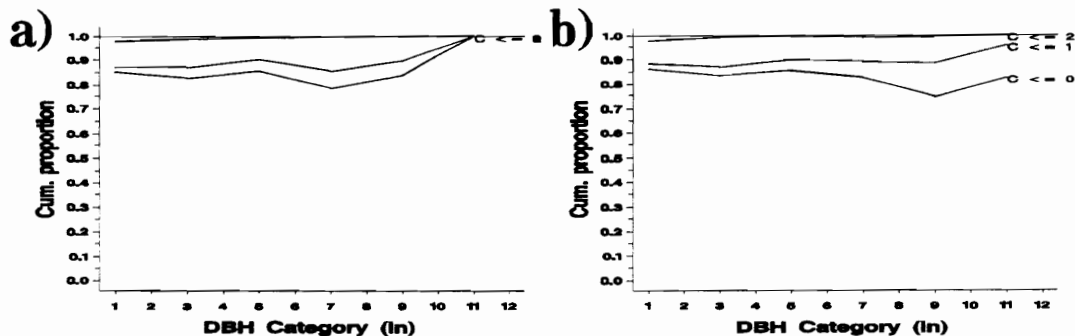
Figure 6.14 contains similar information for loblolly pine data sets. The peak at age 14 in all panels is caused by the sparseness of the data sets. Only one plot falls in this age class (see Table 6.4). Aside from this observation the panels display more stability than the slash pine data. The proportion of dying trees that were stem infected decreases from 10% in slash pine to 3-5 % in loblolly pine. Again, age 5 does not appear to be of particular importance and well on the general trends so that excluding younger observations from the data seems not justifiable.

Diameter at breast height is of some importance as a tree size attribute as well as a time substitute. Intuitively one would expect similar profiles of the transitions by dbh classes as displayed in Figures 6.13 and 6.14. These profiles are plotted in Figures 6.15 and 6.16 for both species. The trajectories are much smoother and regular than the plots by initial age classes. The proportion of stem infected trees that die is larger for small trees and smaller for larger trees. This differs from the profile by ages where this proportion appeared almost constant. The branch infected slash pines are more likely to advance to a stem infection, the smaller their DBH. The reverse relationship was found by age groups. It is obvious that there are confounding factors masking the actual effects. The DBH effect may be confounded by stand density and/or age effects. Comparison of these figures should prevent one from over interpreting the dynamics in either species based on a single variable only.

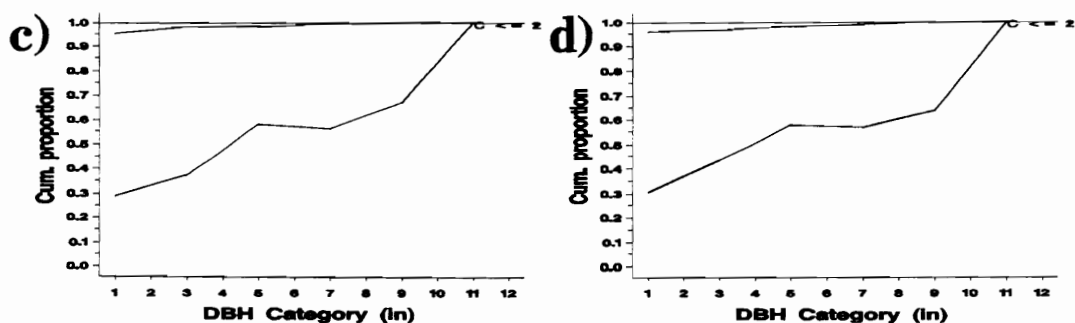
Development data

Evaluation data

Previous state: healthy (0)



Previous state: branch infected (1)



Previous state: stem infected (2)

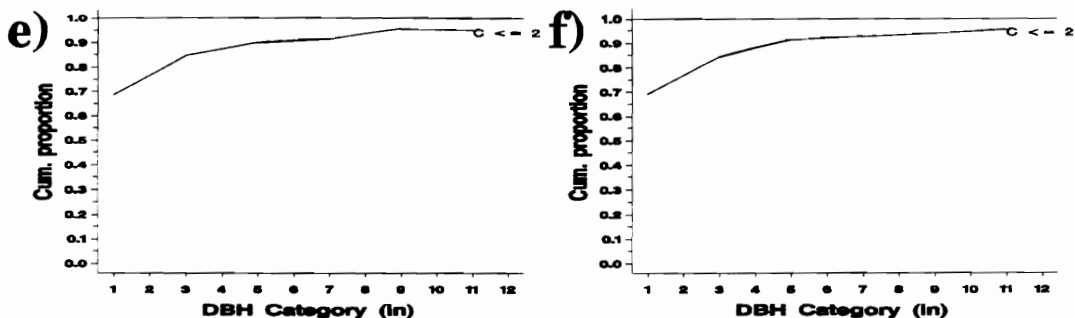
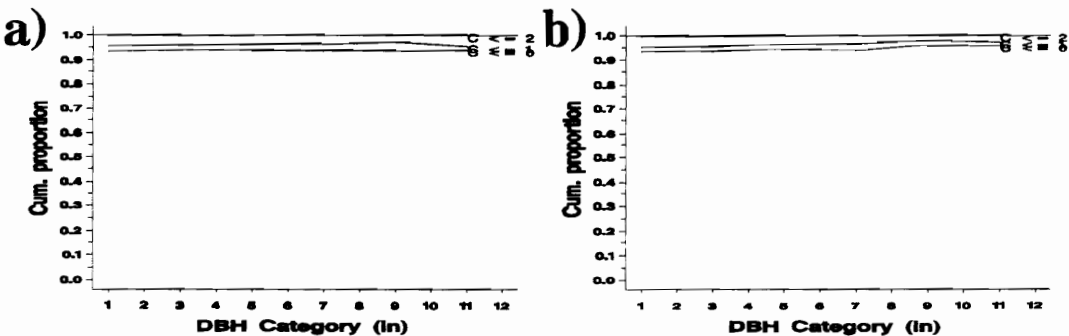


Figure 6.15. Observed cumulative transition proportions in slash pine data sets by (2 in) DBH classes.
C: 0 = healthy, 1 = branch, 2 = stem infection.

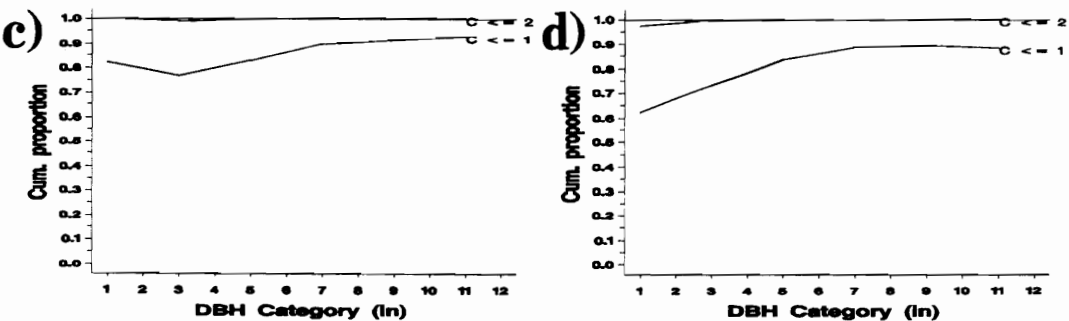
Development data

Evaluation data

Previous state: healthy (0)



Previous state: branch infected (1)



Previous state: stem infected (2)

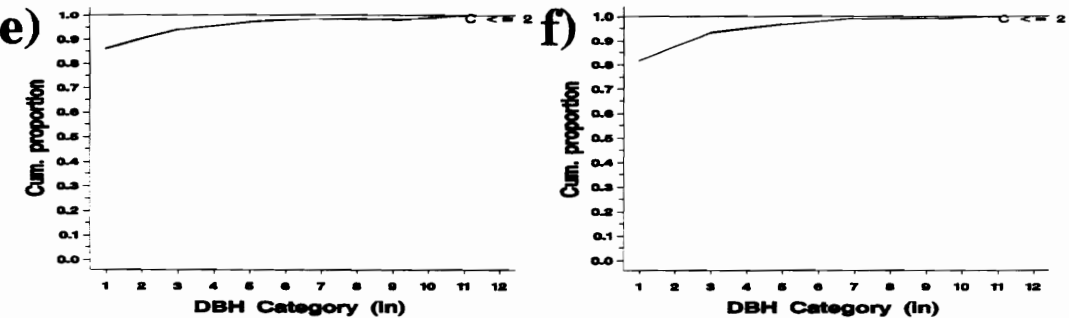


Figure 6.16. Observed cumulative transition proportions in loblolly pine data sets by (2 in) DBH classes.
C: 0 = healthy, 1 = branch, 2 = stem infection.

Chapter 7

Analysis

7.1. Slash Pine

7.1.1. Marginal Plot Models

In this section models are presented to predict the proportion of trees in any fusiform rust category for the entire population (PA) or a plot of trees (SS). Of the many estimation approaches discussed in § 4, GEE0 will serve as means to address the

population averaged problem, mixed models fit via semi-parametric methods have been chosen to yield a subject-specific representation.

The data in the ETPPRP was collected on a tree basis, the unit of interest in this section are plots, however. The tree data has been grouped for this purpose, collecting the number of trees in any category as counts (see § 4.5). Consequently, only plot level covariates will serve as regressors (see Table 6.2). The mean model in all cases will be a proportional odds model.

In § 5 the fusiform rust disease process has been explained in detail and it is obvious that none of the plot level covariates is directly related to the fusiform rust development cycle (see Figure 5.1). The plot level model however is intended to serve as a predictive device, allowing one to relate fusiform rust incidence to typical Forest Inventory data. For models describing biological behavior such as the present ones, it is desirable to make sense based on our knowledge of the biological process under scrutiny. A model displaying satisfactory goodness of fit statistics (see below) alone is not acceptable if the deterministic part of it does not match the actual or anticipated biological relationships. Variable selection will thus not be entirely governed by statistical considerations but also by biological meaningfulness. The analysis of model performance in numerical terms is hampered by (i) the categorical nature of the response, (ii) the multivariate character of the response and (iii) the semi-parametric nature of many approaches to statistical estimation. While measures of model performance are possible and will be discussed in § 7.1.1.1, considerable emphasis in evaluating the predictive capabilities of the models is placed on graphical methods. Several new types of displays will be introduced.

7.1.1.1. Measuring Goodness Of Fit

In the process of identifying parsimonious models that fit well, goodness-of-fit (GOF) measures are typically of great importance. Likelihood estimation is the principle of choice for most generalized linear models if the observations are independent, yielding a suite of possible GOF and diagnostic measures. For example information criteria such as the evaluated likelihood itself, Akaike's information criterion, Schwartz' information criterion, the model deviance, etc. (Bozdogan 1987, McCullagh and Nelder 1989). If semi-parametric routes to estimation are chosen, these measures are not available. Employing Generalized Estimating Equations in the present context thus does not lead to these classical measures, unless one adopts the incorrect assumption of uncorrelated observations, since the procedure then is exact Maximum Likelihood. For subject-specific models the same is true, unless the approximate parametric or fully parametric estimation techniques are chosen. In the approximate parametric approach of § 4.3.2 the resulting measures are of course only as good as the approximations involved.

For these reasons, measures are chosen, that do not depend on any underlying distributional assumptions. Candidates are various forms of residual sums of squares, a great variety of which is discussed in Amemiya (1981). For both, grouped and ungrouped cases, the response vector for subject i is denoted $\mathbf{y}_i$, but it is understood that $\mathbf{y}_i$ consists of proportions in the former and indicators in the latter. The following expressions apply regardless of how the response is coded.

The simplest measure of closeness between observed and predicted values is a straight residual sum of squares:

$$SSR = \sum_{i=1}^K \left(\mathbf{y}_i - \hat{E}(\mathbf{Y}_i) \right)' \left(\mathbf{y}_i - \hat{E}(\mathbf{Y}_i) \right). \quad [7.1]$$

Since the elements of $\mathbf{y}_i$ are correlated and heteroscedastic, a natural improvement of [7.1] is achieved by introducing weights, leading to the weighted sums of squares

$$WSS = \sum_{i=1}^K \left(\mathbf{y}_i - \hat{E}(\mathbf{Y}_i) \right)' \hat{\mathbf{V}}_i^{-1} \left(\mathbf{y}_i - \hat{E}(\mathbf{Y}_i) \right). \quad [7.2]$$

For purposes of comparing observed and predicted values across different estimation methods, [7.2] is not optimal, however. The weight matrix $\mathbf{V}_i$ introduces a scale shift, that depends on the estimation method alone, and not on the quality of fit. When for example, fitting a POM by maximum likelihood and a n -dependent GEE0 approach, the weight matrices will be entirely different. WSS is then not a reasonable measure to compare quality of fit among competing models. Amemiya presents several measures that capture quality of fit regardless of the scale. A probability weighted sums of squares for a J -category response is

$$WSP = \sum_{i=1}^K \sum_{t=1}^{n_i} \sum_{j=1}^J \frac{\left(y_{itj} - \hat{E}(y_{itj}) \right)^2}{\hat{E}(y_{itj})}. \quad [7.3]$$

For grouped data WSP can also be weighted by the observed frequency counts c_{it} on plot i at time t yielding

$$WSF = \sum_{i=1}^K \sum_{t=1}^{n_i} c_{it} \sum_{j=1}^J \frac{\left(y_{itj}^* - \hat{E}(y_{itj}) \right)^2}{\hat{E}(y_{itj})}. \quad [7.4]$$

where y_{itj}^* is the proportion of trees in category j at time t on plot i . WSF is identical to WSS if the repeated measurements are independent. Similarly to WSP, SSR can be weighted by the frequency counts:

$$SSF = \sum_{i=1}^K \sum_{t=1}^{n_i} c_{it} \sum_{j=1}^J \left(y_{itj} - \hat{E}(y_{itj}) \right)^2. \quad [7.5]$$

Although SSR and SSF do not account for the heteroscedasticity of the responses, it has been put forth because of analogies to linear model theory (Efron 1978). Especially it allows to compute measures that resemble coefficients of determination. Amemiya (1981) calls this statistic Efron's R^2 :

$$R_p^2 = 1 - \frac{SSF}{\sum_{i=1}^K \sum_{t=1}^{n_i} \sum_{j=1}^J \left(y_{itj}^* - \bar{y}_{..}^* \right)^2}. \quad [7.6]$$

We use the subscript p to indicate that this is a pseudo R^2 measure. R_p^2 approximately measures the extent of explained variability by the model. In likelihood analysis, another R^2 like measure is the scaled increase of the likelihood compared to a model without covariates. While this measure is constrained between 0 and 1, intermediate values do not have a clearly defined interpretation (Arabatzi 1990). If R_p^2 is applied to ungrouped data, where the responses are indicators, one should not expect values that exceed 0.3 - 0.4, simply because a discrete value and a continuous expectation can only be mapped to a certain degree.

In mixed models additional prediction criteria can be used. A penalized sums of squares criterion has been used repeatedly for continuous responses (Green 1987,

Lindstrom and Bates 1990), and can be adapted similarly.

$$PSS = WSS + \sum_{i=1}^K \hat{\mathbf{b}}_i' \hat{\mathbf{D}}^{-1} \hat{\mathbf{b}}_i \quad [7.7]$$

PSS measures deviations between observed and predicted responses but penalizes for the uncertainty in the random coefficients. Schabenberger (1995b) derived PSS from a Bayesian point of view. But more importantly than penalizing for the random effects variability, the subject specific expectations $\hat{E}(\mathbf{Y}_i | \mathbf{b}_i)$ can be used to draw comparisons between observed and predicted values, accounting for the differences between a subject's and the population's behavior. The sums of squares statistics corresponding to [7.1] - [7.6] when accounting for the BLUPs are

$$SSR_b = \sum_{i=1}^K \left(\mathbf{y}_i - \hat{E}(\mathbf{Y}_i | \mathbf{b}_i) \right)' \left(\mathbf{y}_i - \hat{E}(\mathbf{Y}_i | \mathbf{b}_i) \right), \quad [7.8]$$

$$WSS_b = \sum_{i=1}^K \left(\mathbf{y}_i - \hat{E}(\mathbf{Y}_i | \mathbf{b}_i) \right)' \hat{\mathbf{V}}_i^{-1} \left(\mathbf{y}_i - \hat{E}(\mathbf{Y}_i | \mathbf{b}_i) \right), \quad [7.9]$$

$$WSP_b = \sum_{i=1}^K \sum_{t=1}^{n_i} \sum_{j=1}^J \frac{\left(y_{itj} - \hat{E}(y_{itj} | \mathbf{b}_i) \right)^2}{\hat{E}(y_{itj} | \mathbf{b}_i)}, \quad [7.10]$$

$$WSF_b = \sum_{i=1}^K \sum_{t=1}^{n_i} c_{it} \sum_{j=1}^J \frac{\left(y_{itj}^* - \hat{E}(y_{itj} | \mathbf{b}_i) \right)^2}{\hat{E}(y_{itj} | \mathbf{b}_i)}, \quad [7.11]$$

$$SSF_b = \sum_{i=1}^K \sum_{t=1}^{n_i} c_{it} \sum_{j=1}^J \left(y_{itj} - \hat{E}(y_{itj} | \mathbf{b}_i) \right). \quad [7.12]$$

$$R_{pb}^2 = 1 - \frac{SSF_b}{\sum_{i=1}^K \sum_{t=1}^{n_i} \sum_{j=1}^J (y_{itj}^* - \bar{y}_{..}^*)^2}. \quad [7.13]$$

The subscript b will be used to denote criteria that are based on the conditional expectations $\hat{E}(\mathbf{Y}_i | \mathbf{b}_i)$. Comparing [7.3] - [7.6] to [7.10] - [7.13] gives insight how much can be gained in terms of predictive power by making a model subject-specific rather than population averaged. Table 7.1 summarizes these GOF statistics and highlights their utility.

To compare models of the same size, Thall and Vail's (1990) design based measure $H = \log | \hat{\text{var}}(\hat{\theta})^{-1} |$ is also of some help, where θ stands for the entire parameter vector being estimated. The model with higher H does not necessarily yield the smaller residual sums of squares. H measures can be exploited best by comparing the H value based on the model variance estimator [4.4] with the H value based on the robust estimator [4.5]. Barring influential outliers, closeness of these two measures indicates that the correlations have been modeled properly. Placing too much emphasis on this statistic is not recommended, however, since it is not quite clear how robust the *robust* estimator really is. Differences between [4.4] and [4.5] are not only caused by modeling correlations incorrectly, but also by influential data points.

In the present study where an independent evaluation data set is available, it is useful, to compute all candidates of GOF measures for both data sets. If little in GOF is lost by applying the derived model to different data, a model can be assumed not to be tailored too closely to the development data. This helps to avoid excessive inclusion of covariates.

Graphical displays will play a major role in assessing a models performance. For the ordinal responses, where four observations and responses need to be tracked for each subject and point in time, typical residual plots are not helpful. A marginal model with 84 subjects such as the slash pine data set, requires identification of over 670 points without separating time points. To avoid excessive busyness of display, ray charts will be used instead. These are circular displays, where each subject's observations and predictions are plotted along a single ray. The rays are arranged in a circular fashion. Figure 7.1 depicts a single ray schematically. The entire ray is scaled from 0 to 100% frequency. The subject number is printed at the end of the ray. Arrows indicate cumulative proportions along the ray, predicted frequencies are oriented counter clockwise, observed frequencies clockwise. The actual category probabilities are obtained as differences between the arrows.

For each subject (plot) a series of single rays are concatenated corresponding to times of measurements or transitions. Concentric circles delimit between the measurement points.

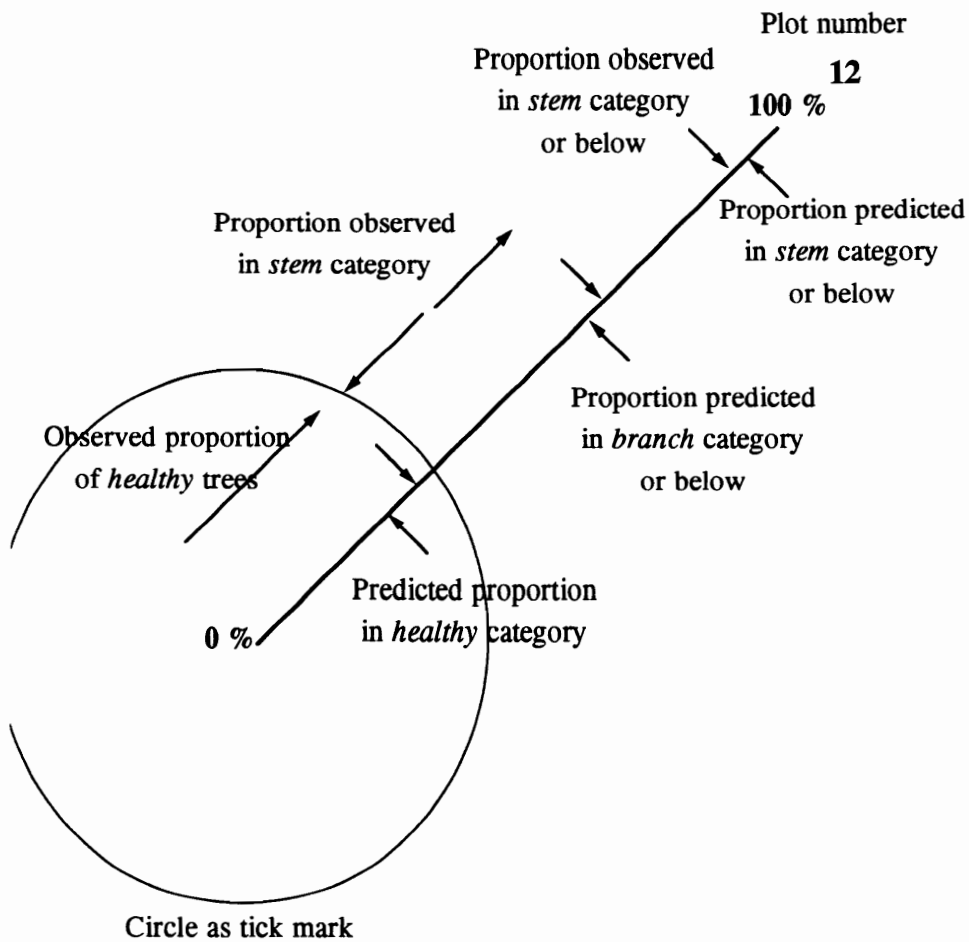


Figure 7.1. Schematic display of ray arrangement as used in the circular prediction plots for marginal ordinal response models.

Table 7.1.
Sums of squares and related criteria and their utility for multivariate
categorical regression models.

Statistic	Formula	Model comparisons		SS models only	Grouped data
		for a given estim. method	across estim. methods		
SSR	7.1		x		
WSS	7.2	x			
WSP	7.3		x		
WSF	7.4		x		x
SSF	7.5		x		x
R_p^2	7.6	x	x		
PSS	7.7	x		x	
SSR_b	7.8		x	x	
WSS_b	7.9	x		x	
WSP_b	7.10		x	x	
WSF_b	7.11		x	x	x
SSF_b	7.12		x	x	x
R_{pb}^2	7.13	x	x	x	

7.1.1.2. Biological patterns

Coefficient interpretation in a proportional odds model is made difficult by the non-linearity of the logistic response profiles. Whether a negative coefficient actually decreases or increases a particular category probability depends on the value of the linear predictor unless all other covariates are held constant (see Figure 3.1). As a precaution, one should not try to infer bearing of a covariate in a cumulative link model based on its sign. A more reasonable approach to see how plot level covariates relate to the fusiform rust incidence rates for the entire sample, is to fit a series of single regressor proportional odds models. GEE0 with a 1-dependence correlation structure has been used here. Figures 7.2 and 7.3 depict how the probabilities in any category change with some of the covariates over the range of values observed in the development data. Stand height, Age, Site index, and Quadratic mean diameter display the same prediction profile. As these covariates increase in value, the proportion of healthy trees declines, the stem infected trees are increasing in frequency correspondingly. A reverse relationship holds for Relative spacing. Branch infections seem to decrease/increase in the same direction as stem infections, the relationship seems to be almost linear, however. The proportion of dead trees is homogeneous for most covariates, except Age, Stand height and Quadratic mean diameter. Terrain Slope and Aspect (not plotted) did not serve as predictors by themselves. The predictions are flat across the range of these covariates.

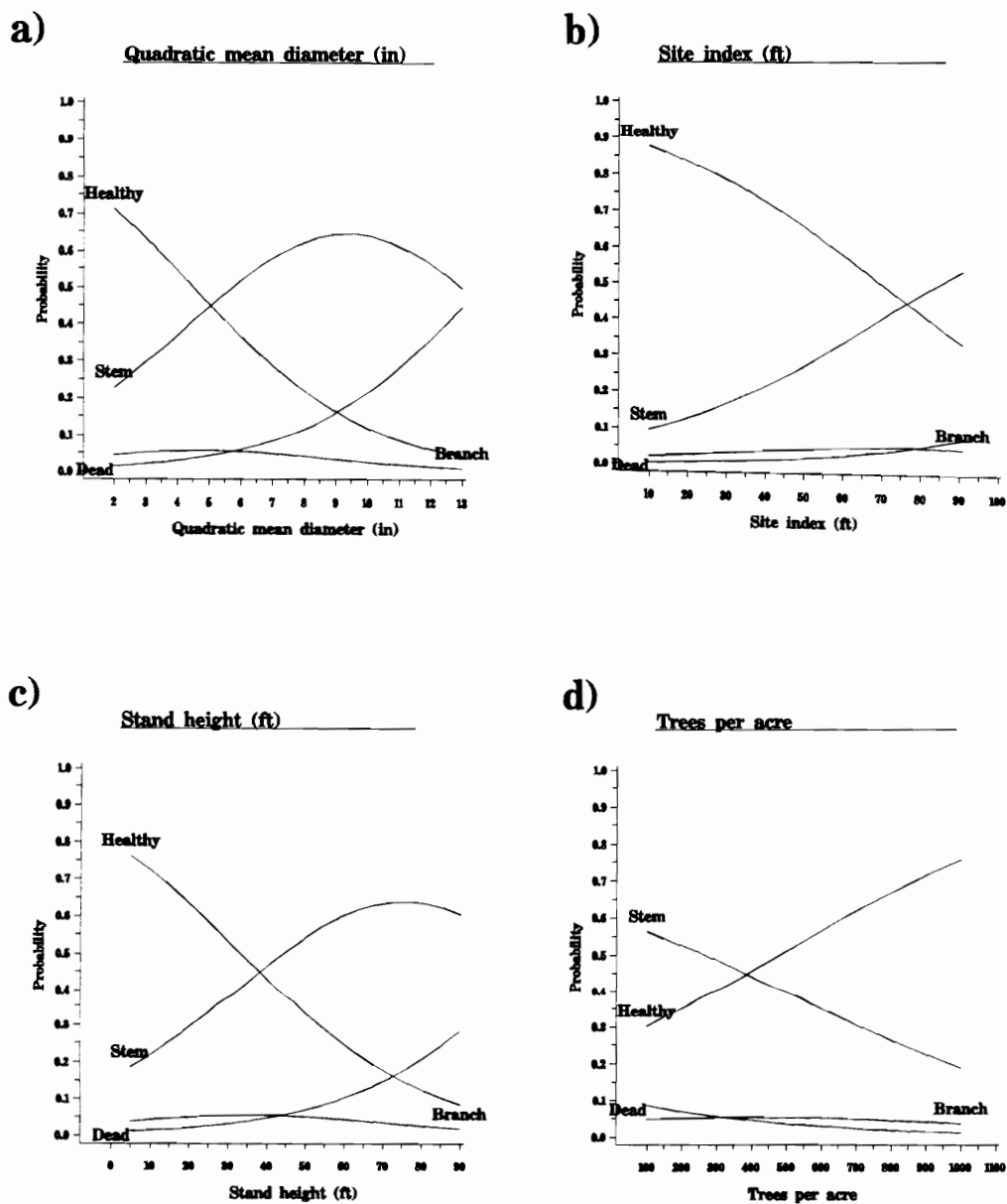


Figure 7.2. Profile plots for single covariate proportional odds models.
Species and data: Slash pine development data.

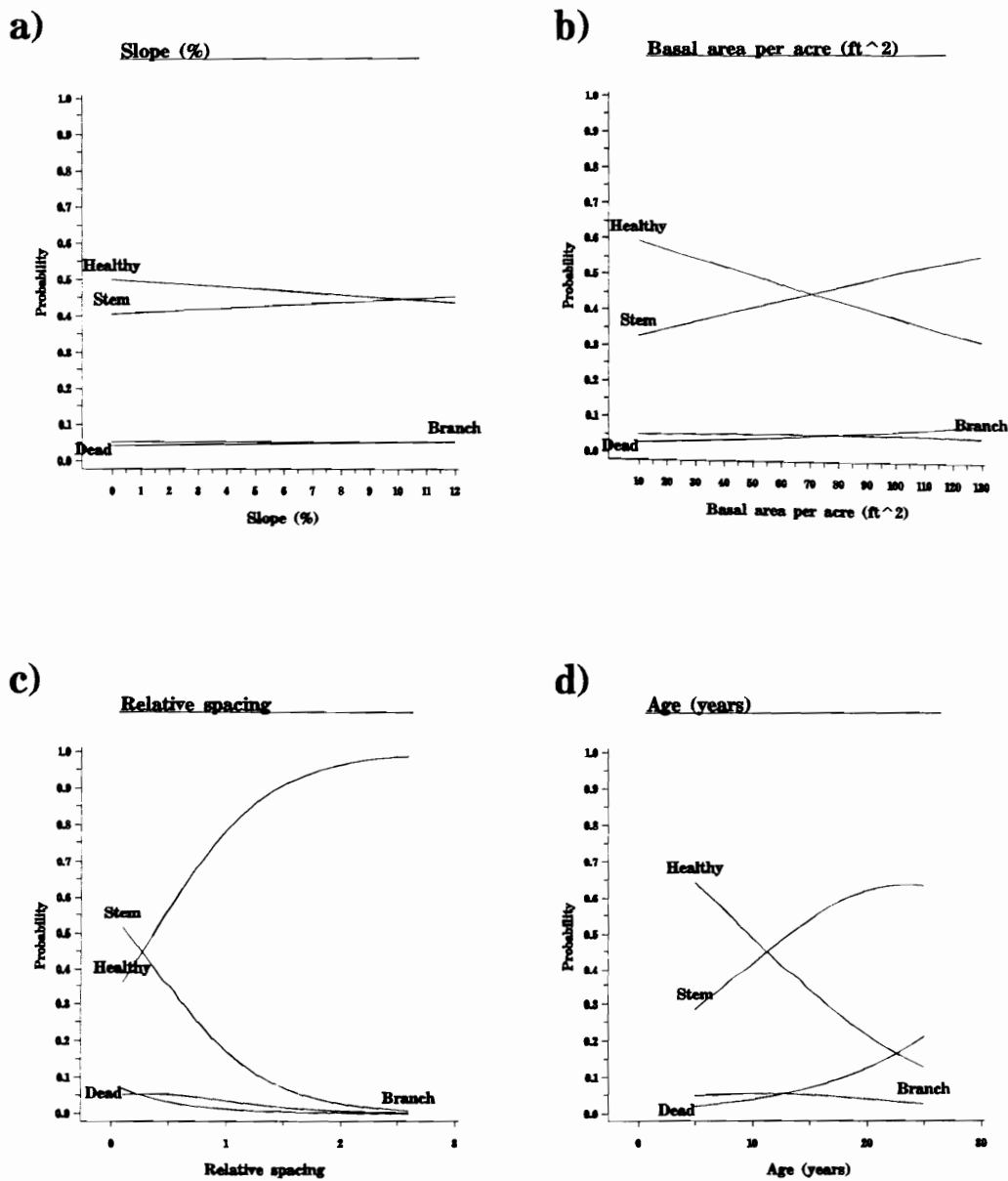
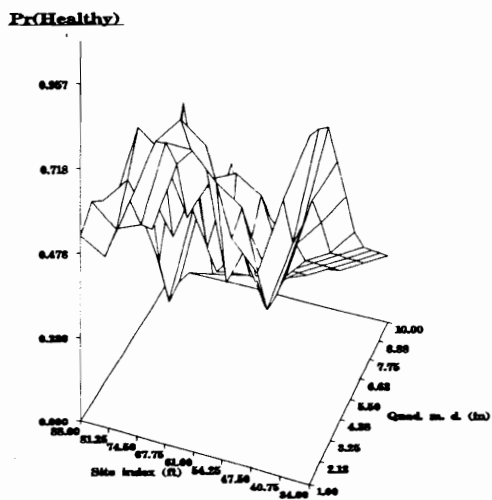
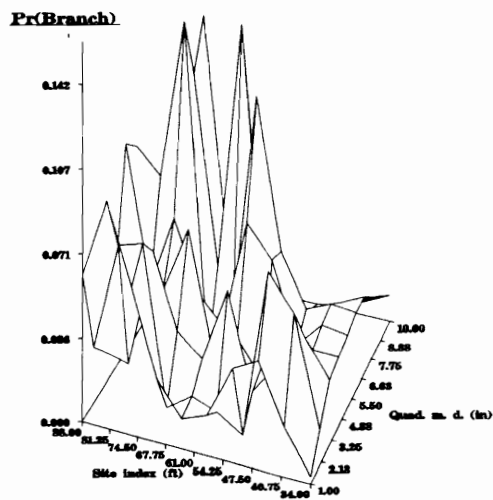


Figure 7.3. Profile plots for single covariate proportional odds models.
Species and data: Slash pine development data.

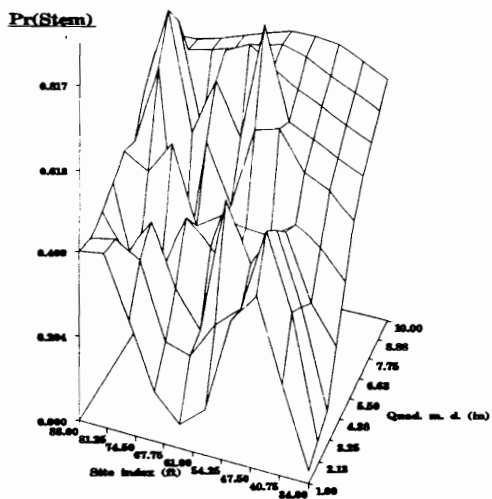
a)



b)



c)



d)

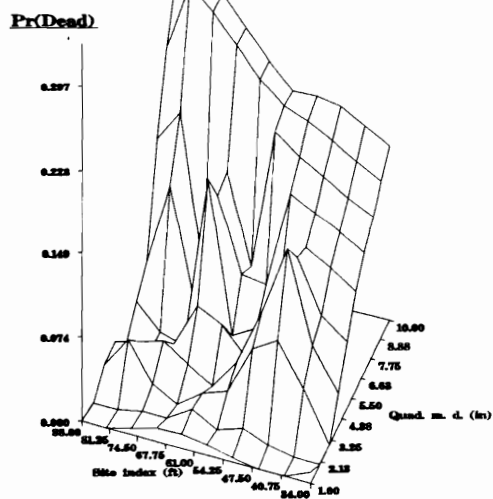
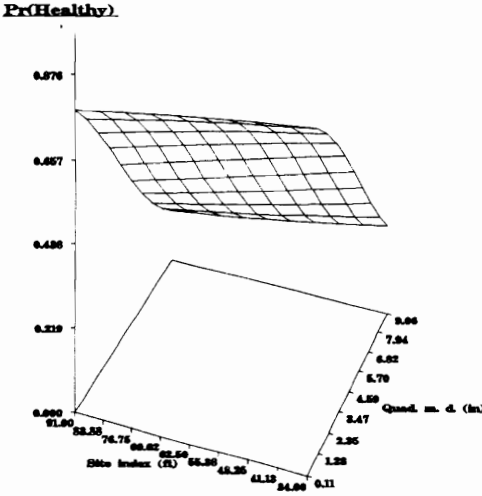
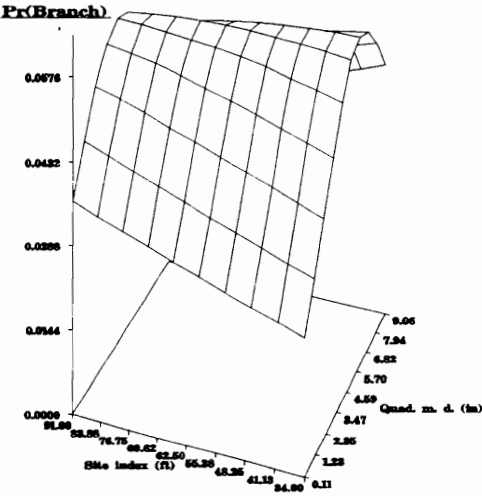


Figure 7.4. Observed probability surfaces for Slash pine.
 Surface interpolated by 5th order polynomials.
 Covariates: Site index (ft), Quadratic mean diameter (in).

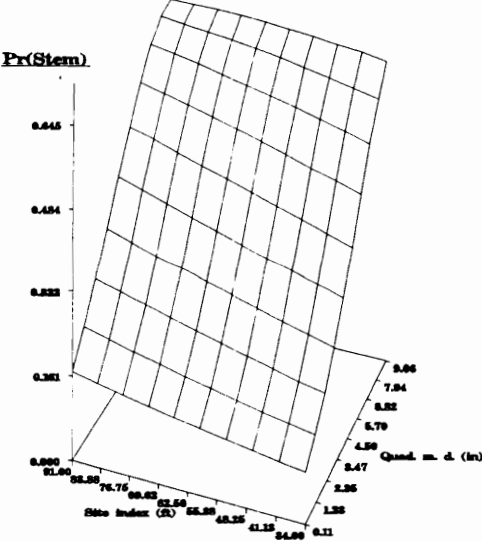
a)



b)



c)



d)

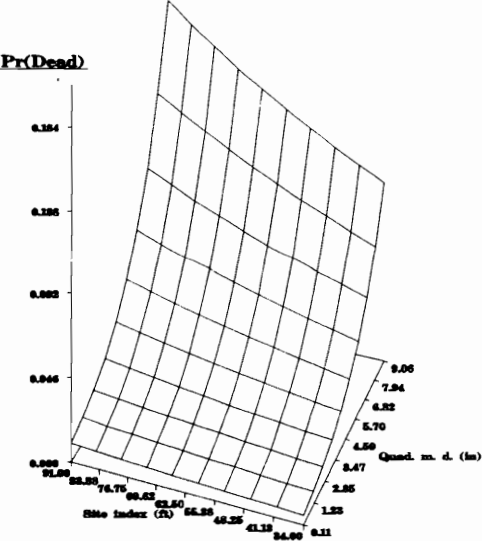


Figure 7.5. Predicted probability surfaces for two covariate proportional odds model. Slash pine development data. Covariates: Site index (ft), Quadratic mean diameter (in).

Are these relationships a good image of the biological trends in the data? To answer this the prediction profiles have to be compared with the observed proportions. For any single variable this is not too meaningful, since Age, Stand height, Quadratic mean diameter, etc. are all interrelated. As a plot increases in age, the height of the trees increases as do the quadratic mean diameter and typically, Relative spacing decreases. Site index and Stand height can be viewed as measures of site quality, Relative spacing, Basal area, Trees per acre, and Quadratic mean diameter as simple measures of stand density.

To select a basic set of covariates to start the modeling process, each site quality measure has been paired with a density measure and the observed proportions for any value combination interpolated by fifth degree polynomials. The same pairs of covariates were then fitted by a proportional odds model (GEE0) and the response profiles as both variables vary were generated. The best agreement for slash pine was produced by a combination of Site index and Quadratic mean diameter. Figure 7.4. displays the observed surface plots, Figure 7.5. the surface generated under the model. The panels in Figure 7.4. clearly display the noise in the data even after smoothing the surface. While clear trends are difficult to discern for healthy trees, stem infections and rust associated mortality (RAM) follow a clear trend. For any given site quality, the proportions in these categories increase rather sharply. At a given mean diameter of the stand, stem infections and RAM are more prevalent on better sites.

7.1.1.3. Ordinal response models

Site index and Quadratic mean diameter provide a reasonable starting point. The erratic nature of the surface plots of course calls for additional regressors. Based on the combined sums of squares, pseudo- R^2 , and H measures the following proportional odds model was selected as a good descriptor of the population averaged behavior.

$$\Pr(C_{it} \leq j) = G^{-1}(\eta_{itj})$$

where

$$\begin{aligned} \eta_{itj} = & \alpha_j + \beta_1 Q_{it} + \beta_2 SI_i + \beta_3 A_{it} + \beta_4 TPA_{it} \\ & + \beta_5 \frac{Q_{it}}{A_{it}} + \beta_6 BU_i + \beta_C C_i \end{aligned} \quad [7.14]$$

and $G(t) = \exp(t) / \{1 + \exp(t)\}$. The pieces of this linear predictor denote

Q_{it}	=	Quadratic mean diameter on plot i at time t
SI_i	=	Site index of plot i
A_i	=	Age of plot i at time t
TPA_{it}	=	Number of alive trees per acre
BU_i	=	Indicator, 1 if site was burned during preparation, 0 otherwise
C_i	=	A design matrix to identify cycles.

For measurement units see Table 6.2. The term $\beta_C C_i$ expands into $\beta_7 C_{i2} + \beta_8 C_{i3} + \beta_9 C_{i4}$, where C_{ik} denotes the k th measurement cycle for plot i . The first cycle was chosen as a baseline, so that the model can be applied to cross-sectional data, if desired.

Model [7.14] was selected as a valuable predictor based on a GEE0 fits with 1-and 3-dependent structures, as well as a mixed model (described below). It was also fit by regular maximum likelihood assuming repeated observations are independent. [7.14] is a population average model. The predictions from such a model for any subject will concur with the observed values only to the degree to which the subject's covariate pattern follows the population trend. While overall there may be a very good agreement between observed and predicted disease propensities, deviations are to be expected for any given subject. These deviations are partly accounted for in the subject-specific approach. Model [7.14] was augmented by various random terms and combinations thereof to accomplish this. The best fitting and predicting model was obtained by varying number of trees per acre across subjects. Since *TPA* is a time varying covariate this not only induces correlations, but lets these vary by time points. *TPA* in the random part of the model has been divided by 100 to scale the variances of the random terms.

Table 7.2. displays coefficients and their asymptotic standard errors for the four estimation approaches. Goodness-of-fit measures are found in Table 7.3. In all cases the reparameterization discussed in § 3.1.6 was applied. Focusing on the mixed model estimates, the proportions in any category can be calculated from the entries in Table 7.2. as follows. First the reparametrized cut-offs are computed as

$$\begin{aligned}\alpha_0 &= 2.7921 \\ \alpha_1 &= \alpha_0 + \exp\{-1.3304\} = 3.0564. \\ \alpha_2 &= \alpha_1 + \exp\{1.2351\} = 6.4952\end{aligned}$$

For a plot at the first cycle that was not burned during site preparation, with covariates

$$\begin{aligned}
Q_{i1} &= 4.95 \\
SI_i &= 66 \\
A_{i1} &= 9 \\
TPA_{i1} &= 283
\end{aligned}$$

the linear predictor becomes: $\hat{\eta}_{i1} = -2.1397$. The cumulative probabilities are then obtained as

$$\begin{aligned}
\hat{\Pr}(\text{healthy}) &= G(2.7921 - 2.1397) = G(0.6524) = 0.657 \\
\hat{\Pr}(\leq \text{branch}) &= G(3.0564 - 2.1397) = G(0.9167) = 0.714 \cdot \\
\hat{\Pr}(\leq \text{stem}) &= G(6.4952 - 2.1397) = G(4.3555) = 0.987
\end{aligned}$$

where $G(t) = \exp(t)/(1 + \exp(t))$ is the inverse of the logistic transform. The category probabilities follow by subtraction

$$\begin{aligned}
\hat{\Pr}(\text{healthy}) &= 0.657 \\
\hat{\Pr}(\text{branch}) &= 0.714 - 0.657 = 0.057 \\
\hat{\Pr}(\text{stem}) &= 0.987 - 0.714 = 0.273 \cdot \\
\hat{\Pr}(\text{dead}) &= 1 - 0.987 = 0.013
\end{aligned}$$

On the particular plot, a randomly selected tree thus has a 66% chance to be healthy, a 6% chance to be branch infected, and the odds are 27% to observe a stem infection. Only 1% of all trees will be dead due to fusiform rust.

Table 7.2.
Results of fitting model [7.14] to slash pine data.
Asymptotic standard errors in parentheses.

Coeff.	Variable	ML	GEE0-1	GEE0-3	Mixed
$\hat{\psi}_0$		3.43958	3.18421	3.03289	2.79215
$\hat{\psi}_1$		-1.44125	-1.37097	-1.39074	-1.33036
$\hat{\psi}_2$		1.13718	1.11984	1.14191	1.23513
$\hat{\beta}_1$	<i>Q</i>	0.03827 (0.04467)	-0.07801 (0.05049)	-0.09432 (0.02723)	-0.11250 (0.06073)
$\hat{\beta}_2$	<i>SI</i>	-0.00804 (0.00218)	-0.00844 (0.00300)	-0.01594 (0.00260)	-0.02128 (0.00683)
$\hat{\beta}_3$	<i>A</i>	-0.17173 (0.01875)	-0.12688 (0.02178)	-0.10238 (0.01229)	-0.10473 (0.02718)
$\hat{\beta}_4$	<i>TPA</i>	0.00112 (0.00008)	0.00090 (0.00011)	0.00141 (0.00011)	0.00240 (0.00049)
$\hat{\beta}_5$	<i>Q/A</i>	-4.12828 (0.43952)	-2.77973 (0.53978)	-1.95409 (0.22883)	-1.39561 (0.57408)
$\hat{\beta}_6$	<i>BU</i>	-1.12300 (0.16834)	-1.14959 (0.27701)	-1.10487 (0.26191)	-1.09473 (0.38425)
$\hat{\beta}_7$	<i>C₂</i>	-0.18452 (0.04190)	-0.26098 (0.02145)	-0.32901 (0.01993)	-0.30615 (0.06564)
$\hat{\beta}_8$	<i>C₃</i>	0.07854 (0.04314)	-0.07596 (0.03951)	-0.19061 (0.03323)	-0.16287 (0.10205)
$\hat{\beta}_9$	<i>C₄</i>	0.35747 (0.04694)	0.24023 (0.05072)	0.08799 (0.04758)	0.19206 (0.14114)
$\widehat{\text{var}}(\mathbf{b}_i) = \hat{D}$					0.035977

Table 7.3.
Goodness-of-fit measures for the estimation methods in
Table 7.2. calculated from evaluation data

GOF Statistic	ML	GEE0-1	GEE0-3	Mixed
SSR	13.62	14.12	14.20	14.04
WSP	65.81	65.27	66.10	67.49
WSF	1230.66	1260.83	1298.84	1321.95
SSF	5743.17	5665.45	5863.64	6072.84
R_p^2	0.78	0.78	0.78	0.78
SSR_b				7.01
WSP_b				48.33
WSF_b				629.31
SSF_b				4322.19
R_{pb}^2				0.89

The four different estimation methods lead to quite different coefficient estimates. Only a few estimates are similar across the columns of Table 7.2. Most similarities are recognizable between the GEE0-3 and the mixed model fit. Estimates of similar magnitude are interesting with respect to comparisons of the estimated standard errors. $\hat{\beta}_6$ for example has the same size and magnitude in all four approaches, but the standard error of the incorrect ML analysis is considerable smaller than for the other approaches that account for serial correlation. Although one can observe repeatedly, that ignoring correlations leads to smaller standard errors, this is not a rule and counter examples are frequent (Diggle *et al.* 1994). It should be kept in mind though, that the precision estimates in the leftmost column of Table 7.2 are inconsistent and biased. Comparisons should thus not be taken too far.

The GOF statistics in Table 7.3. underline the interesting and well documented observation, that ignoring correlations does not impair the predictive capabilities of the model. With respect to the population averaged sums of squares statistics, the ML fit predicts as well as the other two population averaged approaches, GEE0-1 and GEE0-3. The mixed model without the BLUPs for the random effects does not perform better than any other population averaged approach, which is not surprising. Incorporating the BLUPs however, and individualizing the fit, leads to a considerable improvement in predictive power. Compared to the other approaches, SSR_b , and WSF_b are cut into half, by allowing TPA to vary across plots and utilizing predictions of these random variables. The entries in Table 7.3. have been calculated from the evaluation data, not the development data that was used to derive the coefficient estimates. Typically, BLUPs are calculated for the data, from which the coefficients were obtained.

Applying the GOF criteria to the development data resulted in SSR_b of 4.4, and R_{pb}^2 above 0.9.

To assess uncertainty in the predicted probabilities, asymptotic 95% confidence intervals for the cumulative response probabilities were computed for the four estimation methods in Table 7.2. Table 7.4. depicts the results. In the mixed model analysis confidence bounds are based on the fixed effects only, ignoring the variability of the random effects to make the bounds comparable. For each approach, Table 7.4. lists the total length of the interval - which is not symmetric because of the nonlinearity of the logit link - in probability units, the average predicted category probabilities, and the ratio of interval length to cumulative and category probabilities expressed as percentages. Although the coefficient estimates and standard errors in Table 7.2. are quite different, the confidence bounds are very similar across estimation methods. However, the ML bounds should be interpreted with caution as the precision estimates are biased and inconsistent and may not reflect the actual precision of the estimates.

The interval length for healthy and branch infected trees are similar in all approaches. Since branch infected trees are much less frequent, however, there is considerable uncertainty in the prediction of branch infections, as displayed by $l/\Pr(C = j)[\%]$ in Table 7.4. Predicting the population-averaged responses in a mixed model based on only the fixed effects appears to be as precise as in a genuine population-averaged approach that accounts for correlations directly.

Table 7.4.

Asymptotic 95 % confidence intervals for predicted proportions in marginal slash pine model [7.14]. Results for mixed model based on fixed effects only, ignoring random effects variability. Length of intervals in probability units.

		State of fusiform rust infection		
		Healthy $j = 0$	Branch $j = 1$	Stem $j = 2$
ML	length l	0.03378	0.03409	0.00981
	$\Pr(C = j)$	0.46182	0.05238	0.43547
	$l/\Pr(C \leq j)[\%]$	7.31	6.63	1.03
	$l/\Pr(C = j)[\%]$	7.31	65.09	2.25
GEE0-1	length l	0.03909	0.03931	0.01133
	$\Pr(C = j)$	0.45705	0.05623	0.41823
	$l/\Pr(C \leq j)[\%]$	8.20	7.38	1.19
	$l/\Pr(C = j)[\%]$	8.20	69.91	2.71
GEE0-3	length l	0.03548	0.03576	0.00987
	$\Pr(C = j)$	0.45889	0.05499	0.42201
	$l/\Pr(C \leq j)[\%]$	7.45	6.73	1.03
	$l/\Pr(C = j)[\%]$	7.45	65.02	2.34
Mixed	length l	0.03404	0.03449	0.00988
	$\Pr(C = j)$	0.44384	0.05264	0.43693
	$l/\Pr(C \leq j)[\%]$	7.38	6.72	1.04
	$l/\Pr(C = j)[\%]$	7.38	65.52	2.26

How well does the model perform as a classifier? Table 7.5. shows the observed marginal proportions in the development data against the classified marginal proportions. The classifications were carried out according to the supremum rule [3.10] (I), Anderson and Philip's latent class rule [3.9] (II) and the direct classification (III) discussed in § 3.1.7 (p. 64).

Table 7.5.
 Prediction of marginal proportions for model [7.14]

Method	Classif. Rule	F u s i f o r m r u s t c a t e g o r y			
		Healthy	Branch	Stem	Dead
		<i>Observed</i>		<i>Frequencies</i>	
		0.503	0.057	0.399	0.041
		<i>Classified</i>		<i>Frequencies</i>	
ML	I	0.634	0.000	0.366	0.000
	II	0.520	0.131	0.349	0.000
	III	0.499	0.053	0.406	0.043
GEE0-1	I	0.672	0.000	0.328	0.000
	II	0.553	0.145	0.302	0.000
	III	0.511	0.057	0.391	0.042
Mixed model	I	0.579	0.000	0.421	0.000
	II	0.611	0.070	0.319	0.000
	III	0.503	0.049	0.400	0.048

Obviously, both (I) and (II) are performing poorly. This happens, when some categories occur only with low frequencies, such as branch infections and RAM. At least for grouped data, the direct classification should be preferred. Neither (I) nor (II) classify any tree as dead of fusiform rust, and (I) does not classify any branch infections.

Judged by (III), all approaches lead to excellent marginal agreement. The proportion of trees in any category will be exactly represented in the predictions under the model. Note that the marginal observed and classified proportions do not have to be equal, since they are the result of a cross-classification. The actual classification table relays information about the degree of mis-classifications. Table 7.6 shows this table for the 1-dependent GEE0 estimates. The classification tables for the other approaches were similar.

Table 7.6.
Cross-classification table based on direct probability
accumulation in 1-dependent GEE0 fit.

		<i>Predicted</i>				<i>Category</i>	Margin
		0	1	2	3		
<i>Observed</i>	0	0.282	0.028	0.176	0.017	0.503	
	1	0.026	0.003	0.025	0.003	0.057	
	2	0.186	0.023	0.171	0.020	0.399	
	3	0.017	0.002	0.019	0.003	0.041	
Margin		0.511	0.057	0.391	0.042		

Figures 7.6. and 7.7 display circular ray plots for the 1-dependent GEE0 fit, Figures 7.8. and 7.9. for the mixed models utilizing the BLUPs.

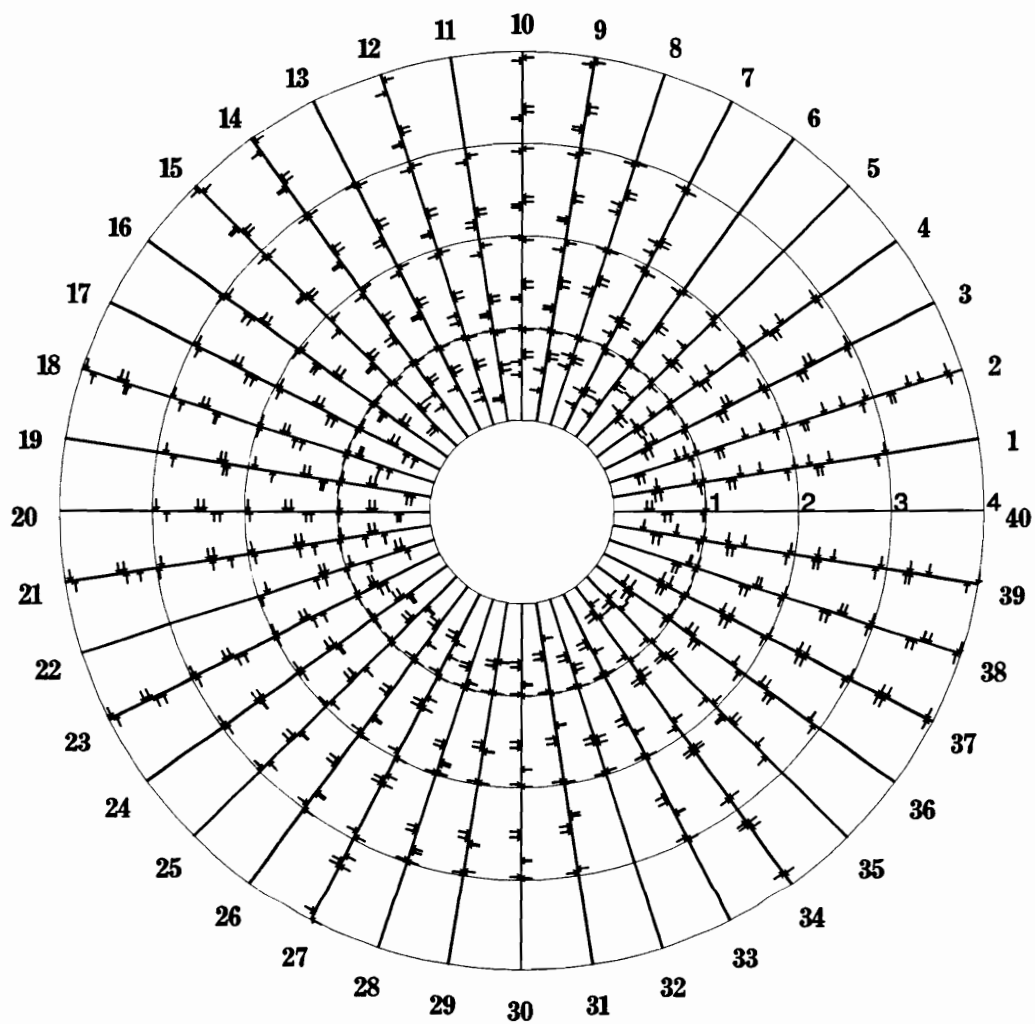


Figure 7.6. Observed and predicted proportions in Slash pine model [7.14].
 Each segment corresponds to one occasion and a 0–100% interval.
 Observed values denoted by clockwise arrows, predicted values
 counter clockwise. GEE0 with 1–dependence structure.

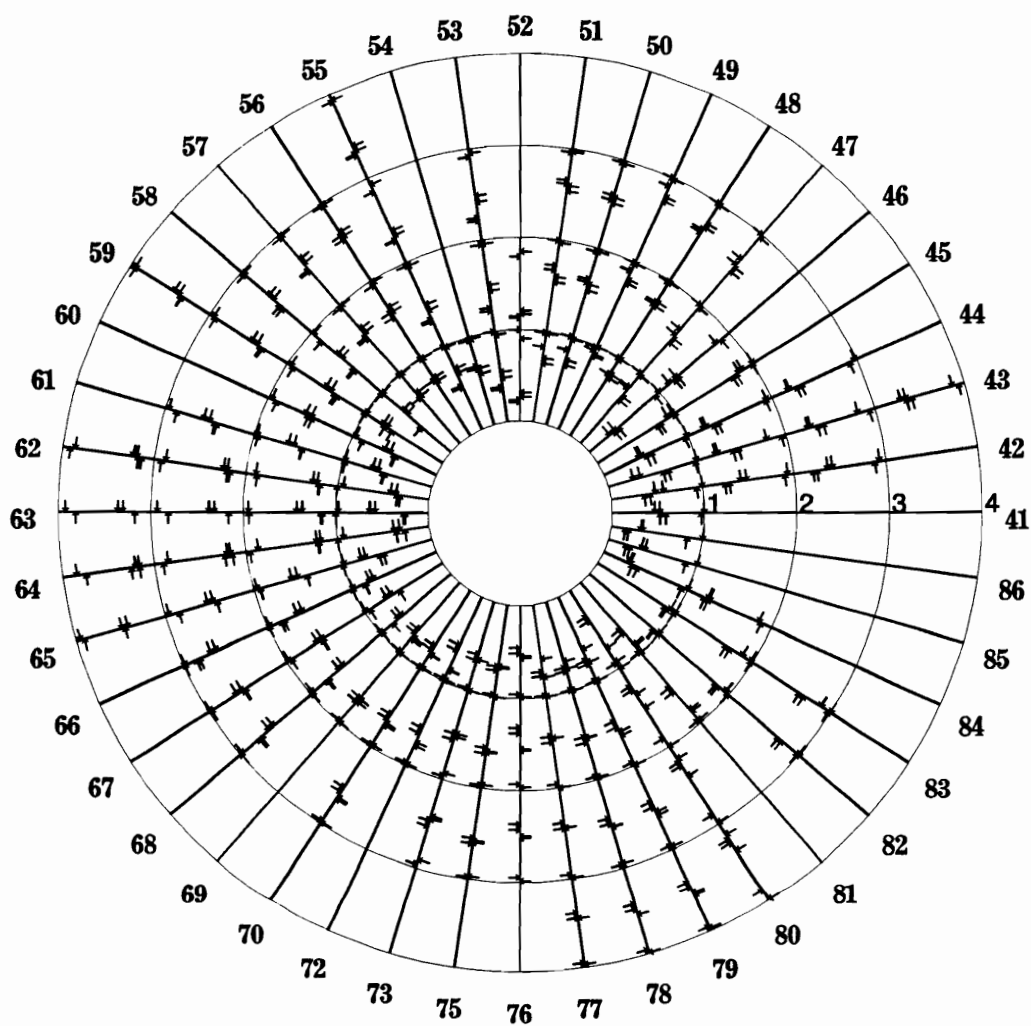


Figure 7.7. Observed and predicted proportions in Slash pine model [7.14].
 Each segment corresponds to one occasion and a 0–100% interval.
 Observed values denoted by clockwise arrows, predicted values
 counter clockwise. GEE0 with 1–dependence structure.

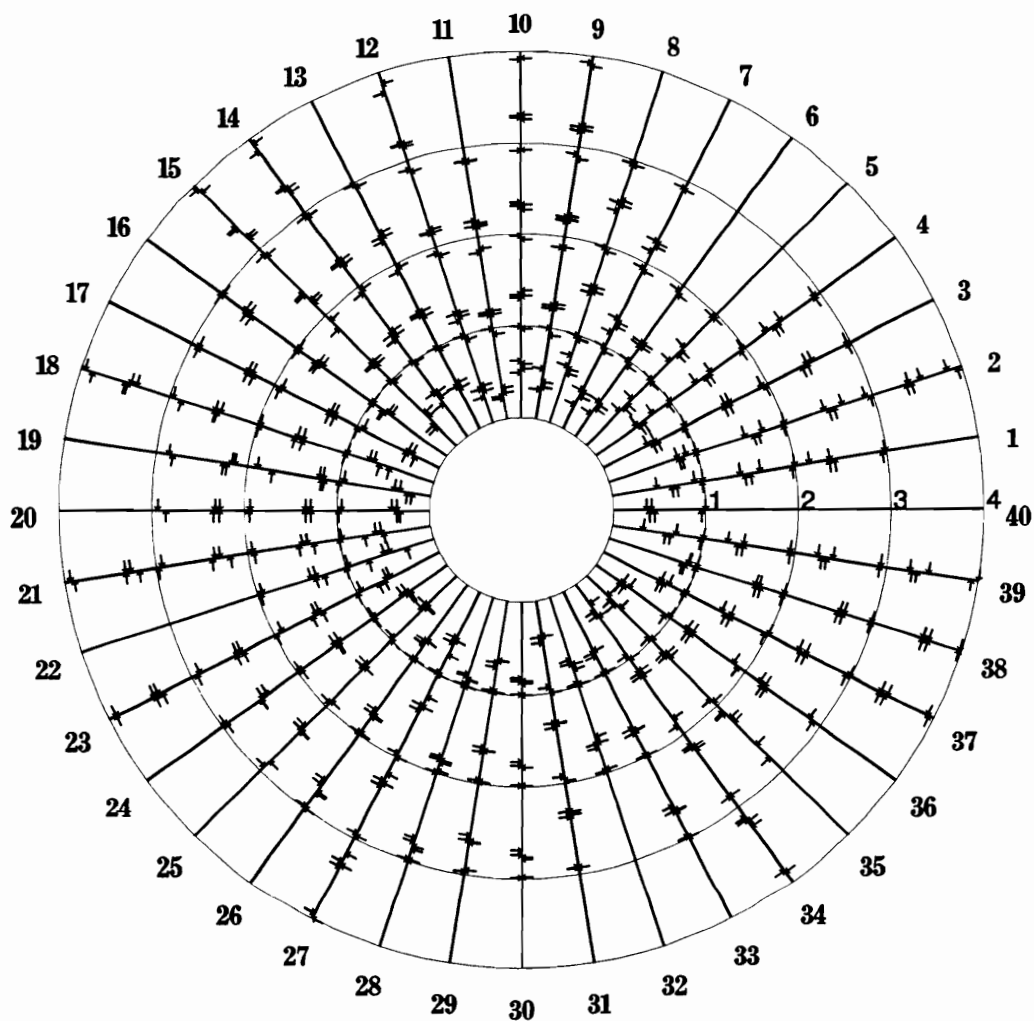


Figure 7.8. Observed and predicted proportions in Slash pine model [7.14].
 Each segment corresponds to one occasion and a 0–100% interval.
 Observed values denoted by clockwise arrows, predicted values
 counter clockwise. Mixed model fit semi-parametrically.

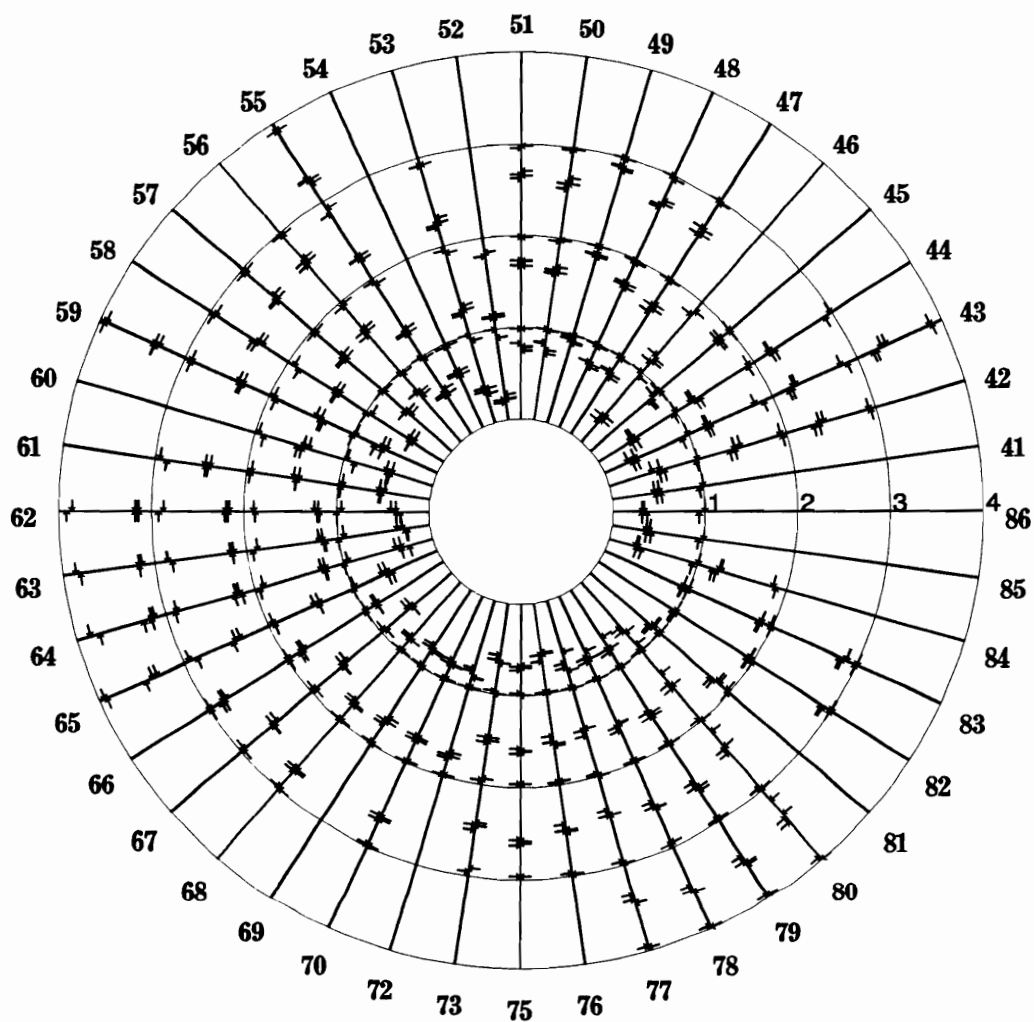


Figure 7.9. Observed and predicted proportions in Slash pine model [7.14].
 Each segment corresponds to one occasion and a 0–100% interval.
 Observed values denoted by clockwise arrows, predicted values
 counter clockwise. Mixed model fit semi-parametrically.

The population averaged model in Figures 7.6. and 7.7 displays lack of fit for several plots and occasions (eg #2, 9). For plot 2, for example, a high observed frequency of healthy trees is grossly underestimated at the first and second occasion. For plot 9, the model overestimated the probabilities in the healthy category at all time points. This is not surprising, since subjects can only be predicted accurately, if they do not deviate from the actual population average.

The gain by using a random effect evidenced through the fit statistics in Table 7.3 is articulated comparing Figures 7.8 - 7.9 to 7.6 - 7.7. In all four plots, however, it appears that the initial occasion is responsible for most deviations between predictions and observations. The predicted proportions for plots 2 and 9 are much closer to the observed frequencies. The subject-specific deviation from the population average was compensated by the inclusion of the random TPA effect. Table 7.5 shows, that although marginally the observed and predicted proportions comply nicely, there is a substantial degree of misclassification. Especially the healthy and stem infected category are frequently incorrectly assigned. Combining this observation with the mentioned lack of fit for the initial occasion suggests that the model has a deficiency for young stands, but performs better as stand age progresses. It is also possible that the second category, branch infections, is not clearly distinguished from the two adjacent states and adds an erratic element to the ordering of the response.

7.1.1.4. Binary response models

If interest is in the marginal fusiform rust probabilities in any of the four categories, a multivariate model like [7.14] is appropriate. In some instances it is reasonable however to reduce the dimensionality of the response. If for example, only the comparison of healthy versus infected trees is of interest, one could either calculate the probabilities from the proportional odds model, or alternatively, fit a new model to a binary response. By choosing the latter, the erratic influence of intermediate categories (branch infections for example) can be diminished and noise reduced.

Three binary response models for slash pine are presented here. The outcomes created are

<i>Healthy:</i>	1 if a tree is free of any fusiform rust infection 0 otherwise
<i>Stemfree:</i>	1 if a tree is free of stem infections 0 otherwise
<i>Alive:</i>	1 if a tree is alive 0 if not.

The generic term *event* will be used for the outcome coded as 1, *no-event* otherwise. Separate models were compared for each response. The approximate parametric subject-specific approach of § 4.3.2 (Pseudo-Likelihood) was used to fit mixed binary models based on a logit link. This link was used since the proportional odds model algorithms apply and collapse to a standard logistic regression. The sums of squares and related criteria for both, development and evaluation data, were used to discriminate among competing models.

In binary models classifications are typically carried out by comparing a predicted proportion against a user-defined cutoff point, ξ . Frequently, $\xi = 0.5$ is chosen as

cutoff and if the predicted probability for a subject is larger than 0.5, it is assumed the model predicted an event (see § 3.1.7). Consequently, there are four possible situations: a subject responded and the model can either predict an event or a no-event, and similarly if a subject did not respond. Two of these cases are misclassifications which should be kept at a minimum. Table 7.7 displays a 2*2 classification table and simple measures to characterize the performance of a binary regression model derived from it are listed in Table 7.8.

Table 7.7.
Classification table for a binary response models.

		<i>P r e d i c t e d</i>		
		Event	No-Event	
<i>Obser- ved</i>	Event	#EE	#ENE	#OE
	No-Event	#NEE	#NENE	#ONE
		#PE	#PNE	Total

Table 7.8.
Measures of classification power for
binary response models.

Measure	Interpretation	Calculation
Sensitivity	Proportion of obs. events actually predicted events	$\#EE / \#OE$
Specificity	Proportion of obs. no-events predicted as no-events	$\#NENE / \#ONE$
False Positive	Proportion of predicted events that are incorrect	$\#NEE / \#PE$
False Negative	Proportion of predicted no-events that are incorrect	$\#ENE / \#PNE$
FCC	Fraction correctly classified	$(\#EE + \#NENE) / \text{Total}$

A model is said to perform well, if its sensitivity is high and its false positive rate low, since then many events are predicted as such without committing classification errors. Unfortunately both measures depend on the cutoff point employed. For example, if $\xi = 0$ is chosen as a cutoff, all observations will be classified as events, the model has sensitivity 100% but a high false positive rate, since all no-events have also been predicted as events (#NEE large). To correctly judge the predictive power of a logistic regression model, it is more reasonable to vary the cutoff points and compute the performance measures of Table 7.8 for each. The choice of the cutoff point depends on the problem at hand. The liability of misclassifications is a major determinant. In the case of response *Healthy*, if it is important to label trees as healthy only if it is very certain that they are, one will use a large ξ . This will decrease the False positive rate. A good model for this purpose will retain a sizeable sensitivity at ξ .

A plot of model sensitivity against the False positive rate for various classification rules (ξ) is known as a Receiver-Operating-Characteristic (ROC) curve. In some applications the area under the ROC curve is calculated as a measure of model performance. We will interpret the ROC curve only graphically.

The models selected for the three binary responses are quite similar. The linear predictors and random terms are as follows:

Healthy:

$$\begin{aligned} \eta_{it} = & \beta_0 + \beta_1 Q_{it} + \beta_2 TPA_{it} + \beta_3 IA_i + \beta_4 WI_i \\ & + \beta_5 BU_i + (\beta_6 + b_{i1}/100) SI_i + (\beta_7 + b_{i2}) C_1 \end{aligned} \quad [7.15]$$

Stemfree:

$$\eta_{it} = \beta_0 + \beta_1 Q_{it} + (\beta_2 + b_{i1}/100)TPA_{it} + \beta_3 IA_i + \beta_4 BU_i \quad [7.16]$$

$$+ \beta_5 D^2 H_{it} + \beta_6 \frac{SH_{it}}{Q_{it}} + b_{i2} C_1$$

Alive:

$$\eta_{it} = \beta_0 + \beta_1 A_{it} + \beta_2 TPA_{it} + \beta_3 C_1 + \beta_4 WI_i \quad [7.17]$$

$$+ \beta_5 BU_i + (\beta_6 + b_{i1}/100)SI_i + \beta_7 D^2 H_{it}$$

The model terms not yet encountered are

$$\begin{aligned} D^2 H_{it} &= \text{a proxy for volume per acre on plot } i \text{ at time } t \\ C_1 &= 1 \text{ if measurement at first occasion, } 0 \text{ otherwise} \\ WI_i &= 1 \text{ if site preparation involved windrowing, } 0 \text{ otherwise} \\ IA_i &= \text{Initial age of plot } i. \end{aligned}$$

If more than one random term was used it was assumed that these are uncorrelated, hence $\text{var}(\mathbf{b}_i)$ diagonal. Table 7.9. lists coefficient estimates, asymptotic standard errors and some GOF statistics.

The models fit very well, but it should be recalled that these are models for proportions, where R_p^2 takes on larger values than for models based on indicators. The initial occasion seems to be of particular importance in all instances. In contrast to the proportional odds model of the previous section, windrowing as a site preparation method improves the predictive powers of the *Healthy* and *Alive* models, although it is not a partially significant regressor. Also in contrast to the four category model initial age emerges as more valuable than plot age, another indication of the relevance of initial conditions in slash pine.

Table 7.9.

Estimates and asymptotic standard errors (in parentheses) for binary response models. Approximate parametric approach (Pseudo-Likelihood).

Model term	<i>R E S P O N S E</i>		
	Healthy	Stemfree	Alive
Intercept	4.50708 (0.52389)	0.66604 (0.49727)	5.67324 (0.94653)
Q_{it}	-0.08058 (0.02442)	-0.07548 (0.04289)	
TPA_{it}	0.00137 (0.00045)	-0.00010 (0.00053)	0.001441 (0.00052)
IA_i	-0.15565 (0.02829)	-0.09181 (0.02254)	
WI_i	-0.18933 (0.18147)		0.21482 (0.15209)
BU_i	-1.40227 (0.68668)	-0.31052 (0.38329)	-0.98466 (0.53721)
SI_i	-0.05202 (0.00715)		-0.03726 (0.00966)
C_1	0.42803 (0.07227)		1.66948 (0.18743)
SH_{it}/Q_{it}		0.10108 (0.03882)	
D^2H_{it}		-0.000059 (0.000029)	0.000159 (0.000067)
A_{it}			-0.12471 (0.03798)
$\text{var}(b_{i1})$	1.074033	0.032864	0.003469
$\text{var}(b_{i2})$	0.193921	0.267876	
SSR_b	1.20	0.99	0.60
WSP_b	3.06	2.53	5.84
WSF_b	98.91	197.66	483.58
R^2_{pb}	0.95	0.95	0.99

Figures 7.10 - 7.12 display the ROC curves for the three models. Sensitivity is depicted in three dimensional space with diamond shaped symbol as False positive rate and the cutpoint ξ vary. The range of false positive rates is important in any model. One will typically not accept False positive percentages over 40-50, unless they occur at the boundaries. A separate box in Figures 7.10 - 7.12 shows the range of cutpoints and False positive rates, that are paired with sensitivities and simplifies the interpretation of the three dimensional plot. Fractions correctly classified (FCC) as a function of the cutpoint are plotted as a marginal distribution against the back wall of the coordinate system. The marginal False negative percentages as a function of False positive rates are plotted against the side wall.

All three models perform well. For the response *Healthy*, False positive rates go up to 0.5, but decrease sharply as the cutpoint moves away from unrealistic values < 0.1 . ROC curves indicate a well classifying model, if Sensitivity does not drop too severely, as False positive rates approaches 0 and the cutpoint approaches 1. If cutpoints between 0.4 and 0.6 are reasonable, the models for *Healthy* and *Stemfree* response are satisfactory and will classify 2/3 to 3/4 of all observations correctly.

The low prevalence of rust associated mortality causes Sensitivity of the model for *Alive* trees to be close to 1 regardless of the cutpoint. False positive rates, i.e. living trees, classified as dead, are minimal. This comes at the cost of a high False negative rate, dead trees attributed alive by the model. But the overall FCC is satisfying. If cutpoint below 0.75 are reasonable to label a tree as dead, model [7.17] performs excellent.

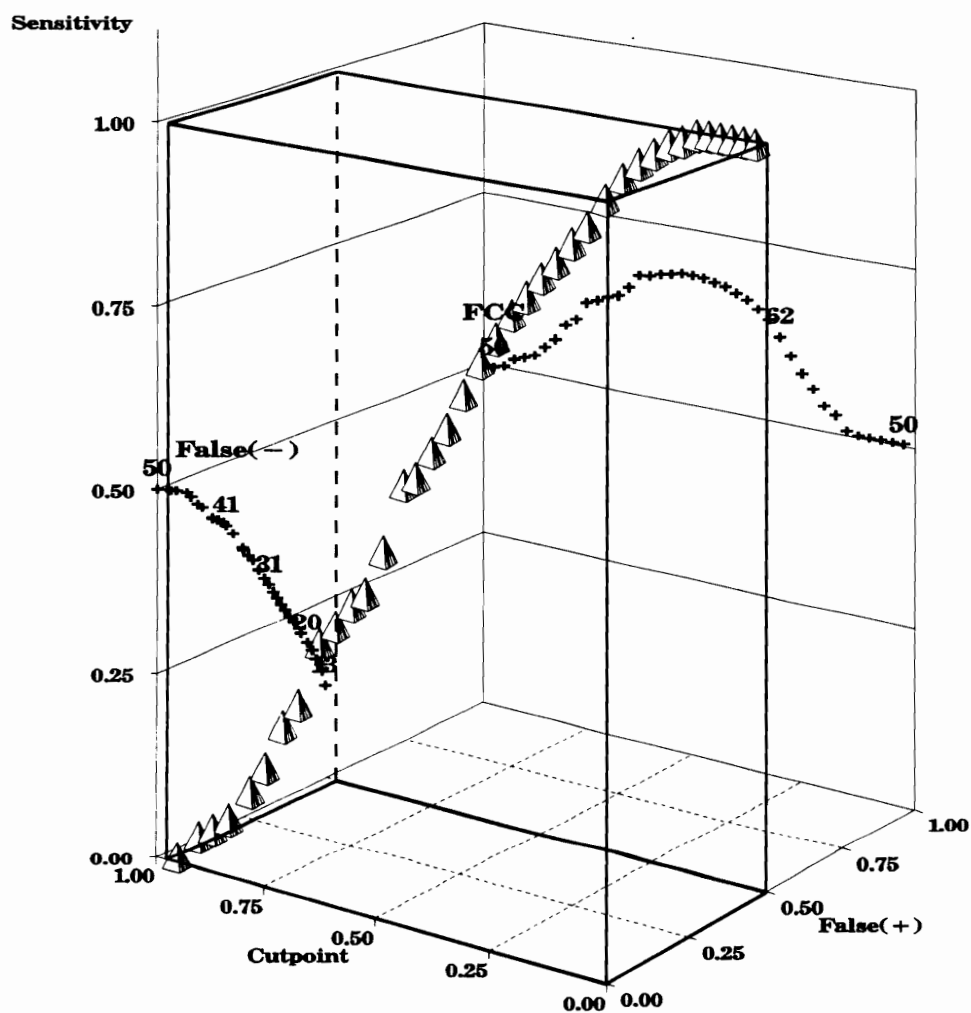


Figure 7.10. Receiver-Operating-Characteristic (ROC) for healthy trees.
Mixed model: Site index and first occasion random.
Sensitivity: Proportions of events, predicted as events.
False(+): Proportions of predicted events, that are incorrect.
False(-): Prop. of predicted no-events, that are incorrect.
FCC: Fraction correctly classified, events and no-events (in %).

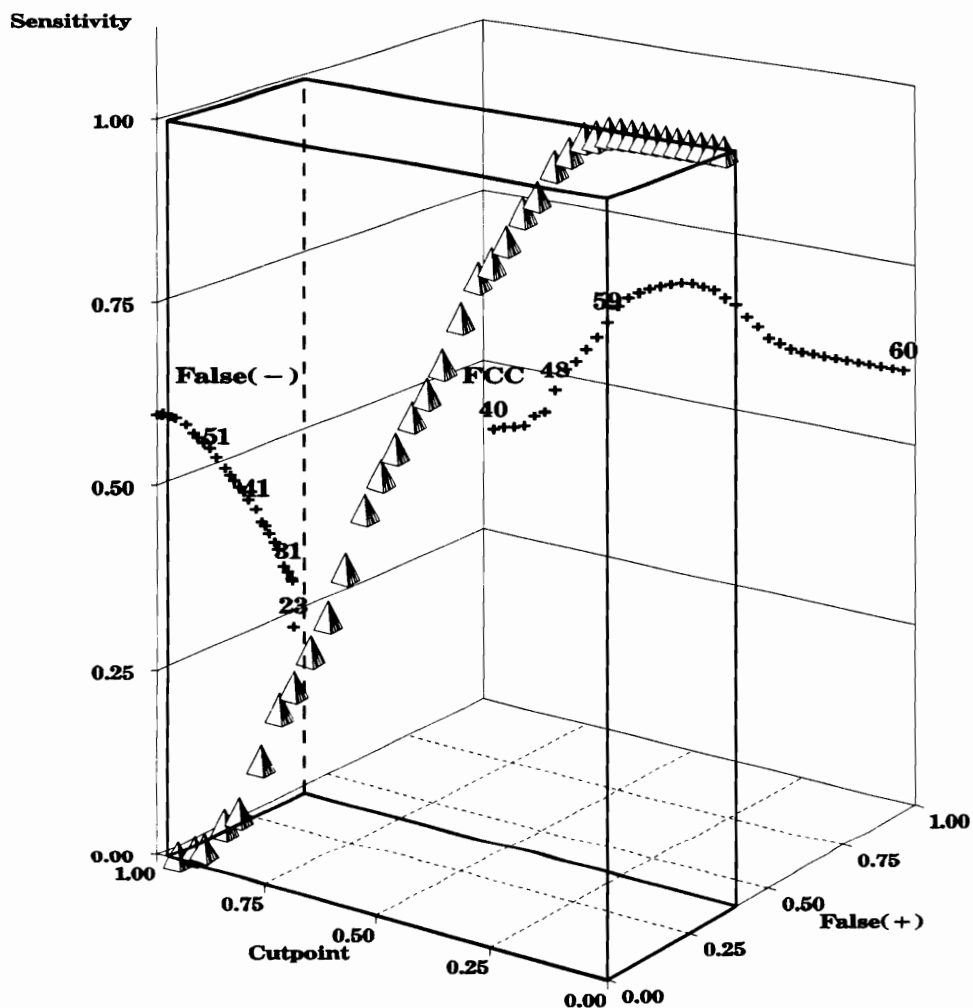


Figure 7.11. Receiver-Operating-Characteristic (ROC) for trees free of stem infect.
Mixed model: Trees per acre and first occasion random.
Sensitivity: Proportion of events, predicted as events.
False(+): Proportion of predicted events, that are incorrect.
False(-): Prop. of predicted no-events, that are incorrect.
FCC: Fraction correctly classified, events and no-events (in %).

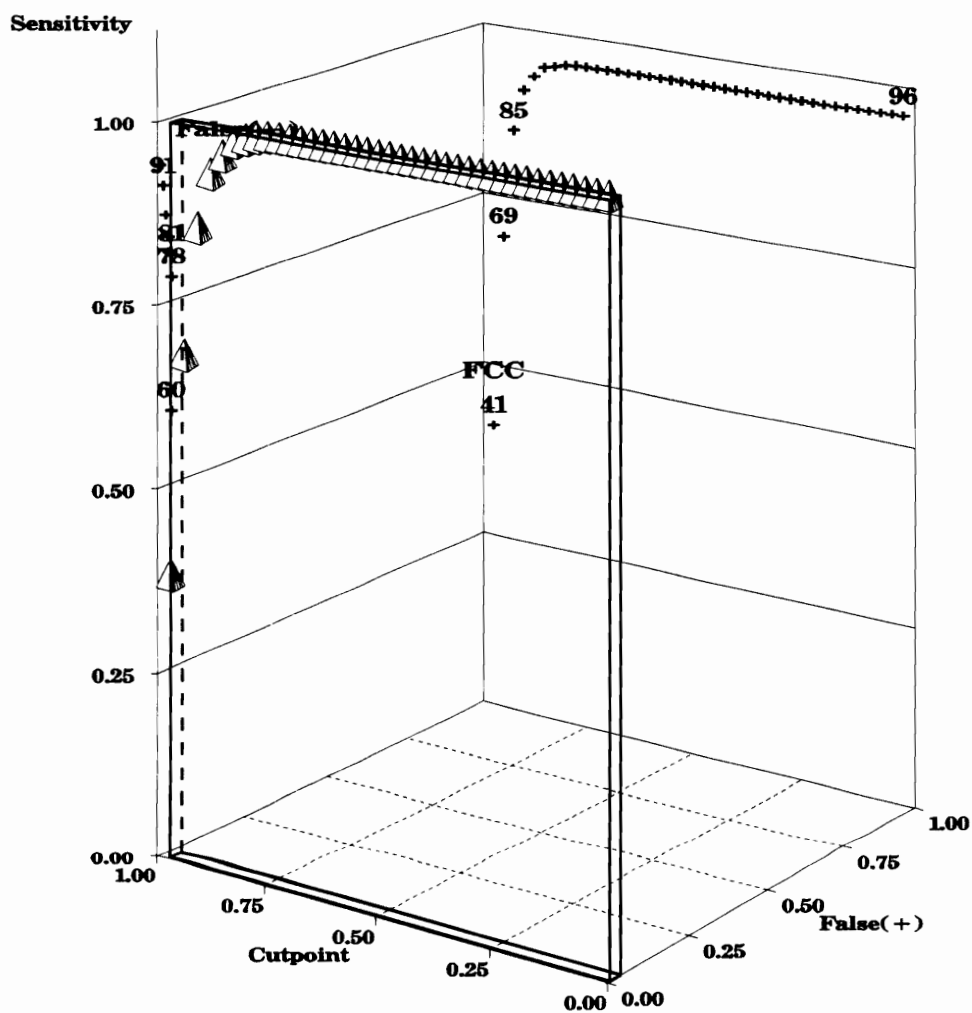


Figure 7.12. Receiver – Operating – Characteristic (ROC) for alive trees.

Mixed model: Site index random.

Sensitivity: Proportion of events, predicted as events.

False(+): Proportion of predicted events, that are incorrect.

False(-): Prop. of predicted no – events, that are incorrect.

FCC: Fraction correctly classified, events and no – events (in %).

7.1.2. Transitional Plot Models

In § 7.1.1. the marginal trend in the population or for any plot was at the center of investigation. The models presented do not allow one to infer about transitional behavior over time. Questions like "*Given that a tree was healthy, what are the odds of staying healthy*" require a transitional focus. Transitional models for slash pine are the topic of this section. From the various approaches discussed in § 4.4 the RC model for conditional events based on a multinomial and a proportional odds parameterization, and Bonney's CC approach adapted for an ordinal response (§ 4.4.3.1) have been chosen.

Goodness-of-fit criteria as discussed in § 7.1.1.1 apply here in the same manner. Two additions can be made however. One, by separating sum of squares criteria according to the conditioning events, the total deviations can be broken up into additive components. This allows a more detailed analysis of potential lack of fit. For plot models, one should draw comparisons based on frequency weighted sums of squares ([7.4], [7.5]) carefully, however, since the actual number of observations in the conditioning classes will differ. Two, conditional models are typically fit by maximum likelihood, hence likelihood based criteria are available. For discrimination among nested models likelihood ratio tests can be used. Whether nested or not, Akaike's information criterion (AIC), the maximized likelihood minus the number of estimated parameters (Bozdogan 1987) applies. Among competing models the one with highest AIC is chosen. Similarly to Mallow's C_p statistic in linear regression, AIC combines a measure of fit (the maximized likelihood) and a penalty to avoid overfitting.

7.1.2.1. RC model for conditional events

For a unidirectional process with 4 categories, 9 transition probabilities have to be modeled. The parameterization [4.24] is based on a multinomial regression model and constrains one coefficient vector for each previous response, leaving 6 parameter vectors to be estimated. The models that refer to different previous states can in principle differ in size. That is, the regressor variables used to model the transitions from the healthy category can differ from the ones that model transitions from the stem infected category, since the systems are independent. In fitting various models to the slash pine development data it turned out however, that little or nothing could be gained in fit or predictive power by varying the regressors. In all cases, breaking up sums of squares showed that most variability was caused by conditioning on the branch infected state. The subset of trees that were branch infected at the last occasion is extremely noisy and erratic. In terms of actual sums of squares, 70 - 80% of squared deviations were attributed to the subset of previously branch infected trees. When weighing by observed frequencies, the influence decreased to about 10 - 20%, since branch infections are overall rather rare, compared to the stem infected and healthy states. RC models have a rather large number of parameters, depending on the number of possible transitions. The final model selected for slash pine entails seven regressors and an intercept term, thus $6 \times 8 = 48$ parameters needed to be estimated.

The probability $p_{jj'}$ to move into state j after j' was observed at the last occasion was assumed not to depend on the actual time point, otherwise the number of

parameters would have been $48 \times 4 = 336$. For the i th plot at time t the model selected was

$$p_{jj} = \frac{\exp\{\mathbf{x}'_{it} \theta_{jj}\}}{\sum_{l=j}^3 \exp\{\mathbf{x}'_{it} \theta_{jl}\}} = \frac{\exp\{\eta_{itjj}\}}{\sum_{l=j}^3 \exp\{\eta_{itjl}\}}$$

where

$$\begin{aligned} \eta_{itjj} = & \theta_{jj0} + IA_{it}\theta_{jj1} + Q_{it}\theta_{jj2} + TPA_{it}\theta_{jj3} + \Delta SH_{it(t-1)}\theta_{jj4} \\ & + \frac{\Delta SH_{it(t-1)}}{Q_{it}}\theta_{jj5} + \frac{\Delta TPA_{it(t-1)}}{A_{it}}\theta_{jj6} + C_{i1}\theta_{jj7} \end{aligned} \quad [7.18]$$

The new variables, not encountered in the marginal models of § 7.1.1. are

$$\begin{aligned} \Delta SH_{it(t-1)} &= \text{Change in stand height on plot } i \text{ between} \\ &\quad \text{measurements at times } t \text{ and } (t-1). \\ \Delta TPA_{it(t-1)} &= \text{Change in trees per acre on plot } i \text{ between} \\ &\quad \text{measurements at times } t \text{ and } (t-1). \end{aligned}$$

The initial occasion C_{i1} again seemed to be of importance and improved the model fit considerably. Model [7.18] was selected based on the overall likelihood, AIC, and sums of squares criteria. Reducing the dimension for any previous occasion worsened the model significantly, additional terms for all previous occasions or only a subset thereof did not improve the model fit. Table 7.10 lists the parameter estimates and their standard errors along with fit and prediction statistics.

Table 7.10.

Parameter estimates and asymptotic standard errors for transition model [7.18]. Fit via Maximum Likelihood. The parameter vector for staying in once category was constrained to zero. Standard errors italicized.

Previous condition	Current condition	Coefficient estimates and standard errors							
		$\hat{\theta}_{fj0}$	$\hat{\theta}_{fj1}$	$\hat{\theta}_{fj2}$	$\hat{\theta}_{fj3}$	$\hat{\theta}_{fj4}$	$\hat{\theta}_{fj5}$	$\hat{\theta}_{fj6}$	$\hat{\theta}_{fj7}$
Healthy	Branch	-4.377 <i>0.9315</i>	0.0484 <i>0.0376</i>	0.1675 <i>0.1528</i>	0.0008 <i>0.0003</i>	0.0619 <i>0.0537</i>	-0.4839 <i>0.2992</i>	0.0941 <i>0.0298</i>	0.8436 <i>0.1540</i>
Healthy	Stem	-3.621 <i>0.5685</i>	0.0357 <i>0.0255</i>	0.1195 <i>0.0972</i>	-0.0001 <i>0.0002</i>	0.0711 <i>0.0306</i>	-0.2523 <i>0.0159</i>	0.0357 <i>0.0162</i>	1.2013 <i>0.1013</i>
Healthy	Dead	-3.087 <i>1.6480</i>	0.0124 <i>0.0862</i>	-0.2384 <i>0.3017</i>	-0.0009 <i>0.0007</i>	0.0836 <i>0.0935</i>	-0.7487 <i>0.4755</i>	-0.2098 <i>0.0358</i>	0.4019 <i>0.2708</i>
Branch	Stem	-0.1395 <i>1.4436</i>	0.0280 <i>0.0393</i>	-0.0417 <i>0.2079</i>	-0.0005 <i>0.0004</i>	-0.0685 <i>0.0904</i>	0.2772 <i>0.5478</i>	0.1685 <i>0.0484</i>	1.8377 <i>0.2084</i>
Branch	Dead	-10.435 <i>5.8678</i>	0.0221 <i>0.1405</i>	0.7460 <i>0.7856</i>	0.0009 <i>0.0019</i>	-0.1028 <i>0.3146</i>	0.8948 <i>1.9685</i>	-0.1393 <i>0.1364</i>	1.7215 <i>0.7669</i>
Stem	Dead	-1.3351 <i>0.5415</i>	0.0432 <i>0.0203</i>	-0.0845 <i>0.0789</i>	-0.0019 <i>0.0003</i>	-0.0064 <i>0.0323</i>	-0.0604 <i>0.1978</i>	-0.1949 <i>0.0149</i>	-0.4110 <i>0.0986</i>
Goodness of fit statistics for the evaluation data									
Previous Condition:		Healthy	Branch	Stem					
SSR		4.26	47.71	2.63					
WSP		35.49	188.48	9.91					
SSF		160.44	113.33	45.09					
WSF		1439.7	581.9	211.9					
R_p^2		0.95	0.31	0.95					
Entire model across all previous responses									
SSR	54.60	-2 Log-Likelihood		14692.5					
WSP	233.89	Akaike's AIC		-7338.27					
SSF	318.87								
WSF	2233.4								
R_p^2	0.73								

The negative influence of the submodel for the previously branch infected trees on the overall fit statistics is apparent. The R^2 for these conditional responses is only 0.31 based on the evaluation data (0.36 for the development data) and decreases the total pseudo R^2 considerably. Based on the frequency weighted statistics SSF, and WSF the influence of this submodel is not as strong, however.

How well does model [7.18] predict the actual transition pattern over time? To address this question, data from the first two measurement points were used to calculate the transition probabilities for each plot. These were assumed to form the transition matrix for a one-dependent, homogeneous Markov chain (see § 4.4.4). The Markov chain allows one to calculate the one-, two-, three-step transition probabilities and so forth. These can be used to calculate the number of trees that are expected in any category at a future time point assuming the model and the Markov assumption are correct. Since each plot provided a series of observations, these numbers can be contrasted with the actual frequencies observed. If these assumptions are reasonable, the observed and predicted numbers should agree well.

Figures 7.13 and 7.14 display ray charts for this situation. Every ray consists of 5 segments. Each reflects a probability scale from 0 to 1 and the arrows within each segment cumulative probabilities. The innermost circle stands for the initial occasion, the next for the one-step transition probability, the two-step probability, and so forth. The outermost circle thus reflects a transition over 4 steps (12 years) calculated only from the covariate data of the first two occasions, the fitted model [7.18], and the Markov chain. As far as actual observations are available, observed cumulative proportions and predicted proportions are plotted on opposite sides of the rays.

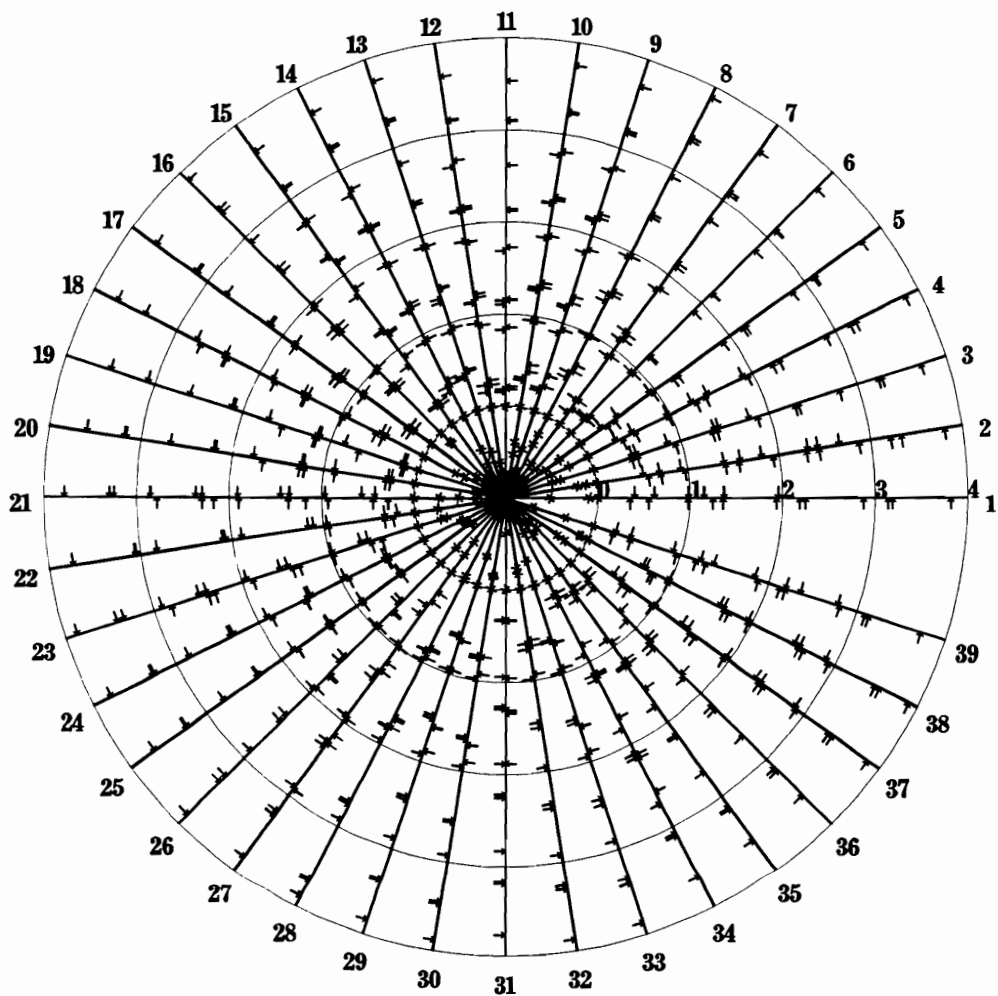


Figure 7.13. Observed and forecasted proportions based on RC model for conditional events. Circles delimit initial occasion and remeasurements. Initial marginal proportions plotted in inner—most segment. Plots with only one measurement are excluded from the analysis.

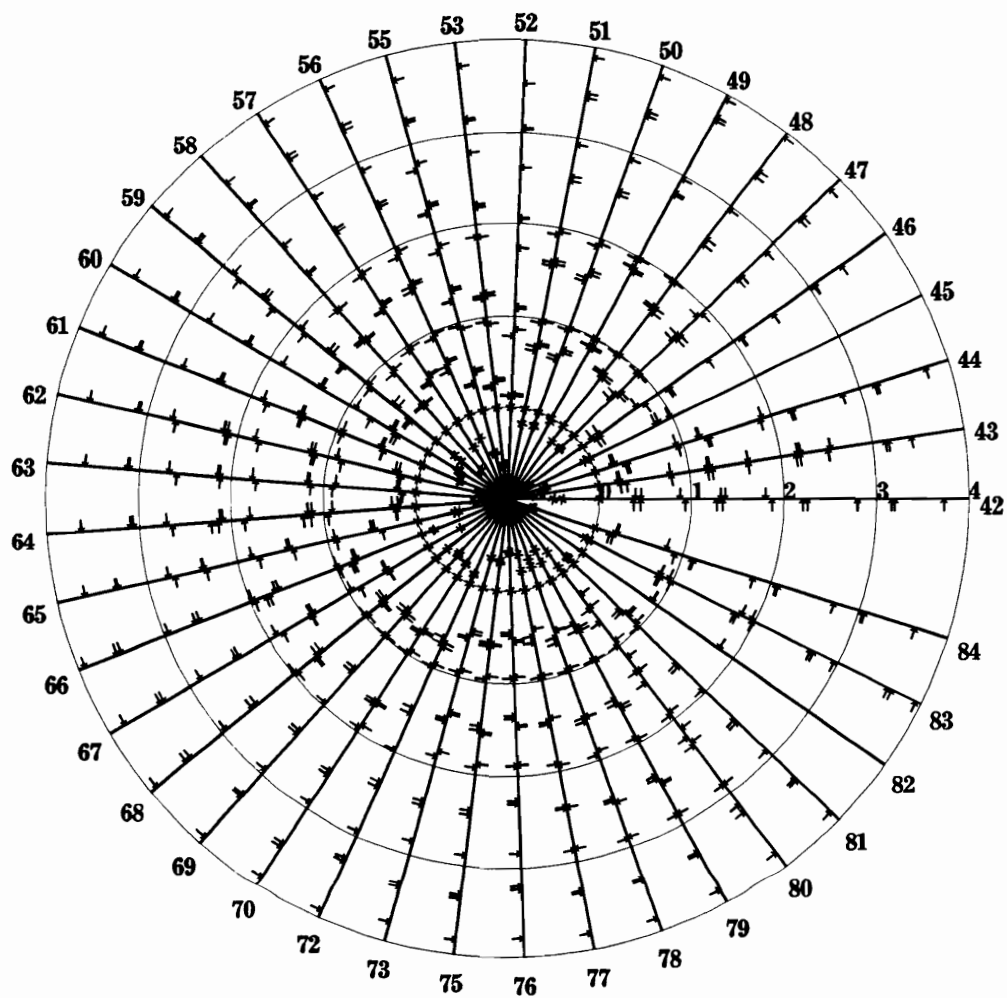


Figure 7.14. Observed and forecasted proportions based on RC model for conditional events. Circles delimit initial occasion and remeasurements. Initial marginal proportions plotted in inner – most segment. Plots with only one measurement are excluded from the analysis.

It should be kept in mind, that the RC transition model is not subject-specific. The estimates in Table 7.10 have a conditional interpretation but are obtained by combining information across the population of plots. In calculating the predictions there is no means of adjusting how a particular plot relates to the overall average like in a mixed model. Having said this, the agreement between observed and predicted proportions in Figures 7.13 - 7.14 is quite remarkable. Observed and predicted arrows are almost opposite of each other for many plots. This instills confidence to use the proportions calculated under the Markov assumption beyond the observed time points. Following the predicted and forecasted arrows along any ray, one finds easily that the proportion of healthy trees decreases continuously, the branch infections are almost constant and an increasing number of trees dies of fusiform rust.

What does model [7.18] imply for the actual transition probabilities? The mean values of the first transition in the slash pine evaluation data was used to produce the transition probabilities for 3 following periods (9 years) and the variables Initial age, Quadratic mean diameter, and Trees per acre were then varied to see how the patterns change with these, assuming other regressors held constant. The results for selected values are listed in Table 7.11. In the first part of this table initial age and quadratic mean diameter are varied, holding trees per acre constant. The second part of Table 7.11 repeats this for a different value of TPA. Impossible transitions such as 2-1, 2-0, etc., and the deterministic 3-3 transition probability are not reported.

Table 7.11.

Transition probabilities predicted from RC model [7.18] and marginal frequencies obtained from applying Markov chain to mean initial frequencies for selected settings of Initial Age, Quadratic mean diameter, and Trees per acre. Three forecasts are listed for each setting.

Transition probabilities									Marginal prob.			
0-0	0-1	0-2	0-3	1-1	1-2	1-3	2-2	2-3	0	1	2	3
<i>Mean of first transition</i>												
0.62	0.05	0.30	0.03	0.08	0.84	0.08	0.82	0.18	0.47	0.04	0.44	0.05
0.49	0.04	0.40	0.07	0.02	0.81	0.17	0.74	0.26	0.37	0.03	0.50	0.09
0.39	0.03	0.47	0.11	0.01	0.75	0.24	0.67	0.33	0.29	0.03	0.54	0.14
<i>Initial age 5 years, Quadr. mean d. 5 in, 200 Trees per acre</i>												
0.64	0.04	0.28	0.04	0.07	0.82	0.10	0.77	0.23	0.48	0.04	0.43	0.06
0.51	0.03	0.37	0.08	0.02	0.77	0.21	0.67	0.33	0.38	0.03	0.48	0.11
0.41	0.03	0.51	0.13	0.01	0.69	0.30	0.59	0.41	0.30	0.02	0.50	0.18
<i>Initial age 9 years, Quadr. mean d. 5 in, 200 Trees per acre</i>												
0.60	0.04	0.31	0.04	0.06	0.82	0.12	0.73	0.27	0.46	0.04	0.44	0.06
0.47	0.04	0.40	0.10	0.02	0.75	0.24	0.63	0.37	0.36	0.03	0.48	0.13
0.36	0.03	0.45	0.16	0.00	0.65	0.35	0.54	0.46	0.28	0.02	0.50	0.20
<i>Initial age 11 years, Quadr. mean d. 5 in, 200 Trees per acre</i>												
0.58	0.05	0.33	0.05	0.06	0.81	0.13	0.72	0.28	0.45	0.04	0.44	0.07
0.44	0.04	0.42	0.10	0.01	0.73	0.26	0.61	0.39	0.35	0.03	0.49	0.14
0.34	0.03	0.46	0.17	0.00	0.63	0.37	0.51	0.49	0.26	0.02	0.50	0.22
<i>Initial age 5 years, Quadr. mean d. 7 in, 200 Trees per acre</i>												
0.58	0.05	0.34	0.03	0.08	0.78	0.14	0.80	0.20	0.45	0.05	0.45	0.05
0.44	0.04	0.44	0.08	0.02	0.75	0.23	0.71	0.29	0.34	0.03	0.52	0.10
0.33	0.03	0.51	0.13	0.01	0.69	0.31	0.64	0.36	0.26	0.03	0.55	0.16
<i>Initial age 9 years, Quadr. mean d. 7 in, 200 Trees per acre</i>												
0.54	0.05	0.37	0.04	0.06	0.78	0.15	0.77	0.23	0.44	0.05	0.46	0.06
0.39	0.04	0.47	0.09	0.02	0.73	0.25	0.67	0.33	0.32	0.04	0.53	0.12
0.29	0.03	0.53	0.15	0.00	0.65	0.34	0.59	0.41	0.23	0.03	0.55	0.19
<i>Initial age 11 years, Quadr. mean d. 7 in, 200 Trees per acre</i>												
0.51	0.06	0.39	0.04	0.06	0.78	0.16	0.75	0.25	0.43	0.05	0.46	0.06
0.37	0.04	0.49	0.10	0.01	0.72	0.27	0.65	0.35	0.31	0.04	0.53	0.13
0.26	0.03	0.53	0.17	0.00	0.63	0.36	0.56	0.44	0.22	0.03	0.55	0.20

Table 7.11.
(Continued).

Transition probabilities									Marginal prob.			
0-0	0-1	0-2	0-3	1-1	1-2	1-3	2-2	2-3	0	1	2	3
<i>Initial age 5 years, Quadr. mean d. 5 in, 400 Trees per acre</i>												
0.64	0.05	0.29	0.03	0.09	0.83	0.08	0.83	0.17	0.48	0.04	0.44	0.04
0.51	0.04	0.39	0.06	0.03	0.82	0.15	0.76	0.24	0.38	0.03	0.50	0.09
0.41	0.03	0.46	0.10	0.01	0.77	0.23	0.69	0.31	0.30	0.03	0.54	0.13
<i>Initial age 9 years, Quadr. mean d. 5 in, 400 Trees per acre</i>												
0.60	0.05	0.32	0.03	0.07	0.83	0.09	0.81	0.19	0.46	0.05	0.45	0.05
0.46	0.04	0.42	0.07	0.02	0.80	0.18	0.72	0.28	0.36	0.03	0.51	0.10
0.36	0.03	0.49	0.12	0.01	0.73	0.26	0.65	0.35	0.28	0.03	0.54	0.15
<i>Initial age 11 years, Quadr. mean d. 5 in, 400 Trees per acre</i>												
0.58	0.05	0.33	0.04	0.07	0.83	0.10	0.79	0.21	0.45	0.05	0.45	0.05
0.44	0.04	0.44	0.08	0.02	0.79	0.19	0.70	0.30	0.34	0.04	0.52	0.10
0.34	0.03	0.50	0.13	0.00	0.72	0.28	0.63	0.37	0.26	0.03	0.55	0.16
<i>Initial age 5 years, Quadr. mean d. 7 in, 400 Trees per acre</i>												
0.58	0.06	0.34	0.03	0.09	0.78	0.14	0.86	0.14	0.45	0.05	0.46	0.04
0.44	0.05	0.46	0.06	0.03	0.78	0.20	0.79	0.21	0.34	0.04	0.54	0.08
0.33	0.04	0.53	0.10	0.01	0.73	0.26	0.73	0.27	0.26	0.03	0.59	0.12
<i>Initial age 9 years, Quadr. mean d. 7 in, 400 Trees per acre</i>												
0.53	0.06	0.37	0.03	0.07	0.78	0.14	0.83	0.17	0.43	0.06	0.47	0.04
0.39	0.05	0.49	0.07	0.02	0.76	0.22	0.76	0.24	0.32	0.04	0.55	0.09
0.28	0.04	0.56	0.12	0.01	0.71	0.29	0.69	0.31	0.23	0.03	0.59	0.14
<i>Initial age 11 years, Quadr. mean d. 7 in, 400 Trees per acre</i>												
0.51	0.07	0.39	0.03	0.07	0.78	0.15	0.82	0.18	0.42	0.06	0.47	0.05
0.36	0.05	0.51	0.08	0.02	0.75	0.23	0.74	0.26	0.30	0.04	0.56	0.10
0.26	0.04	0.57	0.13	0.00	0.69	0.30	0.67	0.33	0.22	0.03	0.60	0.15

Regardless of the covariate values, Table 7.11 reveals some interesting features of model [7.18]. The largest transition probabilities for any given previous state are to remain healthy, to remain stem infected, and to move from a branch to a stem infection. Over time the 0-0 transitions decrease sharply. Based on the mean observation (first panel in Table 7.10), only 39% of the initially healthy trees will still be healthy after 9 years. All transitions that end in the dead state (0-3, 1-3, 2-3) increase over time, simply because the dead state is absorbing. Regardless of the regressors however, the probability to die after a branch infection is considerably smaller than to die after a stem infection, which is of course reasonable based on the unidirectionality of the disease process and the ordering of the response. The marginal proportion of branch infections or rust associated mortality are little responsive to changes in the covariates. On plots of 5-inch quadratic mean diameter the marginal proportions change little with increasing age. If, however trees get thicker or the stand becomes denser the odds of stem infections increase as time progresses. For stands of 7 in quadratic mean diameter an increase in density causes rust associated mortality to decrease. For example after three occasions on a plot initially 9 years of age, increasing the mean diameter from 5 to 7 inches causes stem infections to increase from 50 to 55%. Increasing the density without affecting the mean diameter has a similar effect. If both influences operate simultaneously, the odds increase to 59% (9 years, 7 in, 400 trees per acre).

Overall the dynamics in slash pine seem to be moderate. Over time, they can have a quite sizeable impact, however. The mean initial frequencies in the evaluation data set are $\{0.60, 0.06, 0.33, 0.01\}$ in the four categories. Given these frequencies as a starting point, after 9 years only 29% of the trees will still be healthy, stem infections will have increased from 33 to 54% and 20% of the trees will have died of fusiform

rust. Thus for almost $\frac{3}{4}$ of all trees merchantability is questionable after this period, whereas initially only $\frac{1}{3}$ of the trees are in question.

The RC model [7.18] performs quite well overall and may be acceptable as a tool for forecasting fusiform rust propensities in slash pine. This comes at the cost of a large number of parameters. Although fitting models of this type is not too time consuming, inferences and covariate selection are becoming increasingly difficult, as the number of parameter proliferates. Furthermore, [7.18] does not treat the four response categories as ordered, but only as exclusive. The rationale of an RC model based on the multinomial parameterization can also be applied to a proportional odds model. This results in a decrease of the number of parameters, since now only three parameter vectors need to be estimated, one for each possible previous response, instead of 6. The capability to discriminate categories through a large number of coefficients is traded against a model structure that assumes categories are distinguishable, because they are ordered. The same covariates as in the linear predictor of [7.18] have been used to fit a series of proportional odds models simultaneously. A four category response model is fit to the previously healthy trees, a three category model to the previously branch infected trees, and so forth. Table 7.12 summarizes the results. The linear predictors can be expressed similarly as in [7.18]:

$$\begin{aligned} \Pr(C_{it} \leq j | C_{i(t-1)} = j') &= G(\eta_{itjj}), \\ \eta_{itjj} &= \alpha_{jj} + IA_{it}\theta_{j1} + Q_{it}\theta_{j2} + TPA_{it}\theta_{j3} + \Delta SH_{it(t-1)}\theta_{j4} \\ &\quad + \frac{\Delta SH_{it(t-1)}}{Q_{it}}\theta_{j5} + \frac{\Delta TPA_{it(t-1)}}{A_{it}}\theta_{j6} + C_{i1}\theta_{j7} \end{aligned} \quad [7.19]$$

In contrast to [7.18] there is only one parameter vector for each previous response, but a different number of intercepts, depending on the possible transitions from any

state. Table 7.12 lists coefficient estimates and standard errors along with goodness-of-fit statistics for model [7.19]. As expected the likelihood decreased compared to model [7.18], but the overall fit statistics do not indicate much loss. The difference in AIC is much smaller than the one in log-likelihoods since model [7.18] is heavily penalized for its large number of parameters. The model for previously stem infected trees leads to the same fit statistics in both approaches. This is not surprising, since given a tree was stem infected, only two outcomes are possible. Thus, this submodel is a binary choice model and [7.18] and [7.19] are reparameterizations of the same logistic relationship. The actual coefficient estimates in the two parameterizations differ, however.

Table 7.13 displays asymptotic 95% confidence intervals for the predicted cumulative probabilities in model [7.19]. As in the marginal models the branch infected state emerges as being predicted with high uncertainty.

Figures 7.15 - 7.16 display the forecasted and observed proportions calculated from [7.19] under the assumption of a one-dependent Markov chain. A comparison with Figures 7.13 - 7.14 shows that forecasts under either model are almost indistinguishable. On these grounds use of a more parsimonious model utilizing ordering of categories recommends itself.

Table 7.12.
 Parameter estimates and asymptotic standard errors for transition model [7.19].
 Fit via Maximum Likelihood. Standard errors italicized.

Previously healthy trees, $j' = 0$									
α_{00}	α_{01}	α_{02}	$\hat{\theta}_{j1}$	$\hat{\theta}_{j2}$	$\hat{\theta}_{j3}$	$\hat{\theta}_{j4}$	$\hat{\theta}_{j5}$	$\hat{\theta}_{j6}$	$\hat{\theta}_{j7}$
2.9372	3.2838	5.909	0.0409	-0.0958	0.0003	-0.0642	0.3068	-0.0152	-1.013
			0.0214	0.0817	0.0002	0.0260	0.1360	0.0134	0.0831
Previously branch infected trees, $j' = 1$									
α_{11}	α_{12}		$\hat{\theta}_{j1}$	$\hat{\theta}_{j2}$	$\hat{\theta}_{j3}$	$\hat{\theta}_{j4}$	$\hat{\theta}_{j5}$	$\hat{\theta}_{j6}$	$\hat{\theta}_{j7}$
0.4784	5.0896		-0.0263	0.0012	0.0004	0.0627	-0.2612	-0.1332	-1.8202
			0.0378	0.1952	0.0004	0.0827	0.4981	0.0461	0.2020
Previously stem infected trees, $j' = 2$									
α_{22}			$\hat{\theta}_{j1}$	$\hat{\theta}_{j2}$	$\hat{\theta}_{j3}$	$\hat{\theta}_{j4}$	$\hat{\theta}_{j5}$	$\hat{\theta}_{j6}$	$\hat{\theta}_{j7}$
1.3351			-0.0432	0.0845	0.0019	0.0064	0.0604	0.1949	0.4110
			0.0203	0.0789	0.0003	0.0323	0.1978	0.0148	0.0985
Goodness of fit statistics for the evaluation data									
Previous Condition:		Healthy	Branch	Stem					
SSR		4.34	46.35	2.63					
WSP		37.04	268.16	9.91					
SSF		164.46	109.68	45.09					
WSF		1346.39	1014.82	211.9					
R_p^2		0.95	0.33	0.95					
Entire model across all previous responses									
SSR	53.32	-2 Log-Likelihood	14806.88						
WSP	315.11	Akaike's AIC	-7376.44						
SSF	319.23								
WSF	2573.12								
R_p^2	0.74								

Table 7.13.

Asymptotic 95 % confidence intervals for predicted proportions in slash pine transition model [7.19]. Length of intervals in probability units.

	Current state of fusiform rust infection		
	Healthy	Branch	Stem
	$j = 0$	$j = 1$	$j = 2$
<u>Previously healthy trees, $j' = 0$</u>			
length l	0.04864	0.03945	0.00615
$\Pr(C = j)$	0.82878	0.04240	0.11790
$l/\Pr(C \leq j)[\%]$	5.86	4.53	0.62
$l/\Pr(C = j)[\%]$	5.86	93.05	5.21
<u>Previously branch infected, $j' = 1$</u>			
length l		0.18362	0.03090
$\Pr(C = j)$		0.52511	0.46039
$l/\Pr(C \leq j)[\%]$		34.97	3.13
$l/\Pr(C = j)[\%]$		34.97	6.71
<u>Previously stem infected, $j' = 2$</u>			
length l			0.04441
$\Pr(C = j)$			0.86832
$l/\Pr(C \leq j)[\%]$			5.05
$l/\Pr(C = j)[\%]$			5.05

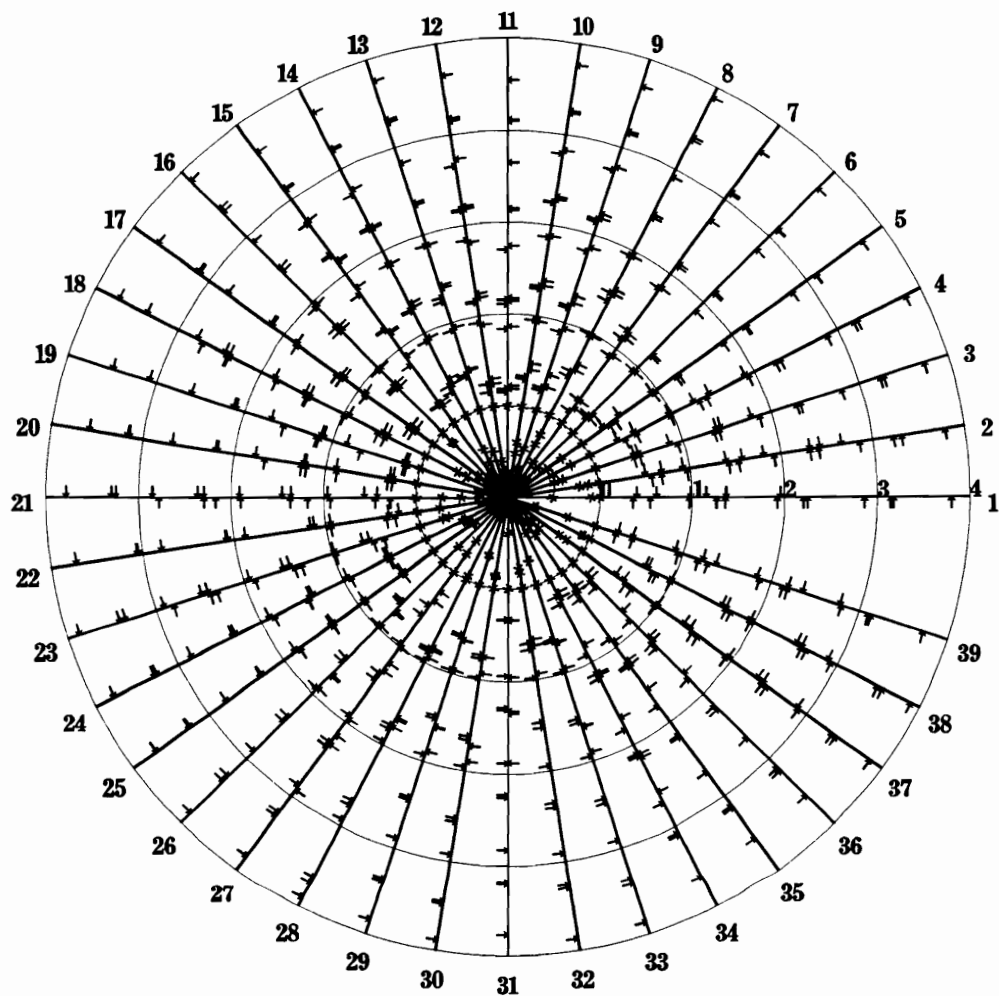


Figure 7.15. Observed and forecasted proportions based on separate proportional odds models. Circles delimit initial occasion and remeasurements. Initial marginal proportions plotted in inner—most segment. Plots with only one measurement are excluded from the analysis.

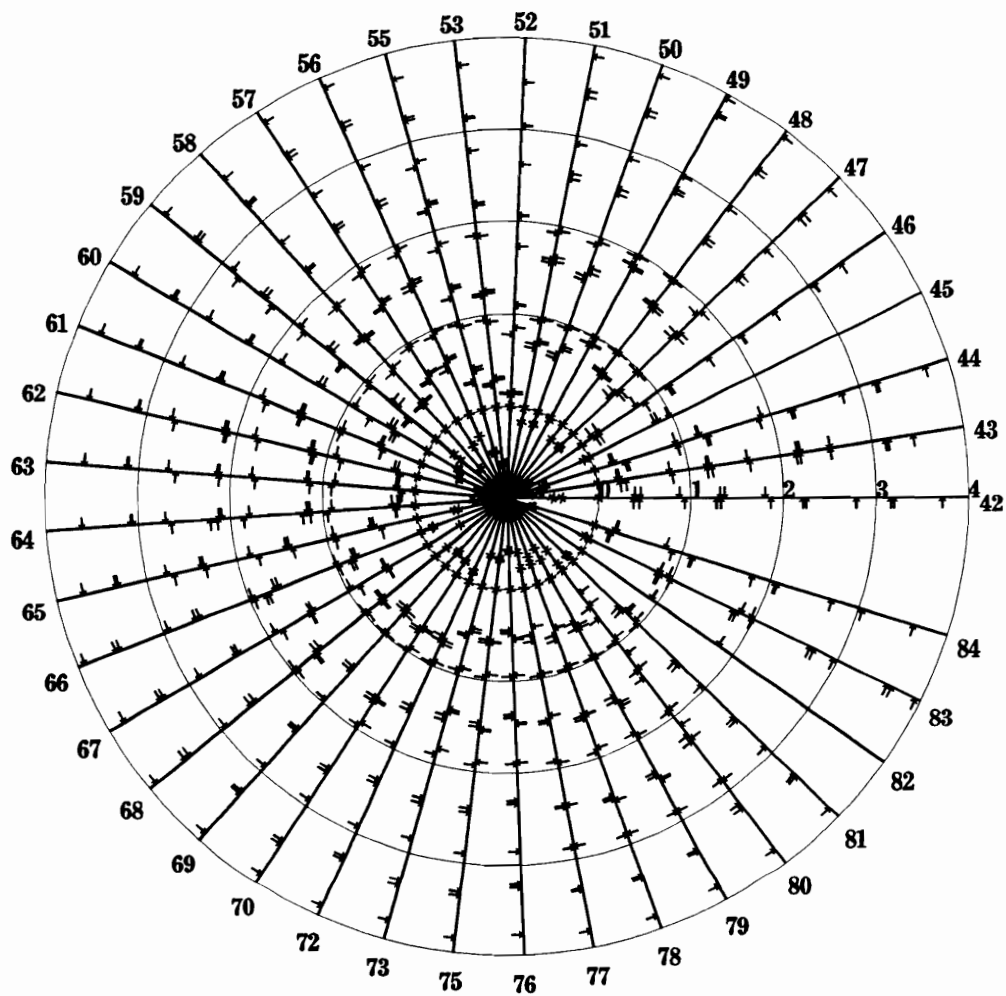


Figure 7.16. Observed and forecasted proportions based on separate proportional odds models. Circles delimit initial occasion and remeasurements. Initial marginal proportions plotted in inner – most segment. Plots with only one measurement are excluded from the analysis.

7.1.2.2. CC model for conditional events

The final slash pine model for conditional events is drawn from Bonney's regressive logistic model as introduced in § 4.4.3.1 and adapted for an ordered response. The linear predictor of the underlying proportional odds model consists of a set of covariates as well as a series of indicators to model events conditional on previous responses. Following the assumption that observations are uncorrelated given the dependence parameters, ordinary maximum likelihood can be employed. The best fitting slash pine model was selected based on the overall likelihood as well as Akaike's information criterion. For simplicity, only the previous response was used as a conditioning event. The linear predictor of the POM is

$$\begin{aligned} \eta_{itj} = & \alpha_j + \beta_1 SI_i + \beta_2 Q_{it} + \beta_3 SH_{it} + \beta_4 IA_i + \beta_5 BU_i \\ & + \beta_6 \frac{SH_{it}}{Q_{it}} + \gamma_0 Z_{i0} + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} \end{aligned} \quad [7.20]$$

The conditioning covariates are defined as

$$Z_{i0} = \begin{cases} 1 & \text{if } j' = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$Z_{i1} = \begin{cases} 1 & \text{if } j' = 1 \\ 0 & \text{otherwise} \end{cases} .$$

$$Z_{i2} = \begin{cases} 1 & \text{if } j' = 2 \\ 0 & \text{otherwise} \end{cases}$$

The interesting feature of model [7.20] is that it allows to generate transition probabilities from purely cross-sectional data. By varying Z_{i0} , Z_{i1} , and Z_{i2} in turn, one can generate the transition probabilities from different previous states $j' = 0, \dots, 2$. If all Z_{ij} are set to zero, the model becomes population-averaged, enabling one to predict the

marginal fusiform rust probabilities. In this sense, covariate conditioned models like [7.20] can be viewed as a mixture of marginal and purely transitional models as discussed in § 7.1.2.1. They typically fit marginal probabilities not as good as a true population-averaged model and transitional probabilities poorer as RC models. However, a well selected suite of measured and conditioning covariates may provide a reasonable compromise between both types of models.

Model [7.20] does not account for the unidirectionality of the response. This would require to simultaneously constrain some of the cut-off points which is not consistent with the development of the proportional odds model. It was decided here, to set transition probabilities for improbable transitions to zero and rescale the remaining probabilities so that they sum to one. The conditioning covariates in [7.20] can be refined further, by using separate sets of variates Z_{ij} for different previous responses or time points. It turned out that little could be gained by such refinements.

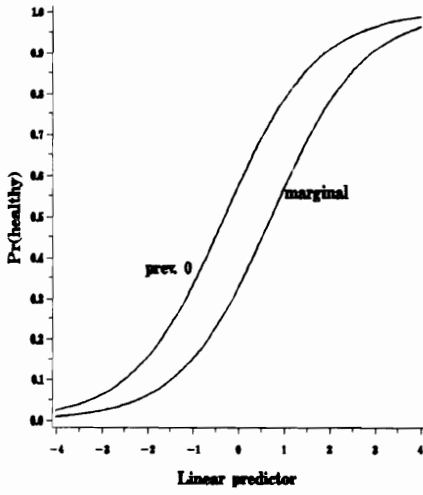
Since we are faced with a grouped case analysis on the plot level the observations need to be rearranged. The grouping procedure collects all observations on a plot that share the same response **and** covariate pattern. For the population averaged models this creates as many observations per plot as there are time points, for the RC model as many as there are transitions ($\#$ time points - 1). In model [7.20] this procedure creates much more groups per plot, depending on the actual responses and transitions observed, since they augment the data matrix through the Z_{ij} . Sums of squares criteria are thus not comparable between CC and RC transitional models. Table 7.14 depicts the coefficient estimates and standard errors for [7.20].

Table 7.14.
 Results of fitting model [7.20] to slash pine development data.
 Standard errors italicized.

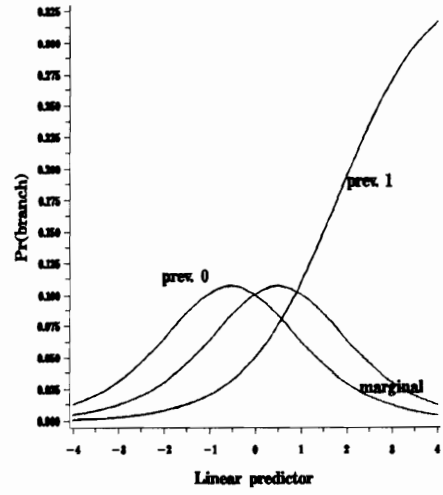
Cut-offs	Covariates					
	SI_i	Q_{it}	SH_{it}	IA_i	BU_i	$\frac{SH_{it}}{Q_{it}}$
$\alpha_0 = -0.7157$	-0.0207	0.2125	-0.0170	-0.1466	-1.0386	0.4400
$\alpha_1 = -0.2839$	<i>0.0026</i>	<i>0.0746</i>	<i>0.0099</i>	<i>0.0105</i>	<i>0.1711</i>	<i>0.0452</i>
$\alpha_2 = 5.9440$						
Dependence parameters						
$\hat{\gamma}_0 : 1.0256 \ (0.0553)$ $\hat{\gamma}_1 : -1.4989 \ (0.0860)$ $\hat{\gamma}_2 : -5.3008 \ (0.1179)$						

The dependence parameters have a very profound effect on the predicted probabilities. To demonstrate, the marginal and transitional probabilities were generated using the parameter estimates in Table 7.12. and varying the linear predictor. The results are plotted in Figure 7.17 that contains one panel for each response category. Comparing the transition probabilities for previous healthy trees with the marginal probabilities shows a shift of the response curves to the left of the marginal profile. That is, knowing a tree was healthy at the last occasion, its probability to remain healthy is larger, to become stem infected or die of fusiform rust is smaller than the respective marginal probability ignoring knowledge about previous states. For currently branch infected trees (Fig. 7.17 b)) the relationship reverses. For small values of the linear predictor $\Pr(\text{branch} \mid \text{healthy}) > \Pr(\text{branch})$ while $\Pr(\text{branch} \mid \text{healthy}) < \Pr(\text{branch})$ for large values.

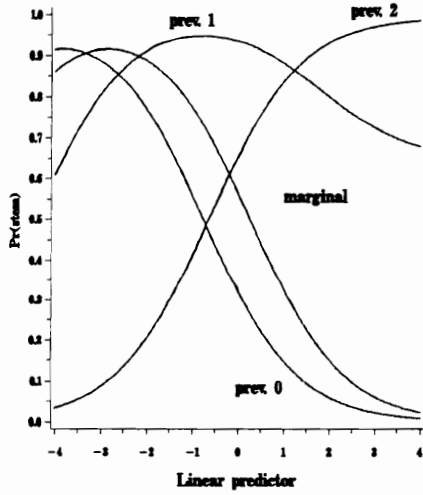
a)



b)



c)



d)

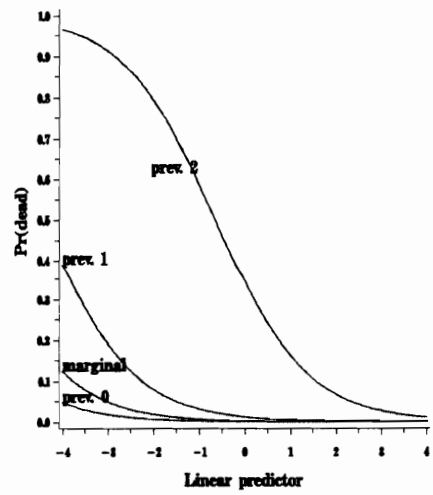


Figure 7.17. Marginal and transition category probabilities as generated by model [7.20]. Horizontal axis denotes the linear predictor without cut-off points. Prev. 0, 1, 2 denotes previously healthy, branch, and stem infected trees, respectively.

For previously branch or stem infected trees we observe not only a shift, since the probabilities were rescaled to eliminate improbable transitions. From panel d) it can be seen how the probability to die of fusiform rust depends on the previous state. RAM occurs most likely after a previous stem infection, less likely after branch infections. To move from the healthy to the dead state in a single transition is more unlikely than overall (marginal) RAM. Model [7.20] appears to capture the dynamics in slash pine discussed in § 6.1.2.3.1 satisfactorily.

Figures 7.18 and 7.19 finally present the results from predicting and forecasting future proportions in the slash pine evaluation data set for sake of comparison with Figures 7.15 and 7.16. The much more parsimonious model [7.20] predicts future proportions not as well as the multinomial or ordinal RC models. Overall the observed and forecasted values agree well, however.

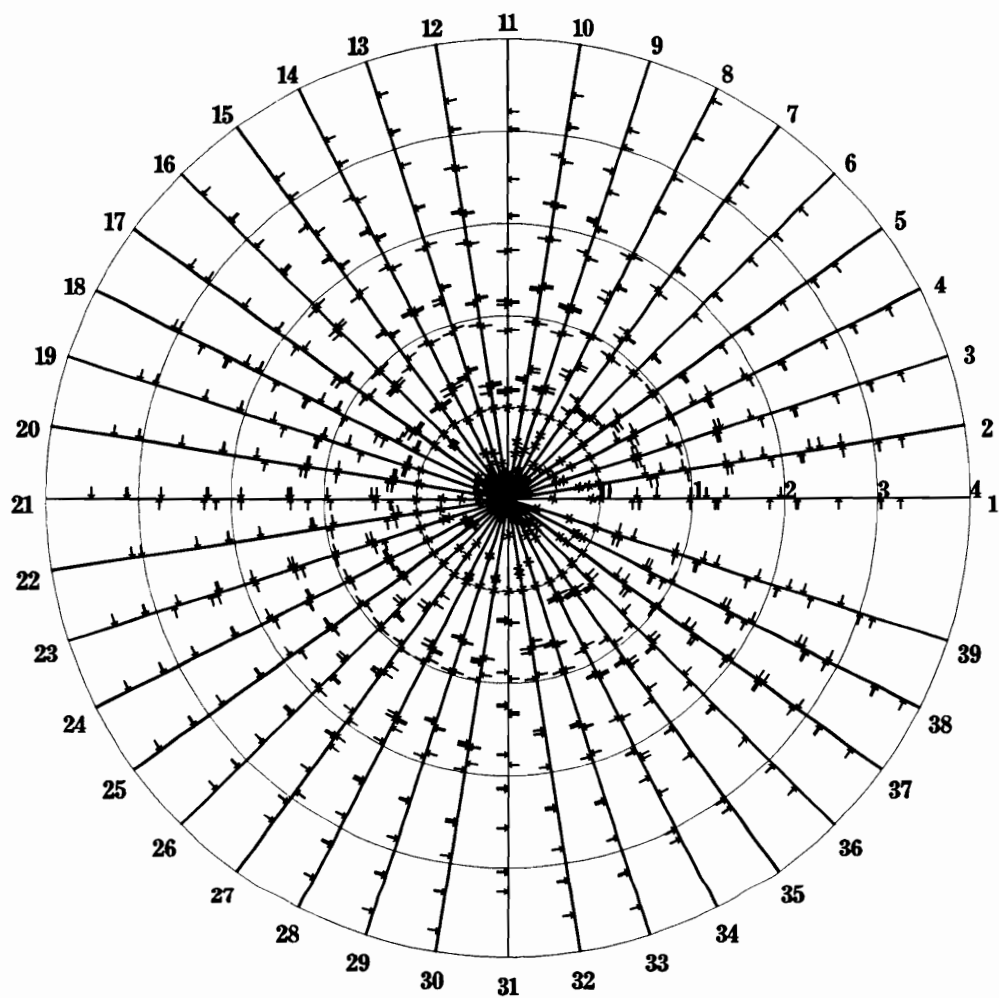


Figure 7.18. Observed and forecasted proportions based on covariate conditioned odds model. Circles delimit initial occasion and remeasurements. Initial marginal proportions plotted in inner—most segment. Plots without remeasurements are excluded.

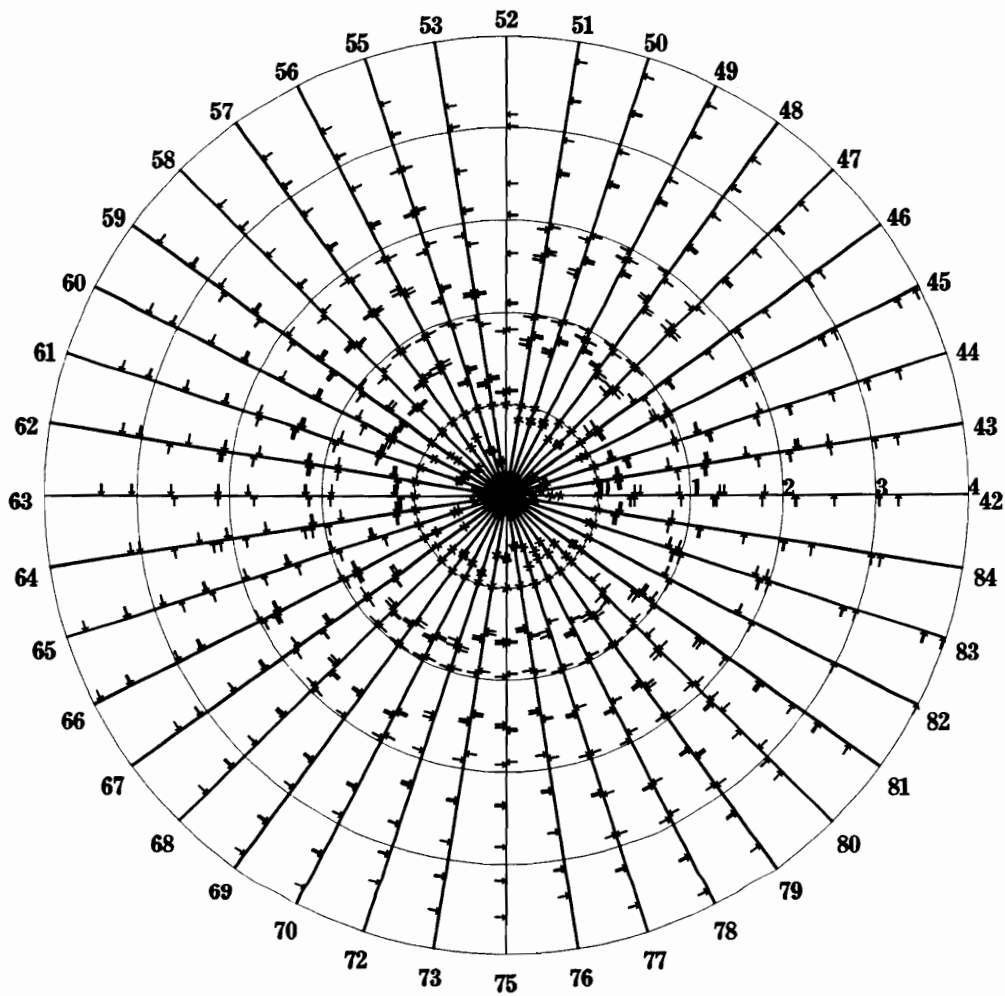


Figure 7.19. Observed and forecasted proportions based on covariate conditioned odds model. Circles delimit initial occasion and remeasurements. Initial marginal proportions plotted in inner-most segment. Plots without remeasurements are excluded.

7.2. Loblolly Pine

7.2.1. Marginal Plot Models

7.2.1.1. Biological patterns

As for slash pine modeling started by generating profile plots from single covariate proportional odds models. Figures 7.20 and 7.21 display these. The profiles are quite different from the ones for slash pine. Site index by itself does not appear to carry much information and the profiles are rather flat. The same applies to terrain slope. In all graphs the high percentage of healthy trees is obvious and stem infections change more slowly with the covariates as in slash pine (see Fig. 7.2 and 7.3). These changes are monotonic for healthy and stem infected proportions in loblolly pine while sigmoidal in slash pine. Within the range of observed covariates stem infections in slash pine show a maximum, which is not the case for loblolly pine.

As for slash pine site quality measures were paired with density measures and observed proportions for each combination fitted by fifth degree polynomials. The best agreement was obtained with quadratic mean diameter and stand height, which thus served as the first covariates to enter in the marginal models. Figure 7.22 shows the interpolated profiles and Figure 7.23 the profiles generated under the proportional odds model. On plots with large mean diameter the proportion of healthy trees increases with stand height and the stem infections decrease accordingly. For small mean diameters, stand height appears to be of little influence on any rust condition.

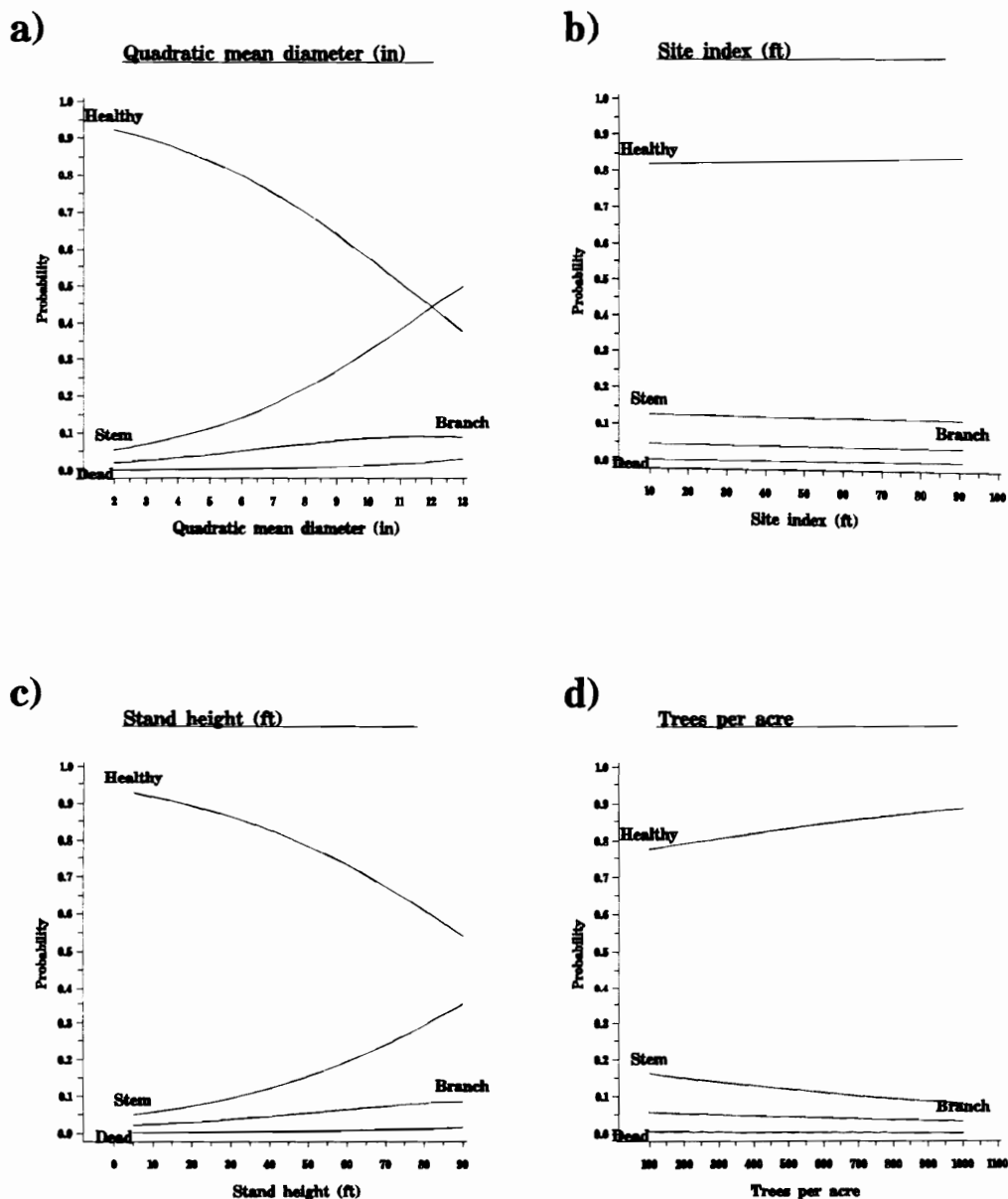


Figure 7.20. Profile plots for single covariate proportional odds models.
Species and data: Loblolly pine development data.

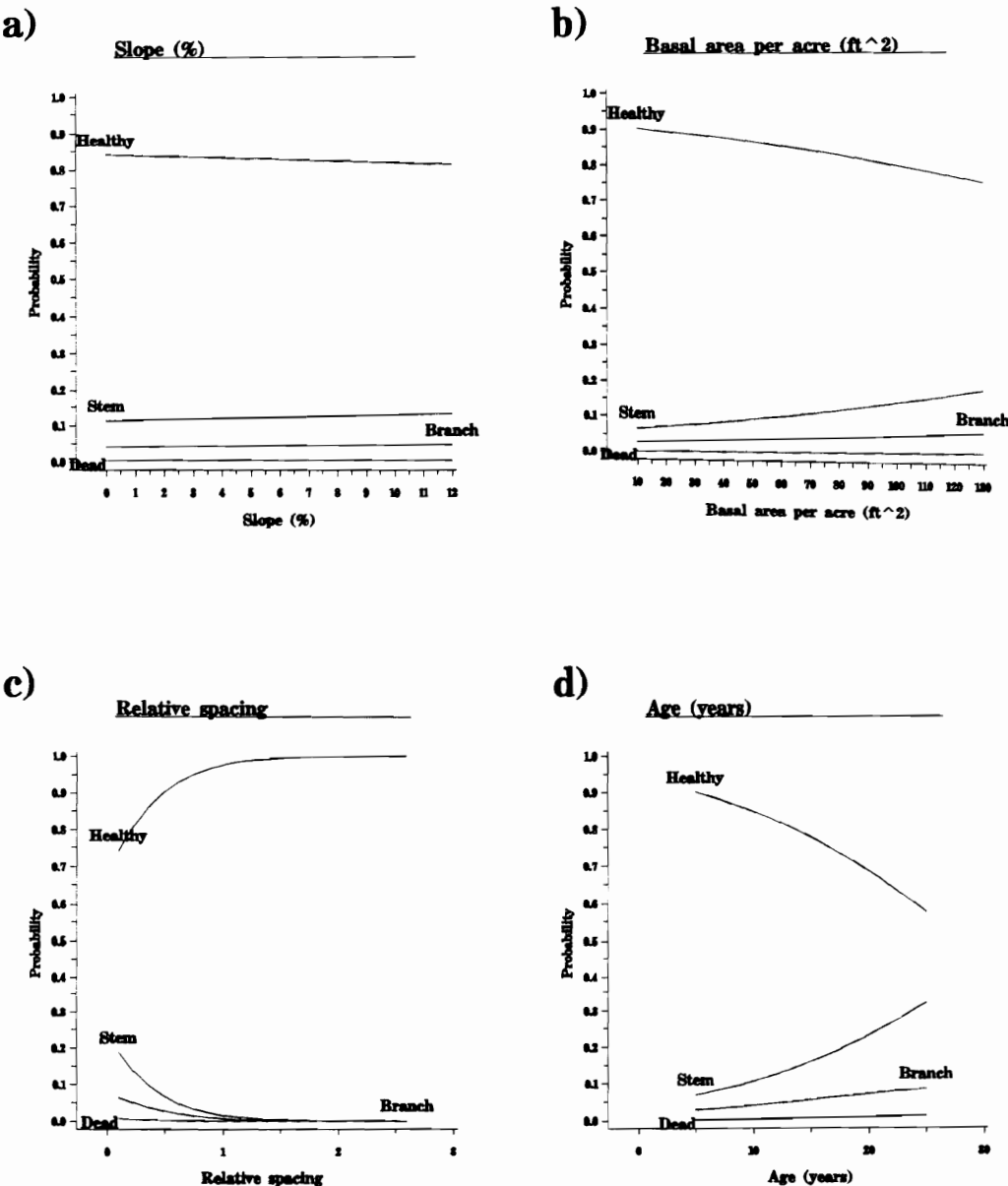
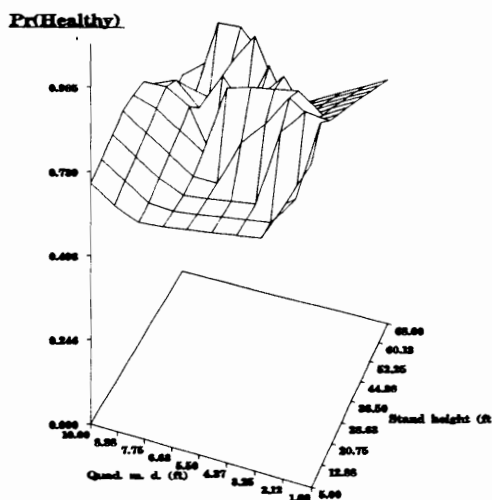
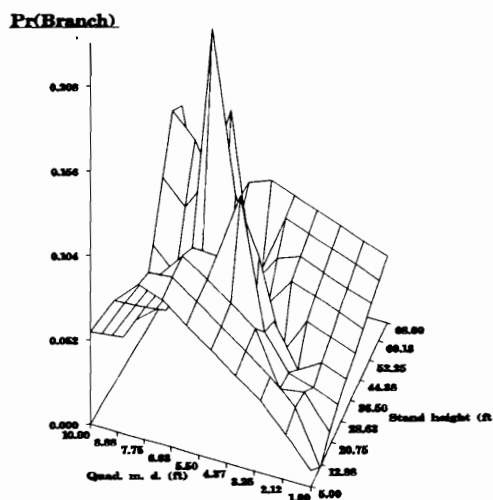


Figure 7.21. Profile plots for single covariate proportional odds models.
Species and data: Loblolly pine development data.

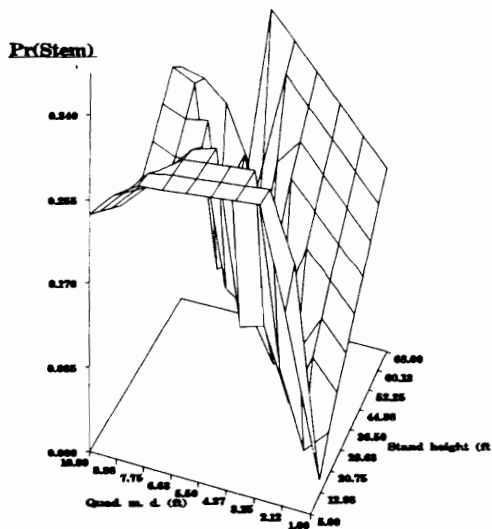
a)



b)



c)



d)

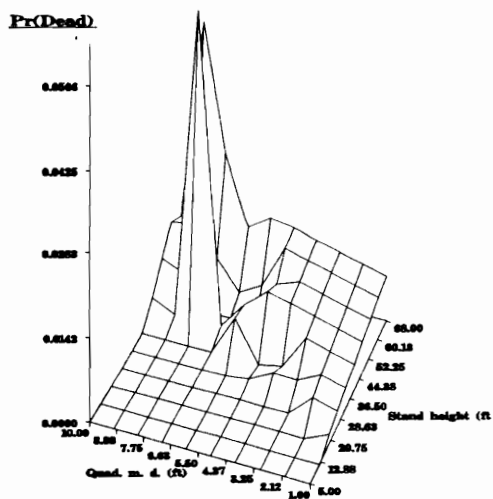
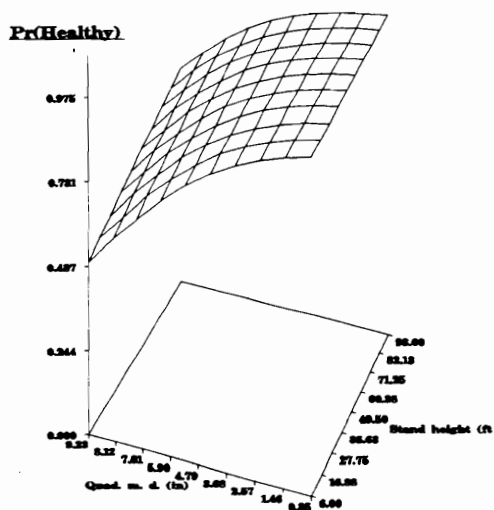
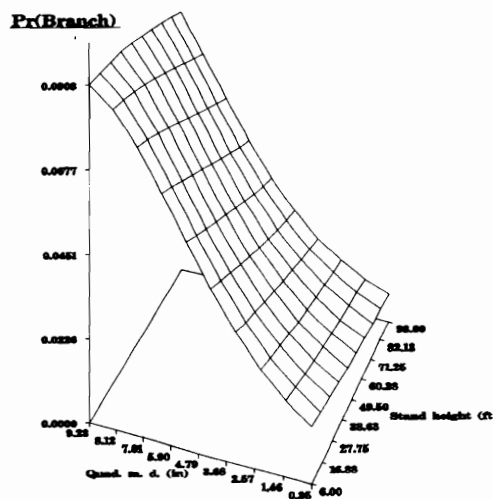


Figure 7.22. Observed probability surfaces for Loblolly pine.
 Surface interpolated by 5th order polynomials.
 Covariates: Quadratic mean diameter (in), Stand height (ft).

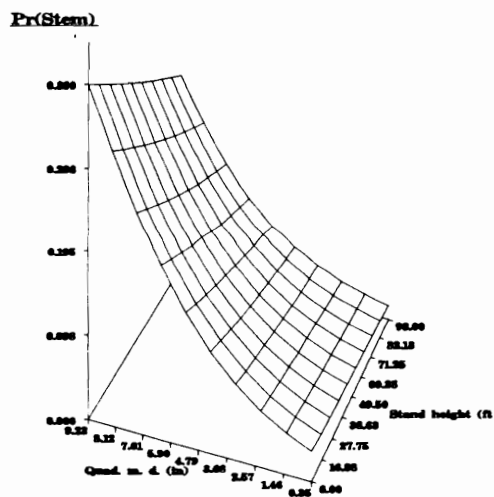
a)



b)



c)



d)

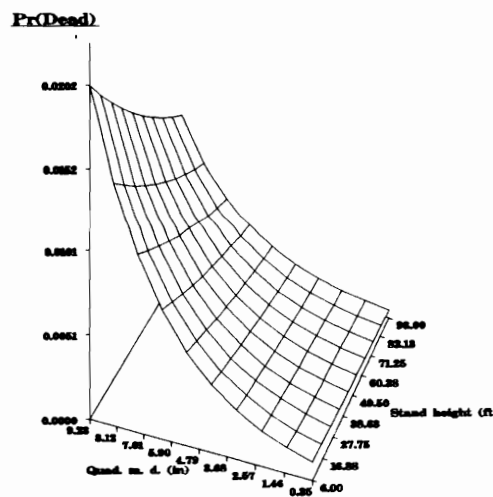


Figure 7.23. Predicted probability surfaces for two covariate proportional odds model. Loblolly pine development data. Covariates: Quadratic mean diameter (in), Stand height (ft).

7.2.1.2. Ordinal response models

Starting with quadratic mean diameter and stand height a series of models was fitted by maximum likelihood under independence and GEE0 with 1-, 2-, and 3-dependence. The best fitting and predicting model was selected based on sums of squares and pseudo- R^2 measures as introduced in § 7.1.1.1. Its components are

$$\Pr(C_{it} \leq j) = G^{-1}(\eta_{itj})$$

where

$$\begin{aligned} \eta_{itj} = & \alpha_j + \beta_1 Q_{it} + \beta_2 SH_{it} + \beta_3 N_i + \beta_4 A_{it} + \beta_5 C_{1it} \\ & + \beta_6 WI_i + \beta_7 \frac{Q_{it}}{A_{it}} + \beta_8 \frac{D^2 H_{it}}{1000} \end{aligned} \quad [7.21]$$

and

Q_{it}	=	Quadratic mean diameter on plot i at time t
SH_{it}	=	Stand height of plot i at time t
N_i	=	Geographic location indicator
A_{it}	=	Age of plot i at time t
C_{1it}	=	Indicator; 1 if first occasion, 0 otherwise
WI_i	=	Indicator; 1 if plot was windrowed, 0 otherwise
$D^2 H_{it}$	=	Tree diameter at breast height squared times tree height; serves as a volume proxy on plot i at time t prorated on a per acre basis. In $ft^3/acre$.

The indicator for geographic location was set to 1 if the plots were located in the north-eastern part of Texas, i.e. counties Cass, Harrison, Marion, Panola, Red River, and Rusk. Otherwise $N_i = 0$. Arabatzis (1990) has found this indicator to be a useful covariate in fitting a subset of the data. This was also found to be true for the complete

data used here. The volume proxy D^2H_{it} made a significant contribution to the model even after stand height and quadratic mean diameter were already in the model. This is interesting, because the same covariate had no bearing in the slash pine models.

For a subject-specific approach model [7.21] was fitted with several combinations of one and more random effects by semi-parametric methods (§ 4.3.1). In this process considerable numerical difficulties were encountered. The algorithm typically failed to converge or the random effects variance was estimated as zero, reducing the model to a population-averaged one. It is believed that the combination of fixed effects in [7.21] combined with the slowly varying response probabilities in loblolly pine and the already good fit of model [7.21] (see Figures 7.24-7.27 later in this section) does not support the assumption that certain coefficients vary across plots. The subject-specific model could be fitted though after reducing the number of covariates in [7.21]. The selected subject-specific model uses linear predictor

$$\eta_{itj} = \alpha_j + \beta_1 Q_{it} + \beta_2 SH_{it} + \beta_3 N_i + \beta_4 A_{it} + \beta_5 C_{1it} + \beta_6 WI_i + b_i \frac{SI_{it}}{100} \quad [7.22]$$

where the effect of site index is assumed to vary across plots.

Gregoire, Schabenberger, and Barrett (1995) encountered similar behavior, in that certain fixed effects are not supported after random effects are included and modified one of their models accordingly. While these authors changed the non-stochastic part of the model to achieve a better fit, here the modification was warranted for numerical reasons.

Table 7.15 shows parameter estimates, asymptotic standard errors and goodness-of-fit statistics for model [7.21], Table 7.16 for model [7.22].

Table 7.15.
Results of fitting model [7.21] to loblolly pine data.
Asymptotic standard errors in parentheses.

Coeff.	Variable	ML	GEE0-1	GEE0-2	GEE0-3
$\hat{\psi}_0$		4.43569	3.21272	3.78632	3.30085
$\hat{\psi}_1$		-1.00216	-0.98980	-1.03947	-0.97015
$\hat{\psi}_2$		1.25934	1.27461	1.24818	1.31443
$\hat{\beta}_1$	Q	-0.08367 (0.03234)	-0.28262 (0.04808)	-0.03028 (0.02690)	-0.10122 (0.02019)
$\hat{\beta}_2$	SH	0.02206 (0.00366)	0.04668 (0.00588)	-0.01594 (0.00295)	0.00053 (0.00206)
$\hat{\beta}_3$	N	0.38655 (0.03744)	0.39100 (0.05679)	-0.10238 (0.05852)	0.29123 (0.05800)
$\hat{\beta}_4$	A	-0.15994 (0.01287)	-0.10664 (0.01872)	-0.11433 (0.01086)	-0.05206 (0.00770)
$\hat{\beta}_5$	C_1	0.27328 (0.03132)	0.17732 (0.03745)	0.05220 (0.02229)	0.32915 (0.01311)
$\hat{\beta}_6$	WI	-0.22901 (0.02516)	-0.23033 (0.03879)	-0.15098 (0.03828)	-0.19132 (0.03775)
$\hat{\beta}_7$	Q/A	-3.20550 (0.31722)	-1.01035 (0.45694)	-1.52906 (0.27767)	-1.29048 (0.17050)
$\hat{\beta}_8$	D^2H	0.02496 (0.01135)	-0.04155 (0.01615)	0.02161 (0.01074)	0.02617 (0.00912)
SSR		16.07	16.62	17.55	16.91
WSP		92.61	98.07	100.63	107.88
WSF		1520.05	1574.38	1640.51	1563.96
SSF		8835.95	9390.10	9392.41	10050.74
R_p^2		0.94	0.94	0.93	0.94

In all four cases displayed in Table 7.15, the coefficients are partially significant at the 5% level, except for stand height in GEE0-3. The actual sign and size of most coefficients changes considerably depending on the order of dependence, while $\hat{\beta}_6$ and $\hat{\beta}_8$ are least affected.

The columns from left to right actually reflect a sequence of increasing dependency. The maximum likelihood fit corresponds to a GEE0 model with 0-dependence. The goodness-of-fit criteria are quite similar, but would point towards favoring the maximum likelihood fit, if predictions alone are important. However, if the precision of the coefficient estimates is of importance, either one of the GEE0 fits should be preferred. As for the slash pine data, a population-averaged approach to modeling correlated categorical data does not guarantee better predictive capabilities than if the correlations are ignored. In fact, the correlations in the loblolly pine data are quite substantial. Between any two time points the correlations decrease as successive observations on the ordinal scale are further apart and largest for staying in once category. The magnitude of these correlations for staying ranges from 0.13 to 0.8 across one occasion, 0.08 to 0.6 across two occasions, and -0.1 to 0.4 across three occasions in the GEE0-3 fit. Correlations of this magnitude should not be ignored.

For the mixed model in Table 7.16, the fit statistics based on the fixed effects of [7.22] only are not much worse than the ones for model [7.21]. Upon including random heterogeneity across plots, the gain in predictive power of the subject-specific model is not as dramatic as for the slash pine model, but sizeable.

This is also evidenced in Table 7.17, which displays asymptotic 95% confidence intervals for the predicted probabilities in [7.22]. The mixed model bounds based on the fixed effects only, are very precise. There appears to be a trend in Table 7.17 with respect to the length of the asymptotic confidence intervals. With the exception of the inconsistent ML bounds, more detailed accounting of correlations improves the precision of the predicted probabilities. The length of the intervals decreases in the order GEE0-1, GEE0-2, GEE0-3. The mixed model bounds appear best. However, they were obtained ignoring random effects variability and will necessarily change, when this variability is accounted for.

Figures 7.24 - 7.27 show the prediction plots for [7.21], 7.28-7.31 for [7.22] based on the evaluation data.

Table 7.16.
Results from fitting model [7.22] semi-parametrically

Coeff.	Variable	Estimate	GOF Statistics	
$\hat{\psi}_0$		3.16133	SSR	16.95
			WSP	101.52
$\hat{\psi}_1$		-0.91378	WSF	1590.48
			SSF	9547.93
$\hat{\psi}_2$		1.35786	R_p^2	0.94
$\hat{\beta}_1$	Q	-0.23749 (0.03059)	SSR _b	8.75
			WSP _b	65.31
$\hat{\beta}_2$	SH	0.01140 (0.00451)	WSF _b	845.50
			SSF _b	6158.33
$\hat{\beta}_3$	N	0.27728 (0.16857)	R_{pb}^2	0.97
$\hat{\beta}_4$	A	-0.04694 (0.01301)		
$\hat{\beta}_5$	C_1	0.26416 (0.04332)	$\widehat{\text{var}}(\mathbf{b}_i) = \hat{D} = 1.06158$	
$\hat{\beta}_6$	WI	-0.21045 (0.11719)		

Table 7.17.

Asymptotic 95 % confidence intervals for predicted proportions in marginal loblolly pine model [7.22]. Results for mixed model based on fixed effects only, ignoring random effects variability. Length of intervals in probability units.

		State of fusiform rust infection		
		Healthy $j = 0$	Branch $j = 1$	Stem $j = 2$
ML	length l	0.01798	0.01440	0.00125
	$\Pr(C = j)$	0.81690	0.03786	0.12420
	$l/\Pr(C \leq j)[\%]$	2.17	1.65	0.13
	$l/\Pr(C = j)[\%]$	2.17	31.84	1.00
GEE0-1	length l	0.02579	0.02052	0.00138
	$\Pr(C = j)$	0.83388	0.04417	0.11797
	$l/\Pr(C \leq j)[\%]$	3.09	2.34	0.13
	$l/\Pr(C = j)[\%]$	3.09	46.45	1.17
GEE0-2	length l	0.02118	0.01732	0.00139
	$\Pr(C = j)$	0.81850	0.04323	0.12280
	$l/\Pr(C \leq j)[\%]$	2.55	1.98	0.14
	$l/\Pr(C = j)[\%]$	2.55	40.07	1.13
GEE0-3	length l	0.01819	0.01449	0.00100
	$\Pr(C = j)$	0.84340	0.04305	0.11040
	$l/\Pr(C \leq j)[\%]$	2.16	1.64	0.10
	$l/\Pr(C = j)[\%]$	2.16	33.67	0.91
Mixed	length l	0.01571	0.01255	0.00117
	$\Pr(C = j)$	0.83346	0.04367	0.11862
	$l/\Pr(C \leq j)[\%]$	1.88	1.43	0.12
	$l/\Pr(C = j)[\%]$	1.88	28.74	0.98

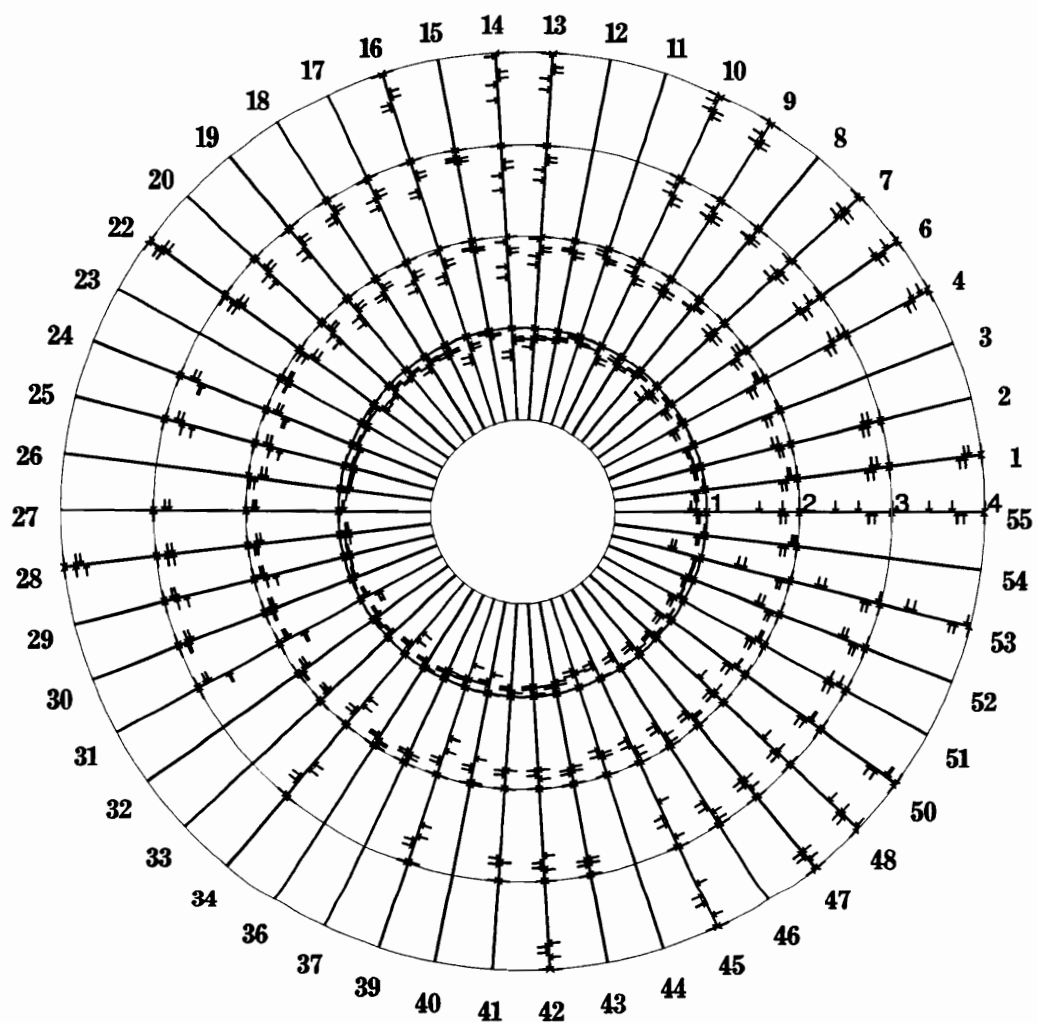


Figure 7.24. Observed and predicted proportions in Loblolly pine model [7.21].
 Each segment corresponds to one occasion and a 0–100% interval.
 Observed values denoted by clockwise arrows, predicted values
 counter clockwise. GEE0 with 1–dependence structure.

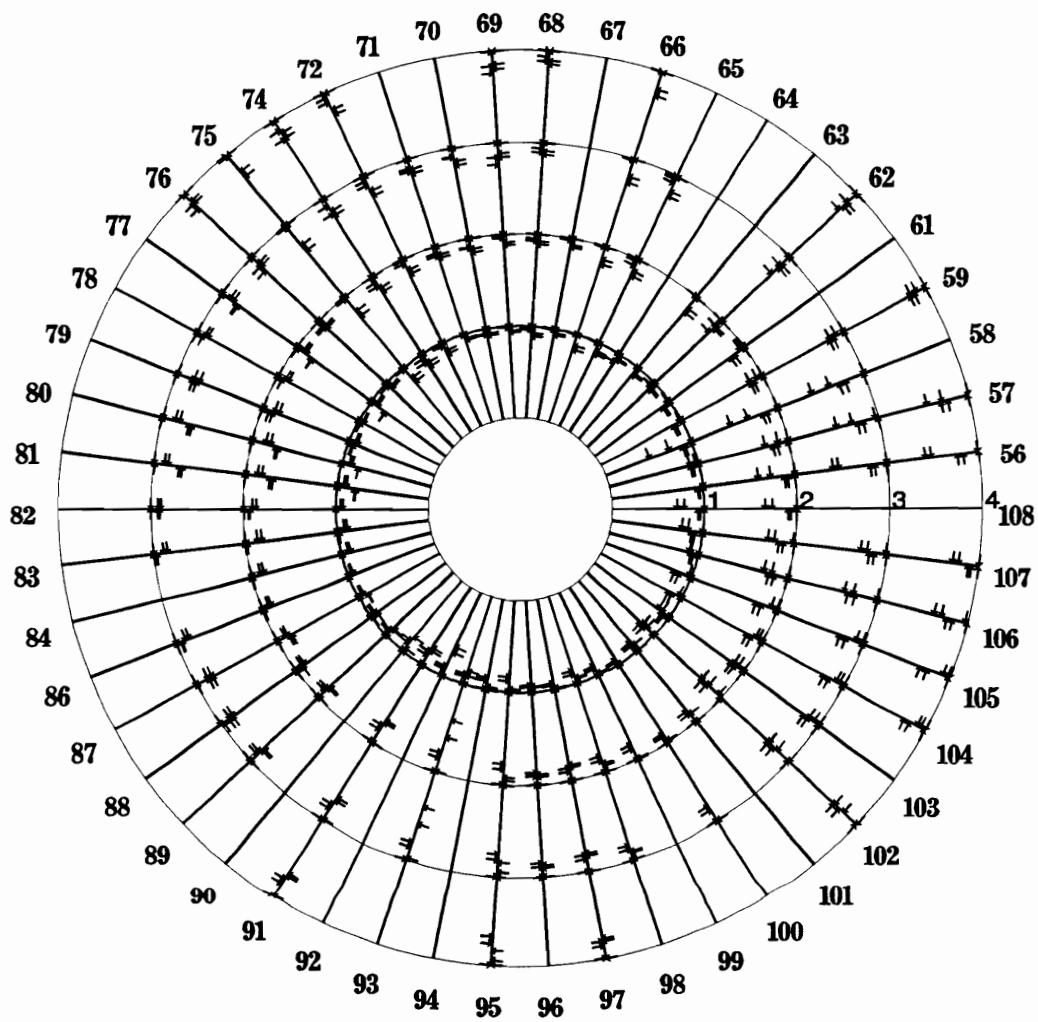


Figure 7.25. Observed and predicted proportions in Loblolly pine model [7.21]. Each segment corresponds to one occasion and a 0–100% interval. Observed values denoted by clockwise arrows, predicted values counter clockwise. GEE0 with 1–dependence structure.

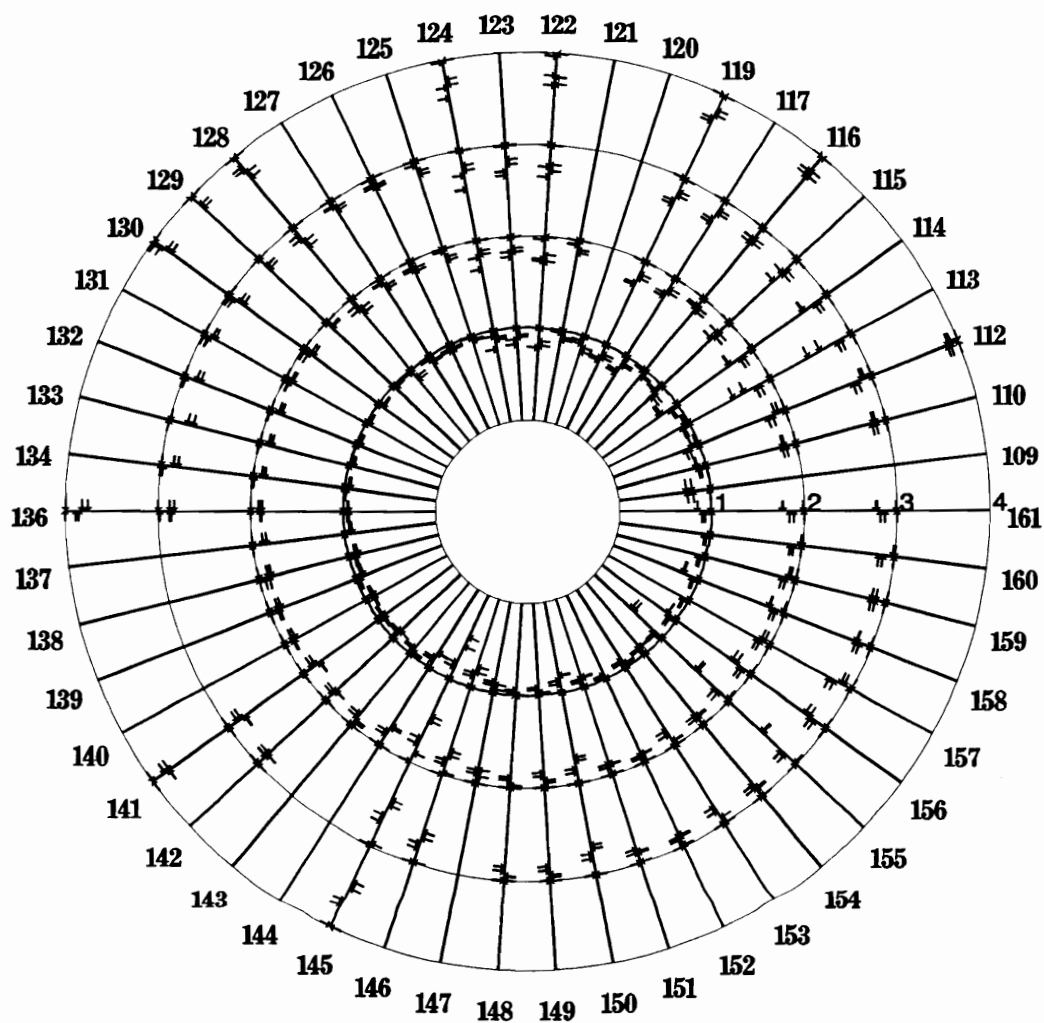


Figure 7.26. Observed and predicted proportions in Loblolly pine model [7.21].
 Each segment corresponds to one occasion and a 0–100% interval.
 Observed values denoted by clockwise arrows, predicted values
 counter clockwise. GEE0 with 1–dependence structure.

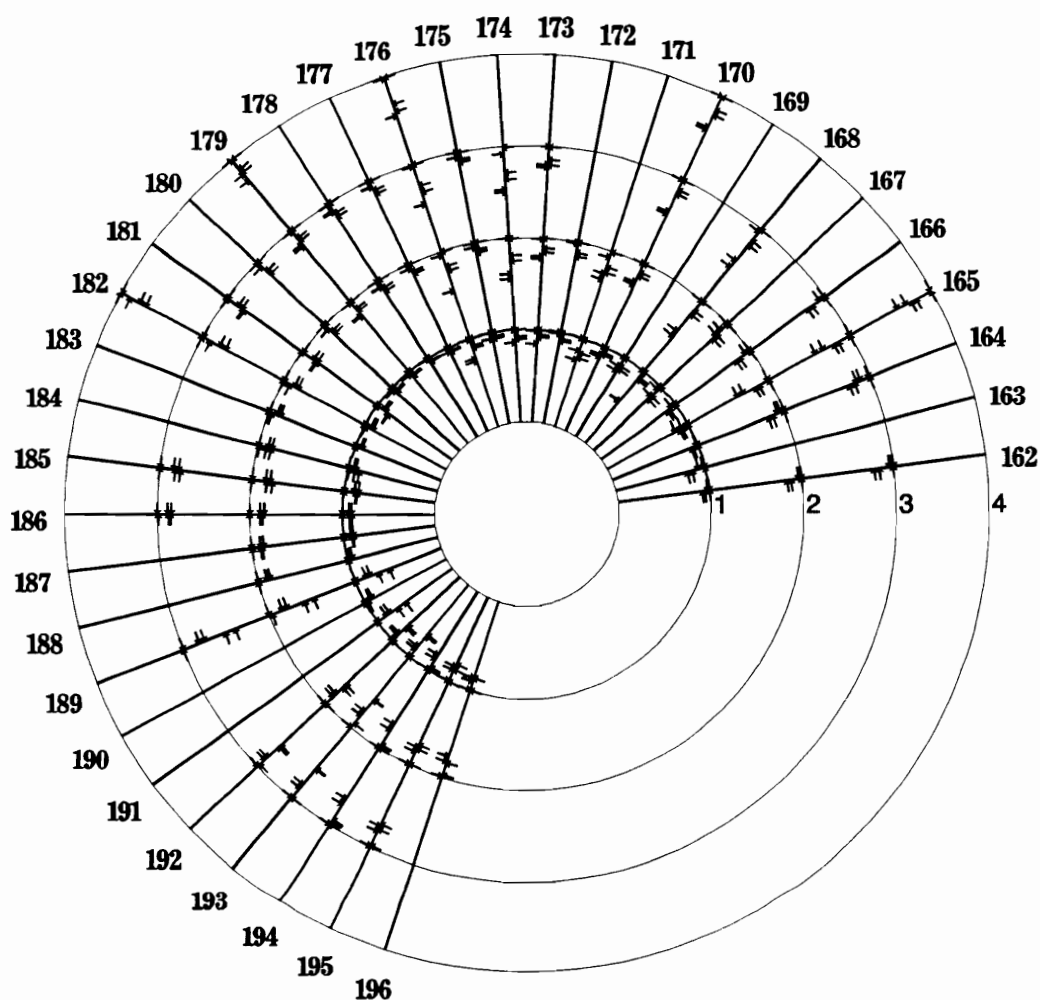


Figure 7.27. Observed and predicted proportions in Loblolly pine model [7.21]. Each segment corresponds to one occasion and a 0–100% interval. Observed values denoted by clockwise arrows, predicted values counter clockwise. GEE0 with 1–dependence structure.

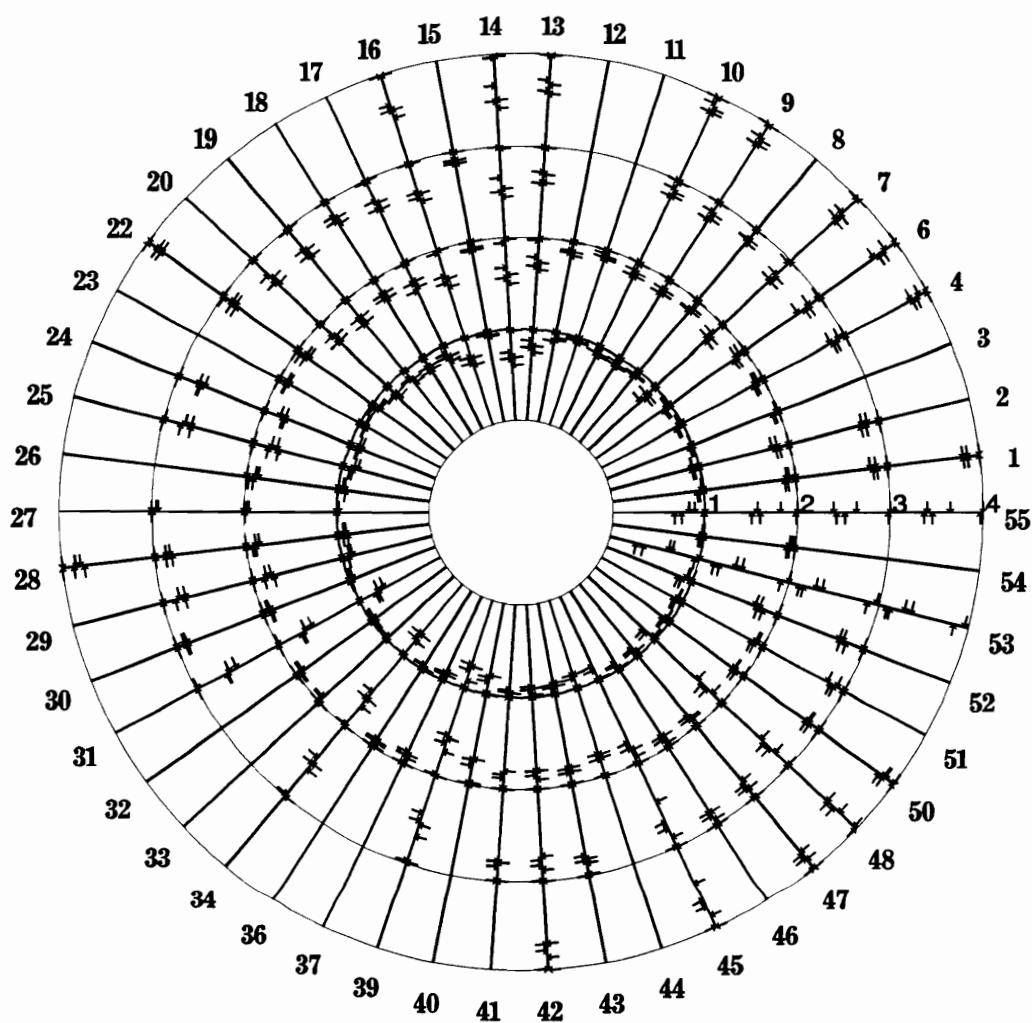


Figure 7.28. Observed and predicted proportions in Loblolly pine model [7.22]. Each segment corresponds to one occasion and a 0–100% interval. Observed values denoted by clockwise arrows, predicted values counter clockwise. Mixed model fit semi-parametrically.

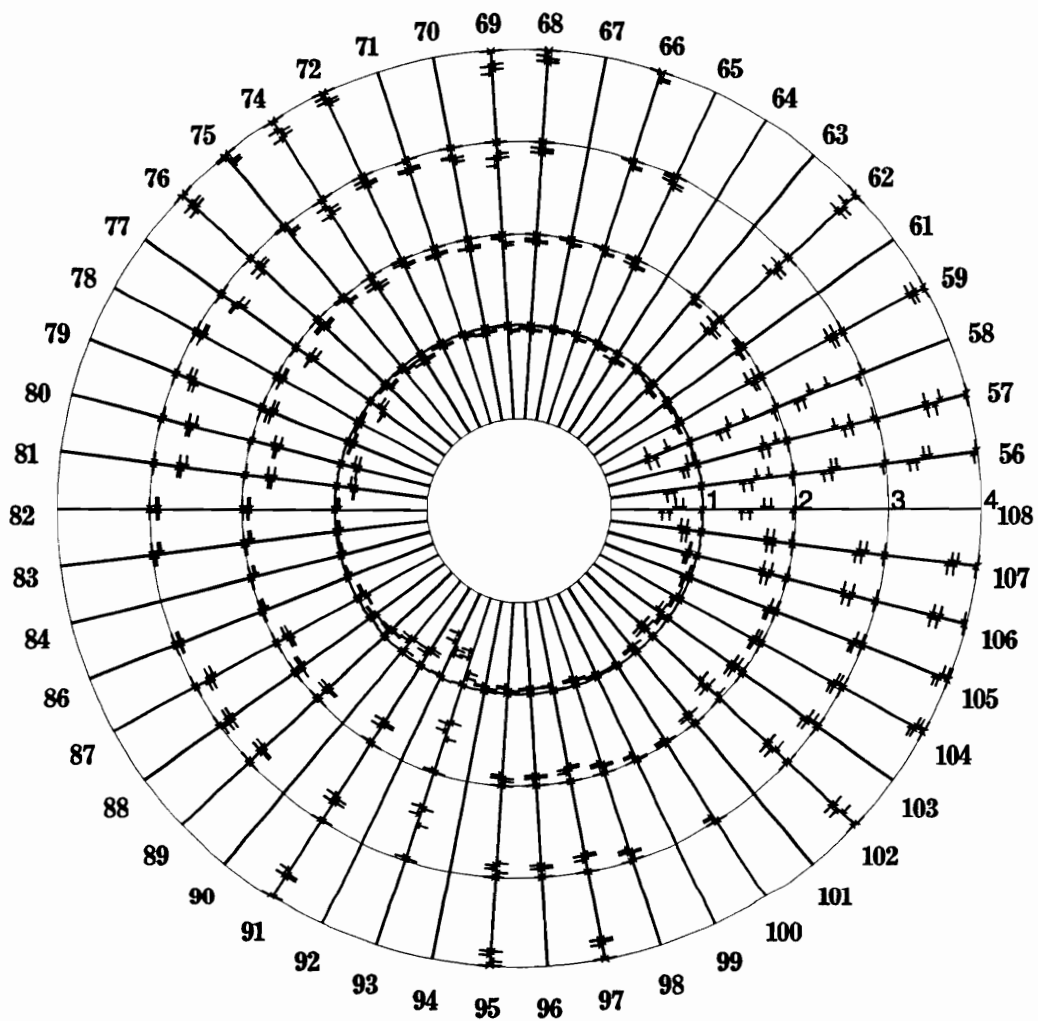


Figure 7.29. Observed and predicted proportions in Loblolly pine model [7.22].
 Each segment corresponds to one occasion and a 0–100% interval.
 Observed values denoted by clockwise arrows, predicted values
 counter clockwise. Mixed model fit semi-parametrically.

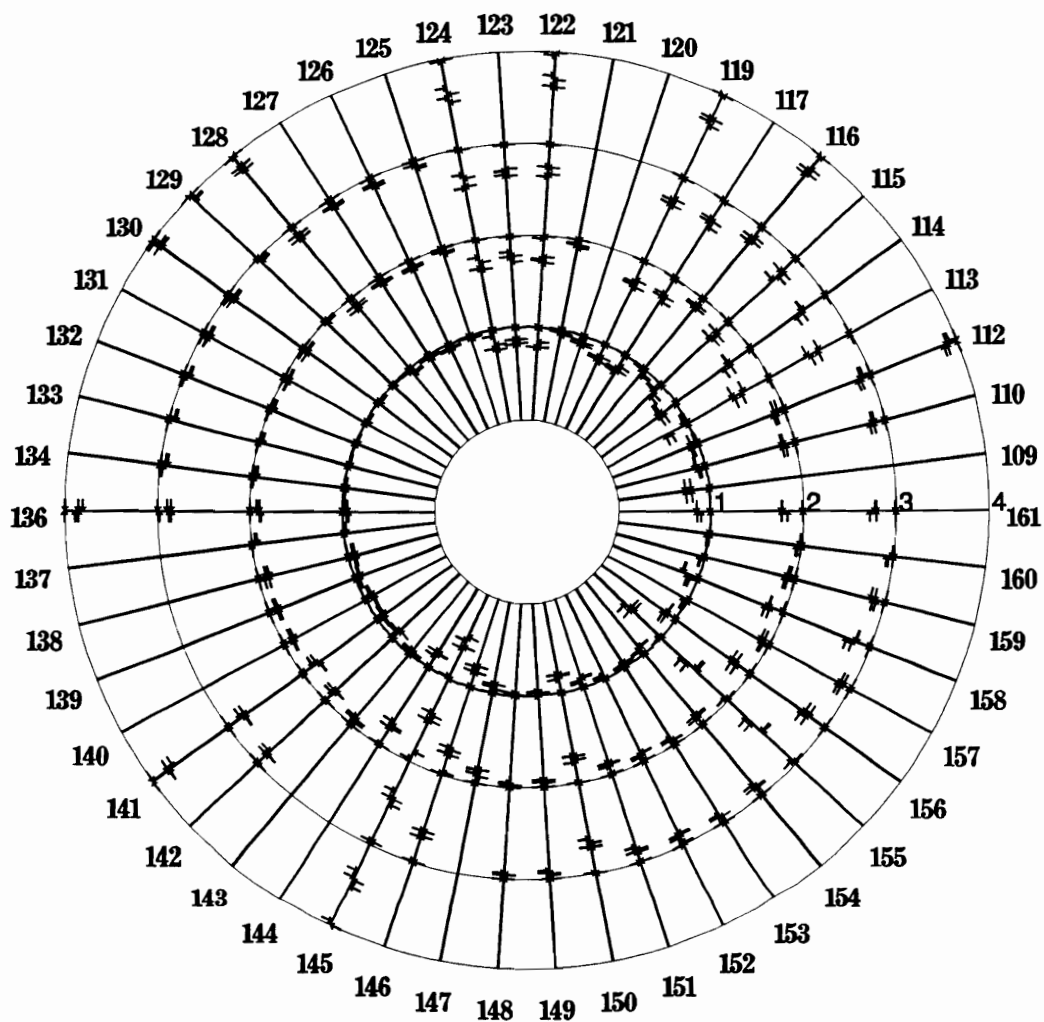


Figure 7.30. Observed and predicted proportions in Loblolly pine model [7.22].
 Each segment corresponds to one occasion and a 0–100% interval.
 Observed values denoted by clockwise arrows, predicted values
 counter clockwise. Mixed model fit semi-parametrically.

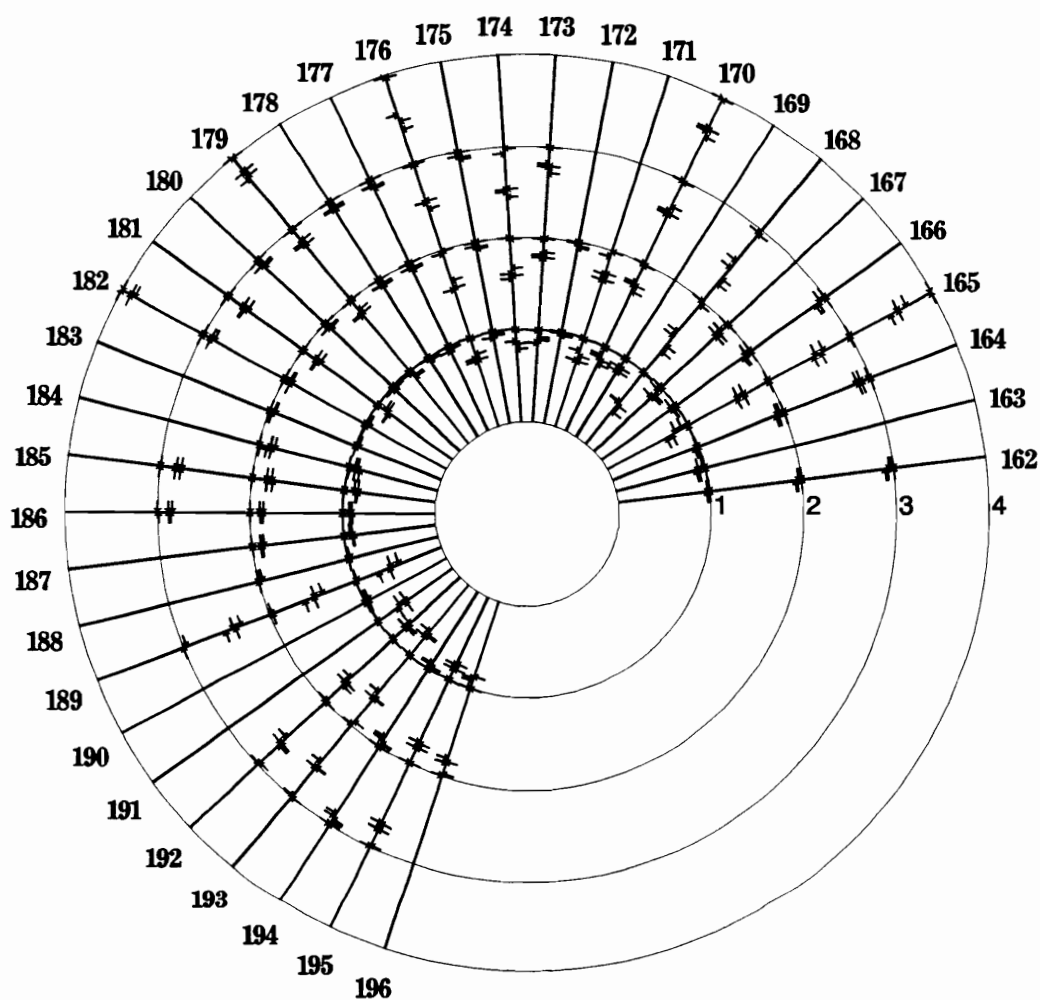


Figure 7.31. Observed and predicted proportions in Loblolly pine model [7.22].
 Each segment corresponds to one occasion and a 0–100% interval.
 Observed values denoted by clockwise arrows, predicted values
 counter clockwise. Mixed model fit semi-parametrically.

Figures 7.24 through 7.27 corresponding to the GEE0-1 fit show remarkable agreement between predicted and observed proportions on the plot level. Only some plots appear to not concur with the overall population trend as described by [7.21] (e.g. plot # 53). From this point it is not surprising that making model [7.21] subject-specific by adding across plot heterogeneity is of less merit than in the slash pine example, where the individual plots did not concur well with the population average. The contribution of the BLUPs to improve the predictions is limited by the already excellent fit on the subject level by the purely fixed effects model. However, the BLUPs are not without effects as Figures 7.28-7.31 exhibit.

7.2.1.3. Binary response models

In cases where the reduced dimensionality of the ordinal response is of interest, binary models are reasonable. Like in the slash pine model three different binary responses were created and coded

<i>Healthy:</i>	1 if a tree is free of any fusiform rust infection 0 otherwise
<i>Stemfree:</i>	1 if a tree is free of stem infections 0 otherwise
<i>Alive:</i>	1 if a tree is alive 0 if not.

Competing models were fit via the approximate parametric likelihood procedure of § 4.3.2. The finally selected models are as follows:

Healthy:

$$\eta_{it} = \beta_0 + \beta_1 Q_{it} + \beta_2 SI_i + \beta_3 C_{1it} + \beta_4 WI_i + \beta_5 \frac{Q_{it}}{SI_{it}} + b_{i1} \frac{A_{it}}{100} + b_{i2} \frac{TPA_{it}}{100} \quad [7.23]$$

Stemfree:

$$\eta_{it} = \beta_0 + \beta_1 Q_{it} + \left(\beta_2 + \frac{bi}{100} \right) SH_{it} + \beta_3 SL_i + \beta_4 WI_i + \beta_5 IA_i \quad [7.24]$$

Alive:

$$\eta_{it} = \beta_0 + \left(\beta_1 + \frac{bi}{100} \right) SI_{it} + \beta_2 BAA_{it} + \beta_3 \frac{Q_{it}}{SI_{it}} + \beta_4 IA_i \quad [7.25]$$

The model for healthy trees utilizes two random effects, which were presumed uncorrelated, models [7.24] and [7.25] a single random effect. Additional covariates not encountered before are the plot slope, SL_i in [7.24] and basal area per acre, BAA_{it} in [7.25]. It appears that initial occasion is only of importance for distinguishing healthy trees from trees that are infected in any way ([7.23]). Using actual plot age instead of initial age in the models for *stemfree* and *alive* trees did not improve these models.

Table 7.18 reports coefficients and goodness-of-fit statistics for the binary models. Figures 7.32-7.34 display the Receiver-Operating-Characteristics. All models fit extremely well as indicated by the last rows of Table 7.18. For models [7.23] and [7.24] the Sensitivity retains a high level as the cut-off increases from 0 to about 0.75. At the same time the False positive rate is quite small. These models detect events with high probability. Unless the cutpoint is chosen large (> 0.9), the fraction correctly classified is greater than 75%. This is a very satisfying result.

Table 7.18.

Estimates and asymptotic standard errors (in parentheses) for binary response models. Approximate parametric approach (Pseudo-Likelihood).

Model term	<i>R E S P O N S E</i>		
	Healthy	Stemfree	Alive
Intercept	4.36074 (0.48286)	3.08492 (0.18360)	8.56721 (1.13920)
<i>Q</i>	0.10176 (0.07740)	-0.32927 (0.04960)	
<i>SI</i>	-0.02120 (0.00638)		-0.01979 (0.01144)
<i>C</i> ₁	0.27609 (0.04958)		
<i>WI</i>	-0.19079 (0.12411)	-0.25350 (0.10805)	
<i>Q/SI</i>	-23.0035 (5.62593)		-16.6146 (7.63841)
<i>SH</i>		0.00923 (0.00632)	
<i>SL</i>		0.02419 (0.01293)	
<i>IA</i>		0.05350 (0.02438)	0.07433 (0.03908)
<i>BAA</i>			-0.01131 (0.00414)
var(<i>b</i> _{i1})	14.98410	0.032864	0.758969
var(<i>b</i> _{i2})	0.025776	0.267876	
SSR _b	2.94	3.99	0.10
WSP _b	10.34	14.89	6.55
WSF _b	269.50	414.03	7.86
R ² _{pb}	0.98	0.98	0.99

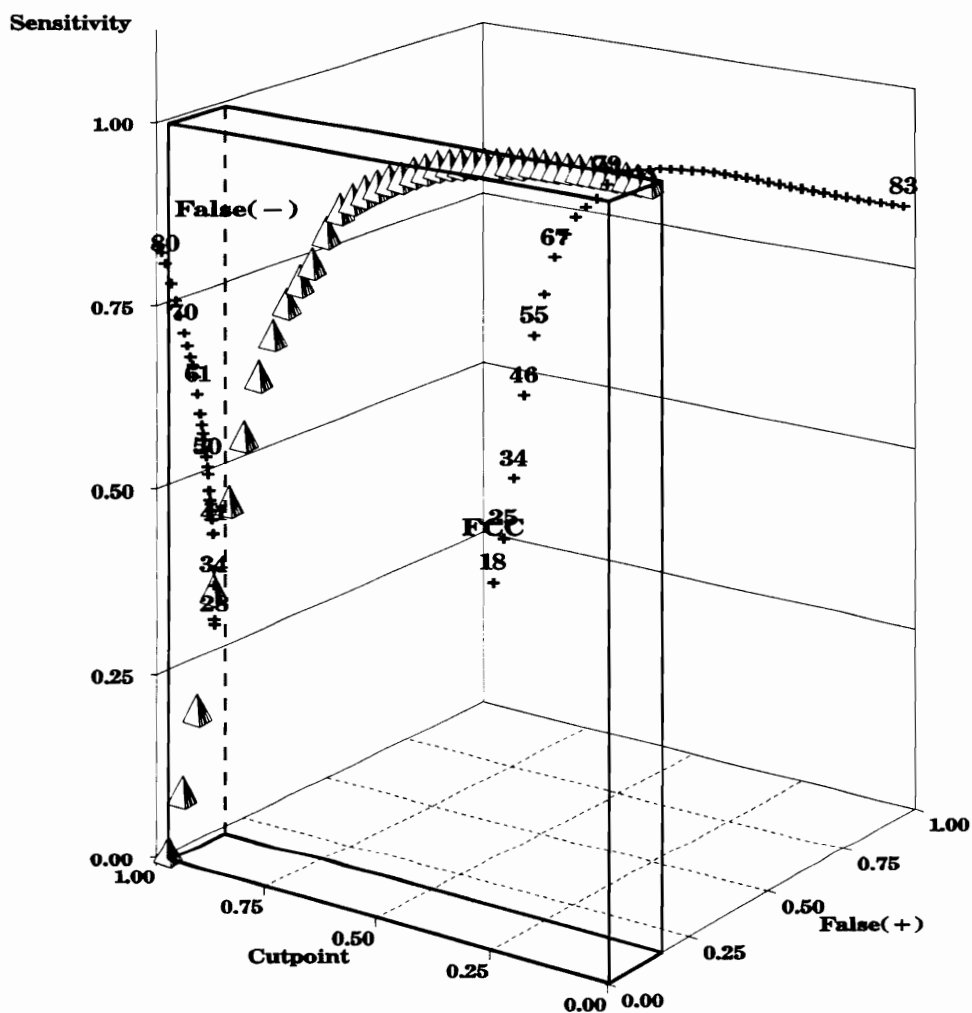


Figure 7.32. Receiver-Operating-Characteristic (ROC) for healthy trees.
Mixed model: Age and Trees per acre random.
Sensitivity: Proportion of events, predicted as events.
False(+): Proportion of predicted events, that are incorrect.
False(-): Prop. of predicted no-events, that are incorrect.
FCC: Fraction correctly classified, events and no-events (in %).

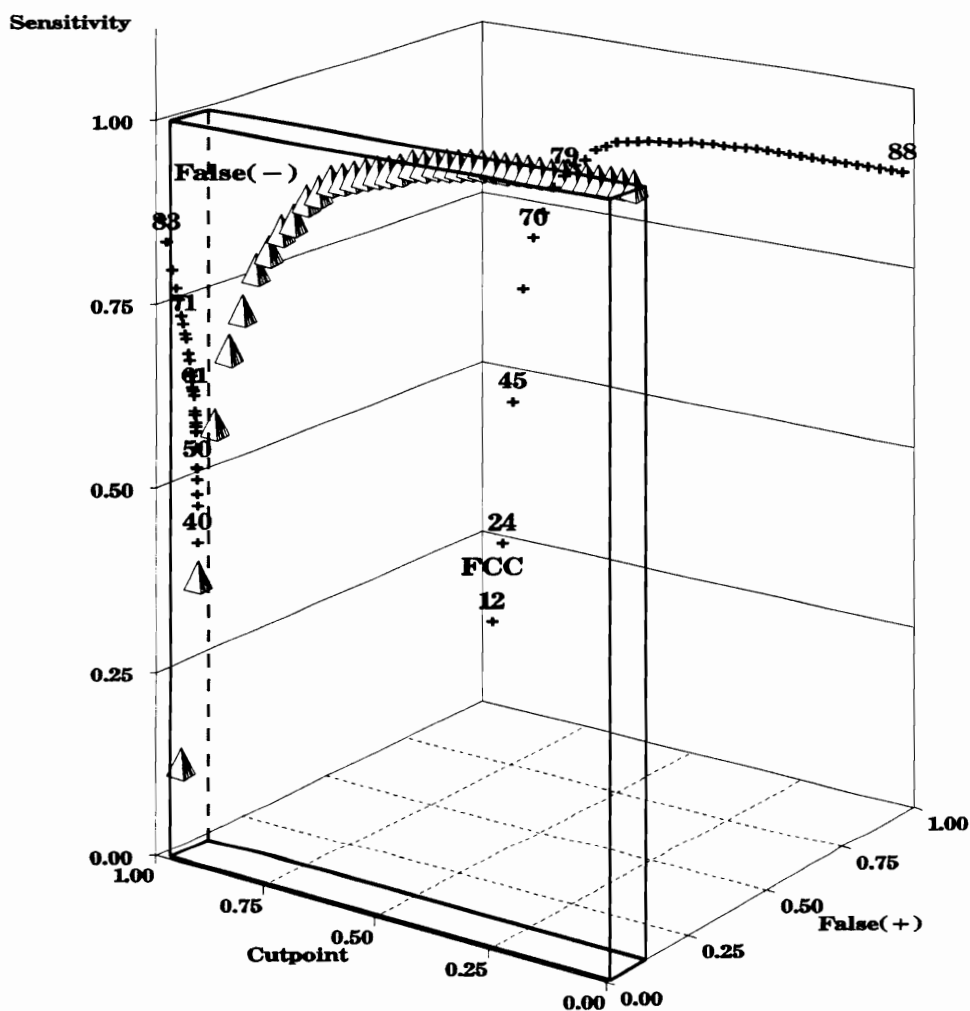


Figure 7.33. Receiver-Operating-Characteristic (ROC) for trees free of stem infect.
Mixed model: Stand height random.
Sensitivity: Proportion of events, predicted as events.
False(+): Proportion of predicted events, that are incorrect.
False(-): Prop. of predicted no-events, that are incorrect.
FCC: Fraction correctly classified, events and no-events (in %).

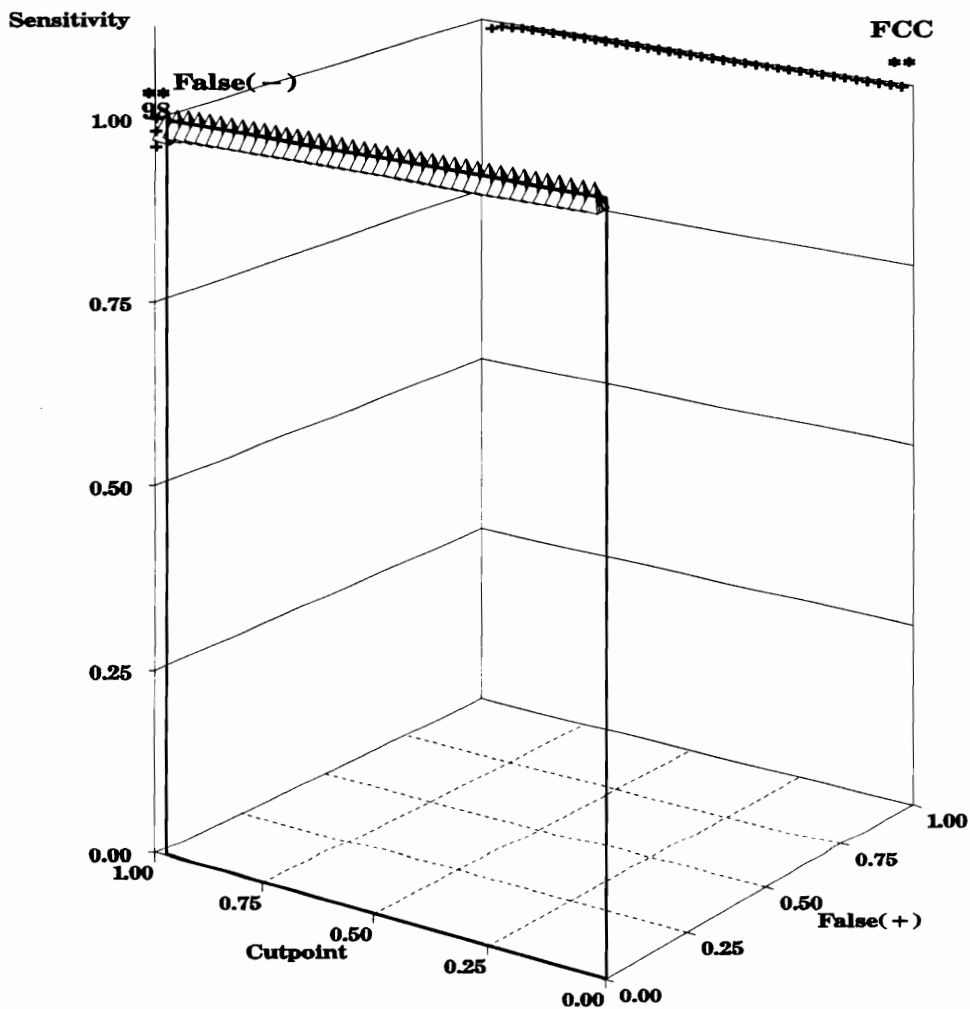


Figure 7.34. Receiver – Operating – Characteristic (ROC) for alive trees.

Mixed model: Site index random.

Sensitivity: Proportion of events, predicted as events.

False(+): Proportion of predicted events, that are incorrect.

False(-): Prop. of predicted no – events, that are incorrect.

FCC: Fraction correctly classified, events and no – events (in %).

The ROC curve for model [7.25] is less informative. The low RAM in loblolly pine causes almost all observations to be responses. Models of this type usually detect events easily. The fraction correctly classified is thus large. In Figure 7.34 it achieves almost 100 %. However, the small proportion of non-responses (dead trees) are almost all classified as alive trees. Thus, if the purpose of the model is to detect dead trees correctly, the model can not be recommended. If only the proportion of alive or dead trees is of importance, the model fairs excellent.

7.2.2. Transitional Plot Models

7.2.2.1. RC models for conditional events

As for slash pine, two different approaches are entertained for modeling one-step transition probabilities; the multinomial parameterization [4.24] and the more parsimonious ordinal representation. As before, the transition probability $p_{jj'}$ from state j' into j was assumed not to depend on the actual time point. Fitting transition models to the loblolly pine data set turned out to be much more difficult than for slash pine. As seen in § 6.1.2.3 loblolly pine is much less dynamic than slash pine and certain transitions occur very rarely. This frequently caused numerical instabilities in the fitting process. While in the slash pine example the same covariates could be used for either previous condition, here the covariates had to be modified. For the multinomial parameterization the linear predictors were:

$$\eta_{it0j} = \theta_{0j0} + A_{it}\theta_{0j1} + SH_{it}\theta_{0j2} + \frac{\Delta SH_{it}}{Q_{it}}\theta_{0j3} + Q_{it}\theta_{0j4} + N_i\theta_{0j5} + C_{1i}\theta_{j06} \quad j = 1, 2, 3$$

for the previously healthy trees,

$$\eta_{it1j} = \theta_{1j0} + IA_{it}\theta_{1j1} + SH_{it}\theta_{1j2} + C_{1i}\theta_{j13} \quad j = 2, 3$$

and

$$\eta_{it2j} = \theta_{2j0} + IA_{it}\theta_{2j1} + SH_{it}\theta_{2j2} + C_{1i}\theta_{j23} \quad j = 3$$

for the previously branch (η_{it1j}) and stem (η_{it2j}) infected trees. That is, for the last two models age was replaced by initial age and the interaction term as well as the geographic indicator were dropped.

For the ordinal model again the same covariates were used to investigate whether information is lost by making a more parsimonious choice and relying on the implicit ordering of the response. The linear predictors are

$$\eta_{it0j} = \alpha_{0j} + A_{it}\theta_{01} + SH_{it}\theta_{02} + \frac{\Delta SH_{it}}{Q_{it}}\theta_{03} + Q_{it}\theta_{0j4} + N_i\theta_{05} + C_{1i}\theta_{06} \quad j = 0, \dots, 2$$

for the previously healthy trees,

$$\eta_{it1j} = \alpha_{1j} + IA_{it}\theta_{11} + SH_{it}\theta_{12} + C_{1i}\theta_{13} \quad j = 1, 2$$

and

$$\eta_{it2j} = \alpha_{2j} + IA_{it}\theta_{21} + SH_{it}\theta_{22} + C_{1i}\theta_{23}. \quad j = 2$$

Notice that the subscript j referring to the current condition has disappeared from the linear predictor in the ordinal model except for the cut-off points α_{jj} .

Table 7.19 contains estimates, standard errors, and goodness-of-fit statistics for the multinomial model, Table 7.20 for the ordinal transition model. Table 7.21 displays asymptotic 95 % confidence intervals in the ordinal model.

Table 7.19.

Parameter estimates and asymptotic standard errors for multinomial loblolly pine transition model. Fit via Maximum Likelihood. The parameter vector for staying in once category was constrained to zero. Standard errors italicized.

Previous condition	Current condition	Coefficient estimates and standard errors for variable							
		Interc.	A	SH	$\frac{\Delta SH}{Q}$	Q	N	C ₁	IA
Healthy	Branch	-3.6116	0.0931	-0.0743	-0.2728	0.3753	-0.4145	0.5174	
		<i>0.4519</i>	<i>0.0244</i>	<i>0.0099</i>	<i>0.0821</i>	<i>0.0740</i>	<i>0.1239</i>	<i>0.0927</i>	
Healthy	Stem	-4.4971	-0.0292	-0.0337	0.1046	0.4224	-0.2861	0.6812	
		<i>0.3738</i>	<i>0.0202</i>	<i>0.0077</i>	<i>0.0624</i>	<i>0.0604</i>	<i>0.0957</i>	<i>0.0794</i>	
Healthy	Dead	-5.3591	-0.3102	0.1243	-0.5714	-0.2895	0.0316	0.2940	
		<i>1.4975</i>	<i>0.0907</i>	<i>0.0308</i>	<i>0.2746</i>	<i>0.2563</i>	<i>0.4129</i>	<i>0.3159</i>	
Branch	Stem	-0.0529		-0.0460				0.6220	0.0276
		<i>0.4108</i>		<i>0.0113</i>				<i>0.1837</i>	<i>0.0351</i>
Branch	Dead	-14.659		0.0938				1.2982	0.2709
		<i>3.7298</i>		<i>0.0524</i>				<i>1.2945</i>	<i>0.2084</i>
Stem	Dead	-3.8255		0.0092				0.3234	0.0077
		<i>0.4224</i>		<i>0.0101</i>				<i>0.2058</i>	<i>0.0354</i>
Goodness of fit statistics for the evaluation data									
Previous Condition:		Healthy	Branch		Stem				
SSR		3.57	90.49		39.56				
WSP		55.58	379.19		129.33				
SSF		248.93	124.71		36.22				
WSF		3928.9	1985.2		473.39				
R _p ²		0.98	0.48		0.76				
Entire model across all previous responses									
SSR	133.62	-2 Log-Likelihood			19896.83				
WSP	564.10	Akaike's AIC			-9940.41				
SSF	409.86								
WSF	6388.2								
R _p ²	0.77								

Table 7.20.

Parameter estimates and asymptotic standard errors for ordinal loblolly pine transition model. Fit via Maximum Likelihood. Standard errors italicized.

Previously healthy trees, $j' = 0$									
α_{00}	α_{01}	α_{02}	A	SH	$\frac{\Delta SH}{Q}$	Q	N	C_1	
3.4421	3.9406	7.0393	-0.0097	0.0416	0.0401	-0.3679	0.3118	-0.5957	
			0.0155	0.0060	0.0497	0.0467	0.0754	0.0600	
Previously branch infected trees, $j' = 1$									
	α_{11}	α_{12}	A	SH	$\frac{\Delta SH}{Q}$	Q	N	C_1	IA
	0.2636	4.6337		0.0421				-0.6540	-0.0319
				0.0111				0.1824	0.0345
Previously stem infected trees, $j' = 2$									
	α_{22}		A	SH	$\frac{\Delta SH}{Q}$	Q	N	C_1	IA
	3.8255			-0.0092				-0.3234	-0.0077
				0.0101				0.2058	0.0354
Goodness of fit statistics for the evaluation data									
Previous Condition:		Healthy		Branch		Stem			
SSR		3.61		90.50		39.56			
WSP		56.22		224.65		129.33			
SSF		251.64		125.27		36.22			
WSF		4002.5		595.28		473.39			
R_p^2		0.98		0.48		0.76			
Entire model across all previous responses									
SSR	133.67	-2 Log-Likelihood		19969.81					
WSP	410.21	Akaike's AIC		-9957.91					
SSF	413.14								
WSF	5071.16								
R_p^2	0.77								

Table 7.21.

Asymptotic 95 % confidence intervals for predicted proportions in ordered
loblolly pine transition model. Length of intervals in probability units.

	Current state of fusiform rust infection		
	Healthy $j = 0$	Branch $j = 1$	Stem $j = 2$
<u>Previously healthy trees, $j' = 0$</u>			
length l	0.01509	0.01003	0.00102
$\Pr(C = j)$	0.93815	0.02323	0.03684
$l/\Pr(C \leq j)[\%]$	1.61	1.04	0.10
$l/\Pr(C = j)[\%]$	1.61	43.20	2.79
<u>Previously branch infected, $j' = 1$</u>			
length l		0.12137	0.00900
$\Pr(C = j)$		0.79555	0.20092
$l/\Pr(C \leq j)[\%]$		15.25	0.90
$l/\Pr(C = j)[\%]$		15.25	4.48
<u>Previously stem infected, $j' = 2$</u>			
length l			0.03450
$\Pr(C = j)$			0.96116
$l/\Pr(C \leq j)[\%]$			3.58
$l/\Pr(C = j)[\%]$			3.58

The quality of fit of the ordinal and the multinomial model is comparable. The ordinal model has a somewhat smaller log likelihood (larger $-2 \log$ likelihood) and larger SSF criteria. Overall the sums of squares and pseudo- R^2 measures are very similar. This is very encouraging in that the more parsimonious ordinal model can be used instead of the multinomial one. Notice also that the estimates and GOF values for the previously stem infected trees in both models are the same, except for the sign of the coefficients. This must be the case and provides a simple check of the correctness of the algorithms used.

The asymptotic confidence intervals in Table 7.21 again identify the branch infected state as the one associated with highest uncertainty. The overall length and coverage proportions are quite good, especially for the previously healthy trees. Healthy-healthy and Healthy-stem infected transitions as well as stem-stem infected transitions can be estimated very precisely. This is important, since these transitions account for the majority of the loblolly pine dynamics.

Figures 7.35-7.38 contain the observed and forecasted proportions calculated from the ordinal model under the assumption of a one-dependent Markov chain. They were found to be indistinguishable from the ones under the multinomial model, which are hence not reported here. The agreement between the observed and predicted proportions is very satisfying on the individual plot level. It is also obvious how the proportion of healthy trees decreases over time, although this decline is less pronounced as for slash pine.

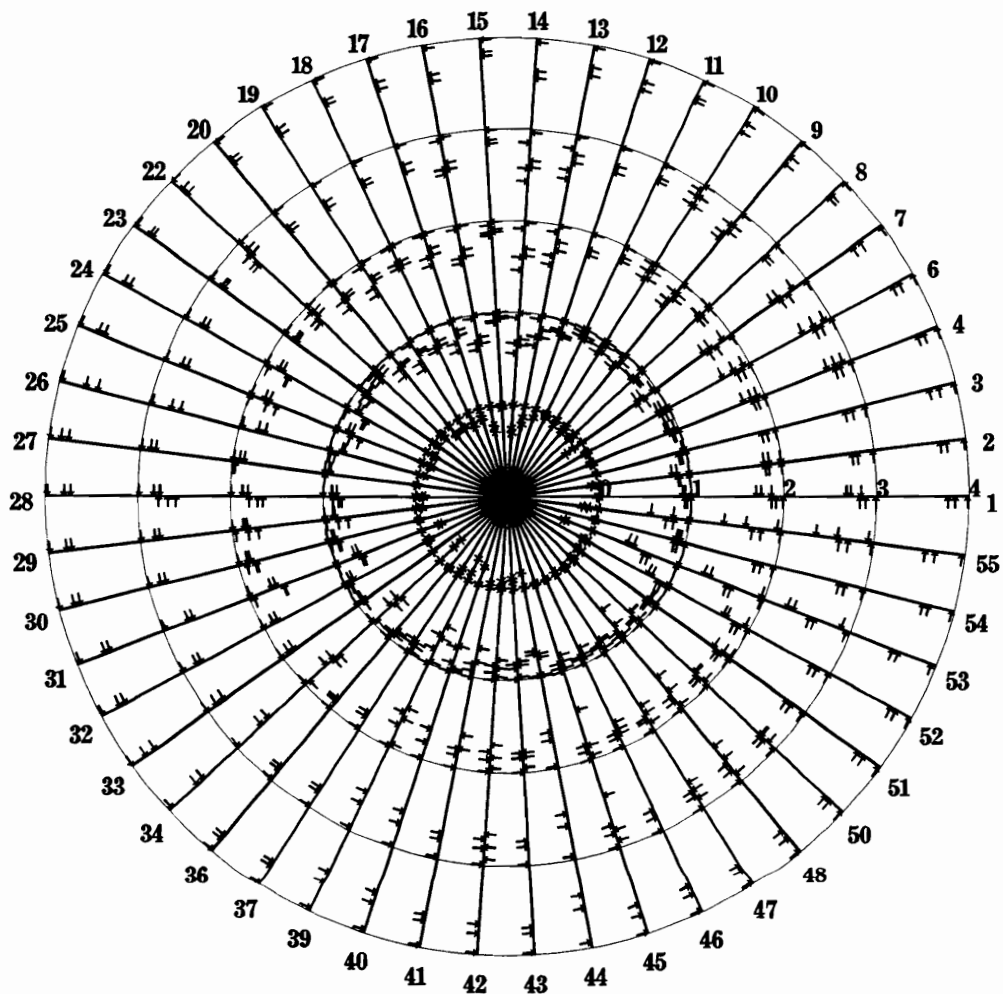


Figure 7.35. Observed and forecasted proportions based on separate proportional odds models. Circles delimit initial occasion and remeasurements. Initial marginal proportions plotted in inner–most segment. Plots with only one measurement are excluded from the analysis.

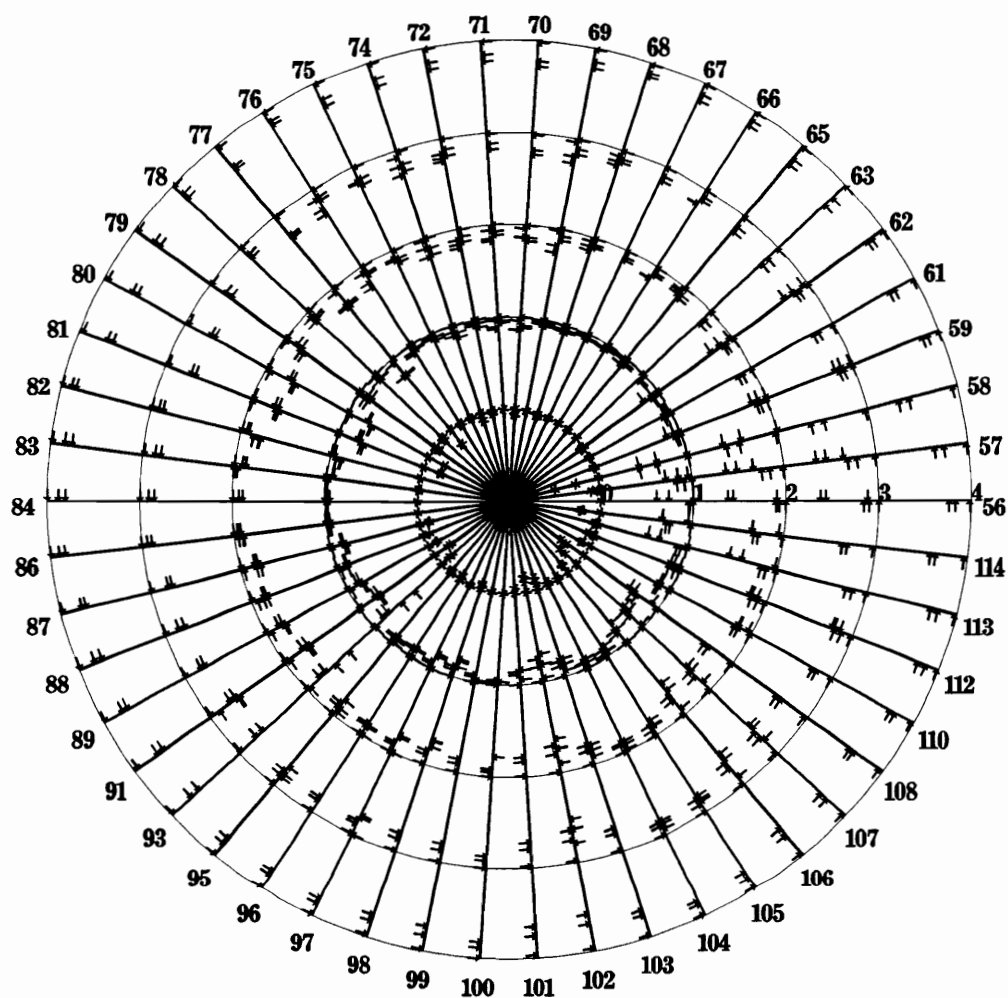


Figure 7.36. Observed and forecasted proportions based on separate proportional odds models. Circles delimit initial occasion and remeasurements. Initial marginal proportions plotted in inner–most segment. Plots with only one measurement are excluded from the analysis.

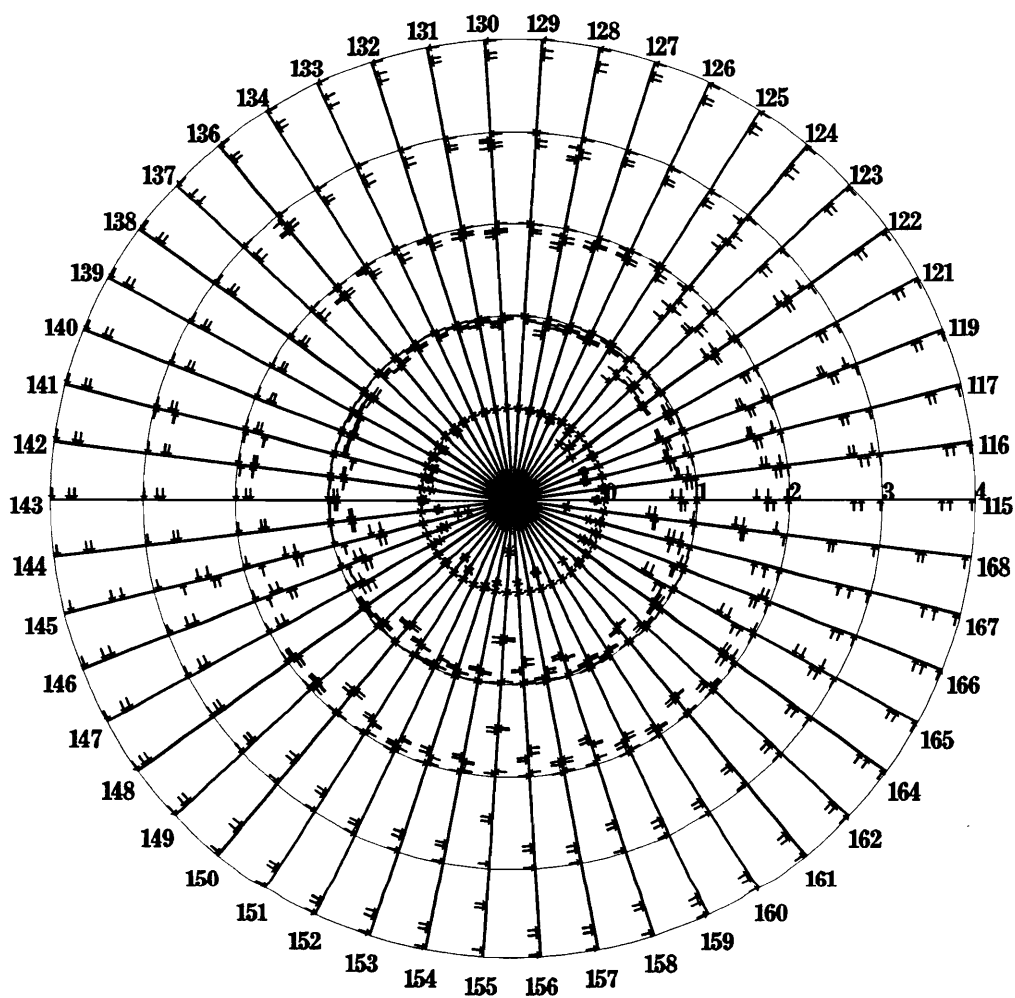


Figure 7.37. Observed and forecasted proportions based on separate proportional odds models. Circles delimit initial occasion and remeasurements. Initial marginal proportions plotted in inner—most segment. Plots with only one measurement are excluded from the analysis.

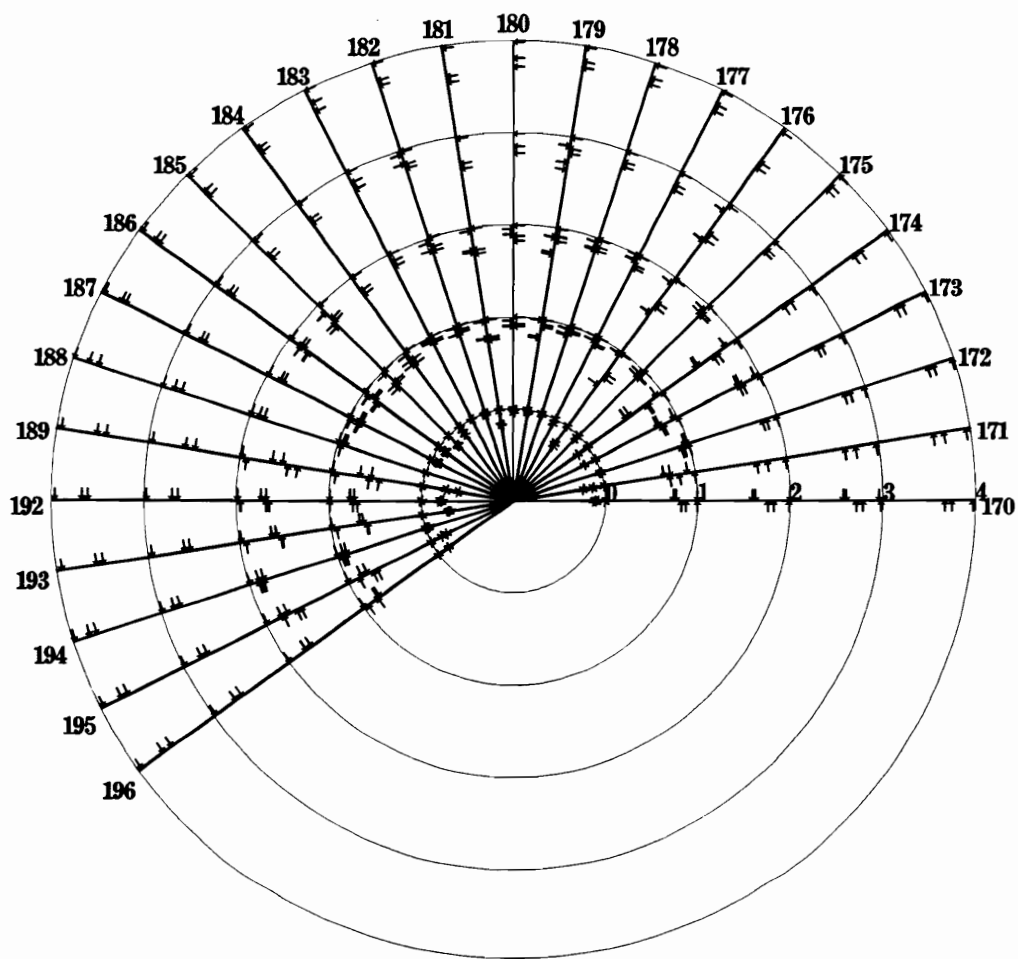


Figure 7.38. Observed and forecasted proportions based on separate proportional odds models. Circles delimit initial occasion and remeasurements. Initial marginal proportions plotted in inner—most segment. Plots with only one measurement are excluded from the analysis.

7.2.2.2. CC model for conditional events

The final model to be discussed in this chapter is a covariate conditioned transition model for loblolly pine. As for slash pine Bonney's CC model for ordinal response has been used. Based on the overall likelihood the best fitting model had linear predictor

$$\eta_{itj'j} = \alpha_j + \beta_1 Q_{it} + \beta_2 IA_{it} + \beta_3 SH_{it} + \beta_4 \frac{SH_{it}}{A_{it}} + \beta_5 N_i + \beta_6 WI_i + \gamma_0 Z_{i0} + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} \tag{7.26}$$

The conditioning covariates are defined again as

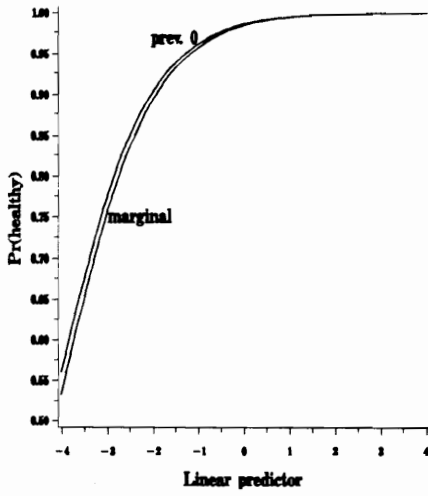
$$Z_{i0} = \begin{cases} 1 & \text{if } j' = 0 \\ 0 & \text{otherwise} \end{cases} \quad Z_{i1} = \begin{cases} 1 & \text{if } j' = 1 \\ 0 & \text{otherwise} \end{cases} \quad Z_{i2} = \begin{cases} 1 & \text{if } j' = 2 \\ 0 & \text{otherwise} \end{cases} .$$

Transition probabilities for improbable transitions are set to zero and the remaining probabilities are rescaled as in § 7.1.2.2. Table 7.22 depicts the coefficient estimates and standard errors for [7.26]. How the dependence parameters γ_0 , γ_1 , and γ_2 affect the transition probabilities compared to the marginal model is depicted in Figure 7.38.

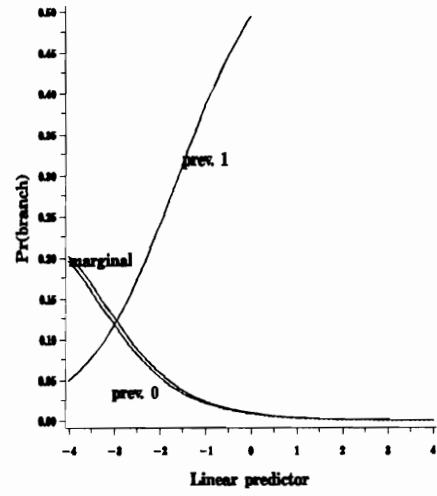
Table 7.22.
Results of fitting model [7.26] to loblolly pine development data.
Standard errors italicized.

Cut-offs	Covariates					
	Q_{it}	IA_{it}	SH_{it}	$\frac{SH_{it}}{A_{it}}$	N_i	WI_i
$\alpha_0 = 4.1344$	-0.2953	-0.1669	0.0710	-0.4201	0.2642	-0.1202
$\alpha_1 = 5.0292$	<i>0.0256</i>	<i>0.0104</i>	<i>0.0036</i>	<i>0.0359</i>	<i>0.0548</i>	<i>0.0348</i>
$\alpha_2 = 12.3717$						
Dependence parameters						
	$\hat{\gamma}_0 : 0.1081 \text{ (0.0510)}$					
	$\hat{\gamma}_1 : -3.4181 \text{ (0.0739)}$					
	$\hat{\gamma}_2 : -7.9968 \text{ (0.1403)}$					

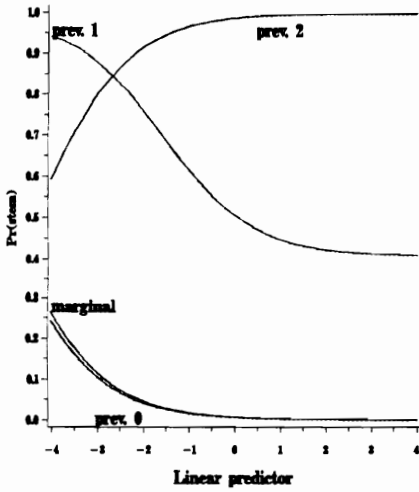
a)



b)



c)



d)

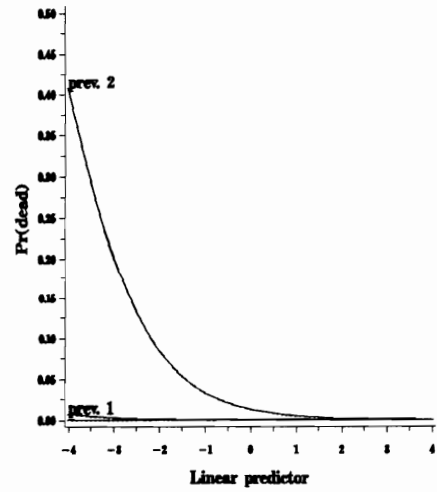


Figure 7.39. Marginal and transition category probabilities as generated by model [7.26]. Horizontal axis denotes the linear predictor without cut-off points. Prev. 0, 1, 2 denotes previously healthy, branch, and stem infected trees, respectively.

The marginal probabilities and the transition probabilities given a tree was previously healthy, are very similar and for dead trees, they are indistinguishable. Even in the most extreme case displayed, the probability to be healthy marginally, or remain healthy in a one-step transition does not fall below 0.5. The most frequent linear predictor values are above -2, however. The probability to observe a healthy tree for such values is already large, and the transition probability healthy-healthy a little greater. In case of the dead trees the same pattern is visible in Figure 7.39 d) as 7.17 d). As previous disease states are more severe, the probability to die increases. Finally, Figures 7.40 - 7.43 display a comparison of observed and forecasted values for model [7.26]. Comparing these plots with Figures 7.36 - 7.40 shows that the CC model [7.26] fits the loblolly pine data not as good as the RC model. For most plots, the agreement is acceptable, however.

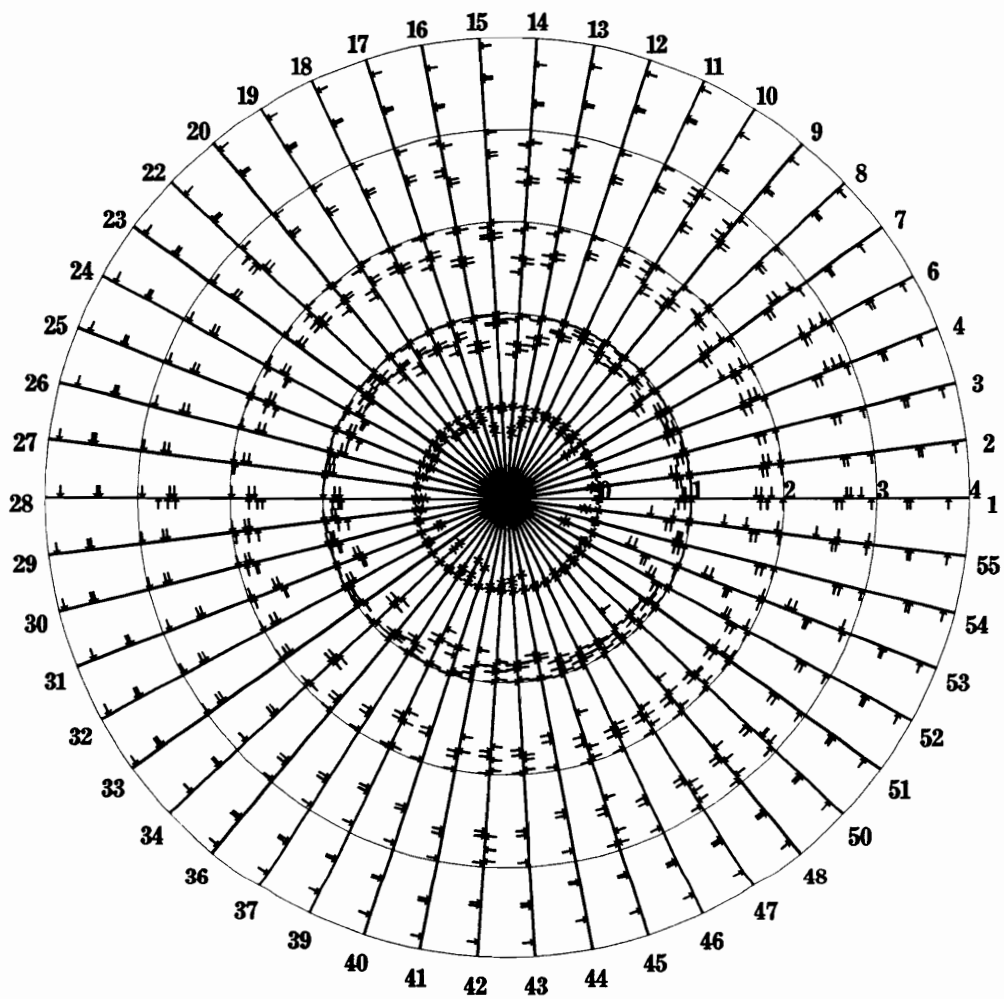


Figure 7.40. Observed and forecasted proportions based on covariate conditioned odds model. Circles delimit initial occasion and remeasurements. Initial marginal proportions plotted in inner – most segment. Plots without remeasurements are excluded.

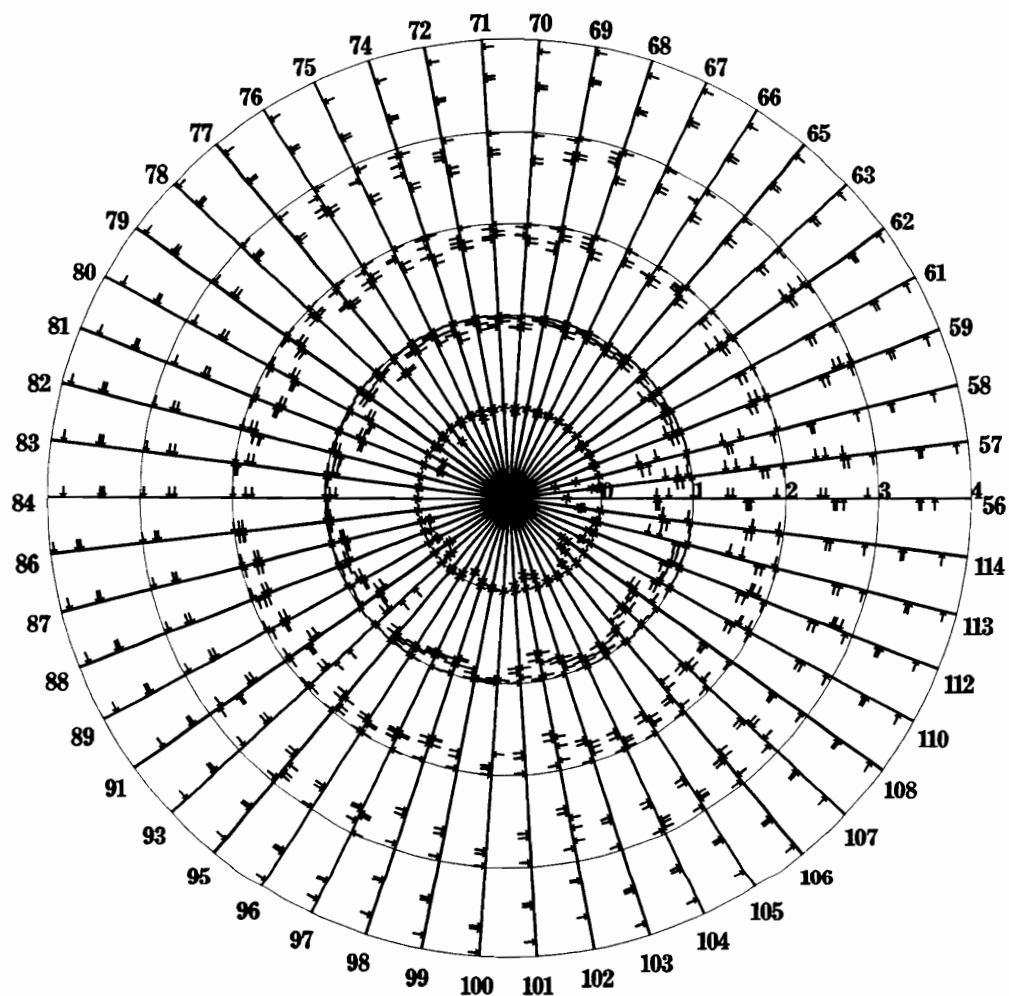


Figure 7.41. Observed and forecasted proportions based on covariate conditioned odds model. Circles delimit initial occasion and remeasurements. Initial marginal proportions plotted in inner – most segment. Plots without remeasurements are excluded.

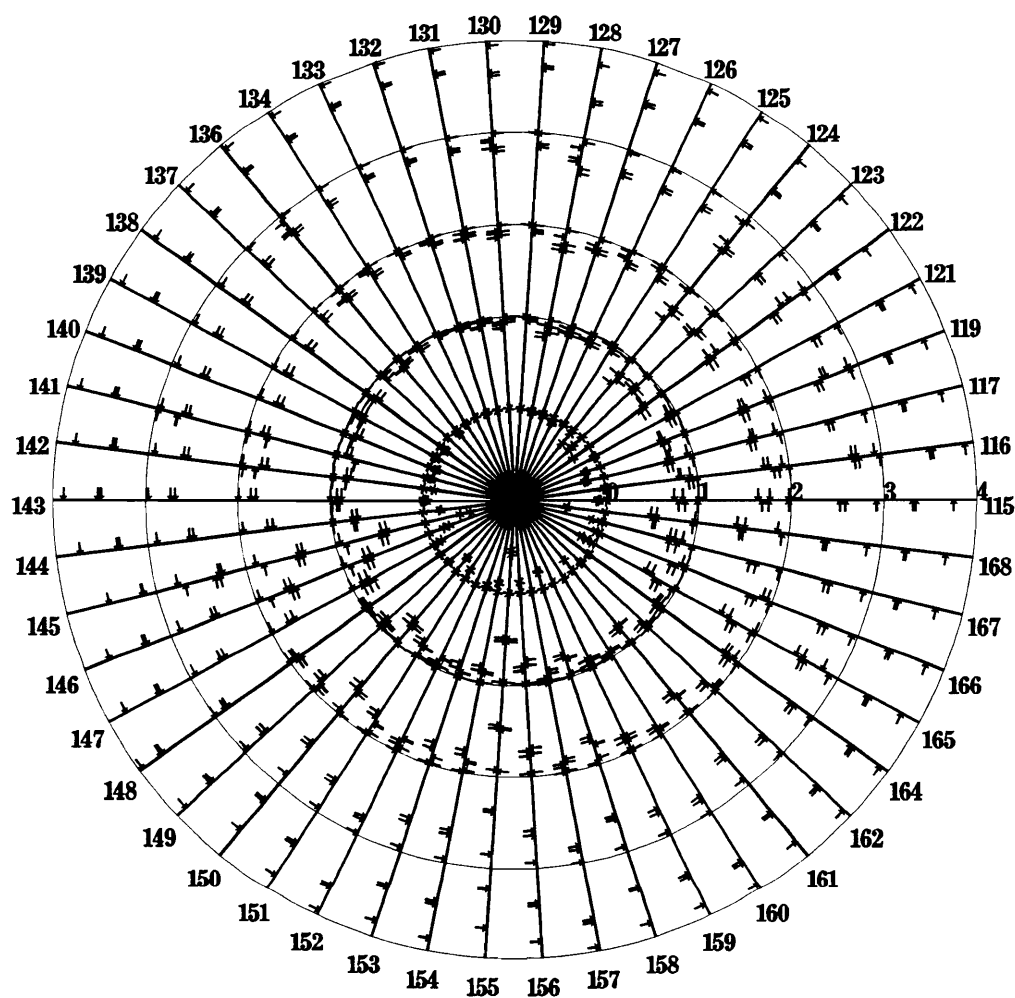


Figure 7.42. Observed and forecasted proportions based on covariate conditioned odds model. Circles delimit initial occasion and remeasurements. Initial marginal proportions plotted in inner—most segment. Plots without remeasurements are excluded.

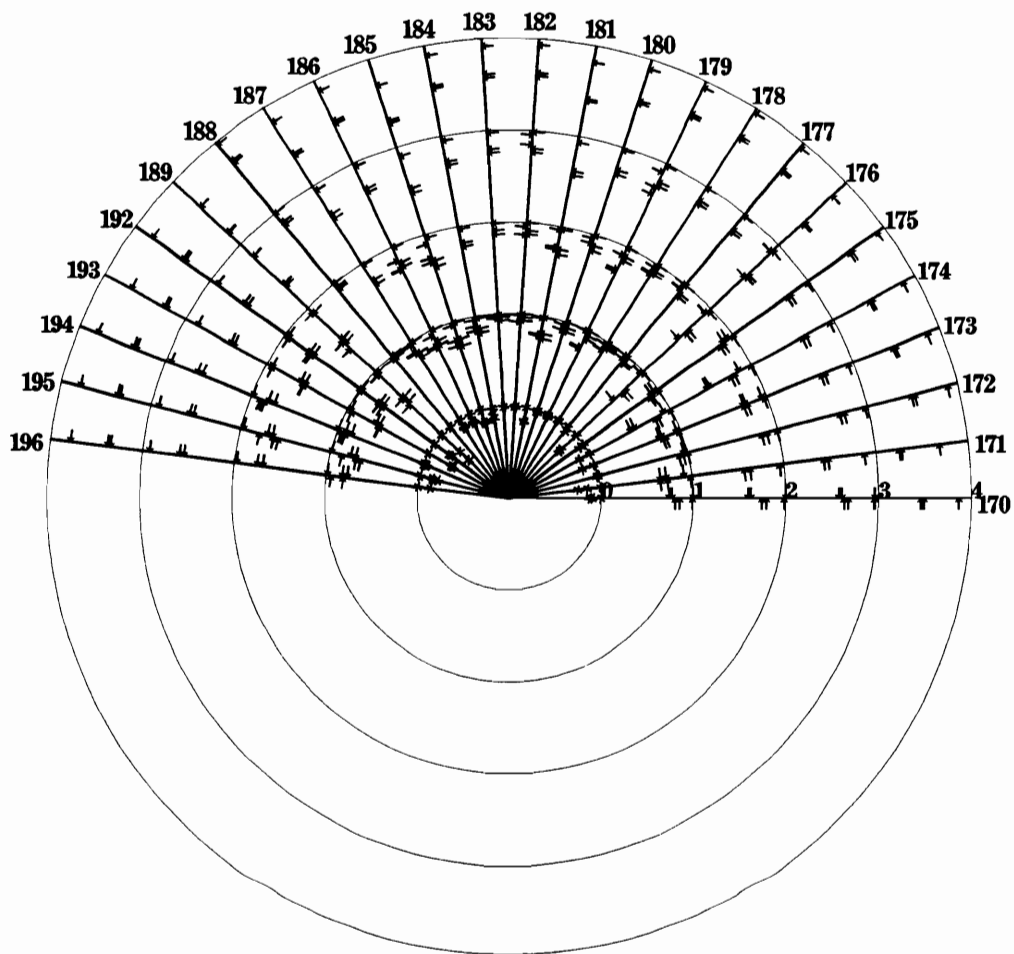


Figure 7.43. Observed and forecasted proportions based on covariate conditioned odds model. Circles delimit initial occasion and remeasurements. Initial marginal proportions plotted in inner—most segment. Plots without remeasurements are excluded.

Chapter 8

Discussion and Summary

The focus of this dissertation is on the analysis of longitudinal categorical data with special emphasis on ordinal and binary experimental outcomes. It is organized essentially in two parts. § 1-4 introduce the theory of both, categorical data and longitudinal data analysis, § 5-7 apply a series of methods discussed earlier to data from the East Texas Pine Plantation Research Project.

Both longitudinal and categorical data confront analysts with specific challenges. Longitudinality results in a series of observations over time for a typically large number of experimental units, which are correlated. Categorical responses require modeling counts and proportions that can take on only a limited or restricted range of values and imply functional dependence between the first and higher order moments of the response distribution. This dissertation attempted to combine the two topics.

Because of the prevalence of longitudinal data (LD) in forestry comparatively more emphasis was placed on the aspect of correlatedness than on the quantal nature of the response. This choice was motivated by the fact that analytic techniques for longitudinal data presently grow much more rapidly than quantal regression techniques, which by themselves are well established. However, there is a natural interaction between choosing a specific LD technique and the response. Whether one is confronted with a Bernoulli, binary, ordinal response, or count data will affect the number of alternatives for modeling LD of the particular nature.

In hindsight of the application in § 5-7 a specific type of ordinal models, referred to as cumulative logit models has received most attention. Certain shortcomings of the models presented in § 7 may be induced by a suboptimal choice of the link function and/or the underlying ordinal regression model. Unfortunately, there was no room to investigate the comparative efficacy of the proportional odds model with different links versus Anderson's stereotype model or a continuation link formulation. Such a comparison would be generally desirable but must be deferred to another study.

As outlined in § 3, longitudinality presents one with intriguing challenges and opportunities. First of all, correlations need to be taken into account if the purpose of modeling is not merely population-averaged prediction, but to find a suite of (partially) significant regressors and to draw correct inferences under the model. Secondly, longitudinality allows research questions that can not be addressed with cross-sectional data. It is thus important to decide first what are the quantities and units of interest. If the population as a whole is at the center of investigation and the response history is not of importance, a population-averaged approach is suitable (§ 2.2.1.2.1, 3.3.3, 4.2).

Models of this type lead to parameter interpretation comparable to classical regression models and are thus appealing to many analysts.

One may, however choose the individual unit (tree, plot, person, ...) as the unit of primary interest. Techniques appropriate in this case were labeled subject-specific and discussed in § 2.2.1.2.2, 3.3.3, 4.3. To make a model subject-specific one allows one or more coefficients to vary with subjects. To keep the model parsimonious as the number of subjects in the study increases, these coefficients are typically assumed to be random variables, rather than parameters. Presuming these are chosen from a hypothetical population of coefficients, one arrives at a mixed model formulation.

Finally, if one is interested in transitional behavior, the subject history has to be included in the model. This can be accomplished in two ways, which were denoted RC and CC models (§ 2.2.1.2.3, 4.4.).

The decision as to which class of model to apply is much more important for nonlinear models, than for linear ones. In the first case, there is no transitivity between the different models, that is, marginal estimates typically can not be obtained from a transitional model formulation. This is only possible in the presence of additional assumptions, such as in § 7.1.2, 7.2.2, where a one-dependent, homogeneous Markov chain was used to generate marginal proportions from the predicted one-step transition probabilities.

In the late 1980's population-averaged techniques for correlated data were in great demand, in particular after the introduction of Generalized Estimating Equations. The limitations of PA models to capture subject behavior then prompted investigations into

subject-specific, mixed models. As of this writing, mixed models are the fastest growing class of longitudinal data models. Conditional models, although important, have always played a comparatively minor role. Reasons hereofore are the rapidly growing number of response profiles when conditioning on a subject's history and the fact that many modeling techniques are custom-designed for a particular application or research problem. Rosner's model (§ 2.2.1.2.3) for example, was developed with respect to the particulars of ophthalmology. PA and SS models are more generic and provide a more coherent body of theory in this regard than transitional models. A very recent development for PA models is to replace moment estimation in GEE's by parametric, in particular pseudo-likelihood methods. There was no room to incorporate these most recent additions into this dissertation.

In principal, the choice of the particular class of models should be guided by subject-matter considerations. In the analysis of the slash and loblolly pine data in § 6, and § 7 it was assumed, that all types of research problems are of interest for these data and a wide variety of techniques was covered. Within each class however, there is typically a large number of alternative approaches requiring a choice.

In § 3 and 4 it was attempted to relate the various approaches within each class and across classes as good as possible with the goal to underline common features, parallels, and differences that can guide selection within each class. The unifying and linking device chosen was Estimating Function Theory, presented in § 3.2. In particular, one estimating function was identified in § 3.2.4 and shown to be optimal in a variety of cases. GEE0, GEE1, GEE2, all subject-specific methods, CC models, and CC models for joint occurrences can be implemented based on this function.

This is important since almost no general purpose software is commercially available for analyzing longitudinal categorical data. The estimating function of § 3.2.4 made it possible to code a variety of approaches with acceptable effort. Nevertheless, the choice of a particular method will frequently be governed by numerical stability and tractability of the estimation problem. Among the PA approaches, computational demand increases in the order GEE0, GEE1, and GEE2. If the estimation of covariance parameters is not carried out by method of moments, but pseudo-likelihood, the estimation is more involved. In toto however, PA methods are typically less demanding in terms of computer resources and more stable during estimation than subject-specific models. For the loblolly pine data, for example, a large number of population-averaged models could be fitted and investigated, while only a small number of mixed models actually converged.

Within the class of SS models, computing time increases as more covariance parameters are modeled parametrically. Hence the methods become more demanding in the order semi-parametric (§ 4.3.1), approximate parametric (§ 4.3.2), and fully parametric (§ 4.3.3). Increasing computing time frequently is paired with less stability of the algorithms. For those estimation procedures that are based on the optimal estimating function of § 3.2.4 it was found, that when the algorithm did not converge in 30 iterations, it did not converge at all. Mostly convergence was achieved with 8 to 15 iterations. This is a promising result, since especially for mixed models, the number of arithmetic operations per iteration is very large.

For ordinal responses, the semi-parametric SS method (§ 4.3.1.) worked most successfully, while the approximate parametric procedure often failed to converge or

was aborted for other numerical reasons. A sound implementation of SS techniques demands automatic ridging of gradient matrices to combat ill-conditioning, step size optimizations, etc. The failure of some mixed models to converge may be due to the author's inability to implement these fine-tuning techniques. The fully parametric approach for ordinal responses based on quadrature was found to be unnecessarily involved and suffers from the fact that it does not produce best linear predictors for the random effects, which are important quantities in mixed model analysis.

With binary responses, both population-averaged and subject-specific methods are much better behaved, typically converge nicely and are an order of magnitude faster. Even adding only one category to a binary response amplifies numerical problems considerably and is wasteful in terms of computing time. It should thus be analyzed beforehand, whether all ordered categories are essential, or whether adjacent categories could be combined. The binary mixed models fit via the approximate parametric procedures of § 4.3.2 presented no numerical problems.

In transitional models one does not estimate covariance parameters and the behavior and speed of the algorithms usually increases as the number of parameters decreases. Bonney's CC model was simplest and fastest to fit, followed by CC models for conditional events based on ordinal (§ 7.1.2.2, 7.2.2.2) and multinomial parameterizations (§ 7.1.2.1, 7.2.2.1). The CC models for joint occurrences (§ 4.4.2) turned out to require many iterations with a large number of operations per iteration, making them the slowest and most troublesome models in this class.

Owing to the recent nature of LD techniques, many approaches are still in their infancy. They are mostly concerned with deriving parameter estimates and consistent asymptotic standard errors. Little work has been done in several important areas such as partial significance testing, evaluating goodness of fit, model diagnostics, and the detection of influential data points. Available results mostly pertain to linear mixed models for continuous response, less is available for non-linear mixed models. Models for repeated categorical response are the least developed but methodological research is beginning to attend to these issues. Because sound model diagnostics are not yet developed, graphical tools have been given substantial attention in this dissertation.

Another area of vibrant interest, as far as categorical responses are concerned, is the analysis of overdispersion. The sampling model for a discrete response in theory dictates the functional relationship between means and variances. The actually observed variation in the sample frequently exceeds this nominal variation, however. In generalized linear models overdispersion is typically accounted for by a multiplicative or additive constant to the nominal variances. Fitting these models is then carried out by the principle of quasi-likelihood. For correlated categorical data, overdispersion is also a concern, albeit of lesser importance. In all PA and SS models, the marginal variance-covariance matrix for a subject's responses only partially consists of nominal variances, while other elements are subject to modeling (the off-diagonal elements in GEE methods, for example). The weight matrix thus already adjusts for heterogeneity not found among independent observations, and the additional inclusion of overdispersion parameters is given less priority. However, all algorithms presented here can be adapted for additional overdispersion parameters, a matter not pursued in

this dissertation. Details can be found in Zeger and Liang (1986), Zeger *et al.* (1988), Wolfinger and O'Connell (1993) for example.

The second part of the dissertation, § 5-7, concerns the analysis of a fungus infection (fusiform rust) in slash and loblolly pine in east Texas. The disease is characterized by uni-directionality, i.e. a tree's health status can not improve over time. The status of a tree was classified as *healthy*, *branch infected*, *stem infected*, and *dead*. The categories were treated as ordered.

The data were recorded as part of the East Texas Pine Plantation Research Project (ETPPRP) on permanent plots on a tree by tree basis in 3 year intervals. The analysis, however, was concerned with plot-wise data only, thereby reducing the effective number of observations to be modeled considerably. Modeling the rust disease on a tree basis is of course possible, but the large number of observations for either species makes a presentation of PA, SS, and conditional models and discrimination of competing models in any class difficult. Moreover, the models are intended to aid in modeling a plantation's status at a given point in time and its behavior over time. Plot data provide the natural scale for this purpose. A third reason to favor plot over tree level analysis is that the actual factors leading to the development of fusiform rust on a pine tree are not reflected in the measured covariates which are the typical forest inventory characteristics that seem more appropriate to relate to a plot's status rather than to a tree's condition.

The rust disease in slash pine is characterized by relatively dynamic behavior over time, infected trees can rapidly move towards more severe states of the infection.

Overall the most prevalent categories are healthy and stem infected trees. The former decreases with age, stand height, mean diameter, basal area and the latter increases accordingly. Trees per acre shows a reverse effect. The proportion of healthy trees in slash pine decreases rather rapidly with age and many plots show only 20-30% healthy trees.

Loblolly pine turned out to exhibit much less dynamic behavior and high proportions of trees stay in the same category over an extended period of time. Most prevalent are healthy and stem infected trees as for slash pine, but the healthy proportion hardly drops below 60%. It is more likely in loblolly than slash pine plantations, although affected by fusiform rust, that they can be carried to the end of the planned rotation period.

The models developed for both species reflect the observed marginal and transitional patterns just described and more fully outlined in § 6. In contrast to former modeling approaches for this disease (§ 5) they allow one to relate fusiform rust to easily measurable covariates typically available for all managed stands. They do not require information unlikely to be available such as the proportion of previously infected trees, the distribution and location of intermediate rust hosts, etc. Although the incorporation of such covariates probably would have improved the overall model fit, this was sacrificed in favor of models that can easily be implemented by forest managers.

Because of the lack of more refined goodness of fit measures and model diagnostics various sums of squares criteria and graphical methods were used for model

discrimination. For reasons outlined in § 5 the interpretation of the model coefficients should not be carried to extremes. For any model presented as well-fitting and/or well-predicting in its particular class, there were other combinations of covariates that provided a comparable quality of fit. The regressors do not imply cause of fusiform rust but only provide a suitable set of predictors for the particular data. For other species, geographical regions, and stock of different rust resistance the models do not apply in the same fashion and need to be recalibrated.

The analysis of transitional behavior for both species commenced from a model for conditional events based on a multinomial parameterization. The model found to fit best for either species was then fit based on a more parsimonious and restrictive ordinal parameterization. It turned out that only little was lost by using the ordinal over the multinomial model. This supports treating fusiform rust response in the ETPPRP as ordinal and not merely multinomial, an important result that naturally leads to more parsimonious and simpler models while relying on the inherent structure of the data.

Graphical diagnostics for longitudinal data have to cope with the two-level nature of the data and need to distinguish not only observations but repeated observations by subject. For the 4-category ordinal response analyzed here the problem is amplified as each observation and prediction is multivariate. To compare observed and predicted proportions for each plot a new type of graphical display was developed and used extensively throughout § 7. This display, called ray plot, depicts 4-dimensional data in two dimensions. For each subject a ray is drawn from the center of the graph (dimension 1) along which occasions are marked by ticks (dimension 2). Between two tick marks the observed cumulative proportions are depicted as arrows pointing

clockwise (dimension 3), predicted arrows are pointing counter-clockwise (dimension 4). If these arrows are exactly opposite of each other the model predicts perfectly for that particular plot.

The fusiform rust models in this dissertation are concerned with the proportion of trees on a plot in any of the four rust categories. Although important information in its own right, to aid in plantation management, the proportions now need to be incorporated into growth and yield models. This is simplified by the fact, that the models all utilize information that is directly related to stand or stock table data, such as mean diameter, stand height, age, etc. For example, by holding all other covariates constant and varying the mean diameter, one can generate the marginal or transitional proportions by diameter classes. These proportion can then be used to adjust a stand table and develop diameter distributions for healthy trees only, for example. Using the models in such a fashion allows plantation managers to convert the proportions into economical values and assess the risk and losses in dollars and cents.

Chapter 9

Bibliography

- [1] Abramowitz, M. and Stegun, I.A. 1965. *Handbook of Mathematical Functions*. Dover, New York
- [2] Agresti, A. 1988. A Model for Agreement Between Ratings On an Ordinal Scale. *Biometrics* **44**, 539-548
- [3] Agresti, A. 1989. A Survey of Models for Repeated Ordered Categorical Response Data. *Statistics in Medicine* **1**, 1209-1224.
- [4] Agresti, A. 1990. *Categorical Data Analysis*. John Wiley & Sons, Inc., New York
- [5] Agresti, A. 1992. Analysis of Ordinal Paired Comparison Data. *Applied Statistics* **41**, No. 2, 287-297
- [6] Agresti, A. 1993. Distribution-Free Fitting of Logit Models With Random Effects for Repeated Categorical Responses. *Statistics in Medicine*, **12**, 1969-1987
- [7] Agresti, A. and Lang, J.B. 1993. A proportional odds model with subject-specific effects for repeated ordered categorical responses. *Biometrika* **80**, 3, 527-534
- [8] Albert, A. and Anderson, J.A. 1984. On the existence of maximum likelihood estimates in logistic regression. *Biometrika* **71**, 1, 1-10
- [9] Amemiya, T. 1981. Qualitative Response Models: A Survey. *Journ. of Economic Literature* Vol **21**, 1483-1536

- [10] Anderson, R.L., Mistretta, P.A. 1982. Management strategies for reducing losses caused by fusiform rust, annosus root rot, and little leaf disease. USDA For. Service Agric. Handbook. 597, 30 p.
- [11] Anderson, R.L., Cost, N.D., McClure, J.P., Ryan, G. 1986. Predicting Severity of Fusiform Rust in Young Loblolly and Slash Pine Stands in Florida, Georgia, and the Carolinas. *Southern Journal of Applied Forestry* **10**, 38-41
- [12] Anderson, R.L., McCartney, T.C., Cost, N.D., Devine, H. Botkin, M. 1988. Fusiform-Rust-Hazard Maps for Loblolly and Slash Pines. Research Note SE-351. U.S. Dep. Agric. For. Service. Southeastern Forest Exp. Station. 8 p.
- [13] Anderson, J.A. and Philips, P.R. 1981. Regression, Discrimination and Measurement Models for Ordered Categorical Variables. *Applied Statistics* **30**, No.1, 22-31
- [14] Anderson, J.A. 1984. Regression and Ordered Categorical Variables. *Journal of the Royal Statistics Society (B)* **46**, 1, 1-30 .
- [15] Arabatzis, A.A. 1990. Qualitative Response Models Theory and its Application to Forestry. Unpublished. PH.D. dissertation, Virginia Polytechnic Institute and SU
- [16] Arabatzis, A.A., Gregoire, T.G., Lenhart, J.D. 1991. Fusiform Rust Incidence in Loblolly and Slash Pine Plantations in East Texas. *South. Journ. of Appl. Forestry* **15**, 2, 79-84
- [17] Arabatzis, A.A. and Gregoire, T.G. 1991. Ordered and unordered multinomial response models: an application to assess loblolly pine merchantability. *Canadian Journal of Forest Research* **21**, 261-268
- [18] Armstrong, B.G. and Sloan, M. 1989. Ordinal Regression Models for Epidemiologic Data. *American Journal Of Epidemiology* **129**, No.1, 191-204
- [19] Ashby, D., Pocock, S.J., Shaper, A.G. 1986. Ordered Polytomous Regression: An Example Relating Serum Biochemistry and Haematology to Alcohol Consumption. *Applied Statistics* **35**, 3, 289-301
- [20] Ashby, D., West, C.R., Ames, D. 1989. The Ordered Logistic Regression Model in Psychiatry: Rising Prevalence of Dementia in Old People's Home. *Statistics in Medicine* **8**, 1317-1326
- [21] Bailey, N.T. 1964. *The Elements of Stochastic Processes*. Wiley and Sons, New York
- [22] Beal, S.L. and Sheiner, L.B. 1988. Heteroscedastic Nonlinear Regression. *Technometrics* **30**, No. 3, 327-338
- [23] Begg, C.B. and Gray, R. 1984. Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika* **71**, 1, 11-18

- [24] Berridge, D.M. and Whitehead, J. 1991. Analysis of Failure Time Data with Ordinal Categories of Response. *Statistics in Medicine* **10**, 1703-1710
- [25] Bhapkar, V.P. 1991. Sufficiency, ancillarity, and information in estimating functions. In: *Estimating Functions*, Ed .V.P.Godambe, Clarendon Oxford, p. 241-254
- [26] Bickel, P.J. and Doksum, K.A. 1977. *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, Oakland
- [27] Biging, G.S. 1985. Improved Estimates of Site Index Curves Using a Varying-Parameter Model. *Forest Science* **31**, 1, 248-259
- [28] Borders, B.E., 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
- [29] Bonney, G.E. 1987. Logistic Regression for Dependent Binary Observations. *Biometrics* **43**, 951-973
- [30] Bozdogan, H. 1987. Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions. *Biometrika* **52**, No. 3, 345-370
- [31] Breslow, N.E. and Clayton, D.G. 1993. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statist. Association* **88**, No.421, 9-25
- [32] Brooks, C.A. and Bonney, G.E. 1987. A Simulation Study of Properties of a Regressive Logistic Model. Technical Report No.87-4, 11-10-87, Department of Statistics, VPI & SU
- [33] Burton, J.D., Shoulders, E., Snow, G.A. 1985. Incidence and Impact of Fusiform Rust Vary with Silviculture in Slash Pine Plantations. *Forest Science* **31**, No.3, 671-680
- [34] Campbell, M.K. and Donner, A. 1989. Classification Efficiency of Multinomial Logistic Regression Relative to Ordinal Logistic Regression. *Journal of the American Statist. Association* **84**, No.406
- [35] Campbell, M.K., Donner, A., Webster, K.M. 1991. Are Ordinal Models Useful for Classifications? *Statistics in Medicine* **10**, 383-394
- [36] Carroll, R.J., and Ruppert, D. 1988. *Transformation and Weighting in Regression*. Chapman and Hall, New York.
- [37] Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **20**, 37-46
- [38] Conaway, M. 1989. Analysis of Repeated Categorical Measurements With Conditional Likelihood Methods. *Journal of the American Statist. Association* **84**, No. 405, 53-62

- [39] Conaway, M. 1990. A Random Effects Model for Binary Data. *Biometrics* **46**, 317-328
- [40] Connolly, M.A. and Liang, K.-Y. 1988. Conditional logistic regression models for correlated binary data. *Biometrika* **75**, 3, 501-506
- [41] Cox, C. 1988. Multinomial Regression based on Continuation Ratios. *Statistics in Medicine* **7**, 435-441
- [42] Cox, D.R. 1970. *Analysis of Binary Data (1st ed.)*. Chapman & Hall, London, New York
- [43] Cox, D.R. and Reid, N. 1987. Parameter Orthogonality and Approximate Conditional Inference. *Journal of the Royal Statistics Society (B)* **49**, 1-39
- [44] Cox, D.R. and Snell, E.J. 1989. *Analysis of Binary Data (2nd ed.)*. Chapman & Hall, London, New York
- [45] Davidian, M. and Gallant, A.R. 1993. The nonlinear mixed effects model with a smooth random effects density. *Biometrika* **80**, 3, 475-488
- [46] Davidian, M. and Giltinan, D.M. 1993a. Some General Estimation Methods for Nonlinear Mixed-Effects Models. *Journal of Biopharmaceutical Statistics*, **3**(1), 23-55
- [47] Davidian, M. and Giltinan, D.M. 1993b. Some Simple Methods for Estimating Intraindividual Variability in Nonlinear Mixed Effects Models. *Biometrics*, **49**, 59-73
- [48] Davidian, M. and Giltinan, D.M. 1995. *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall, New York.
- [49] Davies, R.B. 1988. A Reappraisal of Some Simple Statistical Models. In London Papers in Regional Science 18. Longitudinal Data Analysis: Methods and Applications Ed. M.D. Uncles, Pion, London, 103-115
- [50] Davies, R.B. and Crouchley, R. 1984. Calibrating Longitudinal Models of Residential Mobility and Migration. *Regional Science and Urban Economics* **14**, 231-247
- [51] Davies, R.B. and Crouchley, R. 1985. Control for Omitted Variables in the Analysis of Panel and Other Longitudinal Data. *Geographical Analysis* **17**, No.1, 1-15
- [52] Davis, A.W. and West, P.W. 1981. Remarks on *Generalized least squares estimation of yield functions* by I.S. Ferguson and J.W. Leech. *Forest Science* **27**, 233-239
- [53] Davis, C.S. 1991. Semi-Parametric and Non-Parametric Methods for the Analysis of Repeated Measurements with Applications to Clinical Trials. *Statistics in Medicine* **10**, 1959-1980

- [54] Dell, T.R., Driver, C.H. 1963. The Relationship of D.B.H. to Fusiform Stem Cankering in Old Field Slash Pine Plantations. *Journal of Forestry* **61**, 11,
- [55] Dempster, A.P., Laird, N.M, Rubin, D.B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistics Society (B)* **39**, No.1, 1-38
- [56] Devine, O.J., Clutter, J.L. 1985. Prediction of Survival in Slash Pine Plantations Infected with Fusiform Rust. *Forest Science* **31**, No. 1, 88-94
- [57] Diem, J.E. and Liukkonen, J.R. 1988. A Comparative Study of Three Methods for Analysing Longitudinal Pulmonary Function Data. *Statistics in Medicine* **7**, 19-28
- [58] Diggle, P.J. 1988. An Approach to the Analysis of Repeated Measurements. *Biometrics* **44**, 959-971
- [59] Diggle, P.J. 1990. *Time Series: A Biostatistical Introduction*. Clarendon Press, Oxford
- [60] Diggle, P.J., Liang, K.-Y., Zeger, S.L. 1994. *Analysis of Longitudinal Data*. Clarendon Press, Oxford
- [61] Efron, B. 1978. Regression and ANOVA with Zero-One Data: Measures of Residual Variation. *Journal of the American Statist. Association* **73**, 113-121.
- [62] Engel, J. 1988. Polytomous logistic regression. *Statistica Neerlandica* **42**, 4, 233-252
- [63] Engel, J. 1989. Polytomous Regression: Method and Application. Statistical Modelling: Proceedings of GLIM89 Intern. Conference on GLM. 134-143
- [64] Engel, J. 1990. Quasi-likelihood inference in a generalized linear mixed model for balanced data. *Statistica Neerlandica* **44**, nr. 4, 221-239
- [65] Ezzet, F. and Whitehead, J. 1991. A random effects model for ordinal responses from a crossover trial. *Statistics in Medicine* **10**, 901-907
- [66] Fahrmeir, L. and Tutz, G. 1994. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York.
- [67] Feller, W. 1957. *An Introduction to Probability Theory and Its Applications. Vol 1, 2nd ed.* Wiley and Sons, New York
- [68] Ferguson, I.S. and Leech, J.W. 1978. Generalized least squares estimation of yield functions. *Forest Science* **24**, 27-42
- [69] Fienberg, S.E. 1987. *The Analysis of Cross-Classified Categorical Data*. MIT Press, Cambridge
- [70] Fitzmaurice, G.M. and Laird, N.M. 1993. A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80**, No. 1, 141-151

- [71] Fitzmaurice, G.M., Laird, N.M., Rotnitzky, A.G. 1993. Regression Models for Discrete Longitudinal Responses. *Statistical Science* **8**, No. 3, 284-309
- [72] Gelfand, A.E., Smith, A.F.M. 1990. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statist. Association* **85**, No. 410, 398-409
- [73] Gelfand, A.E., Hills, S.E., Racine-Poon, A., Smith, A.F.M. 1990. Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. *Journal of the American Statist. Association* **85**, No. 412, 972-985
- [74] Geron, C.D., Hafley, W.L. 1988. Impact of Fusiform Rust on Product Yields of Loblolly Pine Plantations. *Southern Journal of Applied Forestry* **12**, No. 4, 226-231
- [75] Gibbons, R., Hedeker, D., Elkin, I., Waternaux, C., Kraemer, H.C., Greenhouse, J.B., Shea, T., Imber, S.D., Sotsky, S.M., Watkins, J.Y. 1993. Some Conceptual and Statistical Issues in Analysis of Longitudinal Psychiatric Data. *Archives of General Psychiatry* **50**, 739-750
- [76] Godambe, V.P. 1960. An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics* **31**, 1208-1211
- [77] Godambe, V.P. 1976. Conditional likelihood estimation and unconditional optimum estimating equations. *Biometrika* **63**, 277-284
- [78] Godambe, V.P. and Kale, B.K. 1991. Estimating functions: an overview. In *Estimating Functions*, Ed .V.P.Godambe, Clarendon Oxford, p.3-20
- [79] Godambe, V.P. and Heyde, C.C. 1987. Quasi-likelihood and Optimal Estimation. *International Statist. Review* **55**, 231-244
- [80] Godambe, V.P., and Thompson, M.E. 1989. An Extension Of Quasi-Likelihood Estimation. *Journal of Statist. Plann. and Inference* **22**, 137-152
- [81] Graybill, F.A. 1969. *Introduction To Matrices With Applications in Statistics*. Wadsworth Publ. Co. Belmont, California
- [82] Green, P.J. 1987. Penalized Likelihood for General Semi-parametric Regression Models. *International Statistical Review* **55**, 3, 245-259
- [83] Greenland, S. 1985. An Application of Logistic Models to the Analysis of Ordinal Responses. *Biom. Journal* **27**, 189-197
- [84] Greenwood, C. and Farewell, V. 1988. A comparison of regression models for ordinal data in an analysis of transplanted-kidney function. *The Canadian Journal of Statistics* **16**, No.4, 325-335
- [85] Gregoire, T.G. 1985. Generalized error structure for yield models fitted with permanent plot data. Unpublished Ph.D. diss., Yale University
- [86] Gregoire, T.G. 1987. Generalized Error Structure for Forestry Yield Models. *Forest Science* **33**, No. 2, 423-444

- [87] Gregoire, T.G., Schabenberger, O., Barrett, J.P. 1995. Linear Modelling of Irregularly Spaced, Unbalanced, Longitudinal Data From Permanent Plot Measurements. *Canadian Journal of Forest Research*, **25**, 1, 137-156
- [88] Gregoire, T.G. and Schabenberger, O. 1995a. Fitting bole-volume equations to spatially correlated within-tree data. Proceedings of the 6th annual KSU conference on *Applied Statistics* in Agriculture, April 24-26, 1994, Manhattan, Kansas
- [89] Gregoire, T.G. and Schabenberger, O. 1995b. Nonlinear Mixed-Effects Modeling of Cumulative Bole Volume With Spatially Correlated Within-Tree Data. *Journal of Agricultural, Biological and Environmental Statistics*. In press.
- [90] Gregoire, T.G. and Schabenberger, O. 1995c. A Nonlinear Mixed-effects Model to Predict Cumulative Bole Volume of Standing Trees. *Journal of Applied Statistics*. In press.
- [91] Grizzle, J.E., Starmer, C.F., Koch, G.G. 1969. Analysis of Categorical Data by Linear Models. *Biometrics* **25**, 489-504
- [92] Hagenaaars, J.A. 1990. *Categorical longitudinal data*. Sage Publications, London
- [93] Haberman, S.J. 1978. *Analysis of Qualitative Data*. Academic Press, New York
- [94] Haberman, S.J. 1982. Analysis of Dispersion of Multinomial Responses. *Journal of the American Statist. Association* **77**, No. 379, 568-580
- [95] Harville, D.A. 1977. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statist. Association* **72**, No. 358, 320-340
- [96] Harville, D.A. and Mee, R.W. 1984. A Mixed-Model Procedure for Analyzing Ordered Categorical Data. *Biometrics* **40**, 393-408
- [97] Hastie, T. and Tibshirani, R. 1987a. Non-parametric Logistic and Proportional Odds Regression. *Applied Statistics* **36**, 2, 260-276
- [98] Hastie, T.J., Botha, J.L., Schnitzler, C.M. 1989. Regression with an Ordered Categorical Response. *Statistics in Medicine* **8**, 785-794
- [99] Hedeker, D. and Gibbons, R.D. 1994. A Random-Effects Ordinal Regression Model for Multilevel Analysis. *Biometrics* **50**, 933-944
- [100] Hedeker, D., Gibbons, R.D., Flay, B.R. 1994. Random-effects Regression Models for Clustered Data: With an Example from Smoking Prevention Research. *Journal of Consulting and Clinical Psychology*. In press
- [101] Hinde, J.P. 1988. Longitudinal Data Analysis and Generalised Linear Models. Some Computational Approaches. In London Papers in Regional Science 18. Longitudinal Data Analysis: Methods and Applications Ed. M.D. Uncles, Pion, London 116-123

- [102] Hirst, K., Zerbe, G.O., Boyle, D.W., Wilkening, R.B. 1991. On Nonlinear Random Effects Models for Repeated Measurements. *Commun. Statist.-Simula* **20**(2&3), 463-478
- [103] Holtbrügge, W. and Schumacher, M. 1991. A Comparison of Regression Models for the Analysis of Ordered Categorical Data. *Applied Statistics* **40**, No.2, 249-259
- [104] Hosmer, D.W.Jr., Lemeshow, S., Klar, J. 1988. Goodness-of-fit Testing for the Logistic Regression Model when the Estimated Probabilities are Small. *Biom. Journal* **30**, 911-924
- [105] Hosmer, D.W.Jr. and Lemeshow, S. 1989. *Applied Logistic Regression*. John Wiley & Sons, Inc., New York
- [106] Hotelling, H. 1936. Relation between two sets of variates. *Biometrika* **28**, 321-377
- [107] Hunt, E.V. Jr., Lenhart, J.D. 1986. Fusiform Rust Trends in East Texas. *Southern Journal of Applied Forestry* **10**, 215-216
- [108] Jansen, J. 1990. On the Statistical Analysis of Ordinal Data when Extravariation is Present. *Applied Statistics* **38**, No.1, 75-84
- [109] Jennrich, R.I. and Schluchter, M.D. 1986. Unbalanced Repeated-Measures Models with Structured Covariance Matrices. *Biometrics* **42**, 805-820
- [110] Jobson, J.D. 1990. *Applied Multivariate Data Analysis Vol II: Categorical and Multivariate Methods*. Springer-Verlag, Heidelberg, New York
- [111] Johnson, N.L. and Kotz, S. 1970. *Discrete Distributions*. Wiley and Sons, New York
- [112] Jones, R.H. 1990. Serial Correlation or Random Subject Effects? *Commun. Statist.-Simula* **19**(3), 1105-1123
- [113] Jones, R.H. 1993. *Longitudinal Data with Serial Correlation: A State-space Approach*. Chapman & Hall, New York
- [114] Jones, R.H. and Ackerson, L.M. 1990. Serial correlation in unequally spaced longitudinal data. *Biometrika* **77**, 721-731
- [115] Jones, R.H. and Boadi-Boateng, F. 1991. Unequally Spaced Longitudinal Data with AR(1) Serial Correlation. *Biometrics* **47**, 161-175
- [116] Kalbfleisch, J.D. and Lawless, J.F. 1985. The Analysis of Panel Data Under a Markov Assumption. *Journal of the American Statist. Association* **80**, No. 392, 863-871
- [117] Karim, M.R. and Zeger, S.L. 1992. Generalized Linear Models with Random Effects; Salamander Mating Revisited. *Biometrics* **48**, 631-644
- [118] Kelley, W.D., Runion, G.B. 1991. Control of Fusiform Rust on Loblolly and Slash Pine Seedlings in Forest Nurseries in the Southeastern United States. 338-

340. In: *Rusts in Pine*. Proceedings of the IUFRO Rusts of Pine Working Party Conference September 18-22, 1989; Banff, Alberta, Canada. H.Y. (Hiratsuka, J.K. Samoil, P.V. Blenis; P.E. Crane, B.L. Laishley, eds.)
- [119] Kiefer, J. 1959. Optimum experimental designs. *Journal of the Royal Statistic Society (B)* **21**, 272-319
- [120] Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H., Lehnen, R.G. 1977. A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* **33**, 133-158
- [121] Korn, E.L. and Whittemore, A.S. 1979. Methods for Analyzing Panel Studies of Acute Health Effects of Air Pollution. *Biometrics* **35**, 795-802
- [122] Kuk, A.Y.C. and Chen, C.-H. 1992. A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79**, 3, 531-541
- [123] Läärä, E. and Matthews, J.N.S. 1985. The equivalence of two models for ordinal data. *Biometrika* **72**, 1, 206-207
- [124] Laird, N. 1978. Nonparametric Maximum Likelihood Estimation of a Mixing Distribution. *Journal of the American Statist. Association* **73**, 364, 805-811
- [125] Laird, N.M. and Ware, J.H. 1982. Random-Effects Models for Longitudinal Data. *Biometrics* **38**, 963-974
- [126] Laird, N.M. 1988. Missing Data in Longitudinal Studies. *Statistics in Medicine* **7**, 305-315
- [127] Laird, N. 1991. Topics in Likelihood-Based Methods for Longitudinal Data Analysis. *Statistica Sinica* **1**, 33-50
- [128] Landis, J.R., Miller, M.E., Davis, C.S., Koch, G.G. 1988. Some General Methods for the Analysis of Categorical Data in Longitudinal Studies. *Statistics in Medicine* **7**, 109-137
- [129] Lenhart, J.D., McGrath, W.T., Hackett, T.L. 1988. Fusiform Rust Trends in East Texas: 1969-1987. *Southern Journal of Applied Forestry* **12**, No. 4, 259-261
- [130] Lenhart, J.D., Konrad, G.D., Fountain, M.S. 1993. Comparison of Planted Loblolly and Slash Pine Performance in Southeast Texas. *Southern Journal of Applied Forestry* **17**, No. 1, 26-31
- [131] Lenhart, J.D., Gregoire, T.G., Konrad, G.D., Holley, A.G. 1994. Characterizing Fusiform Rust Incidence and Distribution in East Texas. *Southern Journal of Applied Forestry* **18**, No. 1, 29-34
- [132] Liang, K.-Y. 1987. Estimating functions and approximate conditional likelihood. *Biometrika* **74**, 4, 695-702
- [133] Liang, K.-Y. 1992. Extensions of Generalized Linear Models in the Past Twenty Years. Overview and Some Biomedical Applications. Proceedings of

XVIth International Biometric Conference, Hamilton, New Zealand, December 7-11, 1992

- [134] Liang, K.-Y. and Zeger, S.L. 1986. Longitudinal analysis using generalized linear models. *Biometrika* **73**, 1, 13-22
- [135] Liang, K.-Y. and Zeger, S.L. 1989. A Class of Logistic Regression Models for Multivariate Binary Time Series. *Journal of the American Statist. Association* **84**, 406, 447-451
- [136] Liang, K.-Y., Zeger, S.L., Qaqish, B. 1992. Multivariate Regression Analyses for Categorical Data. *Journ. Royal Statist. Society (B)* **54**, No.1, 3-40
- [137] Liang, K.-Y. and McCullagh, P. 1993. Case Studies in Binary Dispersion. *Biometrics* **49**, 623-630
- [138] Lindstrom, M.J. and Bates, D.M. Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated Measures Data. *Journal of the American Statist. Association* **83**, No. 404, 1014-1022
- [139] Lindstrom, M.J. and Bates, D.M. 1990. Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics* **46**, 673-687
- [140] Lipsitz, S.R., Laird, N.M., Harrington, D.P. 1991. Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika* **78**, 1, 153-160
- [141] Lipsitz, S.R. 1992. Methods for Estimating the Parameters of a Linear Model for Ordered Categorical Data. *Biometrics* **48**, 271-281
- [142] Lipsitz, S.R., Kim, K., Zhao, L. 1994. Analysis of Repeated Categorical Data Using Generalized Estimating Equations. *Statistics in Medicine*, **13**, 1149-1163
- [143] Little, R.J.A. and Rubin, D.B. 1987. *Statistical Analysis with Missing Data*. J.Wiley & Sons, New York
- [144] Longford, N.T. 1993. *Random Coefficient Models*. Clarendon, Oxford
- [145] Louis, T. 1988. General Methods for Analysing Repeated Measures. *Statistics in Medicine* **7**, 29-45
- [146] Magnus, J.R. and Neudecker, H. 1988. *Matrix Differential Calculus With Applications in Statistics and Econometrics*. Wiley and Sons, New York
- [147] Marshall, A.W. and Olkin, I. 1968. A General Approach to Some Screening and Classification Problems. *Journal of the Royal Statistics Society (B)* **30**, No.3, 407-435
- [148] McCullagh, P. 1980. Regression Models for Ordinal Data. *Journal of the Royal Statistics Society (B)* **42**, 2, 109-142
- [149] McCullagh, P. 1983. Quasi-Likelihood Functions. *The Annals of Statistics* **11**, 1, 59-67

- [150] McCullagh, P. 1984. On the Elimination of Nuisance Parameters in the Proportional Odds Model. *Journ. Royal Statist. Society (B)* **46**, No.2, 250-256
- [151] McCullagh, P. and Nelder, J.A. 1989. *Generalized Linear Models (2nd ed.)*. Chapman & Hall, London, New York
- [152] Miller, M.E. and Landis, J.R. 1991. Generalized Variance Component Models for Clustered Categorical Response Variables. *Biometrics* **47**, 33-44
- [153] Miller, M.E., Davis, C.S., Landis, J.R. 1993. The Analysis of Longitudinal Polytomous Data: Generalized Estimating Equations and Connections with Weighted Least Squares. *Biometrics* **49**, 1033-1044
- [154] Molenberghs, G., Lesaffre, E. 1992. Marginal modelling of Correlated Ordinal Data using a n-way Plackett Distribution. In: *Advances in GLIM and Statistical Modelling*, (Fahrmeier, L., Francis, B., Gilchrist R., Tutz, G. eds.), Proceedings of the GLIM92 Conference, Munich, 13-17 July 1992, Lecture Notes in Statistics 78, Springer Verlag
- [155] Nance, W.L., Froelich, R.C., Shoulders, E. 1981. Effects of fusiform rust on survival and structure of Mississippi and Louisiana slash pine plantations. U.S. Dep. Agric. For. Service Res. Pap. SO-172, 11 p. South For. Exp. Station, New Orleans, La.
- [156] Nelder, J.A., Lee, Y. 1992. Likelihood, Quasi-Likelihood and Pseudolikelihood: Some Comparisons. *Journal of the Royal Statistics Society (B)* **54**, 1, 273-284
- [157] Nelder, J.A. and Wedderburn R.W.M. 1972. Generalized linear models. *Journal of the Royal Statist. Society (A)* **135**, 370-384
- [158] Nelder, J.A., Pregibon, D. 1987. An extended quasi-likelihood function. *Biometrika* **74**, 2, 221-232
- [159] Neuhaus, J.M. and Jewell, N.P. 1990. Some Comments on Rosner's Multiple Logistic Model for Clustered Data. *Biometrics* **46**, 534-534
- [160] Neuhaus, J.M., Kalbfleisch, J.D, Hauck, W.W. 1991. A Comparison of Cluster-Specific and Population-Averaged Approaches for Analyzing Correlated Binary Data. *International Statistical Review* **59**, 25-35
- [161] Paik, M.C. 1992. Parametric Variance Function Estimation for Nonnormal Repeated Measurement Data. *Biometrics* **48**, 19-30
- [162] Park, T. 1993. A Comparison of the Generalized Estimating Equation Approach With the Maximum Likelihood Approach for Repeated Measurements. *Statistics in Medicine* **12**, 1723-1732
- [163] Plackett, R.L. 1981. *The Analysis of Categorical Data, (2nd ed.)*. Charles Griffin & Co. Ltd., London

- [164] Powers, H.R. Jr., McClure, J.P., Knight, H.A., Dutrow, G.F. 1975. Fusiform rust: forest survey incidence data and financial impact in the south. *US Dept. Agric. For. Serv., Southeast For. Exp. Station. Research Paper* SE-127
- [165] Powers, H.R. Jr., Schmidt, R.A., Snow, G.A. 1981. Current Status and management of fusiform rust on southern pines. *Ann. Rev. Phytopathology* **19**, 353-371
- [166] Powers, H.R.. Jr., Kaneko, S., Kuhlman, E.G., La, Y.-J., Yi, C.-K. 1991. Susceptibility of Asian and American Oaks and Pines to *Cronartium Quercuum*. 313-318. In: *Rusts in Pine*. Proceedings of the IUFRO Rusts of Pine Working Party Conference September 18-22, 1989; Banff, Alberta, Canada. HY. (Hiratsuka, J.K. Samoil, P.V. Blenis; P.E. Crane, B.L. Laishley, eds.)
- [167] Powers, H.R. Jr., Miller, T., Belanger, R.P. 1993. Management Strategies to Reduce Losses from Fusiform Rust. *Southern Journal of Applied Forestry* **17**, No. 3, 146-149
- [168] Pregibon, D. 1981. Logistic Regression Diagnostics. *The Annals of Statistics* **9**, 4, 705-724
- [169] Preisler, H.K. 1988. Maximum Likelihood Estimates for Binary Data with Random Effects. *Biom. Journal* **30**, 3, 339-350
- [170] Preisler, H.K. 1989. Analysis of a Toxicological Experiment Using a Generalized Linear Model with Nested Random Effects. *International Statistical Review* **57**, 2, 145-159
- [171] Prentice, R.L. 1988. Correlated Binary Regression with Covariates Specific to Each Binary Observation. *Biometrics* **44**, 1033-1048
- [172] Qu, Y., Williams, G.W., Beck, G.J., Goormastic, M. 1987. A Generalized Model of Logistic Regression For Clustered Data. *Commun. Statist. Theory Meth.* **16**(12), 3447-3476
- [173] Racine-Poon, A. 1985. A Bayesian Approach to Nonlinear Random Effects Models. *Biometrics* **41**, 1015-1023
- [174] Rao, C.R. 1973. *Linear Statistical Inference and its Applications 2ed.*. Wiley and Sons, New York
- [175] Robinson, G.K. 1991. That BLUP Is a Good Thing: The Estimation of Random Effects. *Statistical Science*, **6**, No. 1, 15-51
- [176] Rosner, B. 1984. Multivariate Methods in Ophthalmology with Application to Other Paired-Data Situations. *Biometrics* **40**, 1025-1035
- [177] Rosner, B. 1989. Multivariate Methods for Clustered Binary Data With More Than One Level of Nesting. *Journal of the American Statist. Association* **84**, 406, 373-380

- [178] Rosner, B. and Milton, R.C. 1988. Significance Testing for Correlated Binary Outcome Data. *Biometrics* **44**, 505-512
- [179] Rosner, B., Munoz, A., Tager, I., Speizer, F., Weiss, S. 1985. The Use of an Autoregressive Model for the Analysis of Longitudinal Data in Epidemiologic Studies. *Statistics in Medicine* **4**, 457-467
- [180] Rosner, B. and Munoz, A. 1988. Autoregressive Modelling for the Analysis of Longitudinal Data with Unequally Spaced Examinations. *Statistics in Medicine* **7**, 59-71
- [181] Rosner, B. and Tosteson, T.D. 1990. Response on Neuhaus & Jewell. *Biometrics* **46**, 523-534
- [182] Rotnitzky, A. and Jewell, N.P. 1990. Hypothesis testing of regression parameters in semiparametric generalized linear models for clustered data. *Biometrika* **77**, 485-497
- [183] Rubin, D.R. 1976. Inference and missing data. *Biometrika* **63**, 3, 581-592
- [184] Rutter, C.M., Elashoff, R.M. 1994. Analysis of Longitudinal Data: Random Coefficient Regression Modelling. *Statistics in Medicine* (in press).
- [185] Schabenberger, O. 1993. Loglinear Models for Square Contingency Tables, The Validation of Classifications. Paper presented at the 1993 Spring Statistical Meeting of the Biometric Society (ENAR), Philadelphia. March 21-24, 1993
- [186] Schabenberger, O. and Gregoire, T.G. 1995. A Conspectus on Estimating Function Theory and its Applicability to Recurrent Modeling Issues in Forest Biometry. *Silva Fennica*, **29**, 1. (in press)
- [187] Schabenberger, O. 1995a. The Use of Ordinal Response Methodology in Forestry. *Forest Science*, **41**. (in press)
- [188] Schabenberger, O. 1995b. Nonlinear Mixed Effects Growth Models for Repeated Measures in Ecology. Proceedings of the 1994 JSM, Section on Statistics and the Environment, August 14-18, 1994, Toronto, Canada.
- [189] Schabenberger, O., Gregoire, T.G., Burkhart, H.E. 1995. Commentary: *Multi-state models for monitoring individual trees in permanent observation plots* by Urfer, W., Schwarzenbach, F.H., Kütting, J., Müller, P.. *Environmental and Ecological Statistics*. In press.
- [190] Schmidt, R.A., Squillace, A.E., 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
- [191] Seber, G.A.F. and Wild, C.J. 1989. *Nonlinear Regression*. Wiley and Sons, New York

- [192] Sharples, K., and Breslow, N. 1992. Regression Analysis of Correlated Binary Data: Some Small Sample Results for The Estimating Equation Approach. *J. Statist. Comput. Simula.* **42**, 1-20
- [193] Sheiner, L.B., and Beal, S.L. 1980. Evaluation of Methods for Estimating Population Pharmacokinetic Parameters. I. Michaelis-Menten Model: Routine Clinical Pharmacokinetic Data. *Journal of Pharmacokinetics and Biopharmaceutics* **8**, No. 6, 553-571
- [194] Shoulders, E., Nance, W.L. 1987. Effects of Fusiform Rust on Survival and Structure of Mississippi and Louisiana Loblolly Pine Plantations. U.S. Dep. Agric. For. Service Res. Pap. SO-232. Southern For. Exp. Station, New Orleans, La.
- [195] Snow, G.A. , Wells, O.O., Switzer, G.L. 1986. Fusiform Rust Gradient in a Loblolly Pine Plantation. *Forest Science* **32**, No. 2, 372-376
- [196] Stiratelli, R., Laird, N., Ware, J.H. 1984. Random-Effects Models for Serial Observations with Binary Response. *Biometrics* **40**, 961-971
- [197] Stram, D.O., Wei, L.J., Ware, J.H. 1988. Analysis of Repeated Ordered Categorical Outcomes with Possibly Missing Observations and Time-Dependent Covariates. *Journal of the American Statist. Association* **83**, 403, 631-637
- [198] Tanner, M.A. and Young, M.A. 1985. Modeling Agreement Among Raters. *Journal of the American Statist. Association* **80**, 389, 175-180
- [199] Taylor, J.M.G. 1990. Properties of Maximum Likelihood Estimates of the Ratio of Parameters in Ordinal Response Regression Models. *Commun. Statist.-Simula* **19**(2), 469-480
- [200] Thall, P.F. and Vail, S.C. 1990. Some Covariance Models for Longitudinal Count Data with Overdispersion. *Biometrics* **46**, 657-671
- [201] Thompson, R. and Baker, R.J. 1981. Composite Link Functions in Generalized Linear Models. *Applied Statistics* **30**, 2, 125-131
- [202] Tosteson, T.D., Stefanski, L.A., Schafer, D.W. 1989. A Measurement-Error Model for Binary and Ordinal Regression. *Statistics in Medicine* **8**, 1139-1147
- [203] Uncles, M.D. 1988. Issues in Longitudinal Data Analysis. In London Papers in Regional Science 18. Longitudinal Data Analysis: Methods and Applications Ed. M.D. Uncles, Pion, London 1-12
- [204] Vonesh, E.F. 1993. MIXNLIN: A SAS Procedure for Nonlinear Mixed-effects Models. Technical Report # TR92M-0300A Baxter Healthcare Corporation, Illinois
- [205] Vonesh, E.F. and Carter, R.L. 1992. Mixed-Effect Nonlinear Regression for Unbalanced Repeated Measures. *Biometrics* **48**, 1-17

- [206] Ware, J.H. 1985. Linear Models for the Analysis of Longitudinal Studies. *The American Statistician* **39**, 2, 95-101
- [207] Ware, J.H., Lipsitz, S., Speizer, F.E. 1988. Issues in the Analysis of Repeated Categorical Outcomes. *Statistics in Medicine* **7**, 95-107
- [208] Wedderburn, R.W.M. 1974. Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* **61**, 3, 439-447
- [209] Wei, L.J. and Stram, D.O. 1988. Analysing Repeated Measurements with Possibly Missing Observations by Modelling Marginal Distributions. *Statistics in Medicine* **7**, 139-148
- [210] Wells, O.O., Dinus, R.J. 1987. Early Infections as a Predictor of Mortality Associated with Fusiform Rust of Southern Pines. *Journal of Forestry* **76**, 8-12
- [211] White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**, 4, 817-
- [212] Wolfinger, R. 1992. A Tutorial on Mixed Models. SAS Technical Report TS-260. SAS Institute, Cary, North Carolina
- [213] Wolfinger, R., Miles-McDermott, N., Kendall, J. 1992. Analyzing split-plot and repeated measures designs using mixed models. In: Applied Statistics in Agriculture Proceedings (G.A. Milliken and J.R. Schwenke, eds.) 26-28 April 1992, Manhattan, KS.
- [214] Wolfinger, R. 1993. Covariance Structure Selection in General Mixed Models. *Commun. Statist.-Simula.* **22**(4), 1078-1106
- [215] Wolfinger, R. and O'Connell M. 1993. Generalized Linear Mixed Models: A Pseudo-Likelihood Approach. *Journal of Statistical Computation and Simulation* **48**, 233-243
- [216] Wolfinger, R. 1993. Laplace's approximation for nonlinear mixed models. *Biometrika* **80**, 4, 791-795
- [217] Wolfinger, R., Tobias, R., Sall, J. 1993. Computing Gaussian Likelihoods and Their Derivatives for General Linear Mixed Models, *SIAM Journal of Scientific Computing*. In press.
- [218] Wu, M.C. and Ware, J.H. 1979. On the Use of Repeated Measurements in Regression Analysis With Dichotomous Response. *Biometrics* **35**, 513-521
- [219] Wu, M.C. and Carroll, R.J. 1988. Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process. *Biometrics* **44**, 175-188
- [220] Wu, M.C. and Bailey, K. 1988. Analysing Changes in the Presence of Informative Right Censoring Caused by Death and Withdrawal. *Statistics in Medicine* **7**, 337-346
- [221] Zeger, S.L. 1988a. Commentary. *Statistics in Medicine* **7**, 161-168

- [222] Zeger, S.L. 1988b. A regression model for time series of counts. *Biometrika* **75**, 621-629
- [223] Zeger, S.L., Liang, K.-Y. 1986. Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics* **42**, 121-130
- [224] Zeger, S.L., Liang, K.-Y. 1992. An Overview of Methods for the Analysis of Longitudinal Data. *Statistics in Medicine* **11**, 1825-1839
- [225] Zeger, S.L., Liang, K.-Y., Self, S. 1985. The analysis of binary longitudinal data with time-independent covariates. *Biometrika* **72**, 1, 31-38
- [226] Zeger, S.L., Liang, K.-Y., Albert, P.S. 1988. Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics* **44**, 1049-1060
- [227] Zeger, S.L. and Karim, M.R. 1991. Generalized Linear Models With Random Effects: A Gibbs Sampling Approach. *Journal of the American Statist. Association* **86**, No.413, 79-86
- [228] Zeger, S.L. and Qaqish, B. 1988. Markov regression models for time series: A quasi-likelihood approach. *Biometrics* **44**, 1019-1031
- [229] Zhao, L.P. and Prentice, R.L. 1990. Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 642-648
- [230] Zhao, L.P. and Prentice, R.L. 1991. Use of a quadratic exponential model to generate estimating equations for means, variances and covariances. In: *Estimating Functions*, Ed. V.P. Godambe, Clarendon Oxford, p.103-117

9.1. References by Subject

The following list contains references organized by certain subject areas. The numbers refer to the bibliography above. Some references naturally appear in more than one category.

Estimating Function Theory

25, 76, 77, 78, 79, 80, 132, 133, 186, 208

Fusiform Rust

10, 11, 12, 16, 28, 33, 54, 56, 74, 107, 118, 129, 130, 131, 155, 164, 165, 166, 167, 190, 194, 195, 186

Generalized Estimating Equations

53, 60, 70, 71, 79, 80, 133, 134, 136, 140, 141, 142, 149, 151, 153, 161, 162, 171, 186, 192, 200, 221, 223, 224, 226, 229, 230

Generalized Linear Models

General

9, 42, 44, 60, 66, 91, 117, 149, 151, 156, 157, 158, 186, 201, 208

Logistic regression

8, 15, 16, 17, 18, 23, 29, 32, 83, 104, 105, 135, 168

Longitudinal Data

General Issues

49,50,51,53, 58, 59, 60, 66, 71, 75, 101, 112, 113, 114, 115, 120,125,127, 128, 144, 145, 160, 182, 186, 189, 196, 200, 203, 206, 207, 209, 220, 221

Missing Data: 55, 126, 143, 144, 183, 197, 219, 220

Mixed Models

Linear, Continuous Response

58, 59, 60, 66, 87, 95, 100, 109, 112, 113, 114, 115, 125, 138, 175, 179, 184, 186, 206, 212, 213, 214, 217

Non-linear, Continuous Response

22, 45, 46,47,82, 88, 89, 90, 102, 139, 173, 186, 189, 193, 204,205,215,216

Categorical Response

6, 31, 39, 60, 64, 65, 66, 70, 82, 96, 99, 108, 117, 121, 122, 152, 160, 169, 170, 186, 215, 216, 226, 227

Ordinal Regression Models

2, 3, 5, 14, 18, 24, 41, 66, 69, 83, 84, 94, 99, 108, 123, 141, 142, 148, 151, 154, 186, 197, 199, 202

Proportional Odds Model

7, 19, 20, 60, 62, 63, 66, 69, 84, 96, 97, 98, 123, 148, 150, 151, 186, 187, 189

Classification: 13, 34, 35, 37, 84, 103, 147, 185, 198

Transitional, Conditional Models

29, 32, 38, 39, 40, 43, 60, 66, 77, 159, 172, 176, 177, 178, 181, 182

Textbooks

1, 4, 21, 26, 42, 44, 59, 60, 66, 67, 69, 78, 92, 93, 110, 111, 143, 144, 146, 151, 163, 174, 191

Chapter 10

Appendix

The purpose of this appendix is to list some results and theorems in mathematical statistics without extensive proofs but supportive arguments, which are referred to in Chapters 3 and 4 of this dissertation.

The results are mainly taken from Bickel and Doksum (1977), McCullagh and Nelder (1989), Agresti (1990), and Bhapkar (1991) denoted [BD], [MCN], [A], and [B], respectively.

Exponential family of distributions

[A.1] Definition [BD]: The family of distributions of a model $\{P_\theta : \theta \in \Omega\}$ is said to be a one parameter exponential family, if we can write the density of the random variable X as

$$p(x, \theta) = \{\exp[c(\theta)T(x) + d(\theta) + S(x)]\}I_{A(x)}$$

where $c(\theta), d(\theta)$ are real valued on Ω , T and S are functions on R and the

$I_{A(x)}$ is the indicator of the set $A \subset R$ which does not depend on θ . Notation: We use $X \sim P_{\text{expo}}$ as a shorthand to denote membership in the exponential family.

Example: Let $X \sim \text{Bernoulli}(\theta) \Rightarrow p(x, \theta) = \theta^x(1 - \theta)^{1-x}; x \in \{0, 1\}$.

Rewrite: $p(x, \theta) = \exp[\log(\frac{\theta}{1-\theta})x + \log(1 - \theta)]I_{A(x)}$ and

identify: $c(\theta) = \log(\frac{\theta}{1-\theta})$

$$d(\theta) = \log(1 - \theta)$$

$$T(x) = x$$

$$S(x) = 0$$

$$A(x) = \{x : x \in \{0, 1\}\} \Rightarrow \text{does not depend on } \theta$$

$\Rightarrow \text{Bernoulli}(\theta)$ is a member of the one-parameter exponential family.

Corollaries [BD]:

[A.2] If $X_i \sim \text{iid } p(x_i, \theta), i = 1, \dots, n$ and $p(x_i, \theta)$ is in the exponential family with functions $c_1(\theta), d_1(\theta), T(x_i), S(x_i)$, and $I_{A(x_i)}$, then $p(\mathbf{x}, \theta)$ is in the exponential family with

$$c(\theta) = c_1(\theta)$$

$$d(\theta) = nd_1(\theta)$$

$$T(\mathbf{x}) = \sum_i T(x_i)$$

$$S(\mathbf{x}) = \sum_i S(x_i)$$

$$A(\mathbf{x}) = \{[x_i, \dots, x_n] : x_i \in A, \forall i\}$$

[A.3] [BD] $T(\mathbf{x})$ is a complete, minimal sufficient statistic for θ

[A.4] [BD] $T(\mathbf{x})$ is the unique, minimum variance unbiased estimator of $E(T(\mathbf{x}))$

[A.5] [BD] If X_i is distributed in the exponential family, so is $T(x)$

[A.6] Most classical distributions are in P_{expo} . A short list includes the

(i) continuous distributions: Gaussian, Beta family, Gamma family, Weibull, Pareto, ...

(ii) discrete distributions: Poisson, Bernoulli, Binomial, Multinomial, Negative Binomial, Geometric, ...

[A.7] Natural form [BD]: Let $\eta = c(\theta)$. The exponential family then has the form:

$$p(\mathbf{x}, \eta) = \exp[\eta T(\mathbf{x}) + d_0(\eta) + S(\mathbf{x})]I_{A(\mathbf{x})}$$

where $d_0(\eta) = d(c^{-1}(\eta))$, provided c is a 1-1 function of θ (is usually the case).

Corollary [BD]:

[A.8] If $\mathbf{X}$ is distributed in the exponential family, the moment generating function of $T(\mathbf{X})$ is given by:

$$M_T(s) = E(e^{sT(\mathbf{X})}) = \exp[d_0(\eta) - d_0(s + \eta)]$$

for s close to 0.

[A.9] [BD] $E(T(\mathbf{X})) = -d'_0(\eta)$, $\text{var}(T(\mathbf{X})) = -d''_0(\eta)$

[A.10] Definition [BD]: A family of distributions is said to be k parameter exponential family, if the density can be written

$$p(\mathbf{x}, \theta) = \left\{ \exp \left[\sum_{i=1}^k c_i(\theta) T_i(\mathbf{x}) + d(\theta) + S(\mathbf{x}) \right] \right\} I_{A(\mathbf{x})}$$

where $c_1, \dots, c_k, d$ are real-valued functions of θ ; $T_1, \dots, T_k, S$ are real-valued functions on R^n and set A does not depend on θ .

Corollary [BD]:

[A.11] $T(\mathbf{X}) = [T_1(\mathbf{X}), \dots, T_k(\mathbf{X})]$ is the k -element complete minimal sufficient statistic for θ .

Estimation principles

[A.12] Frequency substitution: Suppose multinomial trials can result in categorical outcomes $C_1, \dots, C_k$ with probabilities $p_1, \dots, p_k$. If X_i denotes the number of instances where C_i has been observed in n trials, a natural estimator for p_i is $\hat{p}_i = X_i/n$. The principle of frequency substitution says to estimate $q(p_1, \dots, p_k)$ by $q(\hat{p}_1, \dots, \hat{p}_k) = q\left(\frac{X_1}{n}, \dots, \frac{X_k}{n}\right)$.

[A.13] Method of Moments: Let $m_j(\theta) = E(X_i^j)$ and $\hat{m}_j = \frac{\sum_i x_i^j}{n}$. The method of moments (MOM) estimates $q(\theta) = h(m_1(\theta), \dots, m_l(\theta))$ by $h(\hat{m}_1(\theta), \dots, \hat{m}_l(\theta))$.

[A.14] Method of Least Squares: If $Y_i = h_i(\theta_1, \dots, \theta_r) + \epsilon_i$, $\epsilon_i \sim (0, \sigma^2)$, $\text{cov}(\epsilon_i, \epsilon_j) = 0 \forall i \neq j$, the least squares estimate (LSE) are the ones that satisfy $\min \left\{ \sum_i (y_i - h_i(\theta_1, \dots, \theta_r))^2 \right\}$.

[A.15] Maximum Likelihood [BD]: If $p(\mathbf{x}, \theta)$ is the density of $\mathbf{X}$ and $\theta \in \Theta$ and we consider $p(\mathbf{x}, \theta)$ as a function of θ for fixed $\mathbf{x}$ yields the likelihood $L(\theta(\mathbf{x}), \mathbf{x})$. The maximum likelihood principle finds the value $\hat{\theta}(\mathbf{x})$ that is most likely to have produced the data. That is, $\hat{\theta}(\mathbf{x})$ satisfies

$$L(\hat{\theta}(\mathbf{x}), \mathbf{x}) = p(\mathbf{x}, \hat{\theta}(\mathbf{x})) = \max\{p(\mathbf{x}, \theta) : \theta \in \Theta\} = \max\{L(\theta, \mathbf{x}) : \theta \in \Theta\}$$

[A.16] Invariance property of MLE's: If a maximum likelihood estimate (MLE) $\hat{\theta}$ of θ exists, and $q(\theta)$ is a measurable function, then the MLE of $q(\theta)$ is $q(\hat{\theta})$.

UMVU Estimation:

[A.17] If $T(\mathbf{X})$ is the complete sufficient statistic for θ , then $h(T(\mathbf{X}))$ is the uniformly minimum variance unbiased estimator of $q(\theta) = E(T(\mathbf{X}))$.

[A.18] If $S(\mathbf{X})$ is an unbiased estimator of $q(\theta)$, then $E(S(\mathbf{X}) | T(\mathbf{X}))$ is the uniformly minimum variance unbiased estimator of $q(\theta)$ if $T(\mathbf{X})$ is complete sufficient for θ .

[A.19] Quasi-Likelihood [MCN]: If $E(\mathbf{X}) = \mu$ and $\text{var}(\mathbf{X}) = \sigma^2 \text{Diag}(h_i(\mu))$ it has been found that $U(\mu, Y) = \frac{Y - \mu}{\sigma^2 h(\mu)}$ behaves like a likelihood score function, namely

$$E(U(\mu, Y)) = 0; \text{var}(U(\mu, Y)) = -E\left(\frac{\partial U}{\partial \mu}\right) = \{\sigma^2 h(\mu)\}^{-1}.$$

Consequently, $Q(\mu, y) = \int_y^\mu \frac{y-t}{\sigma^2 h(t)} dt$ should behave like a log-likelihood. Since no distributional assumption has been made, $Q(\mu, y)$ has been termed the quasi-log-likelihood or quasi-likelihood. The quasi-likelihood principle says to select those values as estimates $\hat{\mu}$, that maximize the quasi-log-likelihood.

Estimating equations

[A.20] All of the above estimation principles solve equations of the form $q(\mathbf{x}, \theta) = 0$, solution of which serve as estimates as long as they are interior points of the parameter space Θ .

Method of Moments: The MOM equations for $q(\theta)$ are $q(\hat{\theta}) - h(\hat{m}_1(\theta), \dots, \hat{m}_l(\theta)) = 0$.

Least Squares: The least squares estimating equation(s), also known as the normal equation(s) can be written algebraically as $\sum_i \{y_i - h_i(\theta_1, \dots, \theta_p)\} \frac{\partial h_i(\theta_1, \dots, \theta_p)}{\partial \theta_j} = 0; j = 1, \dots, p.$

Maximum Likelihood: The likelihood equations are algebraically $\frac{\partial L(\theta, \mathbf{x})}{\partial \theta_j} = 0; j = 1, \dots, p.$ Alternatively one may write $\frac{\partial \log L(\theta, \mathbf{x})}{\partial \theta_j} = 0; j = 1, \dots, p.$ In the latter form, the estimating equations are better known as the score equations. If the X_i are independent with densities $p_i(x_i, \theta)$, the score equations take on the simple form $\sum_i \frac{\partial}{\partial \theta_j} \log p_i(x_i, \theta); j = 1, \dots, p.$

[A.21] Fisher information: The Fisher information number $I_n(\theta)$ is the variance of the likelihood score function:

$$I_n(\theta) = \text{var}\left(\frac{\partial \log L(\theta, \mathbf{x})}{\partial \theta}\right) = -E\left(\frac{\partial^2 \log L(\theta, \mathbf{x})}{\partial \theta^2}\right).$$

Corollary:

By independence of X_i , $I_n(\theta) = n I_1(\theta)$, i.e. there is n times as much information in a random sample of size n than in a single observation.

If θ is a $(p \times 1)$ vector parameter, the Fisher information is a $(p \times p)$ matrix, called the information matrix and defined as $I_n(\theta) = \frac{\partial^2 \log L(\theta, \mathbf{x})}{\partial \theta \partial \theta'}$.

[A.22] [BD] The likelihood score equation takes on a specifically interesting and simple form, if $p(\mathbf{x}, \theta)$ is in the exponential family of distributions provided that $c(\theta)$ is a 1-1 function. Then $\frac{\partial \log L(\theta, \mathbf{x})}{\partial \theta} = 0 \Rightarrow E(T(\mathbf{X})) = T(\mathbf{x})$. That is, if $E(T(\mathbf{X})) = T(\mathbf{x})$ has a solution $\hat{\theta}$, such that $c(\hat{\theta})$ is an interior point of the range of $c(\theta)$, then $\hat{\theta}$ is the unique MLE.

Quasi-Likelihood [MCN]: The quasi-likelihood for a sample of size n under the conditions listed above is $Q(\mu, \mathbf{y}) = \sum_i Q_i(\mu_i, y_i)$. If $G^{-1}(\mu_i) = \mathbf{x}_i' \beta$, the quasi-likelihood estimating equations for β are $\mathbf{D}' \mathbf{V}^{-1}(\mathbf{y} - \mu) / \sigma^2 = 0$, where $\mathbf{D} = \partial \mu / \partial \beta$.

Convergence and limit theorems

Convergence in general refers to a sequence of random variables $X(\omega)$, denoted $\{X_n(\omega)\}$. If the limit of the sequence as n increases is a point $X(\omega_o)$, we say that $\{X_n(\omega)\}$ converges to $X(\omega_o)$, written as $\lim_{n \rightarrow \infty} X_n(\omega_o) = X(\omega_o)$ or $X_n(\omega_o) \rightarrow X(\omega_o)$. Several different modes of convergence exist:

[A.23] Convergence with probability one: If $X_n(\omega) \rightarrow X(\omega)$ for all $\omega \in A \subset \Omega$ and $\Pr(A) = 1$, i.e. $\Pr\{\omega : X_n(\omega) \rightarrow X(\omega)\} = 1$, we say X_n converges to X with probability one or almost surely. Notation: $X_n(\omega) \xrightarrow{wp1} X(\omega)$

[A.24] Convergence in probability: X_n converges in probability to X , if for every $\epsilon > 0$ $\Pr\{(\omega : |X_n(\omega) - X(\omega)| > \epsilon)\} \rightarrow 0$ as $n \rightarrow \infty$. Notation: $X_n(\omega) \xrightarrow{p} X(\omega)$.

[A.25] Convergence in distribution:

If $\Pr(X_n \leq x) = F_{X_n}(x)$ and $\Pr(X \leq x) = F_X(x)$, then we say that X_n converges in distribution to X , if $\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x) \quad \forall x$ and $\Pr(X = x) = 0$. Notation: $X_n \xrightarrow{\mathcal{L}} X$.

[A.26] $X_n(\omega) \xrightarrow{wp1} X(\omega) \Rightarrow X_n(\omega) \xrightarrow{p} X(\omega) \Rightarrow X_n \xrightarrow{\mathcal{L}} X$.

Consistency and asymptotic unbiasedness

[A.27] Weak Consistency: If $T_n(\mathbf{X}) \xrightarrow{p} q(\theta)$, then $T_n(\mathbf{X})$ is weakly consistent for $q(\theta)$. A necessary and sufficient condition for $T_n(\mathbf{X})$ to be weakly consistent is that $\lim_{n \rightarrow \infty} E(T_n(\mathbf{X}) - q(\theta)) = 0$ and $\text{var}(T_n(\mathbf{X})) \rightarrow 0$.

[A.28] Strong Consistency: If $T_n(\mathbf{X}) \xrightarrow{wp1} q(\theta)$, then $T_n(\mathbf{X})$ is strongly consistent for $q(\theta)$.

[A.29] Weak asymptotic unbiasedness: If $\lim_{n \rightarrow \infty} E(T_n(\mathbf{X})) = q(\theta)$, then $T_n(\mathbf{X})$ is weak asymptotically unbiased for $q(\theta)$.

[A.30] Strong asymptotic unbiasedness: If

$$\frac{E(T_n(\mathbf{X}) - q(\theta))}{\sqrt{\text{var}(T_n(\mathbf{X}))}} \rightarrow 0$$

we say that $T_n(\mathbf{X})$ is strong asymptotically unbiased for $q(\theta)$. Note that this definition of asymptotic unbiasedness is stronger than weak consistency.

[A.31] Weak law of large numbers: If $\{X_i\}$ is a sequence of iid random variables with $E(X_i) = \mu < \infty$, then $\frac{1}{n} \sum_i X_i \xrightarrow{p} \mu$.

[A.32] Strong law of large numbers: If $\{X_i\}$ is a sequence of iid random variables with $E(X_i) = \mu < \infty$, then $\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{wp1} \mu$.

[A.33] Central limit theorem: If $\{X_i\}$ is a sequence of iid random variables with $E(X_i) = \mu$, $\text{var}(X_i) = \sigma^2 < \infty$, then

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{\mathcal{L}} Z, \text{ where } Z \sim N(0, 1)$$

Generalized linear models

[A.34] Components [A, MCN]: Every generalized linear model (GLM) consists of three components:

- (i) *random component*: The observables $\mathbf{Y}$ are assumed to follow a distribution law in the exponential family of distributions (see above).
- (ii) *systematic component*: A set of explanatory variables are linearly related to what is called a linear predictor η through parameters β , i.e. $\eta = \mathbf{X}\beta$.
- (iii) *link component*: To allow for unconstrained estimation of β , one must permit the elements of η to range over the entire real line. Since $E(Y_i)$ is possibly confined to a subset of R^1 , a monotonic, differentiable transform is used to map $G^{-1}(\mu_i) = \mathbf{x}'_i\beta$. $G^{-1}()$ is known as the link function and oftentimes corresponds to the inverse of a cumulative distribution function.

[A.35] Canonical link: Every distribution in the exponential family of distributions has one link function known as canonical. It maps the indexing parameter of the distribution into its natural form (see above).

Example:

Let $X \sim \text{Bernoulli}(\theta) \Rightarrow p(x, \theta) = \exp[\log(\frac{\theta}{1-\theta})x + \log(1-\theta)] I_{A(x)}$.

To put $p(x, \theta)$ in natural form requires to set $\eta = c(\theta)$. Hence the canonical parameter is $\eta = \log(\frac{\theta}{1-\theta})$ and the canonical link is the logit link.

[A.36] Generalized mixed linear model (GMLM): A GMLM is a GLM whose systematic component contains fixed effects *and* random effects or coefficients. That is $\eta_i = G^{-1}(\mu_i) = \mathbf{x}'_i\beta + \mathbf{z}'_i\gamma_i$, where $\gamma_i \sim F(0, \mathbf{D})$. $F(\bullet, \bullet)$ usually is the Gaussian distribution. A GMLM thus can involve three distribution functions: (i) the distribution of the random component, $\mathbf{Y} \sim P_{\text{expo}}$; (ii) the distribution of the random part of the linear predictor; (iii) an inverted cumulative distribution function as link. Although the link (iii) may involve an inverted cdf, it is always part of the non-random model components.

[A.37] Logistic Regression: Let $Y_i \stackrel{id}{\sim} \text{Bernoulli}(\theta_i)$. A regression model for the canonical parameter is $\eta_i = \mathbf{x}_i' \boldsymbol{\beta}$ which is linked using the canonical logit $\Rightarrow G^{-1}(E(Y_i)) = G^{-1}(\Pr(Y_i = 1)) = \text{logit}(E(Y_i)) = \text{logit}(\theta(\mathbf{x}_i)) = \eta_i$. Now $\log\left(\frac{\theta}{1-\theta}\right) = \eta \Rightarrow \theta = \frac{\exp(\eta)}{1+\exp(\eta)}$. The likelihood and log-likelihood is simply

$$L(\boldsymbol{\theta}, \mathbf{y}) = \prod_{i=1}^n \theta(\mathbf{x}_i)^{y_i} (1 - \theta(\mathbf{x}_i))^{1-y_i}$$

$$l(\boldsymbol{\theta}, \mathbf{y}) = \sum_{i=1}^n y_i \log \theta(\mathbf{x}_i) + (1 - y_i) \log (1 - \theta(\mathbf{x}_i))$$

and the score equations become $\sum_{i=1}^n (y_i - \theta(\mathbf{x}_i)) x_{ij} = 0$, $j = 1, \dots, p$, solutions of which yield maximum likelihood estimates. Notice that the score equations are of the form $\mathbf{X}'\mathbf{y} = \mathbf{X}'\boldsymbol{\theta}$, where $\mathbf{X}'\mathbf{y}$ is sufficient for $\boldsymbol{\theta}$ and thus amplify the result above, that in the exponential family with canonical links the score equations are of form $T_j(\mathbf{X}) = E(T_j(\mathbf{X}))$.

Estimating functions

[A.38] Definition: An estimating function (EF) is a real-valued function of the random observables $\mathbf{Y}$ and parameters in a $(p*1)$ vector $\boldsymbol{\theta}$, denoted $g(\mathbf{y}, \boldsymbol{\theta})$.

Corollary: EF are random quantities.

[A.39] Definition: An unbiased EF is an EF for which $E(g(\mathbf{y}, \boldsymbol{\theta})) = 0$. The class of unbiased estimating functions is denoted $\mathfrak{E}$.

[A.40] Definition: A standardized estimating function is defined as

$$g_s(\mathbf{y}, \boldsymbol{\theta}) = \frac{g(\mathbf{y}, \boldsymbol{\theta})}{E\left(\frac{\partial g(\mathbf{y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)}$$

in the case of a scalar parameter and

$$g_s(\mathbf{y}, \boldsymbol{\theta}) = \mathbf{G}(\boldsymbol{\theta})^{-1} g(\mathbf{y}, \boldsymbol{\theta})$$

where $\mathbf{G}(\boldsymbol{\theta}) = \frac{\partial g_i(\mathbf{y}, \boldsymbol{\theta})}{\partial \theta_j}$ is a non-singular $(p*p)$ gradient matrix for a vector parameter.

[A.41] Regularity conditions (Bhaskar 1991): Besides regularity conditions on $p(\mathbf{y}, \theta)$ the following regularities are required for $g(\mathbf{y}, \theta)$:

- (i) $\partial g(\mathbf{y}, \theta) / \partial \theta$ exists and for all $\theta \in \Theta$ and almost all $\mathbf{y}$.
- (ii) Integration and differentiation with respect to θ in $\int g(\mathbf{y}, \theta) p(\mathbf{y}, \theta) d\mathbf{y}$ are interchangeable
- (iii) $\mathbf{G}(\theta)$ is non-singular $\forall \theta \in \Theta$.
- (iv) $E(g(\mathbf{y}, \theta) g(\mathbf{y}, \theta)')$ is positive definite $\forall \theta \in \Theta$.

[A.42] Implied estimates: An EF $g(\mathbf{y}, \theta)$ implies estimates $\hat{\theta}_1, \dots, \hat{\theta}_p$ if regularity condition (iii) holds. They are obtained by solving $g(\mathbf{y}, \theta) = \mathbf{0}$ for θ .

[A.43] Information matrix (Bhaskar 1991): Analogous to Fisher information in likelihood inference the information in an estimating function $g(\mathbf{y}, \theta)$ is defined as

$$\mathbf{I}_g(\theta) = \mathbf{G}(\theta)' E(g(\mathbf{y}, \theta) g(\mathbf{y}, \theta)') \mathbf{G}(\theta)$$

[A.44] Variance (Bhaskar 1991): The variance of a standardized estimating function is $\mathbf{I}_g(\theta)^{-1}$.

[A.45] Optimality: An EF $g^*(\mathbf{y}, \theta)$ is optimal among all EFs in $\mathfrak{S}$, if $\mathbf{I}_{g^*}(\theta) - \mathbf{I}_g(\theta)$ is at least positive semi-definite against all other EFs $g(\mathbf{y}, \theta) \in \mathfrak{S}$. Equivalently, $g^*(\mathbf{y}, \theta)$ is optimal if $\mathbf{I}_g(\theta)^{-1} - \mathbf{I}_{g^*}(\theta)^{-1}$ is at least positive semi-definite.

[A.46] Relation to Fisher information ([B]): If $\mathbf{I}(\theta)$ is Fisher information for θ and $g(\mathbf{y}, \theta)$ is an unbiased EF, then $\mathbf{I}(\theta) - \mathbf{I}_g(\theta)$ is positive semi-definite.

Corollary:

Since $\mathbf{I}(\theta) = E\left(\frac{\partial \log L(\theta, \mathbf{y})}{\partial \theta}\right) \left(\frac{\partial \log L(\theta, \mathbf{y})}{\partial \theta}\right)'$ the likelihood score equation $\frac{\partial \log L(\theta, \mathbf{y})}{\partial \theta}$ is the optimal estimating function in $\mathfrak{S}$ in the sense of retaining maximum information about θ and in the sense of smallest standardized variance.

[A.47] Rao-Blackwellization ([BD], [B]): If $T(\mathbf{X})$ is sufficient for $\theta \in \Theta$, and $g^*(\mathbf{y}, \theta) = E(g(\mathbf{y}, \theta) | T(\mathbf{x}))$, then $\mathbf{I}_{g^*}(\theta) - \mathbf{I}_g(\theta)$ is positive semi-definite.

[A.48] Cramér-Rao Bound (Godambe 1991): Let $g(\mathbf{y}, \theta)$ be an unbiased EF in $\mathfrak{S}$. Then

$$\text{var}(g(\mathbf{y}, \theta)) \geq \frac{\left[E\left\{ \frac{\partial g(\mathbf{y}, \theta)}{\partial \theta} \right\} \right]^2}{I(\theta)}$$

and $\partial \log p(\mathbf{y}, \theta) / \partial \theta$, the score equation achieves the bound on $\text{var}(g)$.
If $g_s(\mathbf{y}, \theta)$ is the standardization of $g(\mathbf{y}, \theta)$ then

$$\text{var}(g_s(\mathbf{y}, \theta)) \geq \frac{1}{I(\theta)} = \frac{1}{\text{var}\left(\frac{\partial \log p(\mathbf{y}, \theta)}{\partial \theta}\right)}.$$

Multinomial distribution

[A.49] Genesis: Suppose the elements of a population each possess one of J mutually exclusive characteristics $A_1, \dots, A_J$. The probability that characteristic A_k occurs is its population frequency p_k , $\sum_{k=1}^J p_k = 1$. If a random sample of m elements is taken, the probability to observe y_k elements with characteristic A_k , $k = 1, \dots, J$ is given by

$$p(\mathbf{y}, m, \mathbf{p}) = \binom{m}{\mathbf{y}} p_1^{y_1} p_2^{y_2} \dots p_J^{y_J}, \quad \sum_{k=1}^J y_k = m$$

$$\text{or } p(\mathbf{y}, m, \mathbf{p}) = \binom{m}{\mathbf{y}} p_1^{y_1} p_2^{y_2} \dots p_{J-1}^{y_{J-1}} (1 - p_1 - \dots - p_{J-1})^{(m - y_1 - \dots - y_{J-1})},$$

where $\binom{m}{\mathbf{y}}$ denotes the number of ways, m elements can be partitioned without replacement into J groups of sizes $y_1, \dots, y_J$, respectively, i.e. $\binom{m}{\mathbf{y}} = \frac{m!}{y_1! y_2! \dots y_J!}$. Notation: We denote $\mathbf{Y} \sim MN(m, \mathbf{p})$.

[A.50] Special cases:

$$(i) J = 2: \quad p(y, m, p) = \frac{m!}{y!(m-y)!} p^y (1-p)^{m-y}$$

i.e., the Multinomial distribution reduces to the Binomial distribution.

$$(ii) m = 1: \quad p(\mathbf{y}, m, \mathbf{p}) = \prod_{k=1}^J p_k^{y_k}; \quad y_k = \begin{cases} 1 & \text{if characteristic } A_k \\ 0 & \text{otherwise} \end{cases}$$

$$\sum_{k=1}^J p_k = \sum_{k=1}^J y_k = 1$$

(iii) $J = 2, m = 1$: $p(y, 1, p) = p^y(1-p)^{1-y}$
i.e., the Multinomial distribution reduces to the Bernoulli(p) distribution.

[A.51] Assume $\mathbf{Y} = [Y_1, \dots, Y_J] \sim MN(m, \mathbf{p})$. Then

(i) each Y_i is marginally a Binomial(m, p_i) random variable.

(ii) (Johnson and Kotz, 1970): Any subset of the $Y_1, \dots, Y_J$ is also a multinomial random variable, i.e. if $S < J$,

$$P(Y_1 = y_1, \dots, Y_S = y_s) = \frac{m!}{\left(m - \sum_{k=1}^S y_k\right)! \prod_{k=1}^S y_k!} \left(1 - \sum_{k=1}^S p_k\right)^{m - \sum_{k=1}^S y_k} \prod_{k=1}^S p_k^{y_k}.$$

(iii) (Johnson and Kotz, 1970): The conditional distribution of $Y_1, \dots, Y_S$ given the other Y 's is multinomial and depends on the other Y 's only through their sum T

$$P(Y_1, \dots, Y_S | T) = \frac{(N - T)!}{\prod_{k=1}^S y_k} \prod_{k=1}^S \left(\frac{p_k}{\sum_{k=1}^S p_k} \right)^{y_k}$$

[A.52] Cumulants and Moments: The cumulant generating function $K_{\mathbf{y}}(s)$ is the log of the moment generating function, hence $K_{\mathbf{y}}(s) = m \log \sum_{k=1}^J p_k \exp(s_k)$.

This yields $E(Y_j) = mp_j$, $\text{cov}(Y_j, Y_{j'}) = \begin{cases} mp_j(1-p_j) & j = j' \\ -mp_j p_{j'} & j \neq j' \end{cases}$.

[A.53] Cumulative totals: In case of ordered responses it is oftentimes useful to work with cumulative categories. Let

$$Z_k = \sum_{j=1}^k y_j \quad Z_J = m$$

$$q_k = \sum_{j=1}^k p_j \quad q_J = 1$$

Then $E(Z_k) = m q_k$, $\text{cov}(Z_k, Z_{k'}) = m q_k(1 - q_{k'}) \quad k \leq k'$.

[A.54] G-inverse: The variance-covariance matrix of a multinomial random variable with J categories is singular, unless one category is removed to account for the linear constraint $\sum_{k=1}^J p_k = 1$. A g-inverse of $\text{var}(\mathbf{Y})$ is given by $\text{var}(\mathbf{Y})^- = \text{Diag}\{1/mp_k\}$, $k = 1, \dots, J$.

[A.55] Likelihood: The likelihood contribution of the i th multinomial random vector in a random sample of $i = 1, \dots, n$ vectors is

$$L(\mathbf{y}_i, \mathbf{p}) = \binom{m_i}{\mathbf{y}_i} \prod_{j=1}^J p_{ij}^{y_{ij}} \quad j = 1, \dots, J$$

The likelihood for the entire sample is

$$L(\mathbf{y}, \mathbf{p}) = \prod_{i=1}^n \binom{m_i}{\mathbf{y}_i} \prod_{j=1}^J p_{ij}^{y_{ij}}$$

with kernel $L^*(\mathbf{y}, \mathbf{p}) = \prod_{i=1}^n \prod_{j=1}^J p_{ij}^{y_{ij}}$ and corresponding log-likelihood kernel

$$l^*(\mathbf{y}, \mathbf{p}) = \sum_{i=1}^n \sum_{j=1}^J y_{ij} \log(p_{ij}).$$

If cumulative responses and cumulative probabilities are used instead, one can write

$$l^*(\mathbf{z}, \mathbf{q}) = \sum_{i=1}^n \sum_{j=1}^J (z_{ij} - z_{i(j-1)}) \log(q_{ij} - q_{i(j-1)}).$$

Subject to the constraint $\sum_j p_{ij} = 1 \forall i$ we find the likelihood equations in terms of p_{ij} as

$$\begin{aligned} l^{**}(\mathbf{y}, \mathbf{p}) &= \sum_i \sum_j y_{ij} \log(p_{ij}) + \lambda \left(\sum_j p_{ij} - 1 \right) \\ \frac{\partial l^{**}(\mathbf{y}, \mathbf{p})}{\partial p_{ij}} &= \frac{y_{ij}}{p_{ij}} = -\lambda & \frac{\partial l^{**}(\mathbf{y}, \mathbf{p})}{\partial \lambda} &= \sum_j p_{ij} = 1 \\ \Rightarrow \frac{\partial l^*(\mathbf{y}, \mathbf{p})}{\partial p_{ij}} &= \frac{y_{ij}}{p_{ij}} - m_i \end{aligned}$$

Similarly for cumulative responses one obtains:

$$\frac{\partial l^*(\mathbf{y}, \mathbf{q})}{\partial q_{ij}} = \frac{y_{ij}}{p_{ij}} - m_i - \left(\frac{y_{i(j-1)}}{p_{i(j-1)}} - m_i \right) = \frac{y_{ij}}{p_{ij}} - \frac{y_{i(j-1)}}{p_{i(j-1)}}$$

[A.56] Exponential family: The $MN(m, p_1, \dots, p_k)$ is in the $k - 1$ parameter exponential family. To see this rewrite

$$P(Y_1 = y_1, \dots, Y_J = y_J) = \binom{m}{\mathbf{y}} \prod_{k=1}^J p_k^{y_k}$$

as

$$\exp \left\{ y_1 \log p_1 + \dots + y_J \log p_J + \log m! - \log \left(\prod_{k=1}^J y_k! \right) \right\}$$

and identify

$$T(\mathbf{Y}) = (Y_1, \dots, Y_{J-1})$$

$$c_i(\mathbf{p}) = \log p_i$$

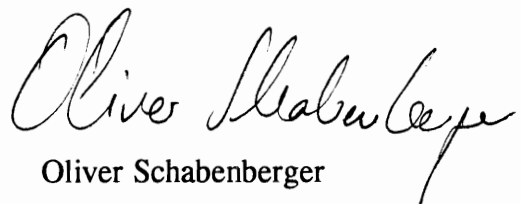
$$d(\mathbf{p}) = \log m!$$

$$S(\mathbf{Y}) = -\log \left(\prod_{k=1}^J y_k! \right) + y_J \log p_J.$$

Note that the $MN(m, p_1, \dots, p_k)$ is only a $(k - 1)$ parameter member of the exponential family, since it permits only a $J - 1$ dimensional sufficient statistic. For this reason $y_J \log p_J$ is absorbed into $S(\mathbf{Y})$.

Vita

The author was born on March 31, 1965 in Pforzheim, Germany as second son of Kurt and Rosemarie Schabenberger. After graduating from Gymnasium in 1983, he was accepted in the Forest Engineering program at the Fachhochschule für Forstwirtschaft, Rottenburg a.N. (Germany). Upon completion of the Dipl.-Ing. (FH) degree in 1987 he enrolled into the Forest Science program at the University of Freiburg, Germany. After four years of study, he was conferred the academic title Diplom-Forstwirt with honors in 1991. Between 1988 and 1992 he was employed by the Department of Forest Biometrics in Freiburg as a research assistant to Prof. Dieter R. Pelz. In the fall of 1992 he commenced graduate studies at Virginia Polytechnic Institute and State University working towards a Ph.D. degree in the Section Forest Biometrics of the Department of Forestry under the guidance of Dr. Timothy G. Gregoire. Concurrently he acquired a Masters of Statistics degree from the same university in December 1994. As of this writing the author is Graduate Research Assistant in the Department of Forestry at Virginia Polytechnic Institute.



Oliver Schabenberger