AN ABSTRACT OF THE DISSERTATION OF

Yonghai Li for the degree of Doctor of Philosophy in Statistics presented on March 2, 2007.
Title: Likelihood Analysis of the Multivariate Ordinal Probit Model for Repeated and Spatial Ordered Categorical Responses

Abstract approved: _____
Daniel W. Schafer

This dissertation is about the likelihood analysis of ordered categorical responses in a longitudinal/spatial study, meaning regression-like analysis when the response variable is categorical with ordered categories, and is measured repeatedly over time or space on the experimental or sampling units. Particular attention is given to the multivariate ordinal probit regression model, in which the correlation between ordered categorical responses on the same unit at different times or locations is modeled with a latent variable that has a multivariate normal distribution. An algorithm for maximum likelihood analysis of this model is proposed and the analysis is demonstrated on several examples. Simulations show that the maximum likelihood estimates can be substantially more efficient than generalized estimating equations (GEE) estimates of regression coefficients. We also propose likelihood analysis of a regression model for spatial-temporal ordered categorical data, and with particular attention to an investigation of determinants of Coho salmon densities in Oregon. This approach avoids defining a neighborhood for each site, which is an awkward step that is required for existing approaches.

©Copyright by Yonghai Li
March 2, 2007
All Rights Reserved

Likelihood Analysis of the Multivariate Ordinal Probit Model for Repeated and Spatial Ordered Categorical Responses

by
Yonghai Li

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented March 2, 2007
Commencement June 2007

Doctor of Philosophy dissertation of Yonghai Li presented on March 2, 2007.

APPROVED:

Major Professor, representing Statistics

Chair of the Department of Statistics

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Yonghai Li, Author

ACKNOWLEDGEMENTS

I would like to express my deep appreciation and sincere thanks to my advisor, Dr. Dan Schafer, for his patience, support, and guidance during the course of this dissertation research. I have benefited greatly from his creative ideas and perspectives. I am also thankful to Dr. Darius Adams, Dr. Alix Gitelman, Dr. Annie Qu, and Dr. Bob Smythe for serving as my committee members.

I wish to thank the entire faculty, staff and my classmates of the Statistics Department at Oregon State University for their contribution to my education and research. Particularly, I am indebted to: 1) Dr. Lisa Madsen for her providing the coho study data and her active involvement in my research in the spatial part of my dissertation; 2) Dr. Annie Qu for her constructive comments on an earlier draft of Chapter two of this dissertation; 3) Dr. Dave Birkes for his generous helps in my statistics training and research. In the meantime, I would like to thank my friend and classmate, Waseem Alnosaier, who helped me get research papers from the Valley Library when I worked at Washington Mutual Bank in Seattle, Washington.

I could not have completed this degree without the love, faith, and support of my family. I thank my lovely wife, Shuzhen Nong, for her devoted love and companionship. Special thanks go to my parents and my brother, who have never lost faith in me.

CONTRIBUTION OF AUTHORS

Dr. Lisa Madsen assisted with the coho study data collection. She is one of the co-authors of Chapter 3.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF APPENDIX TABLES

Likelihood Analysis of the Multivariate Ordinal Probit Model for Repeated and Spatial Ordered Categorical Responses

1. INTRODUCTION

This dissertation is about the likelihood analysis of the multivariate ordinal probit model for ordered categorical responses in a longitudinal study or a spatial study, meaning regression-like analysis when the response variable is categorical with ordered categories, and is measured repeatedly over time or space on the experimental or sampling units. This is an important data structure in medical studies, for example, when patients receiving different treatments and with different covariate values are categorized according to ordered grades of health status or improvement at multiple points in time. The following two examples indicate the types of data problems we have in mind.

Example 1: a Randomized Experiment on Anesthesia Recovery

In a longitudinal study that compared the effects of varying dosages of an anesthetic on post-surgical recovery (Davis, 1991), 60 young children undergoing outpatient surgery were randomized to one of four dosages (15, 20, 25 and 30 mg/kg) of the anesthesia, with 15 children per dose group. Recovery scores on a seven-point scale (0: least favorable; 6: most favorable) were assigned upon admission to the recovery room and at minutes 5, 15 and 30 following admission. In addition to the dosage, other potential covariates were (a) time when the measurement was taken, (b)

age of the patient (in months), and (c) duration of the surgery (in minutes). Researchers wished to study the profile of the categorical response over time and how the response is associated with dose and other covariates. Generalized estimating equations methods are available for this structure of repeated measures with an ordered categorical response. Likelihood analysis, which had not previously been available for such an analysis because of computational obstacles, is a possible improvement due to greater efficiency and use of likelihood ratio inference tools.

Example 2: an Observational Study on Coho Density

With a spatial-temporal study, researchers wished to identify key habitat factors associated with the abundance and distribution of wild and hatchery coho salmon in streams. (http://oregonstate.edu/Dept/ODFW/spawn/pdf%20files/reports/05SSManual.pdf; http://oregonstate.edu/Dept/ODFW/freshwater/inventory/index.htm). Coho densities and habitat covariates data were collected from 206 distinct sites from 1998 to 2004. The actual density was estimated at each site, in units of coho salmon per linear stream mile. The density was categorized by fish biologists on a three-point scale (0 *absent*: density less than 4; 1 *few or some*: density greater than or equal to 4 and less than 40; 2 *full*: density greater than or equal to 40). The habit covariates included gradient, active channel width/depth, percent of pools in the reach, etc.

In this case, a complicated spatial-temporal correlation structure can be modeled with a multivariate normal latent variable. There is no need that such a latent variable has any real meaning, nor is the assumption as strong as it might at first seem.

The model permits a correlation structure that is more realistic than alternative models. Likelihood analysis offers improved efficiency and the use of likelihood-based fitting criteria, such as AIC and BIC, for covariate selection.

1.1 The Multivariate Ordinal Probit Model for Temporal/Spatial Ordinal Data

The primary model for this dissertation is the multivariate ordinal probit model (Fu *et al.*, 2000) or the (multivariate) grouped continuous model (Anderson and Pemberton, 1985). The model can be derived from a latent variable approach. In this approach, each ordinal response y is derived from a continuous latent variable z using a threshold concept. Specifically, the ordinal variable y is thought of as providing incomplete information about the underlying z according to the measurement equation:

$$y = g \qquad \text{if } \alpha_{g-1} < z \le \alpha_g \text{ for } g = 1 \text{ to } G$$

where G is the number of categories of ordinal responses and α's are called *thresholds* or *cutpoints*. The extreme categories 1 and G are defined by open-ended intervals with $\alpha_0 = -\infty$ and $\alpha_G = \infty$. The joint distribution of latent variables z's is assumed to be multivariate normal (MVN). Consequently, we will account for the temporal/spatial correlation of ordinal responses by modeling the variance-covariance matrix of the MVN distribution of the latent variables.

The latent (variable) distribution is similar to the notion of tolerance distribution in quantal bioassay, but we don't require that such a latent variable exists. If there is no real latent variable, it may also be convenient to think of a fictitious latent variable as a "propensity to respond" or "degree of response". In example 1, for

example, a conclusion can be made about the effect of explanatory variables on the degree of recovery from anesthesia.

The multivariate normal latent variable might not be as strong of an assumption as it first seems, though. For one thing, if there really is a latent variable (as there would be for categorizing the coho densities in example 2, for example), it is only necessary that some monotonic function of the latent variable is normally distributed. Furthermore, the actual distribution of the latent variable might not matter much if the extreme response categories are not too strongly tied to the extreme tails of the latent variable distribution.

1.2 Contributions of the Dissertation

While the multivariate ordinal probit model is attractive for modeling the ordinal responses from longitudinal/spatial studies, its maximum likelihood (ML) analysis has been slow to evolve because of computational difficulties in finding the MLE of parameters in the model. As we will show, the direct ML approach requires an evaluation of integrals of multivariate normal density functions. Until recently, this integration was impractical, especially if it involved more than two dimensions. Consequently, most previous applications of this model are limited to bivariate ordinal probit models (e.g., Kim, 1995).

In this dissertation, we propose an algorithm for maximum likelihood analysis of the multivariate ordinal probit model, which incorporates a numerical integration procedure within a numerical maximization procedure. We believe this algorithm

avoids some of the practical problems of previous methods. It can make use of existing routines for optimization and for multivariate normal probabilities, available, for example, in the software package R. The availability of this routine will permit studies of efficiency, accuracy of tests and confidence intervals, and robustness, which will help clarify the relative merits of likelihood analysis and GEE. For example, in a longitudinal study, our simulations show that the maximum likelihood estimates can have substantially smaller variances than generalized estimating equations (GEE) estimates of regression coefficients.

A second contribution of this dissertation is a fully-likelihood analysis of a spatial-temporal ordinal data set of importance in environmental wildlife management. Using the latent error induced dependency among spatial-temporal ordinal responses, our approach avoids defining neighborhood for each site, which sometimes is not clear but is required for some existing approaches (e.g., the Markov Random Field), particularly on an irregular lattice. More importantly, a fully likelihood analysis can be conducted in our approach without resorting to a composite (or pseudo) likelihood or GEE, which can be less efficient in parameter estimation. Moreover, the familiar likelihood-based methods for testing fit, comparing models (with AIC and BIC, for example), making inference about parameters are available with our approach.

1.3 Organization of the Dissertation

The rest of this dissertation proceeds as follows. In Chapter 2, particular attention is given to the multivariate ordinal probit regression model for a longitudinal study, in

which the correlation between ordered categorical responses on the same unit at different times is modeled with a latent variable that has a multivariate normal distribution. An algorithm for maximum likelihood analysis of this model is proposed and the analysis is demonstrated on the anesthesia recovery example. In addition, Chapter 2 documents a simulation study comparing likelihood analysis to generalized estimating equations. Chapter 3 considers a similar model but in a spatial-temporal setting where the ordered categorical data are recorded over space and time. Likelihood analysis of the multivariate ordinal probit model for this type of data structure is provided. The analysis is demonstrated on the coho study data to identify habitat variables associated with coho density. The dissertation ends with a brief discussion of our conclusions and possible directions of future research in Chapter 4.

2. Likelihood Analysis of the Multivariate Ordinal Probit Regression Model for Repeated Ordinal Responses

Yonghai Li* and Daniel W. Schafer**

Department of Statistics, Oregon State University, Corvallis, OR, 97331, U.S.A.
*email: yonghai@science.oregonstate.edu
**email: schafer@science.oregonstate.edu

2.1 Abstract

This paper is about the analysis of longitudinal ordinal data, meaning regression-like analysis when the response variable is categorical with ordered categories, and is measured repeatedly over time (or space) on the experimental or sampling units. Particular attention is given to the multivariate ordinal probit regression model, in which the correlation between ordered categorical responses on the same unit at different times (or locations) is modeled with a latent variable that has a multivariate normal distribution. An algorithm for maximum likelihood analysis of this model is proposed and the analysis is demonstrated on an example. Simulations show that the maximum likelihood estimates can be substantially more efficient than generalized estimating equations (GEE) estimates of regression coefficients.

2.2 Introduction

This paper is about the analysis of longitudinal ordinal data, meaning regression-like analysis when the response variable is categorical with ordered categories, and is measured repeatedly over time (or space) on the experimental or sampling units. This is an important data structure in medical studies, for example, when patients receiving different treatments and with different covariate values are categorized according to ordered grades of health status or improvement at multiple points in time. The following two examples indicate the types of data problems we have in mind.

Example 1: a Randomized Experiment on Anesthesia Recovery

In a longitudinal study that compared the effects of varying dosages of an anesthetic on post-surgical recovery (Davis, 1991), 60 young children undergoing outpatient surgery were randomized to one of four dosages (15, 20, 25 and 30 mg/kg) of the anesthesia, with 15 children per dose group. Recovery scores on a seven-point scale (0: least favorable; 6: most favorable) were assigned upon admission to the recovery room and at minutes 5, 15 and 30 following admission. In addition to the dosage, other potential covariates were (a) time when the measurement was taken, (b) age of the patient (in months), and (c) duration of the surgery (in minutes).

Example 2: an Observational Study on Marijuana Use

The National Youth Survey (Elliot, Huizinga, and Menard, 1989; Lang, McDonald, and Smith, 1999) collected five annual waves (1976-80) data on 'marijuana use in the past year' from the 237 respondents who were 13 years old in 1976. The data is on a trichotomous ordinal scale (1, never; 2, not more than once a

month; 3, more than once a month). One of the objectives of the study is to model the probability of marijuana use status over time as a function of gender and time.

Although there are a variety of models and approaches (see Liu and Agresti, 2005), we believe the currently most useful tool for this type of problem is generalized estimating equations (GEE). This uses a generalized linear model for relating the response means to the explanatory variables, and employs a working correlation structure to account for the non-independence of multiple responses on the same subject. The treatment of correlation is not thought to be realistic with this approach, but sufficient for obtaining estimates of the regression parameters of interest. GEE algorithms for regression models for ordered categorical responses are available (e.g., *ordgee* in package *geepack* from R, *geeDesign* and *gee.fit* in the correlatedData library from S-PLUS, *proc genmod* with the independent working correlation from SAS). An important consideration is that these are easy to use (at least, relative to the alternatives) because of their similarity with more familiar methods for independent responses.

While there is a potentially useful full parametric marginal model for this structure—the multivariate ordinal probit model (Fu *et al.*, 2000) or the (multivariate) grouped continuous model (Anderson and Pemberton, 1985) — maximum likelihood (ML) analysis has been slow to evolve because of computational difficulties in finding the MLE of parameters in the model. As we will show, the direct ML approach requires an evaluation of integrals of multivariate normal density functions. Until recently, this integration was impractical, especially if it involved more than two

dimensions. Consequently, most previous applications of this model are limited to bivariate ordinal probit models (e.g., Kim, 1995). McFadden (1989) and Hajivassiliou, McFadden, and Ruud (1996) used Monte Carlo techniques for the integral evaluation, but many researchers feel this approach is too computer intensive. Fu *et al.* (2000) proposed a *limited information estimator* for approximate likelihood analysis.

In this paper, we propose an algorithm for maximum likelihood analysis of the multivariate ordinal probit model, which incorporates a numerical integration procedure within a numerical maximization procedure. We believe this algorithm avoids some of the practical problems of previous methods. It can make use of existing routines for optimization and for multivariate normal probabilities, available, for example, in the software package R. The availability of this routine will permit studies of efficiency, accuracy of tests and confidence intervals, and robustness, which will help clarify the relative merits of likelihood analysis and GEE.

The rest of this paper is organized as follows. Section 2.3 reviews GEE approaches for longitudinal ordinal responses. Section 2.4 introduces the multivariate ordinal probit model and an algorithm for maximum likelihood analysis. Section 2.5 shows an application to the anesthesia recovery example. Section 2.6 documents a simulation study comparing likelihood analysis to generalized estimating equations, followed by a discussion in Section 2.7.

2.3 GEE for Longitudinal Ordinal Data

A generalized estimating equations (GEE) approach for longitudinal ordinal data is a multivariate generalization of quasi-likelihood. The original GEE methodology was proposed by Liang and Zeger (1986) for marginal models with univariate distributions such as binomial and Poisson generalized linear models. It was extended to multinomial responses using the cumulative logit link and the cumulative probit link functions for longitudinal ordinal responses in the mid 1990's. Cumulative logit models have been studied by Kenward, Lesaffre, and Molenberghs (1994); Lipsitz, Kim, and Zhao (1994); Lumley (1996); Mark and Gail (1994); Qu, Piedmonte, and Medendrop (1995); and Williamson, Kim, and Lipsitz (1995). Cumulative probit models can be found in Qu *et al.* (1995) and Toledano and Gatsonis (1996). GEE inference for these models is based on a generalized linear model specification of the first two marginal moments of a subject's response, and a "working correlation" structure to account for dependencies of responses from the same subject at different occasions.

Suppose, in a longitudinal study, there are T occasions of measurement. Let $\underset{\sim}{y}_{it}$ be a G-1 vector to represent an ordinal response variable with G categories and let $\underset{\sim}{x}_{it}$ be a p-dimensional explanatory variable vector observed on subject i (i = 1, …, n) at time t ($t = 1,...,T_i \le T$). Specifically, $\underset{\sim}{y}_{it} = (y_{it1}, y_{it2}, ..., y_{i,t,G-1})'$, where $y_{itg} = 1$ if subject i falls into response category g at time t, and $y_{itg} = 0$ otherwise, for $g = 1,...,G$. The responses for different subjects are independent, but the responses at different

occasions for a given subject are assumed to be correlated. Let

$\pi_{itg} = \pi_{itg}(\underset{\sim}{\theta}) = E(y_{itg} \mid \underset{\sim}{x}_{it}, \underset{\sim}{\theta}) = \Pr(y_{itg} = 1 \mid \underset{\sim}{x}_{it}, \underset{\sim}{\theta})$ denote the probability of response *g* at time *t,* where $\underset{\sim}{\theta}$ is a parameter vector. In addition, let $\underset{\sim}{\pi}_{it} = (\pi_{it1}, ..., \pi_{i,t,G-1})'$, $\underset{\sim}{\pi}_i = (\underset{\sim}{\pi}_{i1}^{'}, ..., \underset{\sim}{\pi}_{iT_i}^{'})'$, $\underset{\sim}{y}_i = (\underset{\sim}{y}_{i1}^{'}, ..., \underset{\sim}{y}_{iT_i}^{'})'$, $x_i = (\underset{\sim}{x}_{i1}, ..., \underset{\sim}{x}_{iT_i})'$, and $\underset{\sim}{\mu}_i \equiv E(\underset{\sim}{y}_i \mid x_i) = \underset{\sim}{\pi}_i$. The main models for π_{itg} (for $g \le G-1$) are the cumulative logit and the cumulative probit, obtained by letting $\log\left[\dfrac{\pi_{it1} + ... + \pi_{itg}}{1-(\pi_{it1} + ... + \pi_{itg})}\right] = \alpha_g - \underset{\sim}{x}_{it}^{'}\underset{\sim}{\beta}$ and $\Phi^{-1}(\pi_{it1} + ... + \pi_{itg}) = \alpha_g - \underset{\sim}{x}_{it}^{'}\underset{\sim}{\beta}$, respectively, where α_g and $\underset{\sim}{\beta}$ are unknown parameters, and $\Phi^{-1}(\cdot)$ denotes the inverse of the standard normal cumulative distribution function.

The following description of GEE follows that given by Lipsitz *et al.* (1994). The GEE for estimating $\underset{\sim}{\theta}$ takes the form

$$U(\underset{\sim}{\theta}) = \sum_{i=1}^{n}\left(\frac{\partial \underset{\sim}{\mu}_i}{\partial \underset{\sim}{\theta}}\right)^{'} V_i^{-1}(\underset{\sim}{y}_i - \underset{\sim}{\mu}_i) = \underset{\sim}{0}, \qquad (2.1)$$

where V_i, the working covariance matrix, is a function of $\underset{\sim}{\theta}$ and other 'nuisance' parameters, $\underset{\sim}{\gamma}$, associated with the working correlation. Let

$A_{it} = diag[\pi_{it1}(1-\pi_{it1}), ..., \pi_{i,t,G-1}(1-\pi_{i,t,G-1})]$, $A_i = diag[A_{i1}, ..., A_{iT_i}]$,

$V_{it} \equiv \text{var}(\underset{\sim}{y}_{it}) = diag[\underset{\sim}{\pi}_{it}] - \underset{\sim}{\pi}_{it}\underset{\sim}{\pi}_{it}^{'}$. Then, $R_i(\underset{\sim}{\gamma})$, the working correlation matrix, can be expressed as

$$\begin{bmatrix} A_{i1}^{-1/2}V_{i1}A_{i1}^{-1/2} & \rho_{i12} & \cdots & \rho_{i1T_i} \\ \rho_{i21} & A_{i2}^{-1/2}V_{i2}A_{i2}^{-1/2} & \cdots & \rho_{i2T_i} \\ \vdots & \vdots & & \vdots \\ \rho_{iT_i1} & \rho_{iT_i2} & \cdots & A_{iT_i}^{-1/2}V_{iT_i}A_{iT_i}^{-1/2} \end{bmatrix},$$

where $\rho_{it_1t_2}$ for any $t_1 \neq t_2$ is a function of the parameter vector $\underset{\sim}{\gamma}$ (see more details in Lipsitz *et al.*, 1994); and the working variance-covariance matrix V_i has the form $A_i^{1/2}R_i(\gamma)A_i^{1/2}$.

The GEE estimator, $\hat{\underset{\sim}{\theta}}$, is the solution to the generalized estimating equation (2.1) above. A robust estimate of the variance-covariance matrix in the sampling distribution of $\hat{\underset{\sim}{\theta}}$ is

$$Var(\hat{\underset{\sim}{\theta}}) = \left[\sum_{i=1}^{n} D_i' V_i^{-1} D_i\right]^{-1} \left[\sum_{i=1}^{n} D_i' V_i^{-1} \text{var}(\underset{\sim}{y}_i) V_i^{-1} D_i\right] \left[\sum_{i=1}^{n} D_i' V_i^{-1} D_i\right]^{-1} \text{ where } D_i = \frac{\partial \underset{\sim}{\mu}_i}{\partial \underset{\sim}{\theta}}.$$

The variance is estimated by substituting $\hat{\underset{\sim}{\pi}}_i$ from the model fit and replacing $\text{var}(\underset{\sim}{y}_i)$ by an empirical variance of $\underset{\sim}{y}_i$.

The GEE approach requires a working correlation structure, but its estimates are thought to be useful even if this structure is misspecified. The parameter estimates from the GEE are consistent as long as the mean function is correctly specified (Lipsitz *et al.*, 1994).

While the GEE approach is appealing because of the relative ease with which it can be used (especially given that the data structure is a rather complicated one), it has these potential drawbacks: (1) the robust standard errors tend to underestimate the true ones unless the sample size is quite large; (2) while the robust covariance matrix

estimate is consistent under certain conditions when the quasilikelihood model is correct, the estimate is often far more variable than the usual parametric variance estimate; (3) the familiar likelihood-based methods for testing fit, comparing models (with AIC and BIC, for example), and making inference about parameters are not available with this approach (Agresti, 2002, p. 468).

While ordinary GEE models (GEE1) regard the correlation structure within clusters as a nuisance, Heagerty and Zeger (1996) developed a GEE2 approach where the correlation structure in terms of global odds ratios is itself of interest and a set of estimating equations for the covariance parameter is built. The estimates of parameters from this approach, however, are no longer consistent if one misspecifies the model for the correlation (Agresti and Natarajan, 2001).

2.4 Likelihood Analysis for the Multivariate Ordinal Probit Model

2.4.1 The Multivariate Ordinal Probit Model

Consider the model described in Section 2.3. Let $T_1 = T_2 = ... = T_n = T$ for convenience. Suppose the ordinal response $\underset{\sim}{y}_{it}$ is generated from a latent (tolerance) variable z_{it} through a threshold concept. Specifically,

$y_{itg} = 1$ (g = 1, 2, …, G) if and only if $\alpha_{g-1} < z_{it} \le \alpha_g$,

where $-\infty = \alpha_0 < \alpha_1 < \alpha_2 < ... < \alpha_{G-1} < \alpha_G = \infty$ are the thresholds for the continuous latent variable z_{it}. Assume $z_{it} \sim N(\underset{\sim}{x}_{it}^{'}\underset{\sim}{\beta}, 1)$ and $\underset{\sim}{z}_i \equiv (z_{i1}, ..., z_{iT})' \sim N_T(x_i\underset{\sim}{\beta}, \Sigma)$, where Σ is a positive definite variance-covariance matrix, which is assumed to be the same

for all subjects. It is important to note that the variance of z_{it} has been set to 1 for identifiability. Hence, Σ is also a correlation matrix. We will account for the correlation of ordinal responses of a given individual at different time points by modeling the within-subject correlation (Σ) of the latent variable as a function of a vector or scalar parameter, $\underset{\sim}{\rho}$. We make this dependence explicit with the symbol $\Sigma_{\underset{\sim}{\rho}}$.

Note in this setting that the probability of a response in ordered category *g* or less can be written as

$$\Pr\{\max(y_{it1},...,y_{itg})=1\}=\Pr(z_{it}\le\alpha_g)=\Phi(\alpha_g-\underset{\sim}{x}_{it}'\underset{\sim}{\beta}),$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

The latent (variable) distribution is similar to the notion of tolerance distribution in quantal bioassay, but we don't require that such a latent variable exists. It is an abstraction, which can be used to motivate and use the multivariate probit model to investigate the relationship between the ordinal response and the covariates (see, for example, McCullagh, 1980, and Kim, 1995). If there is no real latent variable, Long (1997, p.127) suggested (1) predicted probabilities of the observed outcomes be presented in tables or plots; (2) partial (for continuous covariates) and discrete (for factor covariates) change in probabilities be examined. It may also be convenient to think of a fictitious latent variable as a "propensity to respond" or "degree of response." In example 1, for instance, a conclusion can be made about the effect of explanatory variables on the degree of recovery from anesthesia.

The observed data log likelihood is

$$l(\underset{\sim}{\alpha},\underset{\sim}{\beta},\underset{\sim}{\rho}) = \sum_{i=1}^{n} \log \int_{a_{i1}}^{b_{i1}} \int_{a_{i2}}^{b_{i2}} \cdots \int_{a_{iT}}^{b_{iT}} \varphi_T(\underset{\sim}{z}_i; x_i\underset{\sim}{\beta}, \Sigma_{\underset{\sim}{\rho}}) dz_{i1} dz_{i2} \cdots dz_{iT} \qquad (2.2)$$

where $\underset{\sim}{\alpha} = (\alpha_1, ..., \alpha_{G-1})'$, and $\varphi(\cdot)$ is a multivariate normal density,

$$\varphi_T(\underset{\sim}{z}_i; x_i\underset{\sim}{\beta}, \Sigma_{\underset{\sim}{\rho}}) = (2\pi)^{-T/2} \mid \Sigma_{\underset{\sim}{\rho}} \mid^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\underset{\sim}{z}_i - x_i\underset{\sim}{\beta})' \Sigma_{\underset{\sim}{\rho}}^{-1} (\underset{\sim}{z}_i - x_i\underset{\sim}{\beta}) \right), \quad (2.3)$$

and

$a_{it} = \alpha_{g-1}$, $b_{it} = \alpha_g$ if and only if $y_{itg} = 1$, for $i = 1,...,n$ and $t = 1,...,T$.

The maximum likelihood estimates of parameters ($\underset{\sim}{\alpha}, \underset{\sim}{\beta}, \underset{\sim}{\rho}$) can be obtained by maximizing the log likelihood function in equation (2.2).

2.4.2 An Algorithm for Computing the MLE

We express the log likelihood function as

$$l(\underset{\sim}{\alpha},\underset{\sim}{\beta},\underset{\sim}{\rho}) = \sum_{i=1}^{n} \log \hat{P}_i \{\underset{\sim}{a}_i(\underset{\sim}{\alpha}, \underset{\sim}{y}_i), \underset{\sim}{b}_i(\underset{\sim}{\alpha}, \underset{\sim}{y}_i), \underset{\sim}{\beta}, \underset{\sim}{\rho}\} \qquad (2.4)$$

where $\underset{\sim}{a}_i = (a_{i1}, ..., a_{iT})'$, $\underset{\sim}{b}_i = (b_{i1}, ..., b_{iT})'$, and $\hat{P}_i$ is a numerical approximation to the integral $\int_{a_{i1}}^{b_{i1}} \int_{a_{i2}}^{b_{i2}} \cdots \int_{a_{iT}}^{b_{iT}} \varphi_T(\underset{\sim}{z}_i; x_i\underset{\sim}{\beta}, \Sigma_{\underset{\sim}{\rho}}) dz_{i1} dz_{i2} \cdots dz_{iT}$. This is, of course, an approximation to the probability that a multivariate normal random variable falls in a rectangular region. There are numerous routines available (such as the Genz method; see Genz 1992, 1993). In order to satisfy the constraint $\alpha_1 < \alpha_2 < \cdots < \alpha_{G-1}$, we reparameterize α_g (g = 2, …, G-1) in terms of α_1, d_1, …, and d_{G-2} as follows.

$$\alpha_2 = \alpha_1 + d_1^2,\ \alpha_3 = \alpha_1 + d_1^2 + d_2^2,\ \ldots,\ \alpha_{G-1} = \alpha_1 + \sum_{g=1}^{G-2} d_g^2,$$

where d_1, d_2, …, and d_{G-2} are non-zero.

Therefore, the log likelihood in equation (2.4) can be written as $l(\alpha_1, d_1, ..., d_{G-2}, \underset{\sim}{\beta}, \underset{\sim}{\rho})$. We use a *stepwise ascent* method (see, Gelman *et al.*, 2004, p. 312) to find the parameter values that maximize the log likelihood. This iteratively maximizes with respect to one parameter at a time, with others fixed at their current estimates. To find global maximizers, it may be necessary to run the stepwise ascent routine starting at different initial parameter values spread throughout the parameter space.

An algorithm for the MLE

Step 1:

(a) Obtain initial estimates $\hat{\alpha}_1^{(1)}, \hat{d}_1^{(1)}, ..., \hat{d}_{G-2}^{(1)}$ and $\underset{\sim}{\hat{\beta}}^{(1)}$ by maximizing the log likelihood with respect to $\alpha_1, d_1, ..., d_{G-2}$, and $\underset{\sim}{\beta}$, assuming an independence correlation structure ($\Sigma_{\underset{\sim}{\rho}} = I_T$, a $T \times T$ identity matrix).

(b) Obtain $\underset{\sim}{\hat{\rho}}^{(1)}$ by maximizing the log likelihood with respect to $\underset{\sim}{\rho}$, with $\alpha_1, d_1, ..., d_{G-2}$ and $\underset{\sim}{\beta}$ fixed at $\hat{\alpha}_1^{(1)}, \hat{d}_1^{(1)}, ..., \hat{d}_{G-2}^{(1)}$ and $\underset{\sim}{\hat{\beta}}^{(1)}$.

Step k ($k = 2, 3, \ldots$):

(a) Obtain $\hat{\alpha}_1^{(k)}, \hat{d}_1^{(k)}, ..., \hat{d}_{G-2}^{(k)}$ and $\underset{\sim}{\hat{\beta}}^{(k)}$ by maximizing the log likelihood with respect to $\alpha_1, d_1, ..., d_{G-2}$, and $\underset{\sim}{\beta}$, with $\underset{\sim}{\rho}$ fixed at $\underset{\sim}{\hat{\rho}}^{(k-1)}$.

(b) Obtain $\underset{\sim}{\rho}^{(k)}$ by maximizing the log likelihood with respect to $\underset{\sim}{\rho}$, leaving

$\alpha_1, d_1, ..., d_{G-2}$ and $\underset{\sim}{\beta}$ fixed at $\hat{\alpha}_1^{(k)}, \hat{d}_1^{(k)}, ..., \hat{d}_{G-2}^{(k)}$ and $\hat{\underset{\sim}{\beta}}^{(k)}$.

Repeat the iteration process until convergence.

Note that step 1(a) can be achieved by fitting an ordered probit regression model to the nT observations treated as independent. Many statistical software packages now include this routine for ordered probit regression (PROC GENMOD with *link* = cprobit in SAS; *polr* with *method* = "probit" in S-PLUS or R).

As for step 1(b), we will conditionally maximize the log likelihood with respect to $\underset{\sim}{\rho}$. If $\underset{\sim}{\rho}$ is a vector, we can sequentially maximize its components. If $\underset{\sim}{\rho}$ is a scalar ρ, recall $-1 \le \rho \le 1$, so we are required to search in the interval [-1, 1] for the conditional maximizer of the log likelihood. An efficient approach uses a combination of golden section search and successive parabolic interpolation (see, Brent, 1973), which can be accomplished with the function *optimize* in R, for example.

Step *k*(a) can be implemented with a Newton-type function (using *nlm* in R, for example). See Dennis and Schnabel (1983) and Schnabel, Koontz, and Weiss (1985) for details. Step *k*(b) can be implemented using the same method as that in step 1(b).

Computing the asymptotic standard errors

Let $\underset{\sim}{\theta} \equiv (\alpha_1, d_1, ..., d_{G-2}, \underset{\sim}{\beta}', \underset{\sim}{\rho}')'$ and $\hat{\underset{\sim}{\theta}}$ be the parameter vector and its MLE, respectively. Theoretically, one can compute the asymptotic variance of $\hat{\underset{\sim}{\theta}}$ by

inverting the observed information matrix, $I(\underset{\sim}{\theta})$, or the expected information matrix, $\mathrm{I}(\underset{\sim}{\theta})$, evaluated at the MLE $\hat{\underset{\sim}{\theta}}$. If we use the latter, the standard errors of the MLE are equal to the square roots of the diagonal entries of $\mathrm{I}^{-1}(\underset{\sim}{\theta})$, where

$\mathrm{I}(\hat{\underset{\sim}{\theta}}) = -E_{\hat{\underset{\sim}{\theta}}}(\ddot{l}) = E_{\hat{\underset{\sim}{\theta}}}(\dot{l}\dot{l}')$, $\dot{l} = \dfrac{\partial l}{\partial \underset{\sim}{\theta}}$, and $\ddot{l} = \dfrac{\partial^2 l}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'}$. $E_{\hat{\underset{\sim}{\theta}}}(\dot{l}\dot{l}')$ can be approximated using a Monte Carlo method.

If we use the observed information to calculate the asymptotic variance of the MLE, we have $I^{-1}(\hat{\underset{\sim}{\theta}}) = \left[-\dfrac{\partial^2 l}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'} \right]^{-1} \Bigg|_{\underset{\sim}{\theta}=\hat{\underset{\sim}{\theta}}}$ as the covariance matrix estimator, which can also be approximated using numerical derivatives. Efron and Hinkley (1978) show the inverse of the observed information is superior to that of the expected information as an approximation to the true variance of the MLE in one-parameter families. For multi-parameter families and other discussions, readers can refer to Skovgaard (1985), Pace and Salvan (1997, p. 92-93), and Lindsay and Li (1997). In the anesthesia recovery example below, whenever we experienced computational difficulties of this second method, we resorted to the first method outlined above (i.e., we used the expected information rather than the observed information).

2.4.3 LR Test and LR Confidence Interval

Unlike the GEE, the likelihood approach permits the likelihood ratio (LR) test of parameters and a LR confidence interval through inversion of the LR test. The LR test

statistic is obtained by evaluating the log likelihood function (2.4) at full and reduced models.

2.5 Anesthesia Recovery Example

Our preceding method is demonstrated on a study of the effects of anesthesia dose on post-surgical recovery described in Section 2.2. The data appear in Appendix II in Davis (1991). Figures 2.1 and 2.2 give a rough indication of the relationship between the category of recovery (0 – 6) and the dose of anesthesia and time spent by the children in the recovery room. Since 15 profiles on one plot in each dose group are too cluttered, only profiles of repeated measures of the first 4 children in each dose group are shown in Figure 2.1. Figure 2.2 gives the average recovery profiles from the 15 children in each dose group. Regression analysis of the ordered response was used to explore the effects of dose, time in recovery room, age of child, and duration of surgery.

Statistically significant nonlinearities for the effects of dose, time in recovery room, interactions of dose and age, and interactions of time and duration make simple statements of conclusion infeasible, but a primary result of the analysis is demonstrated in Figure 2.3, where the estimated mean of the latent "recovery" (underlying the ordered categories of recovery) is plotted versus age of child for children in each of the four dose groups. Several normal densities, corresponding to 0, 5, 15, and 30 minutes after admission to the recovery room are sketched to emphasize the role of the latent variable in this demonstration. The horizontal dotted lines

correspond to the estimated cutoffs for recovery categories. The actually observed recovery categories for each child at each of the four observation times are plotted as well, using different symbols. These plots are produced for a surgery time set to 70 minutes.

Notice that there are 15 children per dose group and that each child is measured for recovery at four time points, so that a column of points above a given age in the plots corresponds to the four readings for each child. It is evident that there is a surprising interaction: that for all doses except the largest, older children tend to have lesser recovery scores for given values of the other covariates. For the largest dose, this relationship is reversed (the p-value for interaction from a likelihood ratio test is 0.0029). It is evident from the plots that the effects of child age and the interactive effect of age and dose are of the same order of magnitude as the dose effect.

Our details of fit are as follows. According to our preliminary investigation and suggestions proposed by Tutz and Hennevogl (1996), the covariates in our model are *dosage*, *time*, *age*, *duration*, *dosage*age* and *time*duration*, where *dosage* and *time* are treated as factors through defining the indicator variables: D20 = I(dosage = 20) , D25 = I(dosage = 25), D30 = I(dosage = 30), T5 = I(time = 5), T15 = I(time = 15), and T30 = I(time = 30), where I( ) is an indicator function. The following four different working correlation structures were considered.

Independent: $\Sigma_{\underset{\sim}{\rho}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$; AR(1): $\Sigma_{\underset{\sim}{\rho}} = \begin{bmatrix} 1 & \rho & \rho^3 & \rho^6 \\ \rho & 1 & \rho^2 & \rho^5 \\ \rho^3 & \rho^2 & 1 & \rho^3 \\ \rho^6 & \rho^5 & \rho^3 & 1 \end{bmatrix}$.

Exchangeable: $\Sigma_{\underset{\sim}{\rho}} = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$; Unstructured: $\Sigma_{\underset{\sim}{\rho}} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{bmatrix}$.

Treating the model with unstructured correlation as a full model and others as reduced models, our LR tests showed that AR(1) was adequate while independent and exchangeable were not. The corresponding *p*-values were 0.37 (AR(1)), less than 0.0001 (independent) and 0.0002 (exchangeable).

Table 2.1 lists the estimates and standard errors from our maximum likelihood approach along with those from the GEE approach. Estimates in the GEE column are from the GEE method with an independence working correlation, a multinomial distribution, and a (cumulative) probit link function. We used the S-PLUS (version 6.2 and above) correlatedData library to obtain the GEE estimates. As for the maximum likelihood approach, the standard errors in parentheses were calculated using the expected and/or observed information. Specifically, the standard errors in the third column were calculated through both the observed and expected information, while those in other columns were based on the expected information only, due to computational difficulties in obtaining the observed information-based ones.

As we can see from Table 2.1, the coefficients of the covariates are quite insensitive to the different models we fitted. However, maximum likelihood estimates

generally give smaller standard errors than the GEE does. Furthermore, in this example, the standard errors based on the observed information and the expected information are very similar to each other in the maximum likelihood ignoring the correlation structure.

2.6 Simulation Study

This section compares the MLE and GEE estimators. The version of the GEE used here is the ordinary GEE estimate, sometimes referred to as GEE1 (Thompson, 2006, p. 219; and implemented with the function *ordgee* in the R package *geepack*). We wished to see whether the maximum likelihood estimator of a regression coefficient has better operating characteristics than the corresponding GEE estimator in a setting for which maximum likelihood is expected to work well, meaning with data simulated from the model on which the maximum likelihood estimator is derived. Robustness is a separate question, but it seems appropriate to explore the possible efficiency gains from maximum likelihood first.

2.6.1 Model for Simulation

We based the simulation conditions on the estimated model for the "marijuana use" example described in Section 2.2 and presented in Table A1, (from Hauspie, Cameron, and Molinari, 2004, p. 375). We simulated data from a multivariate ordinal probit regression model by generating a normally distributed latent variable with AR(1) correlation structure for repeated measures, with parameters roughly matching the estimated values from the real example, and identifying ordered response categories

according to the value of the latent variable relative to cutoffs roughly matching the estimated values from the real data. We also manipulated the number of ordinal categories, the number of repeated measures, and the degree of correlation of the within-subject observations.

Inferences from GEE and ML estimators are justified by large *n* asymptotics, and one should be concerned about the operating characteristics for small and moderate sample sizes. In this investigation, though, we compare the two approaches for fairly large samples only, as might be encountered in an important large-scale medical study, again to explore the possible efficiency gains from maximum likelihood in a nearly ideal situation. In addition, we focus on moderate and high correlations of the within-subject responses. If within-subject correlations are small, then GEE and ML estimates of regression parameters are unlikely to differ much from estimates that ignore the correlation structure.

Our comparisons center on the gender effect (i.e., the coefficient β_2). Appendices A1 and A2 outline the details of the multivariate ordinal probit model fitting and the simulation procedures, respectively.

2.6.2 Results

Figure 2.4 and Figure 2.5 show the Monte Carlo sampling distributions of $\hat{\beta}_{2GEE}$ and $\hat{\beta}_{2MLE}$ (sample size = 1000) for eight scenarios. The vertical line indicates the true

value of β_2 (0.38). All the sampling distributions appear to be well approximated by normal distributions, as large-sample statistical theory predicts.

The mean and median values in the Monte Carlo distributions in Table 2.2 are all very close to the true value 0.38, but the MLE variance tends to be smaller than the GEE variance. The superior efficiency of ML is largest with large *G* (number of response categories) , small *T* (number of repeated measures), and large ρ (autocorrelation of lag 1) (see the bottom row in Table 2.2). The greatest disparity occurred when there were 6 response categories observed and 3 repeated measures. The mean square error of the maximum likelihood estimator was only 17% of the mean square error of the GEE estimator in this case. For all scenarios with 6 response categories the MSE of the GEE estimator was more than twice that of the MLE. When there were only three response categories, the MSE of the MLE was 65% to 89% of the MSE of the GEE estimator.

Besides MSE, Table 2.2 reports Monte Carlo SD (MCSD) and Averaged Reported SE (ARSE) from both methods. Patterns of MCSD and ARSE with respect to changes in ρ , *T* and *G* are similar to those of MSE. A detailed discussion can be found in appendix A3. We also investigated the accuracy of tests based on SEs from the GEE and likelihood ratio tests, and found no conclusive results for the sample size we chose, though the likelihood ratio test tends to be more accurate in some scenarios (see appendix A4).

2.7 Discussion

For estimating the various models in the anesthesia recovery example above, the number of iterations required for convergence was never more than 4. Each iteration required about 2 minutes using an R routine on an Intel Pentium 1.60GHz PC. We believe the algorithm is practicable for maximum likelihood estimation and likelihood ratio inference in data analysis of repeated ordinal responses, and for further studying the relative merits of likelihood and GEE analysis as we did in our simulation study.

The proposed method can be extended to fitting a model where (1) each response variable from the same subject has different covariates and/or different thresholds, or (2) thresholds vary across some groups of subjects. It can also be extended to fit a model in which the correlation structure varies across subjects, as for example, when (1) the T_i's differ or (2) part of the repeated response is missing completely at random (MCAR) or missing at random (MAR). The method is also suitable for a spatial study in which data are collected from each unit nested within clusters. Another modification makes the method suitable for a model with known thresholds. In this case, the latent variable is modeled by $z_{it} \sim N(\beta_0 + \underset{\sim}{x}_{it}'\underset{\sim}{\beta}, \sigma^2)$ and we need to estimate the intercept β_0 and the variance σ^2. This would be appropriate for a response like body mass index category, where the categories are based on known cutoffs (i.e., actual body mass index).

In the anesthesia recovery example and the simulation study, the GEE and maximum likelihood estimates were very similar, but the standard errors were not. Our simulations show MLE is more efficient and less sensitive to changes of the

number of categories than GEE. In terms of MSE, more improvement will be attained from using ML if the data has more categories and/or fewer repeated measures, though an increase of the degree of the within-subject correlation from a medium level to a high level gives mixed effects on the ML. It is worth noting that there is no evidence for more improvement of MSE from ML when each subject has more repeated measures. On the contrary, the simulations show less improvement in MSE due to a faster drop of MSE in GEE than in ML when the number of repeated measures increases.

While the maximum likelihood (ML) method based on the multivariate ordinal probit model has a few attractive features mentioned above, one may argue that the GEE has greater robustness. The multivariate normal latent variable might not be as strong of an assumption as it first seems, though. For one thing, if there really is a latent variable (as there would be for categorizing hurricane strength from maximum wind speed, for example), it is only necessary that some monotonic function of the latent variable is normally distributed. Furthermore, the actual distribution of the latent variable might not matter much if the extreme response categories are not too strongly tied to the extreme tails of the latent variable distribution. Future research would be needed, though, to evaluate the robustness of the ML method.

2.8 References

Agresti, A. (2002). *Categorical data analysis*, 2nd edition, John Wiley & Sons, Inc.

Agresti, A. and Natarajan, R. (2001). Modeling clustered ordered categorical data: A survey. *International Statistical Review* **69**, 345-371.

Anderson, J.A. and Pemberton, J.D. (1985). The grouped continuous model for multivariate ordered categorical variables and covariate adjustment. *Biometrics* **41**, 875-885.

Brent, R. (1973). *Algorithms for Minimization without Derivatives*. Prentice-Hall, Englewood Cliffs, NJ.

Davis, C.S. (1991). Semi-parametric and non-parametric methods for the analysis of repeated measurement with applications to clinical trials. *Statistics in Medicine* **10**, 1959-1980.

Dennis, J. E. and Schnabel, R. B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, NJ.

Efron, B. and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika* **65**, 457-482.

Elliot, D.S., Huizinga, D. and Menard, S. (1989). *Multiple Problem Youth: Delinquence, Substance Use and Mental Health Problems*. New York: Springer-Verlag.

Fu, T.T., Li, L.A., Li, Y.M., and Kan K. (2000). A limited information estimator for the multivariate ordinal probit model. *Applied Economics* **32**, 1841-1851.

Gelman, A. G., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*, 2nd edition, Chapman & Hall/CRC.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics* **1**, 141-149.

Genz, A. (1993). Comparison of methods for the computation of multivariate normal probabilities. *Computing Science and Statistics* **25**, 400-405.

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

Hauspie, R.C., Cameron, N. and Molinari, L. (2004). *Methods in Human Growth Research*. Cambridge University Press.

Heagerty, P.J. and Zeger, S.L. (1996). Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association* **91**, 1024-1036.

Kenward, M.G., Lesaffre, E., and Molenberghs, G. (1994). An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics* **50**, 945-953.

Kim, K. (1995). A bivariate cumulative probit regression model for ordered categorical data. *Statistics in Medicine* **14**, 1341-1352.

Lang, J.B., McDonald, J.W. and Smith, P.W.F. (1999). Association modeling of multivariate categorical responses: a maximum likelihood approach. *Journal of the American Statistical Association* **94**, 1161-71.

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

Lindsay B.G. and Li, B. (1997). On second-order optimality of the observed Fisher information. *Annals of Statistics* **25**, 2172-2199.

Lipsitz, S.R., Kim, K., and Zhao, L. (1994). Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine* **13**, 1149-1163.

Liu, I. and Agresti, A. (2005). The analysis of ordered categorical data: An overview and a survey of recent developments. *Sociedad de Estadistica e Investigacion Operativa*. *Test* **14**, no.1, 1-73.

Long, J. Scott (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage publications.

Lumley, T. (1996). Generalized estimating equations for ordinal data: A note on working correlation structures. *Biometrics* **52**, 354-361.

Mark, S.D. and Gail, M.H. (1994). A comparison of likelihood-based and marginal estimating equation methods for analyzing repeated ordered categorical responses with missing data (Disc: p495-498). *Statistics in Medicine* **13**, 479-493.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of Royal Statistical Society*. *Series B* **42**, 109-142.

McFadden, D. (1989). A method of simulated moments for estimation of discrete choice response models without numerical integration. *Econometrica* **57**, 995-1027.

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference: from a Neo-Fisherian Perspective*, World Scientific Publishing Company.

Qu, Y., Piedmonte, M.R., and Medendorp, S.V. (1995). Latent variable models for clustered ordinal data. *Biometrics* **51**, 268-275.

Schnabel, R. B., Koontz, J. E., and Weiss, B. E. (1985). A modular system of algorithms for unconstrained minimization. *ACM Trans. Math. Software* **11**, 419-440.

Skovgaard, I.M. (1985). A second-order investigation of asymptotic ancillarity. *Annals of Statistics* **13**, 534-551.

Thompson, L.A. (2006). *S-PLUS (and R) Manual to Accompany Agresti's Categorical Data Analysis (2002) 2nd edition*. https://home.comcast.net/~lthompson221/#otherresearch

Toledano, A.Y. and Gatsonis, C. (1996). Ordinal regression methodology for ROC curves derived from correlated data. *Statistics in Medicine* **15**, 1807-1826.

Tutz, G. and Hennevogl, W. (1996). Random effects in ordinal regression models. *Computational Statistics & Data Analysis* **22**, 537-557.

Williamson, J.M., Kim, K., and Lipsitz, S.R. (1995). Analyzing bivariate ordinal data using a global odds ratio. *Journal of the American Statistical Association* **90**, 1432-1437.

Tables and Figures

Table 2.1
Estimation results for the anesthesia recovery example

parameters	GEE (independent)	MLE* (independent)	MLE** (AR1)	MLE** (exchangeable)	MLE** (unstructured)
α_1	-2.3618 (0.9404)	-2.3624 (0.6128, 0.6032)	-2.3650 (0.8466)	-2.3624 (0.8803)	-2.3624 (0.8215)
α_2	-1.2278 (0.9469)				
α_3	-0.9220 (0.9387)				
α_4	-0.4576 (0.9267)				
α_5	-0.1240 (0.9341)				
α_6	0.2658 (0.9333)				
d_1		1.0649 (0.0635, 0.0650)	1.0292 (0.0671)	1.0649 (0.0679)	1.0648 (0.0714)
d_2		0.5530 (0.0586, 0.0589)	0.5449 (0.0595)	0.5530 (0.0612)	0.5532 (0.0576)
d_3		0.6815 (0.0569, 0.0556)	0.6869 (0.0572)	0.6813 (0.0557)	0.6816 (0.0557)
d_4		0.5776 (0.0583, 0.0578)	0.5711 (0.0594)	0.5775 (0.0588)	0.5777 (0.0606)
d_5		0.6243 (0.0599, 0.0610)	0.6394 (0.0601)	0.6243 (0.0634)	0.6244 (0.0615)
D20	0.9431 (1.0351)	0.9430 (0.6291, 0.6412)	0.9502 (1.0044)	0.9429 (1.0298)	0.9431 (0.9674)
D25	0.5180 (1.2348)	0.5180 (0.7173, 0.7283)	0.5170 (1.1377)	0.5180 (1.1788)	0.5180 (1.1255)
D30	-2.1161 (1.0579)	-2.1164 (0.5979, 0.5829)	-2.1224 (0.9562)	-2.1165 (1.0109)	-2.1163 (0.9320)
T5	0.1642 (0.1990)	0.1641 (0.4691, 0.4559)	0.1662 (0.1861)	0.1642 (0.2685)	0.1643 (0.2133)
T15	1.0479 (0.3077)	1.0478 (0.4749, 0.4672)	1.0477 (0.2892)	1.0478 (0.2838)	1.0481 (0.3124)
T30	2.1177 (0.4311)	2.1176 (0.5017, 0.4808)	2.1209 (0.3849)	2.1176 (0.3213)	2.1177 (0.3936)
Age	-0.0167 (0.0178)	-0.0168 (0.0116, 0.0116)	-0.0177 (0.0182)	-0.0173 (0.0194)	-0.0169 (0.0176)
Duration	-0.0066 (0.0035)	-0.0067 (0.0039, 0.0038)	-0.0070 (0.0039)	-0.0066 (0.0039)	-0.0066 (0.0040)
*D20*Age*	-0.0378 (0.0227)	-0.0378 (0.0162, 0.0164)	-0.0368 (0.0260)	-0.0379 (0.0268)	-0.0379 (0.0254)
*D25*Age*	-0.0200 (0.0263)	-0.0200 (0.0167, 0.0168)	-0.0209 (0.0267)	-0.0207 (0.0277)	-0.0199 (0.0266)
*D30*Age*	0.0386 (0.0215)	0.0386 (0.0135, 0.0132)	0.0377 (0.0217)	0.0384 (0.0231)	0.0385 (0.0213)
*T5*Duration*	0.0042 (0.0027)	0.0042 (0.0054, 0.0053)	0.0046 (0.0021)	0.0045 (0.0031)	0.0043 (0.0024)
*T15*Duration*	-0.0007 (0.0035)	-0.0007 (0.0054, 0.0053)	-0.0003 (0.0031)	-0.0003 (0.0031)	-0.0007 (0.0033)
*T30*Duration*	-0.0030 (0.0045)	-0.0030 (0.0055, 0.0054)	-0.0021 (0.0040)	-0.0022 (0.0031)	-0.0022 (0.0040)
ρ			0.8863 (0.0243)	0.7316 (0.0523)	

(Continued)					
ρ_{12}					0.8367 (0.0476)
ρ_{13}					0.6191 (0.0908)
ρ_{14}					0.4862 (0.1254)
ρ_{23}					0.8062 (0.0524)
ρ_{24}					0.5228 (0.1176)
ρ_{34}					0.7634 (0.0709)
log likelihood	N.A.	-375.8387	-309.2073	-318.4727	-306.5435

*: the first and second standard errors in each cell from this column are based on the observed information and the expected information, respectively.
**: the standard errors are based on the expected information.

Table 2.2

Descriptive statistics of estimates of β_2 (*true value 0.38*)

	$G = 3$				$G = 6$			
	$T = 3$		$T = 7$		$T = 3$		$T = 7$	
	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$
GEE								
mean	0.3852	0.4038	0.3810	0.3830	0.4257	0.4474	0.3910	0.3939
median	0.3898	0.3978	0.3806	0.3814	0.4035	0.4392	0.3840	0.3869
minimum	0.0252	-0.0407	0.1596	0.0519	-0.2251	-0.3413	0.1279	-0.0821
maximum	0.8244	0.8619	0.5588	0.6962	1.9613	1.5360	0.7476	0.8327
Monte Carlo SD	0.1091	0.1456	0.0668	0.0981	0.2296	0.2550	0.1037	0.1290
average reported SE	0.1076	0.1335	0.0663	0.0972	0.2120	0.2253	0.1017	0.1236
MSE	0.0119	0.0217	0.0045	0.0096	0.0548	0.0695	0.0109	0.0168
MLE								
mean	0.3851	0.3965	0.3798	0.3802	0.3837	0.3974	0.3876	0.3844
median	0.3858	0.3948	0.3807	0.3808	0.3863	0.3981	0.3869	0.3844
minimum	0.0836	0.0407	0.1874	0.1072	0.0766	0.0655	0.1918	0.0498
maximum	0.7481	0.7542	0.5774	0.6723	0.7022	0.7871	0.6069	0.6646
Monte Carlo SD	0.0957	0.1179	0.0632	0.0827	0.0954	0.1075	0.0631	0.0866
average reported SE	0.0883	0.0986	0.0710	0.0868	0.0917	0.1023	0.0828	0.1057
MSE	0.0092	0.0142	0.0040	0.0068	0.0091	0.0118	0.0040	0.0075
MSE improvement (%)*	23.00	34.88	10.34	29.07	83.37	82.96	62.92	55.38

*: (MSE.gee - MSE.mle)/MSE.gee×100%

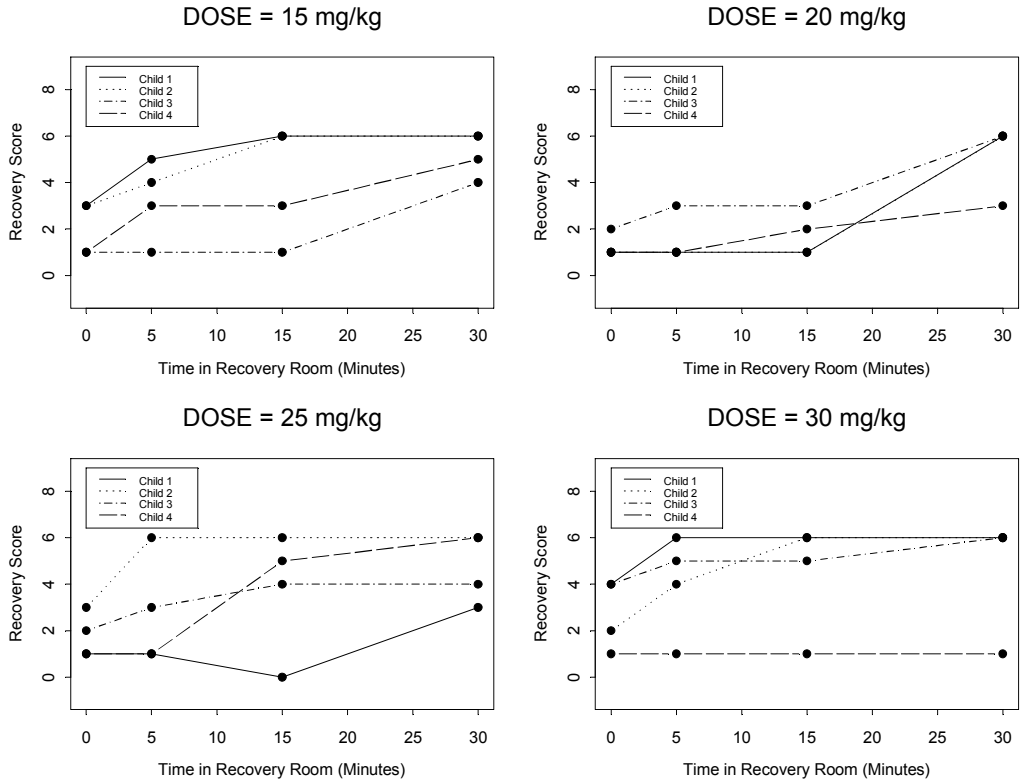

Figure 2.1 Profiles of repeated measures of the first 4 children in each dose group

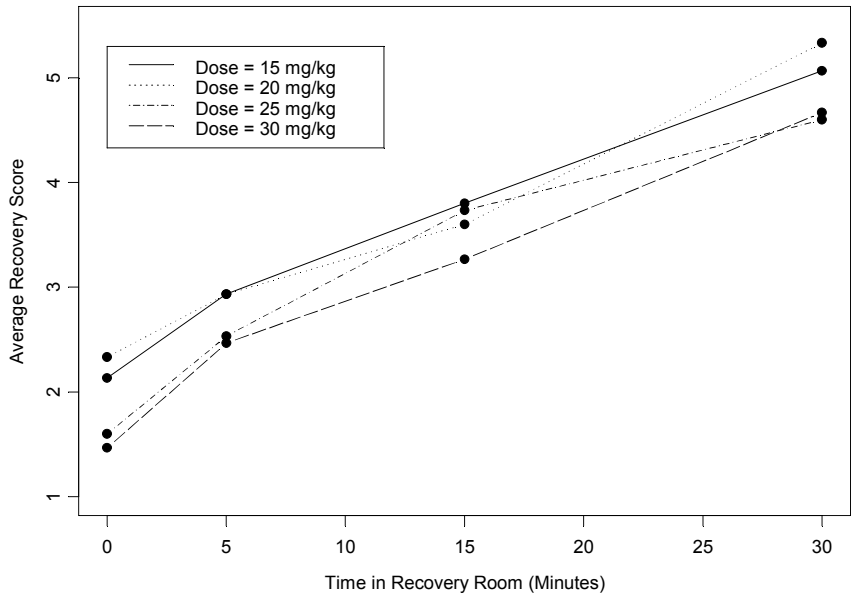

Figure 2.2 Average recovery profiles from 15 children in each dose group

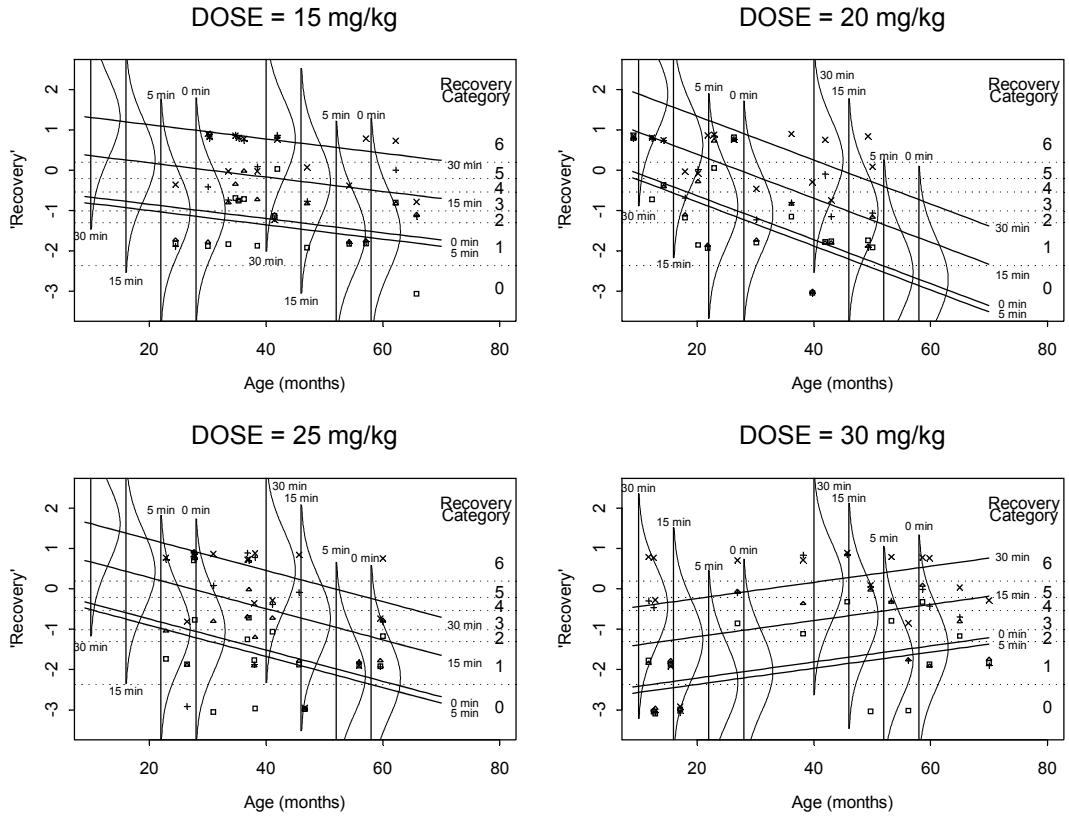

Figure 2.3 Estimated mean of the latent recovery and observed recovery categories

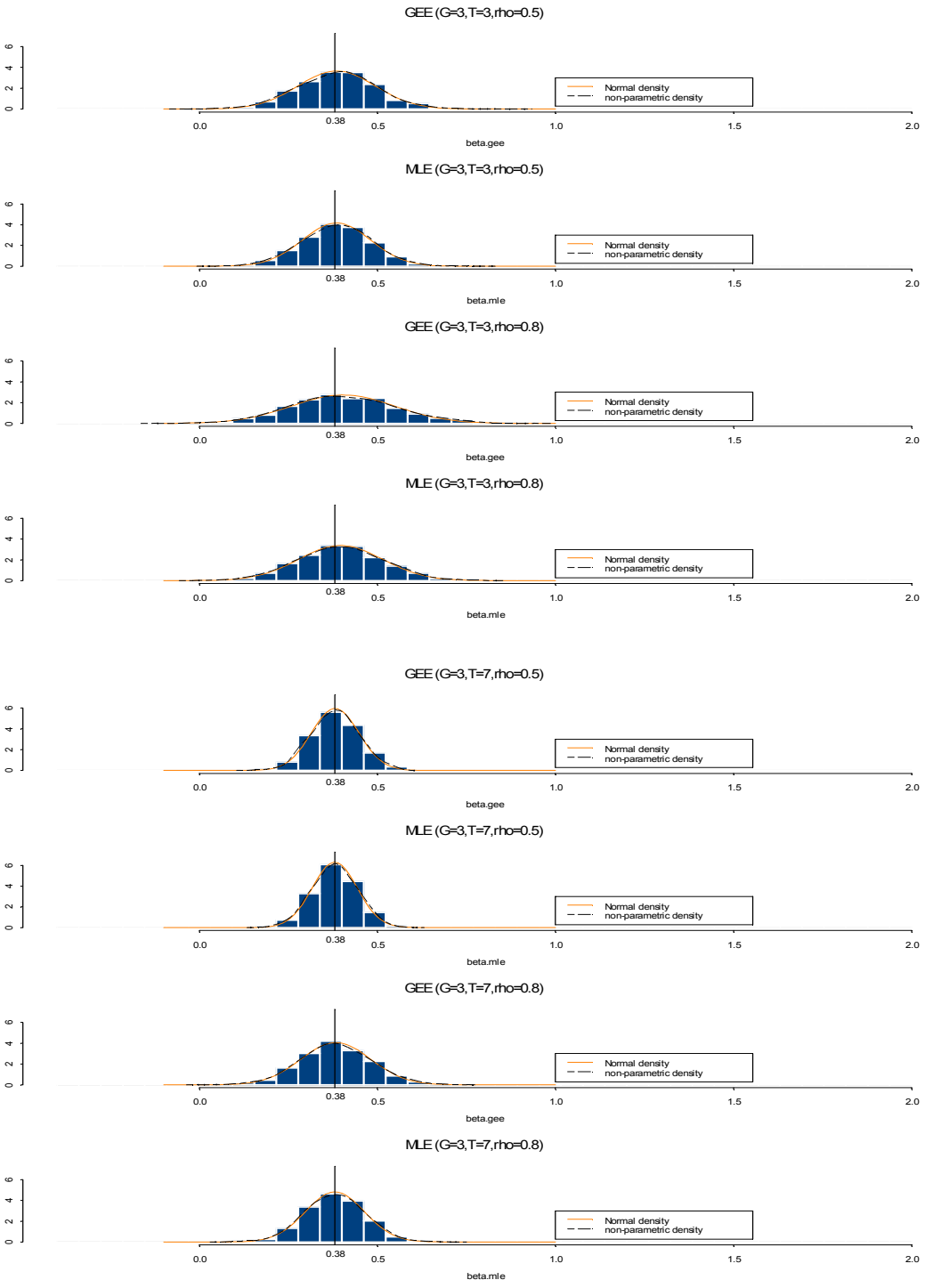

Figure 2.4 Histograms of $\hat{\beta}_{2GEE}$ and $\hat{\beta}_{2MLE}$ with $G = 3$, $T = 3$ or 7, $\rho = 0.5$ or 0.8

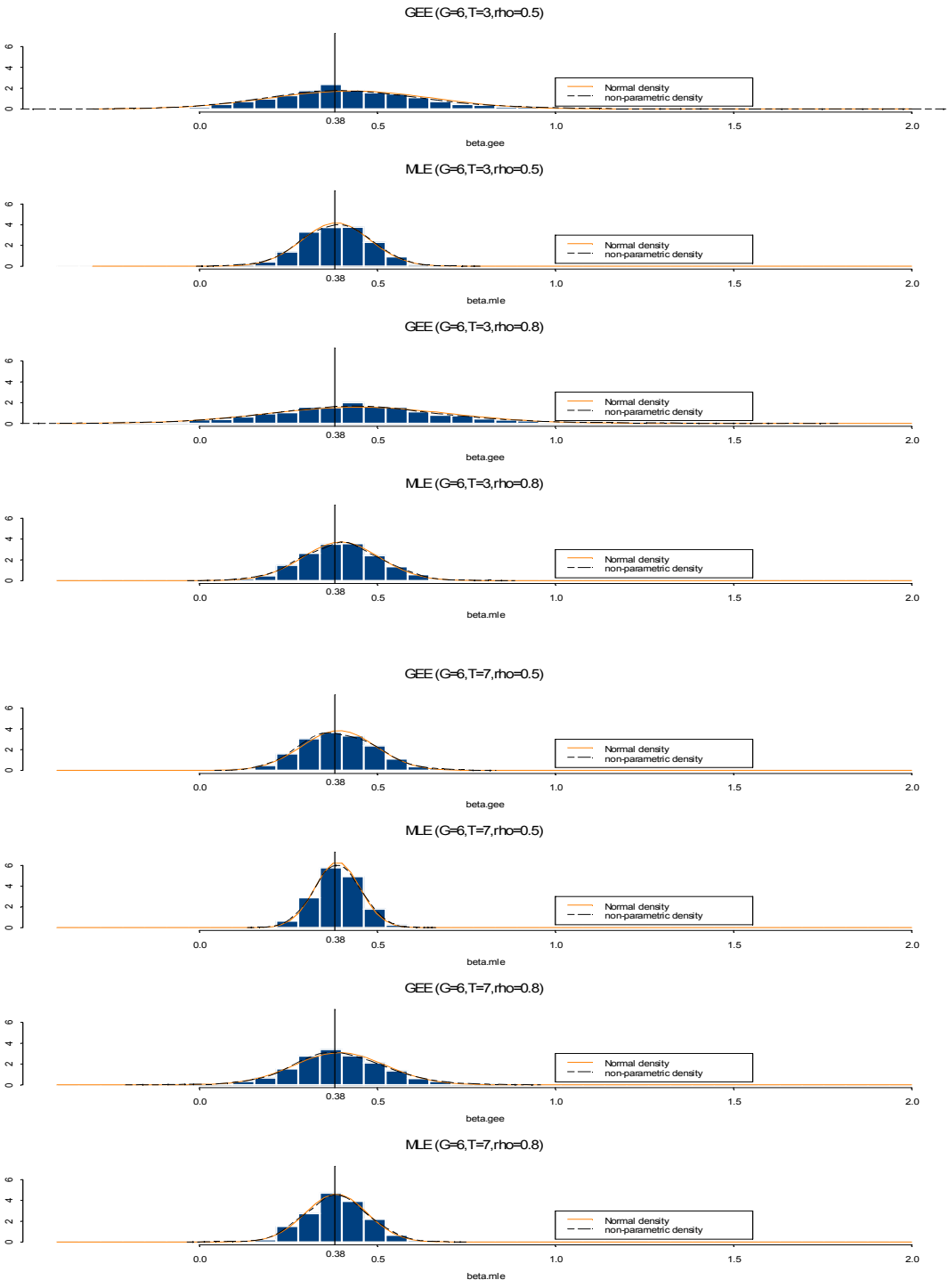

Figure 2.5 Histograms of $\hat{\beta}_{2GEE}$ and $\hat{\beta}_{2MLE}$ with $G = 6$, $T = 3$ or 7, $\rho = 0.5$ or 0.8

3. Regression Analysis for Ordered Categorical Responses with Spatial-temporal Correlation

Yonghai Li* and Daniel W. Schafer**

Department of Statistics, Oregon State University, Corvallis, OR, 97331, U.S.A.
*email: yonghai@science.oregonstate.edu
**email: schafer@science.oregonstate.edu

3.1 Abstract

This paper is about the regression analysis for ordered categorical responses with spatial-temporal correlation. Using the latent error induced dependency among spatial-temporal ordinal responses, we propose likelihood analysis for a regression model for spatial-temporal ordered categorical data. This approach avoids defining a neighborhood for each site, which sometimes is not clear but is required for some existing approaches (e.g., the Markov Random Field), particularly on an irregular lattice. Without resorting to a composite (or pseudo) likelihood or GEE or Bayesian approaches, the proposed approach is demonstrated on a coho study example, in which the coho density at each sample site from each spawning year is classified into 4 categories.

3.2 Introduction

3.2.1 The Coho Data Example

Populations of coho salmon (*Oncorhynchus Kisutch*) that occur in Oregon coastal watersheds between Cape Blanco and the mouth of the Columbia River are being evaluated by the National Oceanic and Atmospheric Administration National Marine Fisheries Service (NOAA Fisheries) for listing under the federal Endangered Species Act (ESA). In 1998, the Oregon Department of Fish and Wildlife (ODFW) and NOAA Fisheries initiated the Coastal Coho Project to address the conservation of coastal coho on the Oregon coast. One of the primary objectives was to identify key habitat factors associated with the abundance and distribution of wild and hatchery coho spawning in streams to help better understand the conditions for conservation management.

Data for assessing this association are shown in Table A5 in the appendices. These come from two different sources. Salmon densities at various stream sites come from the ODFW Costal Salmonid Inventory Project **(**Detailed survey procedures can be found at the following website: http://oregonstate.edu/Dept/ODFW/spawn/pdf%20files/reports/05SSManual.pdf.) Habitat data, on physical characteristics and conditions of the streams are from the Aquatic Inventories Project, a separate ODFW monitoring project (detailed information can be found at the following website: http://oregonstate.edu/Dept/ODFW/freshwater/inventory/index.htm.). The rows in Table A5 correspond to stream sites at which both the salmon density and habitat

variables were available. The first five columns show gene conservation group (GCG), strata (North Coast, Mid Coast, Mid-South Coast, South Coast, and Umpqua), sampling identification number, spawning year, and coho count per mile, respectively. Columns 6-16 are potential habitat covariates, with their meaning included in Table 3.1. The last two columns are the latitude and the longitude of the sampled sites. There are totally 206 distinct sites where 32 sites have repeated measures (see the repeated measure structure in Table 3.2). The data consists of 58, 48, 51, 48, 45, 23 and 22 observations in spawning years 1998, 1999, 2000, 2001, 2002, 2003 and 2004, respectively.

A specific goal is to identify habitat variables associated with salmon density. An actual density is estimated at each site, in units of coho salmon per linear stream mile. Fish biologists categorize density in the following way:

absent: density less than 4 adult coho per mile

few or *some*: density greater than or equal to 4 and less than 40 adult coho per mile (we used 12, the median of the observed data in the interval [4, 12), as an additional cutoff to distinguish between *few* and *some*)

full: density greater than or equal to 40 adult coho per mile

This paper focuses on an analysis using this ordered categorical response variable for assessing the association of salmon density with the explanatory habitat variables. Given that the underlying numerical densities are available, it would also be possible to analyze the data with the uncategorized response (and report the resulting conclusions from that analysis in terms of the categorizations, if desired). However,

there are a few problems with the density as a response variable. The distribution tends to be skewed, but since there are so many zeros a log transformation is out of the question. A square root transformation might work, but its lack of interpretability is too much of a drawback. One could try Poisson log-linear regression (with overdispersion) or Negative Binomial regression, but introducing the spatial/temporal correlation is not quite convenient, especially in the frequentist framework. While the Generalized Estimating Equation (GEE) methods are available (e.g., Zeger (1988) for the temporal count data; McShane, Albert, and Palmatier (1997) for the spatial count data), most other studies employ Bayesian approaches (e.g., Best, Ickstadt, and Wolpert (2000) and Wakefield (2006) for spatial Poisson models; Alexander, Moyeed, and Stander (2000) for a spatial Negative Binomial model). Wakefield (2006) notices there are currently no simple ways of fitting frequentist fixed effects, non-linear models with discrete response and spatially dependent residuals (except for the GEE). As for spatial-temporal count data, to our knowledge, the GEE method for this data structure has not been developed yet.

Our justification for the use of the categorical variable directly is that the analysis is based on weaker assumptions. The spatial/temporal correlation of response variables is addressed with a multivariate normal latent variable. It does not imply that the underlying densities are normal, only that some monotonic transformation of the densities is normally distributed. There is no requirement to specify the transformation.

3.2.2 Literature Review for Spatial/Temporal Ordinal Data

While models for spatially-dependent binary data and non-ordered categorical data have been discussed in the literature for a long time (see Ising (1925) for a first-order auto-logistic model for binary data on a lattice; see Strauss (1977) for non-ordered categorical data), modeling ordered categorical spatial data has not received much attention until more recently. Kutsyy (2001) views ordinal data as arising from a latent continuous-valued spatial process through a threshold concept. Furthermore, a first-order Gaussian Markov random field is used to model the latent data which induces spatial dependence in the ordinal data. Due to computational intractability of maximum likelihood estimator, Kutsyy (2001) considers alternative methods based on a pseudo-likelihood and two other approximations to the likelihood (MnE and MdE). Brewer et al. (2004) also assumes a continuous latent variable, but employs a mixed-effects model, including an exchangeable spatial random effect and a neighborhood based spatial random effect. Their model was estimated through a Bayesian approach for grazing impact data from two areas of Scotland. When estimation of the covariate effect is of primary interest, Baeumler (1995) suggests GEE approaches developed in Miller, Davis, and Landis (1993) and Qu, Piedmonte, and Medendrop (1995).

In contrast with the work on statistical modeling for the purely spatial binary/ordinal data, the literature dealing with the spatial-temporal binary/ordinal data is rather limited and more recent. To name a few, Gumpertz, Wu and Pye (2000) show, in a quasi-likelihood approach, an application of a marginal logistic regression model with spatial and temporal autocorrelations to binary responses. Peraza-Garay

(2004) uses partially ordered Markov models (Cressie and Davidson, 1988; Davidson, Cressie and Hua, 1999; Huang and Cressie, 2000) to handle both spatial and temporal dependencies. Ramos-Quiroga and González-Farías (2005) give a spatial-temporal ordinal model through a pseudo-likelihood approach, where a Markov random field is defined over both the space and the time domains on a regular lattice. Kneib and Fahrmeir (2006) propose a general class of structured additive regression models for categorical responses (including ordinal responses) from a Bayesian perspective. The resulting empirical Bayes method is closely related to penalized likelihood estimation in a frequentist setting. But none of all these methods adopt a fully likelihood approach.

In this paper we propose a likelihood analysis for a regression model for spatial-temporal ordered categorical data. Using the latent error induced dependency among spatial-temporal ordinal responses, our approach avoids defining a neighborhood for each site, which sometimes is not clear but is required for some existing approaches (e.g., the Markov Random Field), particularly on an irregular lattice. More importantly, a fully likelihood analysis can be conducted in our approach without resorting to a composite (or pseudo) likelihood or GEE or Bayesian approaches.

This paper proceeds as follows. In the next section, our model for ordered categorical responses with spatial-temporal correlation is described. Section 3.4 contains an application to the coho study example. The paper ends with a brief discussion of our method and possible directions of future research in Section 3.5.

3.3 Model for Ordered Categorical Responses with Spatial-temporal Correlation

Following Cressie (1991) and Schabenberger and Gotway (2005), we have a spatial-temporal process as follows,

$$\{Z(s,t) : s \in D(t) \subset \Re^q, t \in T\},$$

where $Z(s,t)$ denotes the response of the latent variable at time t, and location s, a $(q \times 1)$ vector of coordinates in the domain D; T is the time domain. The observed ordinal responses $y(s,t)$ are assumed to be generated from the realized $z(s,t)$ through a threshold concept as follows.

$$y(s,t) = g \ \ (g = 1, 2, \ldots, G) \text{ if and only if } \alpha_{g-1} < z(s,t) \le \alpha_g,$$

where $-\infty = \alpha_0 < \alpha_1 < \alpha_2 < \ldots < \alpha_{G-1} < \alpha_G = \infty$ are the thresholds for the latent variable $z(s,t)$. Let $\underset{\sim}{x}_{st}$ denote a $1 \times p$ vector of covariates with a corresponding coefficients vector $\underset{\sim}{\beta}$, assume $Z_{st} \equiv Z(s,t) \sim N(\underset{\sim}{x}_{st}'\underset{\sim}{\beta}, 1)$, a normal distribution with mean $\underset{\sim}{x}_{st}'\underset{\sim}{\beta}$ and variance 1. Given the n observations, we further assume the joint distribution of the latent variables Z_{st} is an n-dimensional multivariate normal $N_n(X\underset{\sim}{\beta}, \Sigma)$, where $X \equiv (\underset{\sim}{x}_{s_1t_1}, \ldots, \underset{\sim}{x}_{s_nt_n})'$ and Σ is the covariance matrix, which is also a correlation matrix because the marginal variances are all 1. We will account for the correlation of ordinal responses at different sites and time by modeling the correlation (Σ) of the latent variables as a function of some unknown parameters. Many different forms of Σ are available and they can be either (time and space) *separable* or *non-separable*. One reasonable form of a separable covariance (correlation) function is

$$\text{cov}[Z(s_i,t_k),Z(s_j,t_l)] = e^{-c_s h_{ij} - c_t|t_k - t_l|} \qquad (3.1)$$

where h_{ij} is the Euclidean distance between s_i and s_j; c_s and c_t are unknown parameters (e.g., Mitchell and Gumpertz, 2003). While separable covariance models such as (3.1) do not incorporate space-time interactions in the covariance, non-separable covariance models allow the interactions and therefore, seem more attractive if the interaction does exist. Gneiting (2002) presented a flexible and elegant approach to construct non-separable covariance functions. In the coho data example, noting coho salmon have a three-year cycle (i.e., fish that hatch one year come back from the ocean after three years to spawn), we expect temporal correlation at lag 3 but not lags 1 and 2. Therefore, we use the following stationary non-separable covariance function,

$$\text{cov}[Z(s_i,t_k),Z(s_j,t_l)] = e^{-c_s h_{ij} - c_t I(|t_k - t_l|=3)}$$

with $\text{cov}[Z(s_i,t_k),Z(s_j,t_l)] = 0$ if (i) $s_i = s_j$, $t_k \neq t_l$, and $|t_k - t_l| \neq 3$ or (ii) $s_i \neq s_j$ and $t_k \neq t_l$. Specifically, the preceding covariance equals (1) 1 if $s_i = s_j$ and $t_k = t_l$; (2) e^{-c_t} if $s_i = s_j$ and $|t_k - t_l| = 3$; (3) 0 if $s_i = s_j$, $t_k \neq t_l$, and $|t_k - t_l| \neq 3$; (4) $e^{-c_s h_{ij}}$ if $s_i \neq s_j$ and $t_k = t_l$; (5) 0 if $s_i \neq s_j$ and $t_k \neq t_l$. It immediately follows that the temporal correlation is restricted to lag 3 within each site while the spatial correlation is restricted within each time (spawning year). We make the dependence of Σ on $\underset{\sim}{c} \equiv (c_s, c_t)'$ explicit with the symbol $\Sigma_{\underset{\sim}{c}}$.

It is straightforward that the observed data log likelihood *l* can be written as:

$$l(\underset{\sim}{\alpha},\underset{\sim}{\beta},\underset{\sim}{\varsigma}) = \log \int_{a_1}^{b_1}\int_{a_2}^{b_2}\cdots\int_{a_n}^{b_n} \varphi_n(\underset{\sim}{z}, X\underset{\sim}{\beta}, \Sigma_{\underset{\sim}{\varsigma}}) dz_{s_1t_1} dz_{s_2t_2}\cdots dz_{s_nt_n} \qquad (3.2)$$

where $\underset{\sim}{\alpha} \equiv (\alpha_1, \alpha_2, ..., \alpha_{G-1})'$ and $\varphi_n(\cdot)$ is an n-dimension multivariate normal density

$$\varphi_n(\underset{\sim}{z}; X\underset{\sim}{\beta}, \Sigma_{\underset{\sim}{\varsigma}}) = (2\pi)^{-n/2} |\Sigma_{\underset{\sim}{\varsigma}}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\underset{\sim}{z} - X\underset{\sim}{\beta})'\Sigma_{\underset{\sim}{\varsigma}}^{-1}(\underset{\sim}{z} - X\underset{\sim}{\beta})\right) \qquad (3.3)$$

with $\underset{\sim}{z} \equiv (z_{s_1t_1}, z_{s_2t_2}, ..., z_{s_nt_n})'$, and

$a_i = \alpha_{g-1}$, $b_i = \alpha_g$ if and only if $y_i \equiv y(s_i, t_i) = g$, for $i = 1, ..., n$.

Note the MLE of parameters $(\underset{\sim}{\alpha}, \underset{\sim}{\beta}, \underset{\sim}{\varsigma})$ can be obtained through maximizing the log likelihood function in equation (3.2). We use a steepest descent approach for maximizing the parameters, using Genz's numerical approximation (Genz, 1992, 1993) to the multivariate normal probability in the log likelihood. Details of the optimization algorithm can be found in Li and Schafer (2006). Basically, the algorithm maximizes the log likelihood function with respect to parameters sequentially and iteratively until convergence. During the optimization, in order to satisfy the constraint $\alpha_1 < \alpha_2 < \cdots < \alpha_{G-1}$, we reparameterize α_g (g = 2, …, G-1) in terms of α_1, d_1, …, and d_{G-2} as follows.

$$\alpha_2 = \alpha_1 + d_1^2,\ \alpha_3 = \alpha_1 + d_1^2 + d_2^2,\ \ldots,\ \alpha_{G-1} = \alpha_1 + \sum_{g=1}^{G-2} d_g^2, \qquad (3.4)$$

where d_1, d_2, …, and d_{G-2} are non-zero.

Statistical inference about the parameters can be conducted through the Wald test and the likelihood ratio (LR) test, using the asymptotic theory of the maximum likelihood estimators in a spatial (-temporal) regression (see, for example,

Schabenberger and Gotway (2005), p. 343-346, for a discussion of the covariance parameters part).

3.4 Analysis of the Coho Data

Figure 3.1 shows scatterplots of the original data. We can see the relationship between the uncategorized response and the habitat covariates is not quite clear except for a few covariates such as "number of pools deeper than 1.0 meter/kilometer of total stream length" and "pieces of large woody debris/100 meters of primary stream length." The figure also implies adding some quadratic terms of covariates might be helpful. In order to avoid multicollinearity in our regression analysis, we investigate the correlation among all 11 habitat covariates (see Table 3.3). It reveals covariates for pieces of large woody debris, volume of large woody debris, and number of key pieces of large wood (i.e., *LWDPIECE1*, *LWDVOL1*, and *KEYLWD1*) are highly correlated. Consequently, our model excludes *LWDVOL1* and *KEYLWD1* in the following analysis.

Besides the habitat covariates, our model also includes time (i.e., spawning years 1998-2004) as a fixed-effect factor with 7 levels. Treating the fixed effect of spawning year 1998 as a reference level, we create yearly dummy variables *D99*, *D00*, *D01*, *D02*, *D03*, and *D04* for spawning years 1999-2004, respectively. For example, *D99*=1 if spawning year = 1999; *D99*=0 otherwise. An important justification for introducing a fixed-effect of time is to take into account other time-varying covariates such as "ocean conditions," which are not available in the data set. Fish biologists

believe a huge factor in the density of spawners is “ocean survival”, which is a yearly effect that integrates ocean conditions.

Recall that our goal is to identify habitat variables associated with salmon density. Therefore, some variable selection procedures are necessary. We start with a model including the first 9 habitat covariates in Table 3.1, their quadratic terms, and yearly dummy variables *D99*, *D00*, *D01*, *D02*, *D03*, and *D04*. The fully likelihood approach in Section 3.3 allows us to choose BIC as a variable selection criterion. One possibility is backward elimination in which one variable at a time is dropped -- based on decreasing BIC -- until BIC can't be decreased any more. This backward elimination method leads to a model including the following explanatory variables: *D99-D04*, *PCTPOOLS*, $(PCTPOOLS)^2$, *PCTBEDROCK*, *POOL1P.KM*, $(POOL1P.KM)^2$, *LWDPIECE1*, $(LWDPIECE1)^2$. Table 3.4 shows how BIC changes in the backward elimination procedure.

Table 3.5 gives parameters estimates for the selected model. The standard errors of the estimates are omitted since our purpose is to identify what habitat covariates are significantly associated with the coho density. In addition, it would be difficult to make sense out of the standard errors and confidence intervals given the complicated sampling scheme used in this study. However, *p*-values from the likelihood ratio test are reported in Table 3.6. The table shows most of the selected habitat covariates are significant at the 0.05 significance level.

The LR test can also be carried out to compare our spatial-temporal model with a model assuming spatial-temporal independence, since the latter is nested within the

former as $c_s \to \infty$ and $c_t \to \infty$. Note the parameters $c_s = c_t = \infty$ lie on the boundary of the parameter space and we simply divide the *p*-value obtained from a χ^2 with 2 degrees of freedom by 2 (Self and Liang, 1987; Schabenberger and Gotway, 2005, p. 344). The adjustment shrinks the *p*-value but it doesn't alter our conclusion that the spatial-temporal model fits significantly better than a model assuming independence (adjusted *p*-value = 0.0014). Similarly, the LR test can be used to test the purely temporal (or spatial) model since the spatial-temporal model reduces to the purely temporal (or spatial) model as $c_s \to \infty$ (or $c_t \to \infty$). It turns out that there exists moderate evidence for the spatial correlation (*p*-value = 0.0241, Table 3.6). In contrast, we found suggestive but inconclusive evidence for the temporal correlation (*p*-value = 0.0626, Table 3.6).

3.5 Discussion

Using the coho study example, we have shown spatial-temporal dependency of the ordinal responses can be simply incorporated in regression analysis through a multivariate normal distributed latent variable. This approach avoids defining neighbors, and is a fully likelihood analysis. Consequently, the familiar likelihood-based methods for testing fit, comparing models (with AIC and BIC, for example), making inference about parameters are available with this approach. In addition, given a correctly specified form of spatial - temporal correlation structure, our parameter estimates tend to be more efficient than those from treating each observation as uncorrelated or other non-fully-likelihood approaches mentioned in

Section 3.2. An example is Ramos-Quiroga and González-Farías (2005). While they claim their pseudo-likelihood approach is particularly useful for spatial-temporal problems, they are concerned about the loss of efficiency of their approach with respect to fully likelihood ones. Consequently, from our point of view, for the data from a regular lattice, future research could be done to compare our method with others to evaluate the gain of efficiency of our method.

A second benefit of our approach is that it could be easily extended to a linear mixed-effects model with correlated errors for the latent variable for ordinal responses data. As a matter of fact, a linear mixed-effects model with i.i.d. errors for the latent variable for ordinal responses in a longitudinal or spatial study is not uncommon (see, e.g., Hedeker and Gibbons, 1994; Crouchley, 1995; Brewer et al., 2004). In most cases, random effects are introduced to account for 1) the correlation among repeated measures from a subject in a longitudinal context; 2) the correlation among the responses from different sites in a spatial context. As far as the error terms themselves are concerned, there are few exceptions, for example, Girard and Parent (2001), which considered an AR(1) for temporal dependence in analyzing on-line quality data. But these exceptions studied fixed effects of covariates only. Consequently, future studies can focus on a linear mixed-effects model along with correlated errors in a spatial-temporal setting. This type of models can be attractive in some cases. For example, in the coho example, if the spawning years in the sample had been randomly selected, we could have treated the time effects as random, and a linear mixed-effects model with correlated errors would have been necessary.

3.6 References

Alexander, N., Moyeed, R., and Stander, J. (2000). Spatial modeling of individual level parasite counts using the negative binomial distribution. *Biostatistics* **1**, 453-463.

Baeumler, Alfred (1995). Marginal regression models for spatial or temporal correlated forest damage data. Spatial and Temporal Modelling in Agricultural Research, Proceedings of the Fourth HARMA Workshop, IACR-Rothamsted 20-21 October 1995, S.32-42.

Best, N.G., Ickstadt, K.I., and Wolpert, R.L. (2000). Spatial Poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American Statistical Association* **95**, 1076-1088

Brewer, M.J., Elston, D.A., Hodgson, M. EA., Stolte, A.M., Nolan, A.J., and Henderson, D.J. (2004). A spatial model with ordinal responses for grazing impact data. *Statistical Modelling* **4**, 127-143.

Cressie, Noel (1991). *Statistics for Spatial Data*. John Wiley & Sons, New York.

Cressie, N. and Davidson, J. L. (1998). Image analysis with partially ordered Markov models. *Computational Statistics & Data Analysis* **29**, 1-26.

Crouchley, Robert (1995). A random-effects model for ordered categorical data. *Journal of the American Statistical Association* **90**, 489-498.

Davidson, J. L., Cressie, N. and Hua, X. (1999). Texture synthesis and pattern recognition for partially ordered Markov models. *Pattern Recognition* **34**, 1475-1505.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics* **1**, 141-149.

Genz, A. (1993). Comparison of methods for the computation of multivariate normal probabilities. *Computing Science and Statistics* **25**, 400-405.

Girard, Philippe and Parent, Eric (2001). Bayesian analysis of autocorrelated ordered categorical data for industrial quality monitoring. *Technometrics* **43**, 180-191.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association* **97**, 590-600.

Gumpertz, M. L., Wu, C. T. and Pye, J. M. (2000). Logistic regression for southern pine beetle outbreaks with spatial and temporal autocorrelation. *Forest Science* **46**, 95-107.

Hedeker, Donald and Gibbons, Robert D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50**, 933-944.

Huang, H. and Cressie, N. (2000). Asymptotic properties of maximum (composite) likelihood estimators for partially ordered Markov models. *Statistica Sinica* **10**, 1325-1344.

Ising, Ernst (1925). Beitrag zur theories des ferromagnetismus. *Zeitschrift ür Physik* **31**, 253-258.

Kneib, Thomas and Fahrmeir, Ludwig. (2006). Structured additive regression for categorical space-time data: a mixed model approach. *Biometrics* **62**, 109-118.

Kutsyy, Vadim (2001). *Modeling and inference for spatial processes with ordinal data*, Ph.D. dissertation, University of Michigan.

Li, Y.H. and Schafer, D.W. (2006) Likelihood analysis of the multivariate ordinal probit regression model for repeated ordinal responses. Working paper, Statistics Department, Oregon State University.

McShane, L.M., Albert, P.S., and Palmatier M.A. (1997). A latent process regression model for spatially correlated count data. *Biometrics* **53**, 698-706.

Miller, M.E., Davis, C.S., and Landis, J.R. (1993). The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares. *Biometrics* **49**, 1033-1044.

Mitchell, M.W. and Gumpertz, M.L. (2003). Spatio-temporal prediction inside a free-air CO_2 enrichment system. *Journal of Agricultural, Biological, and Environmental Statistics* **8** (3), 310-327.

Peraza-Garay, F. (2004). A model for longitudinal and ordinal data with spatial dependency. *Disertación Doctoral*. CIMAT, Mexico.

Qu, Y., Piedmonte, M.R., and Medendrop, S.V. (1995). Latent variable models for clustered ordinal data. *Biometrics* **51**, 268-275.

Ramos-Quiroga, R. and González-Farías G. (2005). A case study: ordinal responses with spatio-temporal dependencies. *Communicación Tècnica*, No. I-05-04/01-02-2005.

Schabenberger, Oliver and Gotway, Carol A. (2005). *Statistical methods for spatial data analysis*. Chapman & Hall/CRC Press.

Self, S.G. and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**, 605-610.

Strauss, D.J. (1977). Clustering on coloured lattices. *Journal of Applied Probability* **14**, 135-143.

Wakefield, J. (2006). Disease mapping and spatial regression with count data. *Biostatistics*, Advance Access published on June 29; also appear as Biostatistics working paper No. 286, University of Washington.

Zeger, S.L. (1988). A regression model for time series of counts. *Biometrika* **75**, 621-629.

Tables and Figures

Table 3.1
Potential covariates for the coho study example

GRADIENT	gradient
ACW	active channel width
ACH	active channel height
PCTPOOLS (%)	percent of pools in the reach
PCTSWPOOL (%)	percent of slackwater pools in the reach
PCTGRAVEL (%)	percent of gravel substrates in reach
PCTBEDROCK (%)	percent of bedrock substrate in the reach
POOL1P.KM	number of pools deeper than 1.0 meter/kilometer of total stream length
LWDPIECE1	pieces of large woody debris/100 meters of primary stream length
LWDVOL1	volume of large woody debris/100 meters of primary stream length
KEYLWD1	number of key pieces of large wood (> 0.59 in diameter and > 10 meters in length)/100 meters of primary stream length

Table 3.2
Repeated measurement structure for 32 sites

site ID	spawning years						
	1998	1999	2000	2001	2002	2003	2004
8056	x	x	x				
4006	x	x	x	x	x		
5323	x	x	x		x		x
2237	x			x			
2556		x	x	x	x		x
1844, 1196, 530, 608, 1984, 2278, 5165		x			x		
978, 1735, 663, 256, 2989, 3231, 2089, 2492, 5338, 5465, 8930	x			x			x
130, 545, 715, 3336, 5638, 4794, 7285, 7999, 9218	x	x	x	x	x		x

'x' indicates measurements were taken for the coho density and the habitat covariates at the site in the spawning year.

Table 3.3
Correlation matrix of all covariates

covariates	1	2	3	4	5	6	7	8	9	10	11
1. *GRADIENT*	1	-0.30	-0.16	-0.50	-0.13	-0.14	-0.11	-0.28	0.25	0.19	0.09
2. *ACW*		1	0.69	0.24	-0.06	-0.20	0.16	0.57	-0.20	-0.10	-0.05
3. *ACH*			1	0.18	-0.05	-0.16	0.03	0.46	-0.22	-0.17	-0.13
4. *PCTPOOLS*				1	0.47	0.02	0.19	0.42	-0.14	-0.15	-0.10
5. *PCTSWPOOL*					1	-0.09	-0.10	0.14	0.07	0.04	0.04
6. *PCTGRAVEL*						1	-0.41	-0.04	0.01	0.01	0.01
7. *PCTBEDROCK*							1	-0.03	-0.31	-0.25	-0.18
8. *POOL1P.KM*								1	-0.13	-0.08	-0.02
9. *LWDPIECE1*									1	0.83	0.65
10. *LWDVOL1*										1	0.91
11. *KEYLWD1*											1

Table 3.4
Variable selection using BIC

step	variable removed	BIC
0	(initial model)	818.52
1	$PCTGRAVEL^2$	811.18
2	*PCTGRAVEL*	807.29
3	$GRADIENT^2$	801.62
4	*GRADIENT*	796.17
5	ACW^2	793.31
6	*ACW*	788.07
7	ACH^2	783.79
8	*ACH*	779.59
9	$PCTBEDROCK^2$	775.71
10	$PCTSWPOOL^2$	772.71
11	*PCTSWPOOL*	768.89

Table 3.5
Parameter estimates

parameters	estimates
α_1	1.6051
d_1	0.7433
d_2	0.8596
D99	0.5258
D00	0.4082
D01	1.0615
D02	1.3933
D03	1.5613
D04	0.9542
PCTPOOLS	0.0247
$PCTPOOLS^2$	-0.0002
PCTBEDROCK	0.0120
POOL1P.KM	0.1247
$POOL1P.KM^2$	-0.0243
LWDPIECE1	0.0461
$LWDPIECE1^2$	-0.0010
c_s	35.8897
c_t	1.1682
log likelihood	-333.2609

Table 3.6
p-values from the LR test

Hypothesis (H_0)	log likelihood (under H_0)	p-value
D99-D04	-352.6399	< 0.0001
$PCTPOOLS^2$	-335.4921	0.0346
PCTPOOLS, $PCTPOOLS^2$	-336.1342	0.0565
PCTBEDROCK	-336.3133	0.0135
$POOL1P.KM^2$	-335.5858	0.0311
POOL1P.KM, $POOL1P.KM^2$	-336.0950	0.0588
$LWDPIECE1^2$	-335.7939	0.0244
LWDPIECE1, $LWDPIECE1^2$	-336.0371	0.0623
$c_s = +\infty$	-335.8059	0.0241
$c_t = +\infty$	-334.9942	0.0626
$c_s = +\infty$, $c_t = +\infty$	-339.1447	0.0014

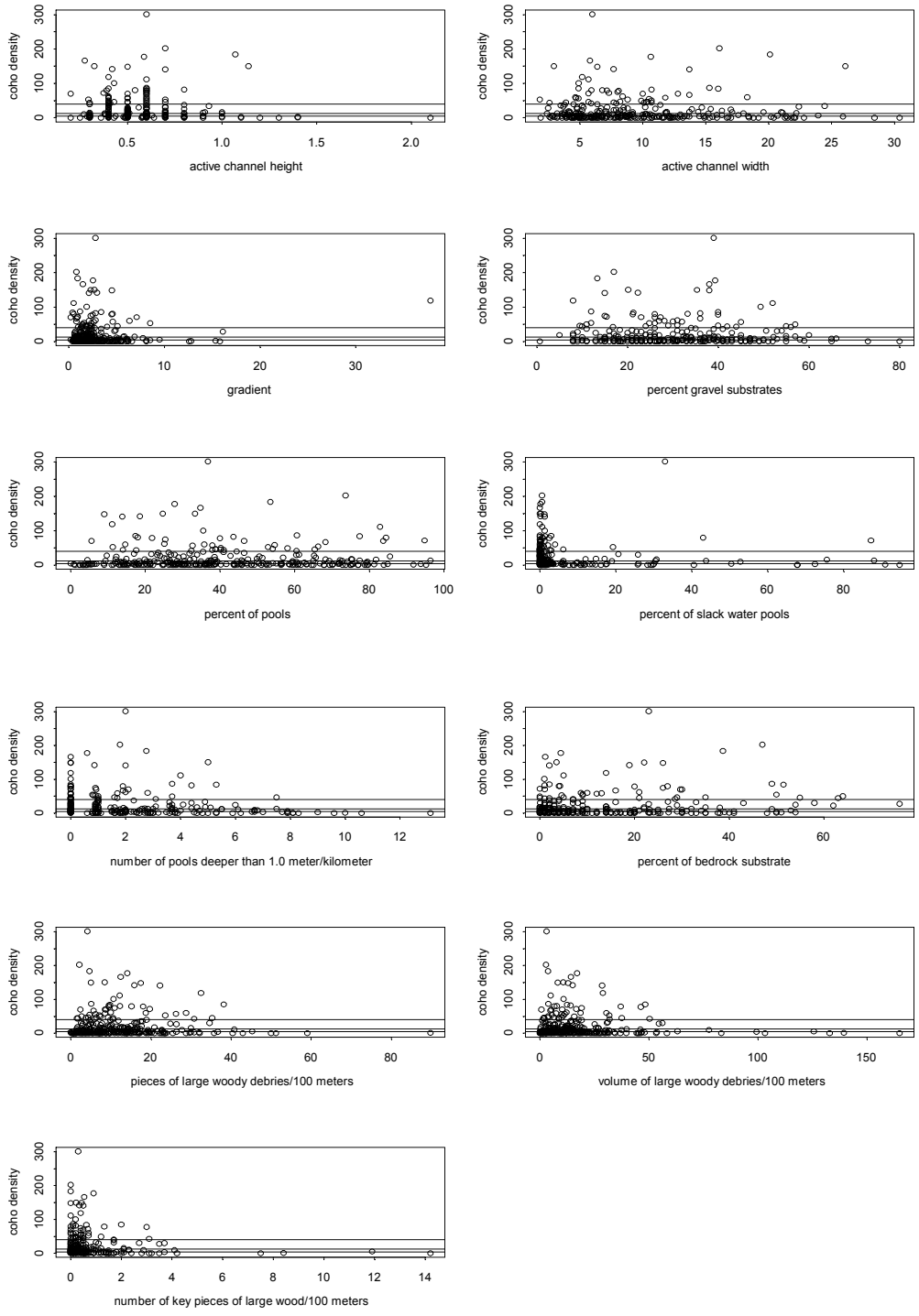

Figure 3.1 Scatterplots of coho density vs. habitat variables at 206 sites (totally 295 observations) (horizontal lines from top to bottom on each graph represent density 40, 12, and 4, respectively)

4. CONCLUSIONS

This dissertation considers the multivariate ordinal probit model for both repeated and spatial ordered categorical data. This model has been largely impractical due to the computational issues in the past. As a result, alternative models and approaches have been proposed for this type of data structure. For example, Agresti and Natarajan (2001) and Liu and Agresti (2005) survey various strategies for modeling ordered categorical data, particularly in a longitudinal study with repeated measures. Among all those alternatives, GEE is currently the most useful tool because it is easy to use (at least, relative to the other alternatives) due to its similarity with more familiar methods for independent responses. As far as the spatial categorical ordinal data are concerned, both GEE and the pseudo-likelihood approaches based on Markov random fields are equally popular (see, Kutsyy, 2001; Baeumler, 1995). But the latter is more important in modeling spatial-temporal ordinal data (see, for example, Ramos-Quiroga and González-Farías, 2005).

While most alternatives can be classified as quasi-likelihood and pseudo-likelihood approaches, fully-likelihood permit sensible models for the correlation structure, are potentially more efficient, permit familiar likelihood ratio inferences, and permit likelihood-based fitting criteria, such as AIC and BIC, for model selection. Potential efficiency gains are supported by our results from chapter 2 for the longitudinal ordinal data. Specifically, our simulations show MLE is more efficient and less sensitive to changes of the number of categories than GEE. In terms of MSE,

more improvement will be attained from using ML if the data has more categories and/or fewer repeated measures, though an increase of the degree of the within-subject correlation from a medium level to a high level gives mixed effects on the ML. It is worth noting that there is no evidence for more improvement of MSE from ML when each subject has more repeated measures. On the contrary, the simulations show less improvement in MSE due to a faster drop of MSE in GEE than in ML when the number of repeated measures increases. We also investigated the accuracy of tests based on SEs from the GEE and likelihood ratio tests, and found no conclusive results for the sample size we chose, though the likelihood ratio test tends to be more accurate in some scenarios.

The simulation study in chapter 2 was accomplished by our algorithm for estimating the multivariate ordinal probit model. Our experiences show the algorithm is practicable for maximum likelihood estimation and likelihood ratio inference in data analysis of repeated/spatial ordinal responses, and for further studying the relative merits of likelihood and GEE analysis as we did in our simulation study.

While chapter 2 focuses on the repeated ordered categorical data in the time domain, chapter 3 extends the model to the spatial-temporal data in both the time and space domains, and its application to spatial data analysis. Using the coho study example, we have shown spatial-temporal dependency of the ordinal responses can be simply incorporated in regression analysis through a multivariate normal distributed latent variable. This approach avoids defining neighbors, and is a fully likelihood analysis. Consequently, the familiar likelihood-based methods for testing fit,

comparing models (with AIC and BIC, for example), making inference about parameters are available with this approach.

Further research on the likelihood analysis of multivariate ordinal probit model for repeated/spatial ordinal categorical data can be continued in the following ways.

Evaluation of the Robustness of the Model

While the maximum likelihood (ML) method based on the multivariate ordinal probit model has a few attractive features mentioned in the previous two chapters, one may suspect its robustness. As we pointed out before, the multivariate normal latent variable might not be as strong of an assumption as it first seems. If there really is a latent variable (as there would be for categorizing hurricane strength from maximum wind speed, for example), it is only necessary that some monotonic function of the latent variable is normally distributed. Furthermore, the actual distribution of the latent variable might not matter much if the extreme response categories are not too strongly tied to the extreme tails of the latent variable distribution. Future research would be needed, though, to evaluate the robustness of the ML method.

Comparing the ML approach and the Pseudo-likelihood Approaches for the Spatial/Temporal Data from a Regular Lattice

Our current study focuses on comparing the GEE and the ML method in the repeated measurement setting. We have not explored the relative performance of the ML method with respect to the Pseudo-likelihood approaches in the spatial/temporal setting. Given ordinal responses from a regular lattice, the pseudo-likelihood approaches based on Markov random fields are particularly useful and popular

(Ramos-Quiroga and González-Farías, 2005). But the loss of efficiency of the pseudo-likelihood approaches with respect to fully likelihood ones may be an issue. Future research could be done to clarify this unresolved question.

Extending the Model to a Linear Mixed-effects One with Correlated Errors for the Latent Variable

The model we proposed contains the fixed-effects of covariates only. But sometimes a mixed-effects model along with correlated errors might be desired. For example, in the coho example, if the spawning years in the sample had been randomly selected, we could have treated the time effects as random, and a linear mixed-effects model with correlated errors would have been necessary. Therefore, we suggest future studies can focus on this type of model in a spatial-temporal setting.

BIBLIOGRAPHY

Agresti, A. (2002). *Categorical data analysis*, 2nd edition, John Wiley & Sons, Inc.

Agresti, A. and Natarajan, R. (2001). Modeling clustered ordered categorical data: A survey. *International Statistical Review* **69**, 345-371.

Alexander, N., Moyeed, R., and Stander, J. (2000). Spatial modeling of individual level parasite counts using the negative binomial distribution. *Biostatistics* **1**, 453-463.

Anderson, J.A. and Pemberton, J.D. (1985). The grouped continuous model for multivariate ordered categorical variables and covariate adjustment. *Biometrics* **41**, 875-885.

Baeumler, Alfred (1995). Marginal regression models for spatial or temporal correlated forest damage data. Spatial and Temporal Modelling in Agricultural Research, Proceedings of the Fourth HARMA Workshop, IACR-Rothamsted 20-21 October 1995, S.32-42.

Best, N.G., Ickstadt, K.I., and Wolpert, R.L. (2000). Spatial Poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American Statistical Association* **95**, 1076-1088

Brent, R. (1973). *Algorithms for Minimization without Derivatives*. Prentice-Hall, Englewood Cliffs, NJ.

Brewer, M.J., Elston, D.A., Hodgson, M. EA., Stolte, A.M., Nolan, A.J., and Henderson, D.J. (2004). A spatial model with ordinal responses for grazing impact data. *Statistical Modelling* **4**, 127-143.

Cressie, Noel (1991). *Statistics for Spatial Data*. John Wiley & Sons, New York.

Cressie, N. and Davidson, J. L. (1998). Image analysis with partially ordered Markov models. *Computational Statistics & Data Analysis* **29**, 1-26.

Crouchley, Robert (1995). A random-effects model for ordered categorical data. *Journal of the American Statistical Association* **90**, 489-498.

Davidson, J. L., Cressie, N. and Hua, X. (1999). Texture synthesis and pattern recognition for partially ordered Markov models. *Pattern Recognition* **34**, 1475-1505.

Davis, C.S. (1991). Semi-parametric and non-parametric methods for the analysis of repeated measurement with applications to clinical trials. *Statistics in Medicine* **10**, 1959-1980.

Dennis, J. E. and Schnabel, R. B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, NJ.

Efron, B. and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika* **65**, 457-482.

Elliot, D.S., Huizinga, D. and Menard, S. (1989). *Multiple Problem Youth: Delinquence, Substance Use and Mental Health Problems*. New York: Springer-Verlag.

Fu, T.T., Li, L.A., Li, Y.M., and Kan K. (2000). A limited information estimator for the multivariate ordinal probit model. *Applied Economics* **32**, 1841-1851.

Gelman, A. G., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*, 2nd edition, Chapman & Hall/CRC.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics* **1**, 141-149.

Genz, A. (1993). Comparison of methods for the computation of multivariate normal probabilities. *Computing Science and Statistics* **25**, 400-405.

Girard, Philippe and Parent, Eric (2001). Bayesian analysis of autocorrelated ordered categorical data for industrial quality monitoring. *Technometrics* **43**, 180-191.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association* **97**, 590-600.

Gumpertz, M. L., Wu, C. T. and Pye, J. M. (2000). Logistic regression for southern pine beetle outbreaks with spatial and temporal autocorrelation. *Forest Science* **46**, 95-107.

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

Hauspie, R.C., Cameron, N. and Molinari, L. (2004). *Methods in Human Growth Research*. Cambridge University Press.

Heagerty, P.J. and Zeger, S.L. (1996). Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association* **91**, 1024-1036.

Hedeker, Donald and Gibbons, Robert D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50**, 933-944.

Huang, H. and Cressie, N. (2000). Asymptotic properties of maximum (composite) likelihood estimators for partially ordered Markov models. *Statistica Sinica* **10**, 1325-1344.

Ising, Ernst (1925). Beitrag zur theories des ferromagnetismus. *Zeitschrift ür Physik* **31**, 253-258.

Kenward, M.G., Lesaffre, E., and Molenberghs, G. (1994). An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics* **50**, 945-953.

Kim, K. (1995). A bivariate cumulative probit regression model for ordered categorical data. *Statistics in Medicine* **14**, 1341-1352.

Kneib, Thomas and Fahrmeir, Ludwig. (2006). Structured additive regression for categorical space-time data: a mixed model approach. *Biometrics* **62**, 109-118.

Kutsyy, Vadim (2001). *Modeling and inference for spatial processes with ordinal data*, Ph.D. dissertation, University of Michigan.

Lang, J.B., McDonald, J.W. and Smith, P.W.F. (1999). Association modeling of multivariate categorical responses: a maximum likelihood approach. *Journal of the American Statistical Association* **94**, 1161-71.

Li, Y.H. and Schafer, D.W. (2006) Likelihood analysis of the multivariate ordinal probit regression model for repeated ordinal responses. Working paper, Statistics Department, Oregon State University.

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

Lindsay B.G. and Li, B. (1997). On second-order optimality of the observed Fisher information. *Annals of Statistics* **25**, 2172-2199.

Lipsitz, S.R., Kim, K., and Zhao, L. (1994). Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine* **13**, 1149-1163.

Liu, I. and Agresti, A. (2005). The analysis of ordered categorical data: An overview and a survey of recent developments. *Sociedad de Estadistica e Investigacion Operativa*. *Test* **14**, no.1, 1-73.

Long, J. Scott (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage publications.

Lumley, T. (1996). Generalized estimating equations for ordinal data: A note on working correlation structures. *Biometrics* **52**, 354-361.

Mark, S.D. and Gail, M.H. (1994). A comparison of likelihood-based and marginal estimating equation methods for analyzing repeated ordered categorical responses with missing data (Disc: p495-498). *Statistics in Medicine* **13**, 479-493.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of Royal Statistical Society*. *Series B* **42**, 109-142.

McFadden, D. (1989). A method of simulated moments for estimation of discrete choice response models without numerical integration. *Econometrica* **57**, 995-1027.

McShane, L.M., Albert, P.S., and Palmatier M.A. (1997). A latent process regression model for spatially correlated count data. *Biometrics* **53**, 698-706.

Miller, M.E., Davis, C.S., and Landis, J.R. (1993). The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares. *Biometrics* **49**, 1033-1044.

Mitchell, M.W. and Gumpertz, M.L. (2003). Spatio-temporal prediction inside a free-air CO_2 enrichment system. *Journal of Agricultural, Biological, and Environmental Statistics* **8** (3), 310-327.

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference: from a Neo-Fisherian Perspective*, World Scientific Publishing Company.

Peraza-Garay, F. (2004). A model for longitudinal and ordinal data with spatial dependency. *Disertación Doctoral*. CIMAT, Mexico.

Qu, Y., Piedmonte, M.R., and Medendorp, S.V. (1995). Latent variable models for clustered ordinal data. *Biometrics* **51**, 268-275.

Ramos-Quiroga, R. and González-Farías G. (2005). A case study: ordinal responses with spatio-temporal dependencies. *Communicación Tėcnica*, No. I-05-04/01-02-2005.

Schabenberger, Oliver and Gotway, Carol A. (2005). *Statistical methods for spatial data analysis*. Chapman & Hall/CRC Press.

Schnabel, R. B., Koontz, J. E., and Weiss, B. E. (1985). A modular system of algorithms for unconstrained minimization. *ACM Trans. Math. Software* **11**, 419-440.

Self, S.G. and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**, 605-610.

Skovgaard, I.M. (1985). A second-order investigation of asymptotic ancillarity. *Annals of Statistics* **13**, 534-551.

Strauss, D.J. (1977). Clustering on coloured lattices. *Journal of Applied Probability* **14**, 135-143.

Thompson, L.A. (2006). *S-PLUS (and R) Manual to Accompany Agresti's Categorical Data Analysis (2002) 2nd edition*. https://home.comcast.net/~lthompson221/#otherresearch

Toledano, A.Y. and Gatsonis, C. (1996). Ordinal regression methodology for ROC curves derived from correlated data. *Statistics in Medicine* **15**, 1807-1826.

Tutz, G. and Hennevogl, W. (1996). Random effects in ordinal regression models. *Computational Statistics & Data Analysis* **22**, 537-557.

Wakefield, J. (2006). Disease mapping and spatial regression with count data. *Biostatistics*, Advance Access published on June 29; also appear as Biostatistics working paper No. 286, University of Washington.

Williamson, J.M., Kim, K., and Lipsitz, S.R. (1995). Analyzing bivariate ordinal data using a global odds ratio. *Journal of the American Statistical Association* **90**, 1432-1437.

Zeger, S.L. (1988). A regression model for time series of counts. *Biometrika* **75**, 621-629.

APPENDICES

A1. Model Fitting for the Marijuana Use Example

Let y_{it} denote the response variable of subject *i* at time *t* (*i* = 1,…, 237; *t* = 1,…, 5), where *t* = 1,…, 5 correspond to year 1976,…, 1980, respectively. The latent variable approach of ML makes use of a (possibly fictitious) multivariate normally distributed latent variable $\underset{\sim}{z}_i \sim N_5(X\underset{\sim}{\beta}, \Sigma)$ for subject *i*, where *X* is a design matrix containing all covariates *time* and *gender*. It may include an interaction term $time \times gender$ and polynomial terms such as $time^2$. $\underset{\sim}{\beta}$ is an unknown parameter vector and Σ is an unknown correlation matrix for the latent variables from repeated measures taken in different years. Given the thresholds $-\infty = \alpha_0 < \alpha_1 < \alpha_2 < \alpha_3 = +\infty$, the latent variable approach assumes $y_{it} = g$ (*g* = 1, 2, 3) if and only if $\alpha_{g-1} < z_{it} \le \alpha_g$. It is convenient to reparameterize the unknown thresholds α_1 and α_2, to α_1 and $d^2 = \alpha_2 - \alpha_1$.

Using the AR(1) correlation structure, a LR test comparing a full model, which includes *time* (treated as a factor), *gender*, and their interaction terms, with a reduced model, which includes *time* (treated as a continuous covariate), *gender* and $time^2$, revealed that the reduced one is adequate (p-value = 0.94). Particularly, when we treated *time* as a continuous covariate, the interaction effect of $time \times gender$ was not significant. The ML estimates from the reduced model with AR(1) correlation structure are: $\hat{\alpha}_1 = 2.2077$, $\hat{\alpha}_2 = 2.8262$ ($\hat{d} = 0.7864$), $\hat{\beta}_1 = 0.6597$ (*time*), $\hat{\beta}_2 = 0.3797$ (*gender*), $\hat{\beta}_3 = -0.0603$ ($time^2$), and $\hat{\rho} = 0.8017$ (a correlation matrix parameter). More estimates can be found in Table A2, where the expected

information-based standard errors and the robust standard errors are reported in parentheses for the ML and the GEE, respectively. Compared with the ML estimates, the GEE approach gave similar estimates of thresholds and covariates coefficients.

A2. Simulation Details

Using the given ρ (0.5 or 0.8), G (3 or 6), and T (3 or 7), we generated the latent variable $z_i \sim N_T(\beta_1 time_i + \beta_2 gender_i + \beta_3 time_i^2, \Sigma)$ for $i = 1, \ldots, n$ (n = 500; 250 females and 250 males), where $\beta_1 = 0.66$, $\beta_2 = 0.38$, $\beta_3 = -0.06$,

$\Sigma = \begin{bmatrix} 1 & \rho & \ldots & \rho^{T-1} \\ \rho & 1 & \ldots & \rho^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \ldots & \rho & 1 \end{bmatrix}$. The cutoff points for the ordinal responses were set as follows.

$G = 3$: $-\infty = \alpha_0 < \alpha_1 = 2.21 < \alpha_2 = 2.83 < \alpha_3 = +\infty$;

$G = 6$:

$-\infty = \alpha_0 < \alpha_1 = 2.21 < \alpha_2 = 2.83 < \alpha_3 = 3.00 < \alpha_4 = 3.25 < \alpha_5 = 3.50 < \alpha_6 = +\infty$.

Note the parameters are $\alpha_1, \alpha_2, ..., \alpha_{G-1}, \beta_1, \beta_2, \beta_3$ and ρ. Suppose we are interested in estimating and testing the *gender* effect (i.e., the slope parameter β_2) only. As far as the testing is concerned, a null hypothesis could be $H_0 : \beta_2 = \beta_{2,0}$, where $\beta_{2,0}$ (= 0.38) is the true value of the parameter for our simulation. An alternative hypothesis could be $H_a : \beta_2 \neq \beta_{2,0}$ or $H_a : \beta_2 > \beta_{2,0}$ or $H_a : \beta_2 < \beta_{2,0}$. We investigated whether the true error rates for falsely rejecting the null hypothesis are 0.05 when (i) a Wald test is

applied in the GEE method and (ii) a LR test is applied in the ML method. Note the one-sided signed version (Z) of the usual likelihood ratio test statistic W is defined as $Z = \text{sgn}(\hat{\beta}_2 - \beta_{2,0})\sqrt{W}$ (Pace and Salvan, 1997, p. 92).

A3. Simulation Results on the Monte Carlo SD and the Averaged Reported SE

Patterns of MCSD and ARSE with respect to changes in ρ, T and G are similar to those of MSE. In other words, for both GEE and MLE, MCSD and ARSE increase if ρ increases from a medium level to a high level, or if T decreases. In contrast, GEE and MLE behavior differently when G changes. With an increase of G, MCSD and ARSE from GEE increase while those (esp. MCSD) from MLE are quite stable. For example, a change from G = 3 to 6 brings an increase of GEE MCSD from 0.1091 to 0.2296 for the case T = 3 and ρ =0.5, but the corresponding change of MLE MCSD is from 0.0957 to 0.0954. Moreover, the MCSD from MLE is smaller than that from GEE, which is also reflected in Figure 2.4 and Figure 2.5 (i.e., a narrower spread of the sampling distribution of MLE). This is generally true for the ARSE as well. Hence, in this sense, MLE is more efficient and less sensitive to changes of the number of categories than GEE.

While MLE outperforms GEE in getting a more accurate estimate, GEE does a better job of reporting the standard error of the estimate. Recall in our simulation, the SE for the GEE is the sandwich estimator, and the SE for the MLE is based on the expected information. Using the MCSD as a proxy for the SD of the sampling distribution of the estimate in Table 2.2, one can find the ARSE deviates from the SD

from -11.65% to -0.75% in GEE, and from -16.37% to 31.22% in MLE. Less accurate estimates of the SD of the sampling distribution of the estimate in MLE could be due to some computational errors from our numerical approximation of the derivatives in calculating the expected information.

A4. Simulation Results on Error Rates for Hypothesis Testing

Table A3 summarizes the actual error rates for falsely rejecting the null hypothesis $H_0 : \beta_2 = \beta_{2,0}$, where $\beta_{2,0}$ (= 0.38) is the true value of the parameter in our simulation. It shows the true error rates of the LR tests from ML tend to be closer to the nominal level 0.05 than those of the Wald tests from GEE. As rows "Either too large or too small" from Table A3 indicate, the 95% confidence interval of the true error rate of the LR test in the ML method contains the nominal level 0.05 in each of eight simulation scenarios. However, the 95% confidence interval of the true error rate of the Wald test in the GEE method excludes the nominal level 0.05 in two scenarios ($G = 3$, $T = 3$, $\rho = 0.8$; $G = 6$, $T = 3$, $\rho = 0.5$).

The preceding error rate is calculated as the chance that the test statistic is either too large or too small. While we reported this overall error rate in Table A3, we also kept track of the error rate due to a too small or a too large test statistic. Table A4 lists the exact p-values of testing whether the error rates in both directions are equal or not (see appendix A5 for the details of the calculation). It suggests both methods have a few occasions where the error rates in two directions are significantly different. It's not clear to us which one is better.

As far as the sensitivity of the true error rate to the changes in G, T and ρ is concerned, the GEE tends to be more sensitive than the ML due to the two extreme values (0.0840 and 0.0330) in the GEE method (see Table A3).

In addition to the two-sided true error rate, we also examined the one-sided rate. For the alternative hypothesis $H_a : \beta_2 > \beta_{2,0}$, the true error rate of falsely rejecting the null hypothesis $H_0 : \beta_2 = \beta_{2,0}$ is more likely to be close to the nominal level 0.05 in ML than in GEE. Table A3 shows there are four scenarios of the GEE where the 95% confidence interval of the error rate doesn't include the nominal level 0.05, compared with the two scenarios of the MLE. On the other hand, for the alternative hypothesis $H_a : \beta_2 < \beta_{2,0}$, all the eight confidence intervals of the error rate from the GEE include 0.05, while one confidence interval of the error rate from the MLE excludes 0.05. Therefore, the simulation results for the one-sided test error rate seem inconclusive.

A5. Calculation of the Exact p-values in Table A4

Take $G = 3$, $T = 3$, $\rho = 0.5$ as an example. In the 1000 iterations of the simulation, there are 26 (28) iterations where the test statistic is too large (small). Conditional on these total 54 (=26+28) iterations, the number of iterations with a too large test statistic follows a Binomial distribution with 54 trials and an unknown probability of "success" p. Therefore, testing whether the error rates in both directions are equal or

not is equivalent to testing $p = 0.5$. The exact p-values were obtained from function *binom.test* in S-PLUS.

A6. R Code for Likelihood Analysis of the Anesthesia Recovery Example with AR-1/Exchangeable Correlation Structure

```
# Anesthesia recovery study [source: Davis (1991)]
# Treat 'dosage' and 'time' as factors
# i.e., create indicator variables for Dosage and Time.
# include dosage*age and duration*time as interaction terms
# Davis (1991) includes only main effects of covariates. Tutz et
al.(1996) point out interaction effects can not be neglected.

# response in a multivariate form
y0 <-  c(3,3,1,1,5,3,6,1,1,2,1,3,1,1,0,1,1,2,1,5,
         1,6,4,1,1,6,0,3,1,2,1,3,2,1,2,6,3,2,1,0,
         0,1,1,1,0,4,2,4,1,2,3,0,0,0,1,0,1,1,3,0) # at time 0
y5 <-  c(5,4,1,3,6,3,6,1,1,2,3,3,1,3,2,1,1,3,1,6,
         1,6,4,4,2,6,0,6,1,2,1,6,3,1,3,6,5,3,1,2,
         0,1,2,1,3,6,4,5,1,3,4,5,0,0,1,0,1,1,5,1) # at time 5
y15 <- c(6,6,1,3,6,6,6,1,4,2,3,5,1,5,2,1,1,3,2,6,
         2,6,6,5,2,6,0,6,5,3,0,6,4,5,6,6,6,3,1,6,
         0,1,6,1,5,6,6,5,1,3,4,5,0,0,1,4,4,4,5,1) # at time15
y30 <- c(6,6,4,5,6,6,6,6,6,2,5,6,4,5,3,6,6,6,3,6,
         4,6,6,5,5,6,4,6,6,5,3,6,4,6,6,6,6,6,3,6,
         0,4,6,1,6,6,6,6,1,5,6,5,4,0,4,6,6,6,6,3) # at time 30
y <- cbind(y0,y5,y15,y30)
y <- y+1 # ordinal response: 1,2,...,7 rather than 0,1,...,6.

# covariates (D20 D25 D30 T5 T15 T30 Age Duration D20Age D25Age
D30Age T5Dur T15Dur T30Dur)
v1 <- c(0,0,0,0,0,0,36,128,0,0,0,0,0,0)
v2 <- c(0,0,0,1,0,0,36,128,0,0,0,128,0,0)
v3 <- c(0,0,0,0,1,0,36,128,0,0,0,0,128,0)
v4 <- c(0,0,0,0,0,1,36,128,0,0,0,0,0,128)
x1 <- c(v1,v2,v3,v4)

v1 <- c(0,0,0,0,0,0,35,70,0,0,0,0,0,0)
v2 <- c(0,0,0,1,0,0,35,70,0,0,0,70,0,0)
v3 <- c(0,0,0,0,1,0,35,70,0,0,0,0,70,0)
v4 <- c(0,0,0,0,0,1,35,70,0,0,0,0,0,70)
x2 <- c(v1,v2,v3,v4)

v1 <- c(0,0,0,0,0,0,54,138,0,0,0,0,0,0)
v2 <- c(0,0,0,1,0,0,54,138,0,0,0,138,0,0)
v3 <- c(0,0,0,0,1,0,54,138,0,0,0,0,138,0)
v4 <- c(0,0,0,0,0,1,54,138,0,0,0,0,0,138)
x3 <- c(v1,v2,v3,v4)
.
```

```
.
. # (x4,…,and x59 are omitted here)

v1 <- c(0,0,1,0,0,0,56,106,0,0,56,0,0,0)
v2 <- c(0,0,1,1,0,0,56,106,0,0,56,106,0,0)
v3 <- c(0,0,1,0,1,0,56,106,0,0,56,0,106,0)
v4 <- c(0,0,1,0,0,1,56,106,0,0,56,0,0,106)
x60 <- c(v1,v2,v3,v4)


x <- rbind(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,
           x11,x12,x13,x14,x15,x16,x17,x18,x19,x20,
           x21,x22,x23,x24,x25,x26,x27,x28,x29,x30,
           x31,x32,x33,x34,x35,x36,x37,x38,x39,x40,
           x41,x42,x43,x44,x45,x46,x47,x48,x49,x50,
           x51,x52,x53,x54,x55,x56,x57,x58,x59,x60)

n <- 60
num <- 4 # four repeated measures for each patient


# define negative log likelihood function assuming the independence
correlation structure
negloglik <- function(theta) {
      alpha1 <- theta[1]
      d1 <- theta[2]
      d2 <- theta[3]
      d3 <- theta[4]
      d4 <- theta[5]
      d5 <- theta[6]
      beta <- theta[7:20]
      a <- matrix(rep(0,n*num),n,num)
      b <- matrix(rep(0,n*num),n,num)
      for (i in 1:n) {
            for (t in 1:num) {
                  if (y[i,t]==1) {
                        a[i,t] <- -Inf
                        b[i,t] <- alpha1
                        }
                  if (y[i,t]==2) {
                        a[i,t] <- alpha1
                        b[i,t] <- alpha1+d1^2
                        }
                  if (y[i,t]==3) {
                        a[i,t] <- alpha1+d1^2
                        b[i,t] <- alpha1+d1^2+d2^2
                        }
                  if (y[i,t]==4) {
                        a[i,t] <- alpha1+d1^2+d2^2
                        b[i,t] <- alpha1+d1^2+d2^2+d3^2
                        }
                  if (y[i,t]==5) {
                        a[i,t] <- alpha1+d1^2+d2^2+d3^2
                        b[i,t] <- alpha1+d1^2+d2^2+d3^2+d4^2
```

```
                        }
                  if (y[i,t]==6) {
                        a[i,t] <- alpha1+d1^2+d2^2+d3^2+d4^2
                        b[i,t] <- alpha1+d1^2+d2^2+d3^2+d4^2+d5^2
                        }
                  if (y[i,t]==7) {
                        a[i,t] <- alpha1+d1^2+d2^2+d3^2+d4^2+d5^2
                        b[i,t] <- Inf
                        }
            }
      }

      value <- 0
      for (i in 1:n) {
            for (t in 1:num) {
            value <- value + log( pnorm(b[i,t],t(x[i,(14*(t-
1)+1):(14*t)])%*%beta,1) - pnorm(a[i,t],t(x[i,(14*(t-
1)+1):(14*t)])%*%beta,1))
            }
      }
      -value
}

# 1)use the GEE estimate (with indep. corr) as starting values
#alph1 <- -2.3618380
#alph2 <- -1.2278329
#alph3 <- -0.9220309
#alph4 <- -0.4575931
#alph5 <- -0.1240333
#alph6 <- 0.2657522
#d1 <- sqrt(alph2-alph1)
#d2 <- sqrt(alph3-alph2)
#d3 <- sqrt(alph4-alph3)
#d4 <- sqrt(alph5-alph4)
#d5 <- sqrt(alph6-alph5)
#result <- nlm(negloglik,c(alph1,d1,d2,d3,d4,d5,0.9431258,0.5180239,-
2.1161054,0.1641910,1.0478657,2.1176630,
#     -0.0167498,-0.0066485,-0.0378423,-
0.0200183,0.0386126,0.0042385,-0.0006770,-0.0029664),iterlim=500)

#negloglik(c(alph1,d1,d2,d3,d4,d5,0.9431258,0.5180239,-
2.1161054,0.1641910,1.0478657,2.1176630,
#     -0.0167498,-0.0066485,-0.0378423,-
0.0200183,0.0386126,0.0042385,-0.0006770,-0.0029664))

# 2) or we could use the MLE from an ordered probit model ignoring
the correlation structure
# could use ‘polr’ or ‘nlm’ (we used the latter in the following)
result <-
nlm(negloglik,c(0,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0),iterlim=500)
#result$estimate
if (result$code==4) {print("ITERATION LIMIT EXCEEDED--NOT
CONVERGE!!!!")}
```

```
alpha1.1 <- result$estimate[1]
d1.1 <- result$estimate[2]
d2.1 <- result$estimate[3]
d3.1 <- result$estimate[4]
d4.1 <- result$estimate[5]
d5.1 <- result$estimate[6]
beta.1 <- result$estimate[7:20]

alpha2.1 <- alpha1.1+d1.1^2
alpha3.1 <- alpha2.1+d2.1^2
alpha4.1 <- alpha3.1+d3.1^2
alpha5.1 <- alpha4.1+d4.1^2
alpha6.1 <- alpha5.1+d5.1^2

D20.1 <- result$estimate[7]
D25.1 <- result$estimate[8]
D30.1 <- result$estimate[9]
T5.1 <- result$estimate[10]
T15.1 <- result$estimate[11]
T30.1 <- result$estimate[12]
Age.1 <- result$estimate[13]
Duration.1 <- result$estimate[14]
D20Age.1 <- result$estimate[15]
D25Age.1 <- result$estimate[16]
D30Age.1 <- result$estimate[17]
T5Dur.1 <- result$estimate[18]
T15Dur.1 <- result$estimate[19]
T30Dur.1 <- result$estimate[20]


# AR(1) Correlation Structure
# note time=0, 5, 15, 30 minutes (unequal spaced)
sigma <- function(rho) {
      mat <- matrix(rep(0,num*num),num,num)
      mat[1,1] <- 1
      mat[1,2] <- rho
      mat[1,3] <- rho^3
      mat[1,4] <- rho^6
      mat[2,1] <- mat[1,2]
      mat[2,2] <- 1
      mat[2,3] <- rho^2
      mat[2,4] <- rho^5
      mat[3,1] <- mat[1,3]
      mat[3,2] <- mat[2,3]
      mat[3,3] <- 1
      mat[3,4] <- rho^3
      mat[4,1] <- mat[1,4]
      mat[4,2] <- mat[2,4]
      mat[4,3] <- mat[3,4]
      mat[4,4] <- 1
      mat
}

# Exchangeable working Correlation Structure
```

```
sigma <- function(rho) {
      mat <- matrix(rep(0,num*num),num,num)
      for (i in 1:num) {
            for (j in 1:num) {
                  if (i==j) {mat[i,j] <- 1 }
                  else {mat[i,j] <- rho}
            }
      }
      mat
}


# Full log likelihood function
loglik.full <- function(thetarho) {
      alpha1 <- thetarho[1]
      d1 <- thetarho[2]
      d2 <- thetarho[3]
      d3 <- thetarho[4]
      d4 <- thetarho[5]
      d5 <- thetarho[6]

      beta <- thetarho[7:20]
      rho <- thetarho[21]
      a <- matrix(rep(0,n*num),n,num)
      b <- matrix(rep(0,n*num),n,num)
      for (i in 1:n) {
            for (t in 1:num) {
                  if (y[i,t]==1) {
                        a[i,t] <- -Inf
                        b[i,t] <- alpha1
                        }
                  if (y[i,t]==2) {
                        a[i,t] <- alpha1
                        b[i,t] <- alpha1+d1^2
                        }
                  if (y[i,t]==3) {
                        a[i,t] <- alpha1+d1^2
                        b[i,t] <- alpha1+d1^2+d2^2
                        }
                  if (y[i,t]==4) {
                        a[i,t] <- alpha1+d1^2+d2^2
                        b[i,t] <- alpha1+d1^2+d2^2+d3^2
                        }
                  if (y[i,t]==5) {
                        a[i,t] <- alpha1+d1^2+d2^2+d3^2
                        b[i,t] <- alpha1+d1^2+d2^2+d3^2+d4^2
                        }
                  if (y[i,t]==6) {
                        a[i,t] <- alpha1+d1^2+d2^2+d3^2+d4^2
                        b[i,t] <- alpha1+d1^2+d2^2+d3^2+d4^2+d5^2
                        }
                  if (y[i,t]==7) {
                        a[i,t] <- alpha1+d1^2+d2^2+d3^2+d4^2+d5^2
                        b[i,t] <- Inf
```

```
                        }
                }
        }

        value <- 0
        for (i in 1:n) {
                xi <- x[i,]
                xinew <- rbind(xi[1:14],xi[15:28],xi[29:42],xi[43:56])
                value <-
value+log(pmvnorm(a[i,],b[i,],as.vector(xinew%*%beta),sigma(rho)))
                }
        value
}

# use optimize to get rho
# define a log likelihood function where alpha1=alpha1.1, d's=d.1's,
beta=beta.1
logl <- function(rho) {
        loglik.full(c(alpha1.1,d1.1,d2.1,d3.1,d4.1,d5.1,beta.1,rho))
        }

rho.1 <- optimize(logl,c(-1,1),maximum=TRUE)$maximum
rho.1


################################
convergence.criterion <- 0.0001
iter <- 2 # iteration number
maxdif <- 1 # initial setting for max abs. value
alpha1.old <- alpha1.1 # starting value of alpha1
d1.old <- d1.1 # starting value of d1
d2.old <- d2.1
d3.old <- d3.1
d4.old <- d4.1
d5.old <- d5.1
beta.old <- beta.1 # starting value of the slopes
rho.old <- rho.1 # starting value of rho

alpha1.iter <- alpha1.old
                # keep track of the estimate of alpha1 during
iteration
d1.iter <- d1.old # keep track of the estimate of d1 during iteration
d2.iter <- d2.old # keep track of the estimate of d2 during iteration
d3.iter <- d3.old # keep track of the estimate of d3 during iteration
d4.iter <- d4.old # keep track of the estimate of d4 during iteration
d5.iter <- d5.old # keep track of the estimate of d5 during iteration
beta1.iter <- beta.old[1]
                # keep track of the estimate of beta during iteration
beta2.iter <- beta.old[2]
beta3.iter <- beta.old[3]
beta4.iter <- beta.old[4]
beta5.iter <- beta.old[5]
beta6.iter <- beta.old[6]
beta7.iter <- beta.old[7]
```

```
beta8.iter <- beta.old[8]
beta9.iter <- beta.old[9]
beta10.iter <- beta.old[10]
beta11.iter <- beta.old[11]
beta12.iter <- beta.old[12]
beta13.iter <- beta.old[13]
beta14.iter <- beta.old[14]
rho.iter <- rho.old
code <- result$code
     # keep track of the occasions when the iteration limit is
exceeded

print(c("iteration", "alpha1", "d1",
"d2","d3","d4","d5","D20","D25","D30","T5","T15","T30","Age","Duratio
n","D20Age","D25Age","D30Age","T5Dur","T15Dur","T30Dur","rho", "log
likelihood","code"),sep="\t")
print(c(round(1,1),round(alpha1.old,4),round(d1.old,4),round(d2.old,4
), round(d3.old,4), round(d4.old,4), round(d5.old,4),
      round(beta.old,4),round(rho.old,4),
      round(loglik.full(c(alpha1.old,d1.old,d2.old,d3.old,d4.old,d5.o
ld,beta.old,rho.old)),5), round(code,1)),sep="\t")
val.old <-
loglik.full(c(alpha1.old,d1.old,d2.old,d3.old,d4.old,d5.old,beta.old,
rho.old))


##############################
while (maxdif>convergence.criterion)  {

      # update alpha1, d's, and beta
      f <- function(theta) {
            -loglik.full(c(theta,rho.old))
            }

      theta.old <- c(alpha1.old, d1.old, d2.old, d3.old, d4.old,
d5.old, beta.old)
      out <- nlm(f,theta.old,iterlim=500)
      alpha1.new <- out$estimate[1]
      d1.new <- out$estimate[2]
      d2.new <- out$estimate[3]
      d3.new <- out$estimate[4]
      d4.new <- out$estimate[5]
      d5.new <- out$estimate[6]
      beta.new <- out$estimate[7:20]
      code <- out$code

      # update rho
      loglikelihood <- function(rho) {
            loglik.full(c(alpha1.new, d1.new, d2.new, d3.new, d4.new,
d5.new, beta.new, rho))
            }
      rho.new <- optimize(loglikelihood,c(-1,1),maximum=TRUE)$maximum
```

```
      val.new <-
loglik.full(c(alpha1.new,d1.new,d2.new,d3.new,d4.new,d5.new,beta.new,
rho.new))


      # print current estimate
      print(c(round(iter,1),round(alpha1.new,4),round(d1.new,4),round
(d2.new,4), round(d3.new,4), round(d4.new,4), round(d5.new,4),
      round(beta.new,4),round(rho.new,4),
      round(loglik.full(c(alpha1.new,d1.new,d2.new,d3.new,d4.new,d5.n
ew,beta.new,rho.new)),4), round(code,1)),sep="\t")

      # keep track of the estimates
      alpha1.iter <- c(alpha1.iter,alpha1.new)
      d1.iter <- c(d1.iter, d1.new)
      d2.iter <- c(d2.iter, d2.new)
      d3.iter <- c(d3.iter, d3.new)
      d4.iter <- c(d4.iter, d4.new)
      d5.iter <- c(d5.iter, d5.new)
      beta1.iter <- c(beta1.iter,beta.new[1])
      beta2.iter <- c(beta2.iter,beta.new[2])
      beta3.iter <- c(beta3.iter,beta.new[3])
      beta4.iter <- c(beta4.iter,beta.new[4])
      beta5.iter <- c(beta5.iter,beta.new[5])
      beta6.iter <- c(beta6.iter,beta.new[6])
      beta7.iter <- c(beta7.iter,beta.new[7])
      beta8.iter <- c(beta8.iter,beta.new[8])
      beta9.iter <- c(beta9.iter,beta.new[9])
      beta10.iter <- c(beta10.iter,beta.new[10])
      beta11.iter <- c(beta11.iter,beta.new[11])
      beta12.iter <- c(beta12.iter,beta.new[12])
      beta13.iter <- c(beta13.iter,beta.new[13])
      beta14.iter <- c(beta14.iter,beta.new[14])

      rho.iter <- c(rho.iter, rho.new)

      # check convergence
      maxdif <- abs(val.new/val.old -1)

      iter <- iter+1
      alpha1.old <- alpha1.new
      d1.old <- d1.new
      d2.old <- d2.new
      d3.old <- d3.new
      d4.old <- d4.new
      d5.old <- d5.new
      beta.old <- beta.new
      rho.old <- rho.new
      val.old <- val.new
}

par(mfrow=c(3,2))
plot(1:length(alpha1.iter),alpha1.iter,xlab="iteration",ylab="Alpha1"
,type="l",xlim=c(0,40))
```

```
title (main="Alpha1")
plot(1:length(d1.iter),d1.iter,xlab="iteration",ylab="d1",type="l",xl
im=c(0,40))
title (main="d1")
plot(1:length(d2.iter),d2.iter,xlab="iteration",ylab="d2",type="l",xl
im=c(0,40))
title (main="d2")
plot(1:length(d3.iter),d3.iter,xlab="iteration",ylab="d3",type="l",xl
im=c(0,40))
title (main="d3")
plot(1:length(d4.iter),d4.iter,xlab="iteration",ylab="d4",type="l",xl
im=c(0,40))
title (main="d4")
plot(1:length(d5.iter),d5.iter,xlab="iteration",ylab="d5",type="l",xl
im=c(0,40))
title (main="d5")

par(mfrow=c(3,3))
plot(1:length(beta1.iter),beta1.iter,xlab="iteration",ylab="D20",type
="l",xlim=c(0,40))
title (main="D20")
plot(1:length(beta2.iter),beta2.iter,xlab="iteration",ylab="D25",type
="l",xlim=c(0,40))
title (main="D25")
plot(1:length(beta3.iter),beta3.iter,xlab="iteration",ylab="D30",type
="l",xlim=c(0,40))
title (main="D30")
plot(1:length(beta4.iter),beta4.iter,xlab="iteration",ylab="T5",type=
"l",xlim=c(0,40))
title (main="T5")
plot(1:length(beta5.iter),beta5.iter,xlab="iteration",ylab="T15",type
="l",xlim=c(0,40))
title (main="T15")
plot(1:length(beta6.iter),beta6.iter,xlab="iteration",ylab="T30",type
="l",xlim=c(0,40))
title (main="T30")
plot(1:length(beta7.iter),beta7.iter,xlab="iteration",ylab="Age",type
="l",xlim=c(0,40))
title (main="Age")
plot(1:length(beta8.iter),beta8.iter,xlab="iteration",ylab="Duration"
,type="l",xlim=c(0,40))
title (main="Duration")
plot(1:length(beta9.iter),beta9.iter,xlab="iteration",ylab="D20Age",t
ype="l",xlim=c(0,40))
title (main="D20Age")

par(mfrow=c(3,3))
plot(1:length(beta10.iter),beta10.iter,xlab="iteration",ylab="D25Age"
,type="l",xlim=c(0,40))
title (main="D25Age")
plot(1:length(beta11.iter),beta11.iter,xlab="iteration",ylab="D30Age"
,type="l",xlim=c(0,40))
title (main="D30Age")
```

```
plot(1:length(beta12.iter),beta12.iter,xlab="iteration",ylab="T5Dur",
type="l",xlim=c(0,40))
title (main="T5Dur")
plot(1:length(beta13.iter),beta13.iter,xlab="iteration",ylab="T15Dur"
,type="l",xlim=c(0,40))
title (main="T15Dur")
plot(1:length(beta14.iter),beta14.iter,xlab="iteration",ylab="T30Dur"
,type="l",xlim=c(0,40))
title (main="T30Dur")
plot(1:length(rho.iter),rho.iter,xlab="iteration",ylab="Rho",type="l"
,xlim=c(0,40))
title (main="Rho")


#################### compute standard errors ########################

### Expected (Fisher) Information Approach ###
alpha1hat <- alpha1.new
d1hat <- d1.new
d2hat <- d2.new
d3hat <- d3.new
d4hat <- d4.new
d5hat <- d5.new
alpha2hat <- alpha1hat+d1hat^2
alpha3hat <- alpha2hat+d2hat^2
alpha4hat <- alpha3hat+d3hat^2
alpha5hat <- alpha4hat+d4hat^2
alpha6hat <- alpha5hat+d5hat^2
betahat <- beta.new
rhohat <- rho.new

M <- 1000 # the larger the better. But it takes more time for a large
M
set.seed(1)
term <- 0
for (m in 1:M) {
      print(m)
      # simulate the latent responses so that we could simulate the
observed y
      z <- c() # null
      for (i in 1:n) {
            xi <- x[i,]
            xinew <- rbind(xi[1:14],xi[15:28],xi[29:42],xi[43:56])
            zi <- rmvnorm(1,as.vector(xinew%*%betahat),sigma(rhohat))
            z <- rbind(z,zi)
            }

      h <- function(alpha1,d1,d2,d3,d4,d5,beta,rho) {
            lower <- matrix(rep(0,n*num),n,num)
            upper <- matrix(rep(0,n*num),n,num)
            for (i in 1:n) {
                  for (t in 1:num) {
                        if (z[i,t]<= alpha1hat) {
                              lower[i,t] <- -Inf
```

```
                              upper[i,t] <- alpha1
                              }
                        if (z[i,t]> alpha1hat & z[i,t]<= alpha2hat) {
                              lower[i,t] <- alpha1
                              upper[i,t] <- alpha1+d1^2
                              }
                        if (z[i,t]> alpha2hat & z[i,t]<= alpha3hat) {
                              lower[i,t] <- alpha1+d1^2
                              upper[i,t] <- alpha1+d1^2+d2^2
                              }
                        if (z[i,t]> alpha3hat & z[i,t]<= alpha4hat) {
                              lower[i,t] <- alpha1+d1^2+d2^2
                              upper[i,t] <- alpha1+d1^2+d2^2+d3^2
                              }
                        if (z[i,t]> alpha4hat & z[i,t]<= alpha5hat) {
                              lower[i,t] <- alpha1+d1^2+d2^2+d3^2
                              upper[i,t] <-
alpha1+d1^2+d2^2+d3^2+d4^2
                              }
                        if (z[i,t]> alpha5hat & z[i,t]<= alpha6hat) {
                              lower[i,t] <-
alpha1+d1^2+d2^2+d3^2+d4^2
                              upper[i,t] <-
alpha1+d1^2+d2^2+d3^2+d4^2+d5^2
                              }

                        if (z[i,t] > alpha6hat){
                              lower[i,t] <-
alpha1+d1^2+d2^2+d3^2+d4^2+d5^2
                              upper[i,t] <- Inf
                              }
                        }
                  }
            value <- 0
            for (i in 1:n) {
                  xi <- x[i,]
                  xinew <-
rbind(xi[1:14],xi[15:28],xi[29:42],xi[43:56])
                  value <-
value+log(pmvnorm(lower[i,],upper[i,],as.vector(xinew%*%beta),sigma(r
ho)))
                  }
            value
      }

      # numerical derivative
      smidge <- 0.01/2
      #smidge2 <- 0.0001/2
      #alpha1,d1,d2,d3,d4,d5,beta,rho
      score1 <- (
h(alpha1hat+smidge,d1hat,d2hat,d3hat,d4hat,d5hat,betahat,rhohat)
                  - h(alpha1hat-
smidge,d1hat,d2hat,d3hat,d4hat,d5hat,betahat,rhohat) )/(2*smidge)
```

```
      score2 <- (
h(alpha1hat,d1hat+smidge,d2hat,d3hat,d4hat,d5hat,betahat,rhohat)
                  - h(alpha1hat,d1hat-
smidge,d2hat,d3hat,d4hat,d5hat,betahat,rhohat) )/(2*smidge)

      score3 <- (
h(alpha1hat,d1hat,d2hat+smidge,d3hat,d4hat,d5hat,betahat,rhohat)
                  - h(alpha1hat,d1hat,d2hat-
smidge,d3hat,d4hat,d5hat,betahat,rhohat) )/(2*smidge)

      score4 <- (
h(alpha1hat,d1hat,d2hat,d3hat+smidge,d4hat,d5hat,betahat,rhohat)
                  - h(alpha1hat,d1hat,d2hat,d3hat-
smidge,d4hat,d5hat,betahat,rhohat) )/(2*smidge)

      score5 <- (
h(alpha1hat,d1hat,d2hat,d3hat,d4hat+smidge,d5hat,betahat,rhohat)
                  - h(alpha1hat,d1hat,d2hat,d3hat,d4hat-
smidge,d5hat,betahat,rhohat) )/(2*smidge)

      score6 <- (
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat+smidge,betahat,rhohat)
                  - h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat-
smidge,betahat,rhohat) )/(2*smidge)

      score7 <- (
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat+c(smidge,0,0,0,0,0,
0,0,0,0,0,0,0,0),rhohat)
                  -
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat-
c(smidge,0,0,0,0,0,0,0,0,0,0,0,0,0),rhohat) )/(2*smidge)
      score8 <- (
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat+c(0,smidge,0,0,0,0,
0,0,0,0,0,0,0,0),rhohat)
                  -
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat-
c(0,smidge,0,0,0,0,0,0,0,0,0,0,0,0),rhohat) )/(2*smidge)
      score9 <- (
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat+c(0,0,smidge,0,0,0,
0,0,0,0,0,0,0,0),rhohat)
                  -
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat-
c(0,0,smidge,0,0,0,0,0,0,0,0,0,0,0),rhohat) )/(2*smidge)
      score10 <- (
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat+c(0,0,0,smidge,0,0,
0,0,0,0,0,0,0,0),rhohat)
                  -
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat-
c(0,0,0,smidge,0,0,0,0,0,0,0,0,0,0),rhohat) )/(2*smidge)
      score11 <- (
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat+c(0,0,0,0,smidge,0,
0,0,0,0,0,0,0,0),rhohat)
```

```
                  -
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat-
c(0,0,0,0,smidge,0,0,0,0,0,0,0,0,0),rhohat) )/(2*smidge)
      score12 <- (
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat+c(0,0,0,0,0,smidge,
0,0,0,0,0,0,0,0),rhohat)
                  -
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat-
c(0,0,0,0,0,smidge,0,0,0,0,0,0,0,0),rhohat) )/(2*smidge)
      score13 <- (
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat+c(0,0,0,0,0,0,smidg
e,0,0,0,0,0,0,0),rhohat)
                  -
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat-
c(0,0,0,0,0,0,smidge,0,0,0,0,0,0,0),rhohat) )/(2*smidge)
      score14 <- (
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat+c(0,0,0,0,0,0,0,smi
dge,0,0,0,0,0,0),rhohat)
                  -
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat-
c(0,0,0,0,0,0,0,smidge,0,0,0,0,0,0),rhohat) )/(2*smidge)
      score15 <- (
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat+c(0,0,0,0,0,0,0,0,s
midge,0,0,0,0,0),rhohat)
                  -
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat-
c(0,0,0,0,0,0,0,0,smidge,0,0,0,0,0),rhohat) )/(2*smidge)
      score16 <- (
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat+c(0,0,0,0,0,0,0,0,0
,smidge,0,0,0,0),rhohat)
                  -
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat-
c(0,0,0,0,0,0,0,0,0,smidge,0,0,0,0),rhohat) )/(2*smidge)
      score17 <- (
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat+c(0,0,0,0,0,0,0,0,0
,0,smidge,0,0,0),rhohat)
                  -
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat-
c(0,0,0,0,0,0,0,0,0,0,smidge,0,0,0),rhohat) )/(2*smidge)
      score18 <- (
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat+c(0,0,0,0,0,0,0,0,0
,0,0,smidge,0,0),rhohat)
                  -
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat-
c(0,0,0,0,0,0,0,0,0,0,0,smidge,0,0),rhohat) )/(2*smidge)
      score19 <- (
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat+c(0,0,0,0,0,0,0,0,0
,0,0,0,smidge,0),rhohat)
                  -
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat-
c(0,0,0,0,0,0,0,0,0,0,0,0,smidge,0),rhohat) )/(2*smidge)
      score20 <- (
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat+c(0,0,0,0,0,0,0,0,0
,0,0,0,0,smidge),rhohat)
```

```
                  -
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat-
c(0,0,0,0,0,0,0,0,0,0,0,0,smidge),rhohat) )/(2*smidge)

      score21 <- (
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat,rhohat+smidge)
                  -
h(alpha1hat,d1hat,d2hat,d3hat,d4hat,d5hat,betahat,rhohat-smidge)
)/(2*smidge)

      score <- c( score1, score2, score3, score4, score5,
                  score6, score7, score8, score9, score10,
                  score11, score12, score13, score14, score15,
                  score16, score17, score18, score19, score20,
score21 )

      # term score%*%t(score)
      term <- term + score%*%t(score)

}

Info <- term/M # Fisher information
stderror <- sqrt(diag(solve(Info)))
stderror


### Observed Information Approach ###
theta.hat <- c(alpha1hat, d1hat, d2hat, d3hat, d4hat, d5hat, betahat,
rhohat)
p <- length(theta.hat)
mat <- matrix(rep(0,p*p),p,p)
for (i in 1:p) {
      for (j in i:p) {
            delta <- 0.006 #
            I <- diag(1,p,p)
            ei <- I[,i]
            ej <- I[,j]
            term1 <- loglik.full(theta.hat+delta*ei+delta*ej)
            term2 <- loglik.full(theta.hat-delta*ei+delta*ej)
            term3 <- loglik.full(theta.hat+delta*ei-delta*ej)
            term4 <- loglik.full(theta.hat-delta*ei-delta*ej)
            mat[i,j] <- (term1-term2-term3+term4)/(4*delta^2)
            mat[j,i] <- mat[i,j]
      }
}
obsInf <- -1*mat
stderror <- sqrt(diag(solve(obsInf)))
stderror
```

Tables in the Appendices

Table A1

Data on marijuana use in the past year and gender, taken from five yearly waves of the National Youth Survey

Gender[a]	1976[b]	1977	1978	1979	1980	Frequency	Gender	1976	1977	1978	1979	1980	Frequency
0	1	1	1	1	1	63	1	1	1	1	3	1	1
0	1	1	1	1	2	10	1	1	1	1	3	3	1
0	1	1	1	1	3	3	1	1	1	2	1	3	1
0	1	1	1	2	1	4	1	1	1	2	2	1	2
0	1	1	1	2	2	2	1	1	1	2	2	2	2
0	1	1	1	3	1	1	1	1	1	2	2	3	1
0	1	1	1	3	2	1	1	1	1	2	3	3	5
0	1	1	1	3	3	3	1	1	1	3	1	2	1
0	1	1	2	1	1	2	1	1	1	3	2	2	1
0	1	1	2	1	2	2	1	1	1	3	3	3	3
0	1	1	2	2	1	3	1	1	2	1	1	2	1
0	1	1	2	2	2	7	1	1	2	1	2	1	1
0	1	1	2	2	3	1	1	1	2	2	1	1	2
0	1	1	2	3	3	1	1	1	2	2	2	1	1
0	1	2	1	1	1	1	1	1	2	2	1	3	1
0	1	2	1	1	2	2	1	1	2	2	3	3	1
0	1	2	2	1	2	1	1	1	2	3	2	2	1
0	1	2	2	2	1	1	1	1	2	3	2	3	1
0	1	2	2	3	3	2	1	1	2	3	3	2	1
0	1	2	3	1	2	1	1	1	2	3	3	3	4
0	1	2	3	3	2	1	1	1	3	1	3	3	1
0	1	2	3	3	3	1	1	1	3	2	2	2	1
0	1	3	3	2	2	1	1	1	3	3	3	3	2
0	2	1	1	3	3	1	1	1	3	3	2	2	1
0	2	1	2	2	2	1	1	2	1	1	1	1	3
0	2	1	3	3	3	1	1	2	2	2	2	2	1
0	2	3	3	3	3	1	1	2	2	3	3	3	1
0	2	3	3	3	2	1	1	2	3	2	1	1	1
0	3	3	3	2	3	1	1	2	3	2	3	3	1
1	1	1	1	1	1	48	1	2	3	3	3	3	2
1	1	1	1	1	2	8	1	3	1	1	1	1	1
1	1	1	1	1	3	4	1	3	2	3	3	3	1
1	1	1	1	2	1	2	1	3	3	3	3	1	1
1	1	1	1	2	2	4	1	3	3	3	3	3	1
1	1	1	1	2	3	1							

[a] 0, female; 1, male.

[b] 1, never; 2, not more than once a month; 3, more than once a month

Table A2

GEE and ML estimates for the marijuana use data

	GEE* independent	GEE* exchangeable	GEE* AR-1	ML Independent	ML Exchangeable	ML AR-1
α_1	2.2775 (0.2558)	2.3032 (0.2519)	2.3134 (0.2599)	2.2075 (0.2256)	2.2084 (0.1594)	2.2077 (0.1436)
α_2	2.8969 (0.2683)	2.9126 (0.2604)	2.9230 (0.2688)			
d				0.7865 (0.0268)	0.7865 (0.0306)	0.7864 (0.0302)
time	0.6798 (0.1434)	0.6517 (0.1288)	0.6590 (0.1311)	0.6591 (0.1572)	0.6602 (0.0976)	0.6597 (0.0910)
gender	0.4418 (0.1427)	0.5513 (0.1807)	0.5425 (0.1768)	0.3797 (0.0807)	0.3806 (0.1193)	0.3797 (0.1096)
$time^2$	-0.0638 (0.0204)	-0.0602 (0.0186)	-0.0609 (0.0185)	-0.0604 (0.0247)	-0.0612 (0.0151)	-0.0603 (0.0142)
ρ		**	**		0.7055 (0.0371)	0.8017 (0.0266)
log likelihood	N.A.	N.A.	N.A.	-833.3125	-680.6168	-661.3866

* Estimates were obtained from using the *ordgee* function from the *geepack* library in R.

** Association among repeated measures was modeled through global odds ratio (see the details in R documentation).

Table A3

True error rate for testing $H_0 : \beta_2 = \beta_{2,0}$ at level 0.05

	$G = 3$				$G = 6$			
	$T = 3$		$T = 7$		$T = 3$		$T = 7$	
	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$
GEE (Wald test)								
(1) $H_a : \beta_2 \neq \beta_{2,0}$								
Test stat. is too large	0.0260 (0.0161, 0.0359)*	0.0570 (0.0426, 0.0714)	0.0320 (0.0211, 0.0429)	0.0340 (0.0228, 0.0452)	0.0100 (0.0038, 0.0162)	0.0310 (0.0203, 0.0417)	0.0330 (0.0219, 0.0441)	0.0350 (0.0236, 0.0464)
Test stat. is too small	0.0280 (0.0178, 0.0382)	0.0270 (0.0170, 0.0370)	0.0200 (0.0113, 0.0287)	0.0250 (0.0153, 0.0347)	0.0230 (0.0137, 0.0323)	0.0290 (0.0186, 0.0394)	0.0220 (0.0129, 0.0311)	0.0210 (0.0121, 0.0299)
Either too large or too small	0.0540 (0.0400, 0.0680)	0.0840 (0.0668, 0.1012)	0.0520 (0.0382, 0.0658)	0.0590 (0.0444, 0.0736)	0.0330 (0.0219, 0.0441)	0.0600 (0.0453, 0.0747)	0.0550 (0.0409, 0.0691)	0.0560 (0.0417, 0.0703)
(2) $H_a : \beta_2 > \beta_{2,0}$	0.0520 (0.0382, 0.0658)	0.0950 (0.0768, 0.1132)	0.0540 (0.0400, 0.0680)	0.0560 (0.0417, 0.0703)	0.0360 (0.0245, 0.0475)	0.0720 (0.0560, 0.0880)	0.0580 (0.0435, 0.0725)	0.0680 (0.0524 0.0836)
(3) $H_a : \beta_2 < \beta_{2,0}$	0.0480 (0.0348, 0.0612)	0.0450 (0.0322, 0.0578)	0.0450 (0.0322, 0.0578)	0.0470 (0.0339, 0.0601)	0.0460 (0.0330, 0.0590)	0.0510 (0.0374, 0.0646)	0.0420 (0.0296, 0.0544)	0.0420 (0.0296, 0.0544)
ML (LR test)								
(1) $H_a : \beta_2 \neq \beta_{2,0}$								
Test stat. is too large	0.0310 (0.0203, 0.0417)	0.0380 (0.0261, 0.0499)	0.0200 (0.0113, 0.0287)	0.0210 (0.0121, 0.0299)	0.0270 (0.0170, 0.0370)	0.0340 (0.0228, 0.0452)	0.0360 (0.0245, 0.0475)	0.0300 (0.0194, 0.0406)
Test stat. is too small	0.0310 (0.0203, 0.0417)	0.0240 (0.0145, 0.0335)	0.0280 (0.0178, 0.0382)	0.0240 (0.0145, 0.0335)	0.0270 (0.0170, 0.0370)	0.0200 (0.0113, 0.0287)	0.0200 (0.0113, 0.0287)	0.0220 (0.0129, 0.0311)
Either too large or too small	0.0620 (0.0471, 0.0769)	0.0620 (0.0471, 0.0769)	0.0480 (0.0348, 0.0612)	0.0450 (0.0322, 0.0578)	0.0540 (0.0400, 0.0680)	0.0540 (0.0400, 0.0680)	0.0560 (0.0417, 0.0703)	0.0520 (0.0382, 0.0658)
(2) $H_a : \beta_2 > \beta_{2,0}$	0.0540 (0.0400, 0.0680)	0.0810 (0.0641, 0.0979)	0.0490 (0.0356, 0.0624)	0.0490 (0.0356, 0.0624)	0.0520 (0.0382, 0.0658)	0.0580 (0.0435, 0.0725)	0.0700 (0.0542, 0.0858)	0.0650 (0.0497, 0.0803)
(3) $H_a : \beta_2 < \beta_{2,0}$	0.0550 (0.0409, 0.0691)	0.0420 (0.0296, 0.0544)	0.0540 (0.0400, 0.0680)	0.0480 (0.0348, 0.0612)	0.0540 (0.0400, 0.0680)	0.0370 (0.0253, 0.0487)	0.0460 (0.0330, 0.0590)	0.0460 (0.0330, 0.0590)

*: numbers in parentheses are 95% confidence intervals.

Table A4

Exact p-values for testing equal chances of rejection in both directions

	$G = 3$				$G = 6$			
	$T = 3$		$T = 7$		$T = 3$		$T = 7$	
	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$
GEE (Wald test)	0.8919	0.0014	0.1263	0.2976	0.0351	0.8974	0.1770	0.0814
MLE (LR test)	1.0000	0.0980	0.3123	0.7660	1.0000	0.0759	0.0440	0.3317

Table A5

The coho study data

GCG	STRATA	ID_NUM	SPAWNINGYE	COHOAUC_MI	GRADIENT	ACW	ACH	PCTPOOLS	PCTSWPOOL	PCTGRAVEL	PCTBEDROCK	POOL1P_KM	LWDPIECE1	LWDVOL1
1-NC	NorthCoast	1832	2001	52.5	8.5	8.8	0.7	11.3	0.4	26.00	9.00	0.90	23.6	31.8
1-NC	NorthCoast	1833	1999	1.9	2.7	8.6	0.5	15.1	2.4	23.00	3.00	2.40	30.8	48.0
1-NC	NorthCoast	1844	1999	12.6	0.9	18.0	0.8	30.6	1.3	14.00	14.00	3.00	4.7	8.8
1-NC	NorthCoast	1844	2002	7.2	1.7	17.6	0.7	39.1	0.0	41.00	6.00	4.00	2.8	5.8
1-NC	NorthCoast	1846	2000	1.4	4.9	5.0	0.6	17.0	0.1	33.00	4.00	0.00	17.0	7.2
1-NC	NorthCoast	1877	2002	44.7	1.8	3.9	0.4	40.2	0.0	55.00	0.00	0.00	10.6	8.6
1-NC	NorthCoast	1884	2000	13.2	1.0	11.9	0.7	77.4	0.3	34.00	10.00	2.60	11.0	12.8
1-NC	NorthCoast	1896	1999	8.5	1.0	8.9	0.4	26.2	1.5	37.00	0.00	2.50	10.8	15.4
1-NC	NorthCoast	902	1998	0.0	4.9	4.5	0.4	13.0	0.0	21.00	0.00	0.00	50.0	165.0
1-NC	NorthCoast	922	1998	0.0	2.1	6.5	0.4	29.7	1.7	26.00	1.00	0.00	9.6	6.7
1-NC	NorthCoast	978	1998	0.0	3.4	5.9	0.6	19.6	1.6	15.00	4.00	1.90	20.9	39.6
1-NC	NorthCoast	978	2001	0.0	4.2	6.0	0.6	19.0	0.0	45.00	4.00	0.00	15.3	19.5
1-NC	NorthCoast	978	2004	0.0	3.2	4.9	0.3	14.0	0.0	39.51	1.73	0.00	23.4	24.6
1-NC	NorthCoast	996	2002	10.4	3.7	8.1	0.4	20.4	0.3	43.00	10.00	0.00	7.0	12.8
1-NC	NorthCoast	1041	2003	30.6	1.5	7.2	0.6	61.2	3.1	50.61	0.08	0.96	15.9	8.5
1-NC	NorthCoast	1045	2001	0.0	2.2	8.2	0.6	48.5	0.7	46.00	0.00	0.00	13.6	9.7
1-NC	NorthCoast	1075	2001	4.5	3.0	11.3	0.9	28.8	1.0	20.00	29.00	0.80	13.6	19.4
1-NC	NorthCoast	1092	2001	0.0	4.2	11.6	0.9	33.4	0.6	35.00	15.00	0.00	13.9	18.8
1-NC	NorthCoast	1123	1999	4.0	1.6	13.3	0.5	36.1	2.1	39.00	4.00	1.90	36.7	99.4
1-NC	NorthCoast	1151	1998	0.0	0.6	21.3	0.8	56.3	1.7	23.00	4.00	4.60	23.0	15.6
1-NC	NorthCoast	1166	2002	58.5	0.8	11.8	0.5	54.6	0.2	30.00	13.00	1.70	8.6	7.4
1-NC	NorthCoast	1196	1999	6.7	6.1	6.8	0.4	11.9	0.0	18.00	19.00	1.50	21.6	7.2
1-NC	NorthCoast	1196	2002	44.5	5.1	7.7	0.4	14.0	0.0	24.00	21.00	1.70	35.3	37.6
1-NC	NorthCoast	1221	2001	2.6	0.7	4.6	0.7	93.8	72.5	50.00	1.00	7.90	10.6	8.5
1-NC	NorthCoast	1226	1998	9.4	0.9	13.0	0.7	74.1	0.4	22.00	2.00	6.70	27.8	37.4
1-NC	NorthCoast	1241	2002	0.0	4.5	10.7	0.3	34.3	28.3	35.00	11.00	0.00	23.8	28.0
1-NC	NorthCoast	1279	1998	0.0	4.7	5.2	0.6	3.3	0.4	9.00	11.00	0.80	13.6	45.5
1-NC	NorthCoast	1337	2000	0.0	2.2	5.1	0.6	77.1	67.8	11.00	0.00	1.10	17.0	17.8
1-NC	NorthCoast	1371	2003	79.0	4.5	7.2	0.4	21.8	2.5	21.20	26.96	0.00	18.1	37.1
1-NC	NorthCoast	1401	1999	0.0	4.1	9.1	0.4	67.8	67.8	13.00	0.00	1.60	51.2	132.9
1-NC	NorthCoast	1468	1999	3.0	0.8	9.3	0.5	45.2	1.0	41.00	3.00	1.80	15.8	18.0
1-NC	NorthCoast	1470	2000	11.3	2.6	10.2	0.6	44.6	19.0	37.00	9.00	2.00	15.7	6.6
1-NC	NorthCoast	1504	2001	35.6	2.4	5.2	0.6	46.5	0.3	39.00	4.00	0.00	17.9	16.4
1-NC	NorthCoast	1517	1998	3.9	1.9	6.5	0.5	43.7	0.3	10.00	6.00	2.30	19.7	14.3
1-NC	NorthCoast	1535	2001	15.2	3.9	3.8	0.5	81.8	75.6	39.00	2.00	0.00	30.0	26.7
1-NC	NorthCoast	1555	1998	6.2	1.9	5.1	0.5	36.2	0.1	38.00	5.00	0.00	15.0	17.9
1-NC	NorthCoast	1569	2002	31.8	1.7	22.4	0.7	21.6	20.7	31.00	33.00	3.60	5.2	14.6
1-NC	NorthCoast	1598	2003	177.6	2.5	10.7	0.6	27.9	0.2	39.33	4.39	0.60	14.1	17.0
1-NC	NorthCoast	1606	2001	78.0	2.5	8.1	0.6	40.0	0.8	40.00	20.00	0.00	9.5	46.3
1-NC	NorthCoast	1640	2000	79.6	0.5	6.8	0.6	84.5	43.0	26.00	4.00	1.90	15.2	6.6
1-NC	NorthCoast	1648	1998	1.0	1.6	5.5	0.5	40.2	0.2	22.00	1.00	0.00	89.7	139.5
1-NC	NorthCoast	1652	1999	3.2	0.4	13.0	0.5	75.6	6.5	30.00	7.00	7.70	14.1	9.4
1-NC	NorthCoast	1677	2000	111.4	0.5	5.7	0.6	82.9	0.6	52.00	5.00	4.00	12.2	4.9
1-NC	NorthCoast	1709	2003	71.7	0.9	7.3	0.4	94.8	87.2	15.38	0.00	0.00	19.7	16.6
1-NC	NorthCoast	1735	1998	3.0	1.1	15.0	0.8	58.5	2.4	17.00	17.00	3.70	12.6	20.8
1-NC	NorthCoast	1735	2001	17.0	1.5	14.4	0.7	68.3	2.0	25.00	22.00	1.60	20.9	14.3
1-NC	NorthCoast	1735	2004	70.0	0.9	12.8	0.5	46.4	0.2	27.30	29.66	0.93	8.8	5.2
1-NC	NorthCoast	107	2001	1.1	1.4	5.9	0.5	31.2	14.7	47.00	0.00	2.70	0.8	0.7
1-NC	NorthCoast	130	1998	0.0	4.3	5.1	0.2	30.7	0.0	42.00	3.00	0.00	6.3	12.1
1-NC	NorthCoast	130	1999	4.0	3.1	4.4	0.3	16.6	1.0	23.00	0.00	0.00	6.5	9.5
1-NC	NorthCoast	130	2000	0.0	5.4	3.2	0.3	13.6	0.5	23.00	7.00	0.00	7.9	9.0
1-NC	NorthCoast	130	2001	4.0	4.9	4.6	0.3	20.1	0.0	43.00	5.00	2.20	4.2	19.1
1-NC	NorthCoast	130	2002	0.0	5.6	4.4	0.4	9.0	0.0	43.00	12.00	0.00	7.1	9.1
1-NC	NorthCoast	130	2004	0.0	5.4	4.0	0.4	3.9	0.0	26.56	5.16	0.00	7.8	13.8
1-NC	NorthCoast	133	2001	2.6	0.5	2.6	0.7	61.1	30.0	38.00	0.00	0.00	0.0	0.0
1-NC	NorthCoast	179	1998	0.0	0.5	17.0	0.3	79.6	0.3	47.00	6.00	2.90	19.4	26.1
1-NC	NorthCoast	210	2000	2.5	3.8	6.2	0.3	36.0	17.4	41.00	15.00	9.00	32.0	46.8
1-NC	NorthCoast	214	2000	0.0	1.9	2.7	0.2	4.2	0.0	65.00	0.00	0.00	2.9	2.7
1-NC	NorthCoast	252	2000	0.0	3.7	5.3	0.4	3.3	0.0	22.00	2.00	0.00	15.0	46.3
1-NC	NorthCoast	256	1998	0.0	1.0	15.9	0.5	61.7	0.7	42.00	4.00	10.60	7.5	10.9
1-NC	NorthCoast	256	2001	1.2	0.7	22.0	1.4	58.7	0.4	49.00	6.00	10.00	3.5	9.4
1-NC	NorthCoast	256	2004	1.2	0.5	18.4	1.0	60.1	0.0	30.62	6.79	8.09	3.7	2.3
1-NC	NorthCoast	276	2001	8.4	2.5	15.9	0.8	23.3	1.0	36.00	7.00	1.60	7.6	31.6
1-NC	NorthCoast	284	1999	1.5	1.0	16.8	0.8	29.7	2.1	27.00	19.00	3.00	0.9	1.3
1-NC	NorthCoast	363	2001	14.7	5.8	21.1	0.9	4.5	0.3	26.00	1.00	0.00	12.1	20.3
1-NC	NorthCoast	378	1999	140.7	2.1	13.7	0.7	13.8	1.3	15.00	2.00	2.00	22.3	28.5
1-NC	NorthCoast	403	2002	0.0	3.6	2.7	0.6	4.4	0.5	73.00	5.00	0.00	7.0	5.1
1-NC	NorthCoast	425	2000	0.0	4.6	5.5	0.6	18.2	0.1	29.00	0.00	1.10	28.8	15.0
1-NC	NorthCoast	476	2000	6.0	3.4	8.6	0.6	18.5	0.1	42.00	1.00	0.00	15.1	7.2
1-NC	NorthCoast	518	2000	1.4	12.8	4.2	0.4	5.4	0.1	46.00	12.00	0.00	17.1	24.9
1-NC	NorthCoast	530	1999	4.2	5.3	14.5	0.9	8.9	1.6	8.00	9.00	0.00	12.1	11.5
1-NC	NorthCoast	530	2002	3.2	6.1	8.0	0.4	5.8	0.0	21.00	4.00	0.00	19.0	53.5
1-NC	NorthCoast	545	1998	10.9	2.4	10.3	0.5	30.4	0.2	34.00	4.00	1.80	40.8	46.1
1-NC	NorthCoast	545	1999	10.9	1.6	9.2	0.5	34.9	0.6	22.00	2.00	6.10	16.1	10.5

(Continued)

GCG	STRATA	ID_NUM	SPAWNINGYE	COHOAUC_MI	GRADIENT	ACW	ACH	PCTPOOLS	PCTSWPOOL	PCTGRAVEL	PCTBEDROCK	POOL1P_KM	LWDPIECE1	LWDVOL1
1-NC	NorthCoast	545	2000	0.0	1.9	8.2	0.6	30.2	0.5	22.00	3.00	5.90	40.4	36.9
1-NC	NorthCoast	545	2001	0.0	1.9	9.9	0.5	33.7	0.0	42.00	5.00	3.00	15.0	31.5
1-NC	NorthCoast	545	2002	15.6	2.4	9.2	0.6	32.0	0.2	55.00	1.00	3.70	13.4	25.7
1-NC	NorthCoast	545	2004	12.5	2.1	7.1	0.4	28.5	0.3	52.19	2.96	3.66	17.4	15.2
1-NC	NorthCoast	588	1998	0.0	5.5	6.1	0.4	26.3	0.2	25.00	8.00	0.00	35.8	58.2
1-NC	NorthCoast	608	1999	0.0	1.8	20.4	0.9	44.0	0.0	8.00	17.00	7.90	5.3	8.2
1-NC	NorthCoast	608	2002	46.4	2.2	10.3	0.7	38.8	0.0	40.00	5.00	7.50	4.4	6.5
1-NC	NorthCoast	624	2002	148.2	4.5	6.4	0.5	9.0	1.1	38.00	26.00	0.00	17.5	12.9
1-NC	NorthCoast	663	1998	0.0	3.0	10.5	0.6	36.0	0.0	21.00	12.00	6.20	11.0	21.3
1-NC	NorthCoast	663	2001	0.0	3.0	13.2	0.8	26.3	0.0	19.00	19.00	5.00	11.8	26.8
1-NC	NorthCoast	663	2004	0.0	3.0	10.8	0.6	17.5	0.1	18.54	19.22	1.80	7.4	10.7
1-NC	NorthCoast	683	1998	0.0	3.1	9.4	0.4	28.0	0.0	30.00	1.00	2.90	59.0	103.3
1-NC	NorthCoast	694	2003	150.0	2.7	26.1	1.1	24.6	0.0	20.17	3.49	5.01	8.5	10.7
1-NC	NorthCoast	699	2003	16.5	2.7	21.0	1.0	12.1	0.0	17.50	2.17	1.46	7.8	6.8
1-NC	NorthCoast	715	1998	0.0	3.5	9.9	0.4	28.4	0.0	30.00	2.00	0.90	25.5	30.1
1-NC	NorthCoast	715	1999	16.7	2.1	11.9	0.5	19.9	1.0	8.00	0.00	1.00	13.3	19.3
1-NC	NorthCoast	715	2000	26.2	3.4	11.0	0.5	24.0	1.3	32.00	2.00	0.00	26.4	25.1
1-NC	NorthCoast	715	2001	2.4	3.2	11.3	0.8	30.5	0.3	34.00	1.00	1.00	23.1	34.0
1-NC	NorthCoast	715	2002	42.9	4.0	10.1	0.6	17.6	3.0	44.00	1.00	0.00	30.8	50.3
1-NC	NorthCoast	715	2004	9.5	4.3	9.2	0.4	21.2	0.3	32.69	2.52	0.00	22.7	30.1
1-NC	NorthCoast	718	1998	0.0	3.1	12.1	0.6	37.1	0.0	18.00	11.00	9.60	2.4	9.1
1-NC	NorthCoast	752	2000	0.0	1.2	19.2	0.5	95.2	94.8	13.00	0.00	5.90	18.8	17.3
1-NC	NorthCoast	807	2003	15.8	1.4	19.6	0.9	38.7	0.0	32.33	0.43	1.92	7.2	12.2
1-NC	NorthCoast	827	2000	20.0	2.1	16.3	0.7	25.0	0.5	24.00	14.00	3.10	9.5	11.9
1-NC	NorthCoast	854	1998	0.0	2.3	6.1	0.3	31.6	0.0	51.00	0.00	0.00	9.7	9.0
1-NC	NorthCoast	856	1999	0.0	2.2	8.9	0.4	21.3	0.5	19.00	0.00	0.00	4.5	3.5
1-NC	NorthCoast	874	1999	5.6	0.7	20.7	0.6	54.2	9.6	36.00	0.00	7.90	45.3	125.6
1-NC	NorthCoast	881	2003	4.0	4.4	13.5	0.5	6.9	0.0	27.29	1.82	0.00	7.4	8.8
2-MC	Mid-Coast	2949	2001	2.0	5.4	6.2	0.6	18.1	0.0	28.00	6.00	0.00	14.3	19.8
2-MC	Mid-Coast	2989	1998	0.0	0.8	7.1	0.4	77.0	1.5	24.00	32.00	2.80	5.4	2.9
2-MC	Mid-Coast	2989	2001	2.5	0.5	8.8	0.5	75.5	0.6	32.00	30.00	0.90	9.6	7.2
2-MC	Mid-Coast	2989	2004	0.0	0.5	7.8	0.4	69.4	10.7	36.62	25.44	1.61	6.7	2.8
2-MC	Mid-Coast	3067	1999	14.1	0.7	11.8	0.6	72.2	1.8	33.00	2.00	4.40	8.5	12.0
2-MC	Mid-Coast	3130	2001	8.7	2.3	10.0	0.6	17.1	0.0	17.00	14.00	0.00	1.8	3.5
2-MC	Mid-Coast	3231	1998	19.4	2.5	6.6	0.5	25.7	0.1	24.00	14.00	0.90	11.5	10.2
2-MC	Mid-Coast	3231	2001	74.5	2.3	8.0	0.6	25.4	0.0	15.00	26.00	0.90	11.0	19.1
2-MC	Mid-Coast	3231	2004	141.8	3.0	7.7	0.4	18.5	0.2	22.29	18.92	0.87	15.9	15.1
2-MC	Mid-Coast	2860	2000	0.0	1.9	5.2	0.4	32.3	0.0	49.00	0.00	1.00	8.7	11.9
2-MC	Mid-Coast	2737	2002	4.0	0.9	21.4	0.9	45.6	4.3	42.00	1.00	6.70	26.4	29.6
2-MC	Mid-Coast	2740	2000	0.0	8.1	6.8	0.4	11.0	0.9	31.00	2.00	1.00	22.2	44.4
2-MC	Mid-Coast	2800	2002	0.0	2.0	17.0	0.6	36.0	0.0	27.00	2.00	3.80	19.4	44.9
2-MC	Mid-Coast	3794	1998	7.1	2.6	6.3	0.5	13.8	2.9	53.00	4.00	1.70	6.2	4.6
2-MC	Mid-Coast	3823	1999	6.1	0.8	23.9	0.6	56.5	3.3	50.00	5.00	2.00	14.3	63.1
2-MC	Mid-Coast	3830	2002	60.0	2.1	18.3	0.6	35.8	4.0	27.00	10.00	2.80	28.8	30.5
2-MC	Mid-Coast	3544	1998	0.0	1.1	6.4	0.3	75.2	16.1	31.00	30.00	0.80	9.8	9.1
2-MC	Mid-Coast	3574	2000	46.6	1.8	8.2	0.4	54.1	0.0	31.00	35.00	1.60	6.4	3.8
2-MC	Mid-Coast	3587	1999	2.6	0.6	8.0	1.4	70.2	0.0	48.00	23.00	0.00	8.4	6.7
2-MC	Mid-Coast	3616	1998	1.4	3.4	12.4	0.5	30.6	0.1	21.00	14.00	2.40	17.1	20.9
2-MC	Mid-Coast	3724	1998	2.9	3.0	8.7	0.5	27.7	3.6	32.00	15.00	0.90	12.0	21.5
2-MC	Mid-Coast	1981	1999	27.0	1.3	6.7	0.5	42.8	0.0	25.00	23.00	0.90	4.9	6.8
2-MC	Mid-Coast	1984	1999	4.5	1.6	8.8	0.5	37.1	0.0	31.00	14.00	0.00	7.2	20.0
2-MC	Mid-Coast	1984	2002	61.7	2.8	8.4	0.4	25.1	0.0	31.00	21.00	2.00	10.0	12.2
2-MC	Mid-Coast	2008	1998	1.1	1.6	8.1	0.5	53.2	0.6	34.00	3.00	3.20	19.8	14.0
2-MC	Mid-Coast	2089	1998	7.2	2.0	8.3	0.5	62.6	2.4	65.00	1.00	2.70	13.7	12.9
2-MC	Mid-Coast	2089	2001	21.6	2.1	5.2	0.4	56.9	0.2	30.00	3.00	0.00	15.0	12.7
2-MC	Mid-Coast	2089	2004	3.6	2.0	3.2	0.3	52.4	4.3	36.70	2.44	0.90	16.9	22.3
2-MC	Mid-Coast	2117	2000	9.7	3.7	7.6	0.5	14.4	0.5	47.00	20.00	0.00	19.3	77.6
2-MC	Mid-Coast	2120	1999	0.0	12.6	3.6	0.4	14.5	0.0	11.00	15.00	0.00	8.8	5.2
2-MC	Mid-Coast	2127	2002	2.6	2.6	5.8	0.5	21.4	1.0	27.00	2.00	0.00	11.7	9.1
2-MC	Mid-Coast	2130	1999	0.0	5.1	6.1	0.4	27.4	1.2	23.00	1.00	0.00	17.2	15.0
2-MC	Mid-Coast	2167	2000	18.5	1.8	7.1	0.5	49.9	14.7	60.00	5.00	0.90	12.4	17.1
2-MC	Mid-Coast	2172	2003	166.7	1.5	5.8	0.3	34.8	0.0	38.01	1.12	0.00	12.5	14.3
2-MC	Mid-Coast	2183	2002	66.7	0.8	13.7	0.6	68.3	0.1	36.00	16.00	1.90	5.7	4.8
2-MC	Mid-Coast	2193	2000	11.8	5.0	2.4	0.3	18.3	0.0	37.00	1.00	0.00	16.5	39.8
2-MC	Mid-Coast	2237	1998	0.0	1.5	7.7	0.5	59.5	10.8	55.00	4.00	3.90	14.0	10.6
2-MC	Mid-Coast	2237	2001	41.3	1.3	5.9	0.4	60.5	0.0	56.00	7.00	3.10	10.5	32.0
2-MC	Mid-Coast	2259	2000	2.8	1.9	6.8	0.4	41.7	3.5	40.00	1.00	1.00	8.1	13.6
2-MC	Mid-Coast	2265	2002	39.2	1.4	6.4	0.4	35.5	1.1	26.00	7.00	1.00	9.9	19.1
2-MC	Mid-Coast	2278	1999	0.0	3.7	5.2	0.4	45.9	0.0	26.00	4.00	0.00	42.9	83.2
2-MC	Mid-Coast	2278	2002	85.1	3.4	4.9	0.4	17.3	3.0	40.00	2.00	0.00	38.2	48.3
2-MC	Mid-Coast	2319	2000	0.0	5.6	4.4	0.4	17.2	1.5	36.00	0.00	0.00	15.0	18.2
2-MC	Mid-Coast	2344	2003	41.7	1.0	7.9	0.4	32.3	0.0	16.78	28.57	0.00	6.0	8.7
2-MC	Mid-Coast	2432	2002	12.0	0.8	4.7	0.6	65.0	43.8	10.00	1.00	5.50	14.2	7.2
2-MC	Mid-Coast	2438	1998	0.0	0.9	6.3	0.6	45.8	6.2	31.00	5.00	0.00	17.2	13.1

(Continued)

GCG	STRATA	ID_NUM	SPAWNINGYE	COHOAUC_MI	GRADIENT	ACW	ACH	PCTPOOLS	PCTSWPOOL	PCTGRAVEL	PCTBEDROCK	POOL1P_KM	LWDPIECE1	LWDVOL1
2-MC	Mid-Coast	2451	1998	1.1	1.8	12.5	0.5	26.8	0.2	18.00	20.00	0.00	13.8	10.5
2-MC	Mid-Coast	2492	1998	0.0	1.7	13.2	0.6	15.0	0.1	16.00	48.00	1.00	1.2	1.7
2-MC	Mid-Coast	2492	2001	1.3	2.0	11.7	0.7	14.4	0.3	8.00	50.00	2.00	4.9	7.1
2-MC	Mid-Coast	2492	2004	1.3	1.2	11.5	0.6	44.7	0.2	15.04	37.59	0.00	7.5	4.8
2-MC	Mid-Coast	2514	1999	0.0	0.4	13.2	0.7	70.1	8.7	33.00	15.00	2.90	7.2	23.7
2-MC	Mid-Coast	2556	1999	0.9	1.5	5.7	0.4	51.1	0.1	27.00	35.00	0.00	6.5	12.2
2-MC	Mid-Coast	2556	2000	6.1	1.8	6.3	0.5	37.4	1.9	44.00	31.00	0.00	7.8	11.3
2-MC	Mid-Coast	2556	2001	13.0	1.8	6.5	0.6	37.0	1.2	26.00	32.00	0.00	10.1	15.9
2-MC	Mid-Coast	2556	2002	46.1	2.3	5.7	0.5	38.9	6.2	29.00	29.00	0.00	9.3	10.0
2-MC	Mid-Coast	2556	2004	10.4	1.2	4.1	0.3	39.2	0.0	44.95	23.20	0.90	4.0	6.7
2-MC	Mid-Coast	2580	2000	3.9	1.1	5.9	0.6	72.0	6.0	40.00	20.00	0.00	11.5	16.3
2-MC	Mid-Coast	2597	1998	2.4	1.3	4.5	0.5	63.1	0.9	59.00	15.00	0.00	16.0	23.0
2-MC	Mid-Coast	2627	2001	12.0	1.1	5.0	0.5	72.7	30.5	33.00	7.00	1.00	14.7	13.9
2-MC	Mid-Coast	2651	2003	0.0	1.9	4.4	0.6	60.8	29.8	23.20	0.03	0.81	23.5	39.6
2-MC	Mid-Coast	2839	2000	4.8	2.2	8.7	0.5	51.6	25.9	27.00	20.00	2.00	3.8	7.6
2-MC	Mid-Coast	3262	2002	18.3	1.4	4.1	0.5	33.1	17.1	5.00	0.00	1.00	12.9	13.4
2-MC	Mid-Coast	3276	2002	1.0	0.6	3.9	0.3	91.8	91.0	25.00	0.00	0.90	20.2	31.0
2-MC	Mid-Coast	3280	1999	43.6	2.4	2.6	0.3	41.0	0.0	42.00	2.00	0.00	1.6	1.7
2-MC	Mid-Coast	3289	1999	8.6	0.6	10.4	0.7	60.5	2.8	66.00	0.00	6.80	5.4	7.5
2-MC	Mid-Coast	3300	2001	23.2	0.8	7.9	0.5	38.5	0.0	17.00	29.00	0.00	9.3	12.3
2-MC	Mid-Coast	3311	2000	2.6	1.9	6.7	0.5	38.8	2.8	31.00	4.00	0.00	4.5	15.6
2-MC	Mid-Coast	3317	2002	301.0	2.8	6.0	0.6	36.8	32.9	39.00	23.00	2.00	4.2	3.1
2-MC	Mid-Coast	3336	1998	7.9	1.5	9.0	0.4	59.4	0.0	15.00	53.00	0.00	3.5	2.9
2-MC	Mid-Coast	3336	1999	27.2	1.2	8.8	0.6	65.2	0.0	9.00	76.00	0.90	3.0	4.3
2-MC	Mid-Coast	3336	2000	21.9	1.5	8.5	0.4	65.7	0.0	8.00	62.00	0.90	2.8	2.8
2-MC	Mid-Coast	3336	2001	54.4	1.6	10.5	0.6	66.3	1.0	12.00	50.00	0.80	4.7	5.5
2-MC	Mid-Coast	3336	2002	43.9	2.4	10.7	0.5	38.0	2.6	10.00	63.00	1.00	1.8	1.9
2-MC	Mid-Coast	3336	2004	45.6	1.4	10.4	0.4	65.4	0.0	9.52	54.93	0.00	3.8	1.9
2-MC	Mid-Coast	3382	2003	52.5	2.3	1.8	0.3	52.9	19.2	24.82	4.61	0.82	7.2	10.8
2-MC	Mid-Coast	3402	2003	149.5	2.3	3.0	0.3	33.3	0.2	35.35	22.03	0.00	5.1	8.1
3-MS	Mid-SouthCoast	5274	1998	17.7	0.6	4.2	0.5	75.2	12.0	27.00	0.00	4.90	1.7	0.3
3-MS	Mid-SouthCoast	5323	1998	57.1	4.6	4.7	0.4	38.0	2.1	32.00	12.00	0.00	26.3	11.4
3-MS	Mid-SouthCoast	5323	1999	14.3	2.7	4.2	0.3	45.8	2.2	57.00	4.00	0.90	16.7	27.0
3-MS	Mid-SouthCoast	5323	2000	29.9	4.8	5.3	0.4	26.8	0.0	53.00	8.00	0.00	34.6	56.2
3-MS	Mid-SouthCoast	5323	2002	39.5	2.7	4.5	0.3	44.5	1.7	34.00	4.00	0.00	18.9	17.7
3-MS	Mid-SouthCoast	5323	2004	100.7	1.9	4.9	0.4	35.5	1.3	49.36	0.93	0.00	10.7	10.8
3-MS	Mid-SouthCoast	5334	1999	7.9	1.6	4.1	0.5	26.8	9.9	39.00	0.00	0.00	4.5	11.2
3-MS	Mid-SouthCoast	5338	1998	1.6	6.2	3.0	0.3	29.1	0.5	41.00	23.00	0.00	18.7	21.6
3-MS	Mid-SouthCoast	5338	2001	32.7	4.9	3.7	0.4	17.0	0.0	21.00	16.00	0.00	16.6	31.2
3-MS	Mid-SouthCoast	5338	2004	12.5	5.7	3.6	0.4	9.3	0.0	19.86	27.33	0.00	16.1	15.2
3-MS	Mid-SouthCoast	5342	2000	2.8	8.1	4.8	0.4	5.0	0.0	26.00	14.00	0.00	30.0	31.1
3-MS	Mid-SouthCoast	5348	2003	60.0	6.4	4.7	0.4	14.6	0.4	28.14	8.65	0.00	18.7	18.6
3-MS	Mid-SouthCoast	5370	2001	8.8	5.2	3.7	0.5	12.6	0.0	55.00	1.00	1.00	22.8	15.9
3-MS	Mid-SouthCoast	5389	2001	25.0	0.6	15.4	0.8	85.5	0.1	20.00	3.00	4.10	5.6	10.2
3-MS	Mid-SouthCoast	5405	1998	2.0	0.6	20.2	1.1	81.3	13.8	25.00	24.00	5.30	14.8	9.1
3-MS	Mid-SouthCoast	5465	1998	0.0	15.8	4.5	0.7	1.2	0.0	22.00	35.00	0.00	12.8	25.4
3-MS	Mid-SouthCoast	5465	2001	4.5	9.5	2.8	0.6	0.0	0.0	9.00	12.00	0.00	2.1	2.8
3-MS	Mid-SouthCoast	5465	2004	2.3	15.3	6.2	0.5	0.9	0.0	14.36	27.18	0.00	11.3	13.8
3-MS	Mid-SouthCoast	5480	2003	84.0	0.4	16.0	0.6	77.4	1.4	21.32	51.45	5.31	9.8	12.5
3-MS	Mid-SouthCoast	5552	2003	0.0	3.0	4.9	0.5	36.2	1.2	18.47	9.50	0.00	32.1	53.4
3-MS	Mid-SouthCoast	5620	2003	33.3	1.1	24.4	0.9	31.8	0.4	18.34	27.07	4.89	3.8	2.6
3-MS	Mid-SouthCoast	5638	1998	16.5	1.3	17.1	0.7	47.5	2.0	21.00	38.00	5.10	3.9	4.7
3-MS	Mid-SouthCoast	5638	1999	11.3	1.4	15.2	1.0	59.4	0.1	15.00	49.00	5.50	3.0	8.2
3-MS	Mid-SouthCoast	5638	2000	24.7	1.2	13.9	0.5	50.5	0.0	13.00	54.00	3.70	4.5	7.0
3-MS	Mid-SouthCoast	5638	2001	202.1	0.8	16.1	0.7	73.6	0.5	17.00	47.00	1.80	2.2	2.8
3-MS	Mid-SouthCoast	5638	2002	86.6	1.2	15.3	0.6	60.6	0.0	12.00	49.00	3.70	5.1	3.9
3-MS	Mid-SouthCoast	5638	2004	183.5	0.9	20.1	1.1	53.5	0.6	13.35	38.66	2.75	4.7	3.7
3-MS	Mid-SouthCoast	5645	1998	13.7	1.5	22.2	0.8	26.5	0.6	18.00	22.00	2.70	4.9	4.1
3-MS	Mid-SouthCoast	5652	2002	118.9	37.8	5.2	0.4	11.2	0.0	8.00	14.00	0.00	32.5	28.9
3-MS	Mid-SouthCoast	5655	2001	82.1	2.8	10.7	0.8	43.6	0.0	36.00	1.00	4.40	9.5	14.1
3-MS	Mid-SouthCoast	5731	1998	2.2	2.3	11.1	0.6	63.2	13.1	11.00	54.00	4.50	7.4	2.5
3-MS	Mid-SouthCoast	5737	2003	12.9	0.7	8.0	0.5	69.3	6.1	39.25	14.30	2.81	8.3	7.2
3-MS	Mid-SouthCoast	5806	1998	21.9	2.3	7.7	0.5	28.9	1.4	36.00	9.00	0.00	9.9	12.9
3-MS	Mid-SouthCoast	4356	2002	35.7	1.8	3.5	0.4	23.5	0.0	41.00	9.00	1.00	17.5	14.0
3-MS	Mid-SouthCoast	4459	2002	28.2	16.1	2.5	0.4	9.7	0.0	23.00	1.00	0.00	24.9	54.6
3-MS	Mid-SouthCoast	4525	2002	9.1	3.2	3.8	0.3	49.9	0.8	47.00	2.00	2.00	14.1	15.0
3-MS	Mid-SouthCoast	4539	2001	70.0	0.2	10.0	0.2	83.7	1.2	26.00	30.00	2.10	2.4	0.7
3-MS	Mid-SouthCoast	4669	2000	0.0	1.3	6.0	0.4	38.4	0.0	51.00	15.00	0.00	22.6	34.6
3-MS	Mid-SouthCoast	4739	1999	0.0	2.6	16.7	0.8	52.5	0.3	37.00	2.00	13.10	4.2	9.8
3-MS	Mid-SouthCoast	4757	2002	4.3	4.3	18.4	0.8	34.8	3.3	28.00	2.00	6.60	13.7	26.0
3-MS	Mid-SouthCoast	4794	1998	13.8	6.9	4.7	0.5	37.7	0.1	40.00	5.00	0.00	14.4	11.5
3-MS	Mid-SouthCoast	4794	1999	0.0	6.2	4.6	0.4	43.7	0.9	33.00	7.00	0.00	14.2	29.5
3-MS	Mid-SouthCoast	4794	2000	4.6	3.1	5.0	0.5	21.2	0.0	32.00	1.00	0.00	19.4	13.5
3-MS	Mid-SouthCoast	4794	2001	25.3	2.6	4.8	0.6	30.2	0.0	43.00	2.00	0.00	10.2	6.1

(Continued)

GCG	STRATA	ID_NUM	SPAWNINGYE	COHOAUC_MI	GRADIENT	ACW	ACH	PCTPOOLS	PCTSWPOOL	PCTGRAVEL	PCTBEDROCK	POOL1P_KM	LWDPIECE1	LWDVOL1
3-MS	Mid-SouthCoast	4794	2002	6.9	6.0	4.8	0.3	25.9	1.5	15.00	6.00	0.00	17.3	14.7
3-MS	Mid-SouthCoast	4794	2004	80.5	4.6	5.9	0.5	18.0	0.6	33.25	4.22	0.00	12.3	7.8
3-MS	Mid-SouthCoast	4811	1999	0.0	0.8	9.6	0.4	67.9	9.0	37.00	38.00	4.80	3.0	2.5
3-MS	Mid-SouthCoast	4828	1998	21.3	2.3	8.2	0.5	51.3	0.7	26.00	3.00	3.60	17.7	18.5
3-MS	Mid-SouthCoast	4901	1999	0.0	2.5	1.9	0.4	34.7	0.7	52.00	0.00	0.00	8.9	27.5
3-MS	Mid-SouthCoast	5003	1998	0.0	4.9	5.9	0.5	27.4	1.1	23.00	3.00	0.00	16.6	23.0
3-MS	Mid-SouthCoast	5008	2000	1.2	4.6	7.4	0.5	11.5	0.0	31.00	0.00	0.00	29.7	18.7
3-MS	Mid-SouthCoast	5165	1999	49.2	1.4	9.9	0.6	58.1	0.5	57.00	0.00	3.70	7.5	15.8
3-MS	Mid-SouthCoast	5165	2002	30.5	2.4	10.2	0.4	27.9	25.9	54.00	1.00	4.40	16.3	30.7
3-MS	Mid-SouthCoast	4075	1999	5.6	1.2	2.9	0.3	15.5	0.3	58.00	0.00	0.00	8.3	6.0
3-MS	Mid-SouthCoast	4006	1998	0.0	0.6	30.4	2.1	78.8	1.4	21.00	5.00	7.90	1.5	1.4
3-MS	Mid-SouthCoast	4006	1999	0.0	0.9	21.5	1.3	73.7	0.1	0.70	7.00	5.30	0.7	2.5
3-MS	Mid-SouthCoast	4006	2000	2.9	0.7	25.9	1.0	84.9	0.0	45.00	6.00	7.00	1.4	0.9
3-MS	Mid-SouthCoast	4006	2001	0.0	0.6	22.8	1.1	79.3	0.1	41.00	2.00	8.10	0.2	0.2
3-MS	Mid-SouthCoast	4006	2002	0.0	1.0	22.1	0.8	74.9	1.5	26.00	6.00	8.30	2.1	1.4
4-UMP	Umpqua	6465	2002	0.0	0.6	5.0	0.5	64.6	0.0	49.00	0.00	1.80	0.9	0.9
4-UMP	Umpqua	6469	2002	18.5	3.2	6.4	0.4	14.9	0.0	45.00	0.00	0.00	14.2	11.5
4-UMP	Umpqua	6506	1999	29.5	0.7	14.7	0.7	60.5	2.2	22.00	43.00	0.00	7.8	16.9
4-UMP	Umpqua	6628	2000	50.0	1.5	10.7	0.4	40.3	0.0	11.00	64.00	1.00	2.7	3.8
4-UMP	Umpqua	6639	1998	5.2	1.0	4.6	0.5	47.2	0.0	43.00	41.00	0.00	11.6	7.6
4-UMP	Umpqua	6757	1998	2.4	0.8	10.9	0.8	56.5	11.7	24.00	39.00	3.80	4.3	4.5
4-UMP	Umpqua	6841	2000	7.9	2.3	16.3	0.7	37.0	5.9	30.00	9.00	6.70	3.4	1.4
4-UMP	Umpqua	6894	2000	1.2	1.7	10.1	0.4	48.4	0.4	14.00	40.00	0.00	4.9	7.2
4-UMP	Umpqua	6912	2000	3.8	1.4	10.1	0.5	71.7	1.7	27.00	50.00	0.00	10.9	10.9
4-UMP	Umpqua	6929	2001	0.0	1.4	7.7	0.3	49.2	0.0	19.00	41.00	1.90	4.7	12.0
4-UMP	Umpqua	6160	2003	5.0	5.0	3.6	0.4	17.8	0.0	32.37	5.92	0.00	23.1	25.9
4-UMP	Umpqua	6963	2003	5.0	0.2	14.8	0.6	66.7	0.2	34.31	7.80	1.82	6.1	6.2
4-UMP	Umpqua	6986	1998	2.1	0.6	12.7	0.6	24.9	0.0	10.00	51.00	1.00	3.2	2.2
4-UMP	Umpqua	7202	2000	13.3	0.8	4.9	0.5	96.4	88.0	15.00	0.00	7.50	7.5	7.8
4-UMP	Umpqua	7251	2001	29.9	2.1	8.5	0.5	33.9	0.0	16.00	58.00	1.00	2.6	1.6
4-UMP	Umpqua	7258	1998	4.7	1.0	12.1	0.5	55.1	1.8	20.00	25.00	0.00	11.6	12.5
4-UMP	Umpqua	7285	1998	0.0	1.1	12.2	0.5	72.8	29.2	17.00	25.00	2.80	8.7	6.1
4-UMP	Umpqua	7285	1999	0.0	0.5	14.9	0.7	83.9	40.6	22.00	33.00	3.60	2.6	5.1
4-UMP	Umpqua	7285	2000	3.9	2.0	12.2	0.6	56.3	50.4	21.00	30.00	2.00	3.9	6.0
4-UMP	Umpqua	7285	2001	9.7	0.9	10.6	0.5	79.5	52.8	23.00	25.00	3.70	3.1	3.3
4-UMP	Umpqua	7285	2002	16.5	0.9	13.7	0.5	75.4	30.7	16.00	35.00	4.60	5.8	3.8
4-UMP	Umpqua	7285	2004	6.8	0.8	10.8	0.6	72.0	0.0	19.22	34.08	4.70	6.9	4.7
4-UMP	Umpqua	7431	1999	3.5	1.5	13.7	0.7	28.2	0.2	13.00	8.00	2.80	1.9	3.7
4-UMP	Umpqua	7543	1998	0.0	1.1	15.3	0.7	34.8	1.3	20.00	9.00	2.80	1.1	0.2
4-UMP	Umpqua	7699	2002	9.7	1.1	7.3	0.5	46.2	0.7	21.00	30.00	3.80	3.4	3.8
4-UMP	Umpqua	7725	2001	7.2	0.5	8.6	0.3	79.2	0.2	48.00	21.00	0.00	1.6	0.9
4-UMP	Umpqua	7778	1998	4.2	1.7	8.3	0.5	27.1	10.9	29.00	0.00	2.00	10.1	3.7
4-UMP	Umpqua	7960	2001	0.0	4.6	1.9	0.4	2.3	0.0	28.00	9.00	0.00	3.8	0.7
4-UMP	Umpqua	7976	2003	1.0	4.0	4.4	0.4	36.6	1.9	37.25	23.26	0.95	9.6	3.6
4-UMP	Umpqua	7999	1998	1.0	2.9	8.2	0.8	43.7	25.8	28.00	4.00	1.60	9.5	12.0
4-UMP	Umpqua	7999	1999	3.9	2.7	6.8	0.7	29.1	2.9	25.00	4.00	1.50	7.3	7.4
4-UMP	Umpqua	7999	2000	3.9	2.8	5.4	0.5	28.7	2.3	30.00	9.00	0.00	10.6	10.2
4-UMP	Umpqua	7999	2001	56.3	2.6	4.9	0.5	32.6	0.0	45.00	9.00	0.00	6.2	9.9
4-UMP	Umpqua	7999	2002	11.7	2.3	5.1	0.4	22.4	0.0	38.00	6.00	0.00	8.0	10.6
4-UMP	Umpqua	7999	2004	29.1	2.1	4.8	0.6	23.2	0.0	33.10	6.98	0.00	6.4	2.5
4-UMP	Umpqua	8037	2000	3.3	3.5	8.8	0.6	6.4	0.2	25.00	1.00	0.00	1.8	1.4
4-UMP	Umpqua	8056	1998	38.0	2.2	3.8	0.6	41.0	0.0	35.00	2.00	1.00	4.7	1.0
4-UMP	Umpqua	8056	1999	5.1	2.4	3.9	0.7	50.5	11.4	47.00	3.00	1.70	0.8	0.1
4-UMP	Umpqua	8056	2000	12.7	1.8	3.2	0.6	37.6	2.8	34.00	0.00	0.00	1.5	0.2
4-UMP	Umpqua	8071	2000	0.0	2.4	6.5	0.6	24.9	1.9	23.00	11.00	0.00	7.1	15.7
4-UMP	Umpqua	8149	1999	0.0	0.7	10.1	0.6	65.1	11.8	80.00	4.00	1.00	3.7	1.2
5-SC	SouthCoast	8850	2000	0.0	6.0	12.8	1.2	22.1	0.5	15.00	20.00	5.90	5.7	4.2
5-SC	SouthCoast	8930	1998	3.9	8.5	6.2	0.7	4.7	0.0	22.00	0.00	0.00	4.2	22.6
5-SC	SouthCoast	8930	2001	70.2	7.1	5.7	0.6	5.6	0.1	23.00	1.00	0.00	8.6	10.3
5-SC	SouthCoast	8930	2004	9.8	7.8	5.9	0.4	13.4	2.4	28.90	0.23	0.00	6.8	18.9
5-SC	SouthCoast	9111	2000	0.0	6.6	8.2	0.7	13.1	0.0	17.00	10.00	1.00	13.3	10.1
5-SC	SouthCoast	9218	1998	0.0	1.4	22.8	0.8	20.5	0.5	47.00	10.00	3.00	0.5	0.0
5-SC	SouthCoast	9218	1999	0.0	1.2	21.8	0.8	37.1	0.4	35.00	16.00	2.50	0.6	0.3
5-SC	SouthCoast	9218	2000	0.0	0.9	28.4	1.2	15.4	0.0	35.00	19.00	1.00	1.3	0.3
5-SC	SouthCoast	9218	2001	0.0	0.6	12.3	0.3	3.4	0.0	33.00	18.00	0.60	0.3	0.1
5-SC	SouthCoast	9218	2002	0.0	0.7	18.8	0.7	37.5	1.9	27.00	18.00	0.90	0.2	0.1
5-SC	SouthCoast	9218	2004	2.6	0.8	15.6	0.6	38.9	5.7	33.77	16.33	3.49	0.1	0.0
5-SC	SouthCoast	9346	2003	0.0	1.8	7.0	0.5	2.5	0.0	36.15	3.85	0.00	0.7	0.3
5-SC	SouthCoast	9425	1999	0.0	3.1	6.8	0.5	24.6	7.7	36.00	4.00	0.00	3.4	1.8
5-SC	SouthCoast	9587	2003	3.4	2.2	7.2	0.4	33.4	17.6	30.43	5.27	2.94	5.3	0.8
5-SC	SouthCoast	9698	2001	24.3	1.9	17.0	0.5	12.3	0.5	38.00	1.00	6.00	7.5	4.1
5-SC	SouthCoast	9769	2000	1.1	2.0	7.9	0.5	47.4	0.0	17.00	14.00	2.90	7.5	5.1
5-SC	SouthCoast	9863	2002	37.1	3.4	11.4	0.8	26.3	0.0	11.00	3.00	2.30	11.9	14.4