1993

A Contribution To Longitudinal Data Analysis: Maximum Quasi-Likelihood Generalized Estimating Equations.

John David Bass
University of Alabama at Birmingham

Order Number 9333171

A contribution to longitudinal data analysis: Maximum quasi-likelihood generalized estimating equations

Bass, John David, Ph.D.

University of Alabama at Birmingham, 1993

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

A CONTRIBUTION TO LONGITUDINAL DATA ANALYSIS: MAXIMUM QUASI-LIKELIHOOD GENERALIZED ESTIMATING EQUATIONS

by

JOHN DAVID BASS

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Graduate School, The University of Alabama at Birmingham

BIRMINGHAM, ALABAMA

1993

ABSTRACT OF DISSERTATION
GRADUATE SCHOOL, UNIVERSITY OF ALABAMA AT BIRMINGHAM

Degree Ph.D. Major Subject Biostatistics

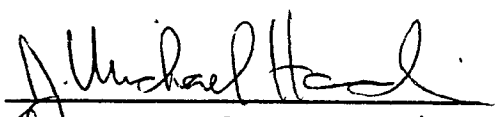
Name of Candidate John David Bass

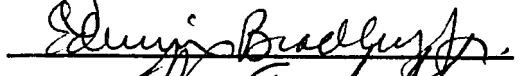
Title A Contribution to Longitudinal Data Analysis: Maximum Quasi-Likelihood Generalized Estimating Equations

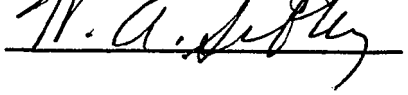
The continued growth of longitudinal data analysis techniques evolved into the formulation of generalized estimating equations (GEEs). The methodology was originally adapted from the univariate quasi-likelihood approach. It is confirmed that an identical formulation is possible through differentiation of the Mahalanobis distance D^2 with respect to the parameter vector $\underline{\beta}$. Unlike the original model, we establish an optimality property for estimation based on minimization of the quadratic form. This differs from the original formulation in which the variance-covariance matrix is considered functionally independent of $\underline{\beta}$. The ensuing system is denoted as the maximum quasi-likelihood estimating equations (MQL GEEs). This framework is further extended to incorporate three specific correlation scenarios. Of particular note are the matrix-vector differentiation lemmas used in the derivations.

Properties of the MQL GEE estimator are explored and contrasted with its predecessor. It is shown that both techniques coincide for normally distributed responses. In the general case, however, the MQL GEEs yield inconsistent and biased estimates. A direct repercussion is the lack of an asymptotic theory for the estimator. To gauge the severity of these consequences several data sets are modeled using both

methods, with jackknife and bootstrap estimates being generated for purposes of comparison. Monte Carlo simulation results highlight the effect of bias in modeling binary data.

Abstract Approved by: Committee Chairman 

Program Director 

Date _____ Dean of Graduate School 

ACKNOWLEDGMENTS

I would like to begin by thanking Dr. Mike Hardin, my advisor, for directing me to such an interesting area of research. If it had not been for his insight this work would not exist. Even though he has been very busy with his own work he maintained confidence that this project was feasible.

The remaining members of the committee were likewise most helpful. Considerable theoretical assistance was afforded by Dr. Edwin Bradley, whom must also be applauded for providing ego support during the trying times that inevitably kept cropping up. Additionally, Dr. Chuck Katholi is thanked for furnishing guidance about numerical approximation methods. Also, Dr. Jeffrey Roseman kept me on the straight-and-narrow concerning longitudinal concepts. Dr. Laura Perkins, who has since moved on to other activities, supplied useful critiques that improved the legibility of my writing. Finally, Dr. Al Bartolucci provided initial financial support which permitted my graduate education. To these individuals I am highly indebted.

Many other persons also assisted in this endeavor. Dr. Ed Meydrech and Dr. Mike Andrew provided a most congenial environment conducive to research in addition to serving as personal sounding boards during periods of ideation. Finally, I must thank my fellow classmates and all other Department of Biostatistics staff for the support I received. This experience will never be forgotten.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF TABLES

LIST OF ABBREVIATIONS

ANOVA	Analysis of variance
ANSI	American National Standards Institute
BP	Blood pressure
CDC	Centers for Disease Control
CRN	Chronic renal failure history
EBV	Extra-binomial variation
FEF	Forced expiratory flow
FEV	Forced expiratory volume
FVC	Forced vital capacity
GEE	Generalized estimating equation
GLM	General linear model
ICU	Intensive care unit
IID	Independent and identically distributed
IRLS	Iteratively reweighted least squares
MANOVA	Multivariate analysis of variance
MD	Minimum distance
ML	Maximum likelihood
MQL	Maximum quasi-likelihood
MVN	Multivariate normal
OC	Oral contraceptive
OPRC	Ordinary polynomial regression coefficient
PED	Polya-Eggenberger distribution

LIST OF ABBREVIATIONS (Continued)

REML	Restricted maximum likelihood
SDU	Standard deviation unit

CHAPTER I

Introduction

Longitudinal data arise from repeated observation of a sample over time. Frequently the evolution of a process is of interest, and this necessitates the acquisition of serial responses. Because time alone may not be a factor deterministic of outcome, other covarying factors collectively known as concomitant information are similarly recorded. Ensuing inquiries are addressed by modeling response behavior through functional relationships of the covariates. In regression terminology responses form dependent variables and concomitant information constitute explanatory, or predictor, variables. The procedure of investigation as characterized above is designated longitudinal data analysis. A synonymous term also present in the literature is event history analysis (Allison, 1984).

Data of this type are commonly found in many disciplines, with examples available pointing well into the past. This has been especially true in the social sciences. More recently, researchers in the biological sciences have shown considerable interest in this methodology. For illustrative purposes, we next support this contention by examining three reports from the field of epidemology. In each case blood pressure (BP) measurements are the responses of interest. The applicability and flexibility of longitudinal techniques for addressing different issues using similar data motivates this demonstration.

Zinner, Martin, Sacks, Rosner, and Kass (1975) reported results pertaining to observed changes in BP levels between initial and follow-up visits. In a prior report (Zinner, Levy, and Kass, 1971), 721 children between the ages of two and fourteen from 190 families in the Boston area were evaluated to determine hypertensive status, and it was confirmed that BP levels in children tend toward familial aggregation. Because hypertension is a potentially devastative disease, treatment is indicated once symptoms appear. However, this point of onset cannot be isolated given data from an initial visit only. The purpose of this study was to contrast BP readings across the periods to determine when hypertensive status changed and if the observed familial aggregation persisted. The sample consisted of 549 subjects (from 163 families) recruited from the earlier study, together with an additional 60 participants (from 44 families). Similar data was collected at each of two visits separated by a period of four years. Both systolic and diastolic readings were adjusted for age and sex, and then normalized into dimensionless standard deviation units (SDU).

Analysis of variance (ANOVA) techniques were used to confirm familial aggregation among children. ANOVA essentially compares sources of variation; here, these are within- and between-family. To determine if familial aggregation persisted, the ANOVA model was applied to the follow-up data:

$$(SDU_2)_{ij} = \mu + FAMILY_i + \varepsilon_{ij}$$

where i indexes through families, j indexes through subjects within family, SDU_2 identifies response at follow-up, μ is the mean population response, $FAMILY_i$ is the effect due to membership in the ith family, and

ε_{ij} is the random deviation of the (ij)th subject due to other unexplained sources of variation.

The comparison of responses across periods was useful for determining whether or not subjects varied individually to a significant degree. One approach useful for gaining insight was through the examination of scatter plots. Structural (deterministic) relationships were postulated based on this empirical evidence. Given initial and follow-up responses, the simple linear regression model was used to assess this relationship formally:

$$SDU_2 = \beta_0 + \beta_1 \cdot SDU_1 + \varepsilon$$

where β_1 represents the magnitude of change expected for the follow-up score SDU_2, given an initial score SDU_1. The significance of β_1 is a measure of the degree of association between the sets of scores.

Another way of exploring this relationship was through categorical data analysis. Natural groupings were created using the rank-orders: (a) SDU < -1; (b) -1 < SDU < 0; (c) 0 < SDU < 1; and (d) SDU > 1. It was then possible to produce four different graphs, each based on the initial score category. These charts consisted of histograms representing the frequencies of follow-up score categories, thus allowing visual inspection of the empirical distributions. Arrangement of these values into a contingency table format permitted a test for significant change in BP status based on the chi square statistic.

Rosner, Hennekens, Kass, and Miall (1977) investigated the determination of age-specific tracking correlations for BP. These are defined for an individual as the correlation between two readings taken at different times. In order to produce a spectrum of these, it was

necessary to include as many subjects of varying age as feasible. The sample, consisting of subjects aged between five and seventy-four years at the time of study initiation, was selected from two separate areas of Wales. From the Vale of Glamoran, 863 subjects were reviewed during 1956, 1960, 1964, and 1971, and from Rhondda Fach, 734 were evaluated in 1954, 1958, 1964, and 1971. Due to the large size of the sample it was impractical to provide separate analyses for each age. For this reason, and to eliminate the possibility of sparse cell counts, individuals were categorized by age at study initiation to be assigned into a stratum based on five-year age increments. Once this was accomplished, sample tracking correlations were generated for each gender-within-stratum combination.

Before formal analyses began, the correlations were converted into approximately standard normal deviates via the inverse hyperbolic tangent transformation. The reason for doing so was to allow the use of available ANOVA techniques for examining both age and gender effects. Even though age was of primary interest, other effects (e.g. gender) were investigated if the possibility of their influence existed. Furthermore, age stratification was easily adjustable and this allowed other partitionings of the data to be analyzed if warranted.

Cook, Scherr, Evans, Laughlin, Chalman, Rosner, Kass, Taylor, and Hennekens (1985) examined the effect of oral contraceptives (OC) on BP. They cited several references which connect OC use with severe health problems. Their study attempted to quantify this association. In 1973 (Survey 1), 5,802 women were recruited for participation. Each was visited at home, interviewed for health and related information, and provided three BP readings (the average of which is used in the

analyses). A second visit (Survey 2) was performed in 1976-77, in which 3,729 subjects agreed to continue. The last visit occurred in 1978 (Survey 3). Instead of polling the remaining cohort, a two-to-one sample stratified by OC-use status (non-use versus use), matched on age, was identified. Of the 1,503 so chosen, 1,111 consented to participate. Having collected the data, eligibility requirements were established before starting the analyses. To alleviate possible confounding problems, subjects classified as postmenopausal, pregnant, or under medication for hypertension at any survey point were exempted from consideration. After adjusting for this and assuring that consecutive data were available, 2,673 subjects were identified for analyses involving Surveys 1 and 2, while 927 were available for Surveys 2 and 3 comparison.

Two specific questions were posed. The first examined whether or not BP changes significantly between visits in light of OC status reversal. To begin, let (t-1) and t represent the prior and current visits, respectively. Also, allow the sample to be partitioned into two sets according to OC status at (t-1). To analyze the consequences of reversal, two separate regressions were proposed (one for each of the sets) using the model:

$$(BP_t - BP_{t-1}) = \beta_0 + \beta_1 \cdot BP_{t-1} + \beta_2 \cdot OC_t + \sum_{i=1}^{n} \psi_i \cdot X_i + \varepsilon$$

where OC_t is an indicator denoting the use (= 1) or non-use (= 0) at the current visit, and X_i represents one of n potential supplementary response modifiers.

Inference on β_2 formed the emphasis of these regressions. For previous non-users β_2 represented the adjusted change in BP due to

initiation of OC. Similarly, β_2 was the adjusted change in BP due to cessation of OC for previous users.

Their second inquiry explored the direct relationship between OC and BP level. Combining Surveys 2 and 3 data, a regression was proposed using the model

$$BP_t = \beta_0 + \beta_1 \cdot OC_t + \sum_{i=1}^{n} \psi_i \cdot X_{it} + \varepsilon_t$$

where t indexes the particular survey data source. The coefficient β_1 represents the adjusted effect on the level of BP due to the use or abstinence of OC.

The above studies have demonstrated that it is not the data that drives the analysis, but rather the questions posed and the models used to implement the structural relationships under investigation. While the thrust of this work is oriented towards the development of longitudinal methodologies, it may be worthwhile to step back and compare these with cross-sectional techniques. Cross-sectional data is also collected at varying times, the difference being that for each observation time the population is resampled. One obvious relevance is that no long term commitment on behalf of the subject is required. This is in direct contrast to the longitudinal practice in which attrition is always a concern. Because it will always be more difficult to recruit and maintain the sample, a longitudinal study must possess benefits which are not available through the cross-sectional avenue. Excellent reviews which detail many of the advantages and/or deficiencies are found in Cook and Ware (1983) and Ware (1985). Having contrasted longitudinal and cross-sectional data acquisition methods, a critical comparison will serve to accentuate proper usages of such data.

Glindmeyer, Diem, Jones, and Weill (1982) contrasted estimates of annual change in lung function. The sample consisted of 52 white males which had never been exposed to noxious chemicals. At the time of recruitment, ages of subjects ranged between 30.7 and 58.0 years and height varied between 62.8 and 73.5 inches. Five examinations were scheduled at one-year intervals to monitor pulmonary function. Spirometry data (responses) recorded included forced vital capacity (FVC, ml), forced expiratory volume in one second (FEV_1, ml), forced expiratory flow rate between 25% and 75% of the FVC ($FEF_{25-75\%}$, ml/s), forced expiratory flow rate at 50% of the FVC (FEF_{50}, ml/s), and forced expiratory flow rate at 75% of the FVC (FEF_{75}, ml/s).

Each response category was analyzed using both cross-sectional and longitudinal techniques. Cross-sectional methods centered on data obtained at each examination period. This required five sets of regression analyses to be performed using age and height as predictors. Conversely, the longitudinal approach concentrated on the individual. For this reason 52 separate regressions were needed using age as the lone covariate (height was assumed constant for each subject across the study period). The slope coefficients from these were then averaged to yield the longitudinal estimate of yearly change. An additional analysis was also performed excluding the initial visit data. This was done to avoid bias possibly induced due to learning effects.

To contrast the results between these two methodologies, the coefficient of mean annual change in FEV_1 was chosen for illustrative purposes. The cross-sectional values were -42.6, -49-2, -44.9, -50.5, and -44.6 ml for successive observation periods, whereas the longitudinal estimates were -12.4 ml (based on five visits) and -17.4 ml

(excluding the first). The cross-sectional coefficients impart estimates of decrease in FEV_1 level between consecutive years of age within the population. Notice that these values remained relatively stable across observation periods. These were in obvious contrast to the longitudinal coefficients which reflect the expected declination in FEV_1 level for an individual per year of aging. This example affirms that direct comparison of cross-sectional versus longitudinal estimates is inappropriate.

Thus far, basic concepts have been presented along with several examples for illustrative purposes. A theme common in each of these reports is that proper modeling and analysis is required to answer hypothesized questions. A problem inherent in the analysis of longitudinal data, regardless of the model, concerns the incorporation of the correlation structure presumed to exist between serial observations on the same subject. There are two primary issues involved, and these are addressed next.

The purpose of modeling is to relate responses to possible explanatory information. It is from this practice that insight is gained for the system under investigation. If a model is defined to be purely deterministic, no room for erroneous response is allowed. Because the idea of perfect prediction is absurd, a model should also incorporate a random, or stochastic, component. Another way to envision this concept is from the response side. A model can propose a set of outcomes from some defined set of possibilities, so one may conceive of a probability function governing this process. The problem faced in either scenario is the dearth of probability distributions which allow the inclusion of correlation structure presumed to exist in serial

observations. One notable exception is the multivariate normal (MVN) distribution for which a substantial body of literature in theory and application exists. However, the MVN is not appropriate in many circumstances (e.g. multivariate discrete responses). This precipitates the need for creating techniques that circumvent usual distributional assumptions.

Even if distributional problems did not exist, estimation of the unknown correlation matrix presents yet another barrier. Typically, the coefficients comprising this matrix are not of direct interest. But their inclusion necessarily affects the modeling process and, hence, cannot be ignored. When the number of observations per subject is sufficiently large (roughly > 20), it may be possible to use time series techniques to create an autocorrelation function, and through it estimates of the correlation coefficients are available. Unfortunately, longitudinal studies typically do not allow for the collection of data with this frequency.

These issues present formidable challenges, but continued research has led to the development of an encouraging technique of analysis. Liang and Zeger (1986) proposed the generalized estimating equations (GEEs) which allow for wide varieties of response modeling while providing consistent estimates of structural parameters even with misspecification of the correlation matrix. As a result of its flexibility, a unifying framework for the analysis of longitudinal data is being established.

By virtue of its relatively short existence many properties of the GEEs have yet to be considered. For this reason an area of rich research potential exists. We propose to extend this body of knowledge

by deriving a modification of the original GEEs from a different perspective.

Chapter 2 begins with definitions, notations, and terminology. This is followed by a retrospection of statistical techniques used to model continuous and discrete serial responses. The conclusion provides a detailed description of the steps leading to the original GEEs formulation.

Chapter 3 provides the development and theory of the original research. Throughout the chapter it is assumed that the correlation structure is arbitrary but constant across subjects. Some statistical properties of the derived estimator are examined.

Chapter 4 extends the research by postulating specific forms for the correlation structure. Coupled with the previous derivations, a generalized estimation framework is established.

Chapter 5 applies the aforementioned framework to several possible response distributions from the exponential family. This is the final step needed for construction of a flexible programmed implementation of the methodology.

Chapter 6 examines further properties of the derived estimator using simulation techniques. Jackknife, bootstrap, and Monte Carlo procedures are used to direct comparison between estimates based on Chapter 3 results versus those obtained from the original GEE formulation. Motivating the discussion is reexamination of several case studies from the literature.

Chapter 7 summarizes this work and suggests areas for continued research effort.

CHAPTER II

Literature Review

It has been mentioned that the history of longitudinal data analysis extends well into the past. Nonetheless, detailed examination of the statistical implications of modeling is only of recent origin. Before making assessments of these contributions we first establish definitions, notations, and terminology used throughout this work. In depth reviews of the general linear model, maximum and quasi-likelihoods, and GEE formulation are presented afterwards.

Definitions, Notations, and Terminology

A study is the scientific examination of a question posed or phenomenon observed. A population is the set of subjects towards whom the study is directed. A sample is a set of subjects participating in the study. The sample size is the number of participants, and is denoted by the lowercase k. Each subject will be assigned an integer between 1 and k, with an arbitrary subject from the sample being identified by the lowercase i.

A response is the objective measure within the study. The response for the ith subject observed on the jth occasion is represented as y_{ij}. The total number of responses for the subject is denoted n_i. The chronologically ordered set of responses is depicted in vector notation by: $\underset{\sim}{Y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^T$. If each subject is observed on the same occasions the design is said to be "balanced on time"

(Ware, 1985), and this allows the above representation to simplify to: $\underline{Y}_i = (y_{i1}, y_{i2}, \ldots, y_{in})^T$.

Concomitant information is additional data that is associated with the subject at the time a response is recorded. Typically one or more factors are available (either through design or happenstance), and these are represented in vector notation for the ith subject on the jth occasion by: $\underline{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})^T$. The set of concomitant information for the subject recorded over all observation periods is represented in matrix notation through augmentation of the individual vectors: $X_i = (\underline{x}_{i1}, \underline{x}_{i2}, \ldots, \underline{x}_{in_i})^T$.

It is often the case that we speculate the existence of a functional relation between the response and its concomitant information. More specifically, we postulate this relation to exist through a linear combination of the latter. We define the linear predictor for the ith subject on the jth occasion as the scalar product of the concomitant data and the parameter vector: $\eta_{ij} = \underline{x}_{ij}^T\underline{\beta}$, where $\underline{\beta}$ is a p×1 vector of unknown population parameters. Inference on $\underline{\beta}$ is the objective of the present research effort.

Responses are the primary elements of interest within the study. The characteristic inherent in these is that it is not possible to predict with certainty the value each will attain, and this implies our models cannot be purely deterministic. For this reason we consider a response to be a random variable. It is clear that this definition is expandable to encompass more complex aggregations such as random vectors and matrices.

We next consider terminology associated with random entities. Let Z_1 and Z_2 represent two arbitrary random entities. Every random entity

is presumed to be associated with a given distribution function F. For example, it may be that Z_1 follows the distribution function F_1. This relation is expressed by the notation $Z_1 \sim F_1$, where "~" is pronounced "is distributed as." If Z_1 and Z_2 exist such that for every possibility it is true that $f_3(Z_1,Z_2) = f_1(Z_1)\cdot f_2(Z_2)$ (where f_1, f_2, and f_3 are probability functions associated with distributions F_1, F_2, and F_3, respectively), Z_1 and Z_2 are defined as being independent. Moreover, if Z_1 and Z_2 are independent in addition to having the same distribution function, they are said to be independent and identically distributed (IID). In many cases a specific distribution will be postulated for the random entity under discussion and will be identifiable from the context.

Statistical functions of random entities are of special importance. The most common of these (when they exist) are the expectation, covariance (i.e. variance or dispersion), and correlation. The definitions of these are available in standard statistical texts. Due to their special status, the following functional representations are assigned: $E\{\cdot\}$, $V\{\cdot\}$, and $R\{\cdot\}$, respectively. More often, we refer to individual elements of these contructs. Let $\underline{y}_i$ be a random vector, and y_{im} and y_{in} be any two of its elements. We denote $E\{y_{im}\} = \mu_{im}$, $V\{y_{im},y_{in}\} = \sigma_{imn}$, and $R\{y_{im},y_{in}\} = \rho_{imn}$. When m equals n, we write $V\{y_{im}\} = \sigma^2_{imm}$ to distinguish the variance from its square root σ_{imm}, the standard deviation.

Lastly, we consider two special matrix-vector functions. Let $\underline{v}$ be a n×1 vector. We define diag($\underline{v}$) as the n×n diagonal matrix formed by placing each jth (j = 1 to n) element of $\underline{v}_i$ onto the corresponding (j,j)th diagonal position of the matrix. Next, let M be a n×n square

(not necessarily diagonal) matrix. We define vec(M) as the 1×n transposed vector formed by placing each (j,j)th (j = 1 to n) diagonal element onto the corresponding jth position of the vector.

Models for Serial Observations

Due to the preponderance of "nearly normal" data found in experimental environments, the bulk of the literature has dealt with means to provide adequate analyses under the MVN assumption. While it may be possible to transform various measured quantities into approximately normal deviates, qualitative (including count) data do not fair quite as favorably. This does not imply, however, that techniques to analyze the latter forms of data have not been forthcoming. Considerable effort has been expended to produce data-analytic methods in both categories of response.

Correlated quantitative responses.

One generally accepted starting point is in the area known as growth curve analysis. As is implied by the name, measurements (responses) are recorded over time during the maturation process for subjects within the study. It is then conceivable that growth could be hypothesized to be a function time. More specifically, if n responses are recorded for each subject, it is possible to construct a polynomial of degree (n-1) for each subject.

While there is nothing inherently wrong in the above scenario, the only thing that can be inferred are properties of the individual subject. In order to study processes relevant to the population under investigation, there must exist some method by which information for all subjects are pooled. Several such strategies will be explored next. In each case, we assume k subjects are observed on each of n occasions.

Potthoff and Roy (1964) provided an early attempt at describing the nature of the problem. In this seminal work, the multivariate analysis of variance (MANOVA) model was used as a basis for analysis:

$$E\{X\} = A\xi$$

where X is a k×n matrix with mutually independent rows, A is a known k×p design matrix, and ξ is a p×n matrix of unknown parameters. They assumed row $\underline{x}_i^T$ was distributed $N_n(\underline{a}_i^T\xi, \Sigma)$. Solution of this system was accomplished via ordinary least squares. The MANOVA model was augmented via post-multiplication by a constant matrix P to produce the growth curve model:

$$E\{X_0\} = A\xi P$$

where P is a n×t (t<n) known matrix of orthogonal polynomials of desired degree. Using the estimated parameter matrix obtained previously, tests of hypotheses for the degree of polynomial growth and equality of growth between subjects were possible.

Rao (1965) considered the model for an individual vector of responses to be of the form:

$$\underline{y}_i = A\underline{t}_i + \underline{\varepsilon}_i$$

where $\underline{y}_i$ is a n×1 vector of responses for the ith subject, A is a known n×p matrix, $\underline{t}_i$ is a p×1 vector of parameters unique to the subject such that $\underline{t}_i \sim N_p(\underline{\tau}, \Lambda)$, and $\underline{\varepsilon}_i \sim N_n(\underline{0}, \Sigma)$. Assuming $\underline{t}_i$ and $\underline{\varepsilon}_i$ were uncorrelated, $\underline{y}_i$ followed $N_n(A\underline{\tau}, A\Lambda A^T + \Sigma)$. Further inspection noted that the subject-specific parameter distribution was centerable: $\underline{\gamma} = \underline{t}_i - \underline{\tau}$. This reparameterization expressed the model as:

$$\underline{Y}_i = A[\underline{\tau} + \underline{\gamma}] + \underline{\varepsilon}_i = A\underline{\tau} + A\underline{\gamma} + \underline{\varepsilon}_i$$

where $\underline{\tau}$ is a p×1 vector of fixed effects associated with the process under study, and $\underline{\gamma}$ is a p×1 vector of random effects that encompasses intersubject variability. This was a form of the mixed effects model, and its application to growth curves was illustrated through the use of orthogonal polynomials premultiplying the response vector. This process generates independent (under normality) random variables known as orthogonal polynomial regression coefficients (OPRCs). By combining these across subjects, significance testing of the degree of polynomial growth was possible.

Grizzle and Allen (1969) proposed a model that permitted subjects to be either followed over time or observed under differing experimental conditions. Subjects were assigned to one of r treatment groups, and p responses were recorded on each. For a sample of size n, this model was expressed by:

$$X = B\xi A + E$$

where X is a p×n response matrix, B is the p×q within-subject design matrix, ξ is a q×r matrix of unknown parameter coefficients, A is the r×n design matrix across individuals, and E is a p×n random matrix whose rows are IID for some arbitrary multivariate distribution. As seen previously, A was usually a matrix of orthogonal polynomials. This construction was more flexible in that it allowed each group of subjects to follow its own unique growth curve as determined by the vector of coefficients ξ_j. Nonetheless, growth curves for all treatment categories were necessarily of the same degree by design.

Geisser (1970) extended the above approach by applying Bayesian techniques to the pseudo-augmented model:

$$E[Y] = (X,Z)(\tau,0)^T A = X\tau A$$

where Y is a p×n response matrix such that its columns are independent MVN vectors which share a common dispersion matrix Σ, X is the p×m within-subject design matrix, Z is a p×(p-m) matrix such that $Z^TX = 0$, τ is a m×r matrix of unknown parameters, and A is the r×n design matrix across subjects. Given MVN and independence for the response vectors, specification of the likelihood function $l(\tau,\Sigma)$ was possible. Using this and specification of non-informative prior distributions for Σ^{-1} and τ, a joint posterior density was formed . Upon integrating out Σ^{-1}, the posterior density remaining was a function of τ only. It was shown that this function is independent of Z and thus provided the estimate:

$$\hat{\tau} = (X^TS^{-1}X)^{-1}X^TS^{-1}Y\cdot[A^T(AA^T)^{-1}]$$

where $S = Y(I - A^T(AA^T)^{-1}A)Y^T$. It was noted that the result bore similarity to the least squares derived covariance-adjusted estimate of Rao (1965).

Lindley and Smith (1972) also proposed a Bayesian approach, but for the linear model in general. They cited the validity of the method based on earlier findings that had shown "least squares estimates are typically unsatisfactory" when viewed from the loss function standpoint. The usual linear model assumes the form: $E\{\underline{y}\} = A\underline{\theta}$, where $\underline{y}$ is a n×1 response vector, A is the n×p design matrix, and $\underline{\theta}$ is a p×1 vector of unknown parameters. For the actual formulation of the general Bayesian linear model, they assumed:

1) given $\underline{\theta}_1$, $\underline{y} \sim N_n(A_1\underline{\theta}_1, C_1)$

2) given $\underline{\theta}_2$, $\underline{\theta}_1 \sim N_p(A_2\underline{\theta}_2, C_2)$

3) given $\underline{\theta}_3$, $\underline{\theta}_2 \sim N_q(A_3\underline{\theta}_3, C_3)$.

This cascading design led to development of a three-stage model. The posterior distribution of $\underline{\theta}_1$, given the above conditions, was shown to equal $N_p(D\underline{d}, D)$, where:

$$\underline{d} = A_1^T C_1^{-1}\underline{y} + (C_2 + A_2C_3A^T)^{-1}A_2A_3\underline{\theta}_3$$

and

$$D^{-1} = A_1^T C_1^{-1}A_1 + (C_2 + A_2C_3A_2^T)^{-1} .$$

Next, they assumed vague prior knowledge surrounding the final stage by allowing C_3 to be unbounded. To view the implication of this premise, it was necessary only to examine the subsequent distribution of $\underline{\theta}_1$. The inverse of its dispersion, D^{-1}, diminished to $A_1^T C_1^{-1}A_1$, while the expression for $\underline{d}$ reduced to $A_1C_1^{-1}\underline{y}$. Together these imply the expectation of $\underline{\theta}_1$ attained the familiar form $(A_1^T C_1^{-1}A_1)^{-1}A_1C_1^{-1}\underline{y}$.

Fearn (1975) furthered the work of Lindley and Smith by applying Bayesian techniques to the growth curve model. He hypothesized separate curves for each subject, but posited observations within subject were IID normal. This allowed the vector of responses for the ith individual to be written: $\underline{y}_i|\underline{\beta}_i, \sigma_i^2 \sim N_n(X_i\underline{\beta}_i, \sigma_i^2 I)$, where $\underline{y}_i$ is a n×1 vector of responses, X_i is the n×p design matrix, $\underline{\beta}_i$ is a p×1 vector of unknown parameters unique to the subject, and σ_i^2 is the variance common to all responses for the subject. This first stage marked a departure from earlier works by postulating independence. Next, a second stage was introduced by considering the actual vectors of parameters to be exchangeable across individuals, implying $\underline{\beta}_i|\underline{\mu}, C \sim N_p(\underline{\mu}, C)$, where $\underline{\mu}$ is

the p×1 vector of the mean population parameters, and C is the p×p covariance matrix. Upon combining these stages, the distribution of the response compounded to:

$$\underline{y}_i | \underline{\mu}, \sigma_i^2, C \sim N_n(X_i\underline{\mu}, X_i C X_i^T + \sigma_i^2 I) \ .$$

Following Lindley and Smith (1972) a vague prior distribution for $\underline{\mu}$ was specified, with the inverse of its dispersion matrix vanishing in the analysis. Fearn continued with the additional work of estimating the components of dispersion for $\underline{y}_i$: C and σ^2. He assumed C^{-1} and σ^2 follow Wishart and non-central chi square distributions, respectively, conditional on the parameters needed for complete specification of these forms: $C^{-1} | \rho, R \sim W_p(\rho, R)$ and $\sigma^2 | \lambda, \omega \sim \chi^2(\lambda, \omega)$. Assumption of vague prior knowledge here was not helpful because [p(p+1)/2 + 2] parameters were involved. Instead of tackling this problem directly, an approach based on the following identity was attempted:

$$V[\underline{\mu}] = E[\ V[\underline{\mu} | \sigma^2, C]\ ] + V[\ E[\underline{\mu} | \sigma^2, C]\ ] \ .$$

By MVN the expectation is independent of its dispersion, implying the second term of the expression above was zero if the conditioning parameters were known. In practice they were not and neglecting this term underestimated the actual variance. However, a simpler estimation procedure resulted. The choices for σ^2 and C given were based on an approximation to marginal posterior distribution, and further investigation was continuing.

Rao (1975) employed empirical Bayes techniques for simultaneous estimation of parameter vectors across k sets of linear models of the form:

$$\underline{y}_i = X\underline{\beta}_i + \underline{\varepsilon}_i \ , \ i = 1, \ \ldots, \ k$$

where $\underline{y}_i$ is a n×1 vector of responses for the ith subject, X is the constant n×m design matrix, $\underline{\beta}_i$ is the m×1 vector of parameters, and $\underline{\varepsilon}_i$ is a n×1 vector of random deviations such that $E\{\underline{\varepsilon}_i|\underline{\beta}_i\} = \underline{0}$, $V\{\underline{\varepsilon}_i|\underline{\beta}_i\} = \sigma^2 V$, $E\{\underline{\beta}_i\} = \underline{\beta}$, $V\{\underline{\beta}_i\} = F$, and $V\{\underline{\beta}_i, \underline{\beta}_j\} = O$. He described several scenarios, but each involved estimation of the parameter vector for a current or future individual. One criterion on which to base the simultaneous estimation was through minimizing the mean square error. For this, let $\underline{p}$ be some known m×1 vector. The estimates were solutions that minimize the quadratic form $E\{(\underline{p}^T\underline{\beta}_i - a_0 - a_1{}^T\underline{y}_i)^2\}$. He proved that the estimator $\underline{\beta}^{(b)}$, formed from a weighted linear combination of least squares and ridge estimates, was optimum.

Laird and Ware (1982) combined empirical Bayes and maximum likelihood methods for analysis using the mixed model. Subjects were presumed independent, and the model for the ith subject was expressed:

$$\underline{y}_i = X_i\underline{\alpha} + Z_i\underline{b}_i + \underline{\varepsilon}_i$$

where $\underline{y}_i$ is a n_i×1 vector of responses, X_i is the n_i×p design matrix associated with the p×1 population parameter vector $\underline{\alpha}$, Z_i is the n_i×k design matrix associated with the k×1 subject-specific effects vector $\underline{b}_i$, and $\underline{\varepsilon}_i$ is a n_i×1 vector of error terms with distribution $N_{n_i}(\underline{0}, R_i)$. At the first stage both $\underline{\alpha}$ and $\underline{b}_i$ were considered fixed. The second stage introduced variation between subjects through allowing $\underline{b}_i$ to be distributed $N_k(\underline{0}, D)$. Together these implied the distribution of $\underline{y}_i$ was $N_{n_i}(X_i\underline{\alpha}, R_i + Z_iDZ_i^T)$, and the problem expanded to estimate these components. Let $\underline{\theta}$ be a q×1 vector defined through augmentation of the elements of D and every R_i. Maximum likelihood techniques could be used

to estimate both $\underline{\alpha}$ and $\underline{\theta}$; however, a Bayesian formulation was constructed by placing a noninformative prior on $\underline{\alpha}$. Upon removal of $\underline{\alpha}$ through integration, the ensuing likelihood function was maximized with respect to $\underline{\theta}$ only. This union of maximum likelihood and Bayesian methods produced restricted maximum likelihood (REML) estimates. This technique has been extensively studied by Cook (1982), among others.

A common factor in each of the above techniques is that responses are assumed to follow the MVN distribution. As it may be plausible to propose models with these characteristics for quantitative (i.e. measured) outcomes, one may not infer that proper analyses would result if the outcomes are discrete. Because studies are often undertaken in which qualitative responses are of interest, methods for modeling this category of outcome are explored next.

Correlated qualitative responses.

The lack of distribution functions which incorporate structure for correlation presumed to exist between serial observations has impeded methodology development for discrete modeling. The work that has been accomplished deals primarily with multiple serial binary outcomes, and it is on this subject that the discussion centers.

Korn and Whittemore (1979) described a procedure for the analysis of repeated binary outcomes. This technique prescribes outcomes for an individual to follow a unique logistic response curve. Each response z_t is postulated to be function of both current covariates and prior outcomes. The probability of observing the response vector $\underline{z}$ for a given subject takes the form:

$$\exp(\alpha_0 \Sigma z_j + \alpha_1 \Sigma z_j z_{j-1} + \underline{\beta}^T \Sigma z_j \underline{x}_j) \Big/ \prod_{j=1}^{n} [1 + \exp(\alpha_0 + \alpha_1 z_{j-1} + \underline{\beta}^T \underline{x}_j)]$$

where z_j and $\underline{x}_j$ are the response and concomitant data vector on the jth occasion, and $\underline{\beta}$ is the p×1 subject-specific parameter vector. The intercept terms α_0 and α_1 were regarded as nuisance parameters, but must be present for modeling purposes. They concluded the discussion by examining methods of combining the $\underline{\beta}$ parameters derived from each subject.

Stiratelli, Laird, and Ware (1984) extended this concept by proposing a two-stage approach. In the first stage, modeling of the individual response likelihood proceeded as above. Use of the logit transformation produces the linearization: $\underline{\lambda}_i = Z_i\underline{\nu}_i$, where Z_i is a n×m matrix of concomitant information and $\underline{\nu}_i$ is a m×1 vector of subject-specific parameters. The matrix Z_i was quite general in that it may contain elements from $\underline{y}_i$, further generalizing the technique of Korn and Whittemore (1979). In the second stage, variability between subjects is introduced by assuming $\underline{\nu}_i \sim N_m(W_i\underline{\alpha}, D)$, where W_i is a time-independent m×p constant matrix unique to the subject, and $\underline{\alpha}$ is a p×1 vector of population parameters. By observation, the difference $(\underline{\nu}_i - W_i\underline{\alpha})$, call it $\underline{b}_i$, was distributed $N_m(\underline{0}, D)$. Upon rearrangement of terms followed by substitution, the logit of response probabilities vector is expressed by the random effects model:

$$\underline{\lambda}_i = X_i\underline{\alpha} + Z_i\underline{b}_i$$

where X_i equals Z_iW_i. The remainder of their discussion centered on methods for estimating $\underline{\alpha}$ and D. Approaches discussed include maximum likelihood, REML, and Bayesian techniques.

Bonney (1987) suggested a mathematical approach for the modeling process suitable to any multivariate distribution. Conditional

probability is defined by the expression: $Pr(A|B) = Pr(A,B)/Pr(B)$, where $Pr(B)$ is nonzero. Upon rearrangement the joint probability is written: $Pr(A,B) = Pr(A|B) \cdot Pr(B)$. Now consider the generalized joint probability $Pr(\underline{y}|X)$ where $\underline{y}$ is a $n \times 1$ vector of responses and X is a $n \times p$ matrix of known constants. Because X is constant, it was simply carried along throughout the following inductive process:

$$\begin{aligned} Pr(\underline{y}|X) &= Pr(y_1, \ldots, y_n|X) \\ &= Pr(y_1|X) \cdot Pr(y_2, \ldots, y_n|y_1, X) \\ &\quad \vdots \\ &= Pr(y_1|X) \cdot \ldots \cdot Pr(y_n|y_1, \ldots, y_{n-1}, X) \\ &= \prod_{j=1}^{n} Pr(y_j|\{y_m; m<j\}, X) \; . \end{aligned}$$

The final decomposition consisted of n separate factors. In the case of binary outcomes, it may be suitable to hypothesize logistic functions for each of those. The canonical parameter of the logit link for the jth response of the subject was expressed by:

$$\theta_j = \alpha + z_1\gamma_1 + \ldots + z_{j-1}\gamma_{j-1} + 0\gamma_j + \ldots + 0\gamma_n + \underline{x}_j^T\underline{\beta}$$

where $\underline{z}$ is a linear transformation of $\underline{y}$ that ensures zero values for entries j through n, and $\underline{\gamma}$ is the vector of coefficients juxtaposing $\underline{z}$. The vector of ordered logits $\underline{\theta} = [\theta_1, \ldots, \theta_n]^T$ is model through the equation $\underline{\theta} = A\underline{\lambda}$, where $\underline{\lambda} = [\alpha, \gamma_1, \ldots, \gamma_{n-1}, \beta_1, \ldots, \beta_p]^T$ is a vector of unknown parameters, and A is the design matrix imposed on $\underline{\lambda}$.

Qu, Williams, Beck, and Goormastic (1987) proposed a logistic regression model based on the properties of an alternative distribution.

Suppose that $\underline{y}_i$ was a n×1 response vector for an individual, or unit. They postulated that the probability of a positive response on the jth subunit ($y_{ij} = 1$) was influenced by positive responses within the remaining subunits through the sum c ($= \Sigma\, y_{im}$, m = j to n). One choice available was the Polya-Eggenberger distribution (PED):

$$\Pr(\underline{y}_i; a, b, c, s) = \prod_{m=0}^{c-1}(\pi + m\theta) \cdot \prod_{m=0}^{n-c-1}(1 - \pi + m\theta) \Big/ \prod_{m=0}^{n-1}(1 + m\theta)$$

where $a, b > 0$, $\pi = a/(a + b)$, $\theta = s/(a + b)$, and $s \in \{-1, 0, 1\}$. It was noted that the PED reduced to the hypergeometric, binomial, or beta-binomial distribution when the value of s attains -1, 0, or +1, respectively. They showed the correlation coefficient $R\{y_{ij}, y_{ik}\}$ equals $s/(a + b + s)$. This supported the validity of using the beta-binomial in cases of presumed positive intraclass correlation.

Connolly and Liang (1988) reexamined the work of Qu et al. (1987) with respect to properties of the model and estimation efficiency. They considered a class of conditional logistic models written:

$$\text{logit}\,[\Pr(y_j | \underline{y}_{(-j)}, \underline{x}_j)] = F_n(w_j; \underline{\theta}) + \underline{\beta}^T \underline{x}_j$$

where y_j is the response, $\underline{y}_{(-j)}$ is the vector of responses excluding y_j, w_j is the sum across $\underline{y}_{(-j)}$, $\underline{\theta} = (\theta_1, \theta_2)^T$ [$= (\pi, \theta)^T$ in the notation of Qu et al, 1987], F_n is an arbitrary scalar function, $\underline{\beta}$ is the p×1 parameter vector, and $\underline{x}_j$ is a p×1 vector of concomitant data from the jth observation.

General Linear Models

Introduction of the general linear model (GLM) by Nelder and Wedderburn (1972) provided a integrated framework for the inclusion of wide varieties of univariate response and their relation to linear

predictors. Before this, much of the literature was devoted to the ordinary linear model, although specialized models for logistic and probit analyses, and other nonlinear relationships, were also being developed. With the advent of a unifying methodology, strengths were borrowed from many of the prior investigations.

The GLM is represented by $h(\mu) = \underline{x}^T\underline{\beta} = \eta$, where μ is the expected response of y, $\underline{x}$ is the vector of concomitant information, $\underline{\beta}$ is the vector of unknown structural coefficients, η is the linear predictor, and $h(\mu)$ is some monotonic differentiable transformation known as the link function. Choosing $h(\mu)$ to be the identity function obviously reduces the GLM to the ordinary linear model $\mu = E\{y\} = \underline{x}^T\underline{\beta}$. This may or may not be appropriate, however, for a particular application. In the ordinary linear model the range of the response is the entire real line, whereas responses may in actuality be restricted by design. Because no inherent limitations exist, per se, on the predictor, the mapping performed by $h(\mu)$ is of considerable interest. The existence of a particular class of link functions, known as canonical links, and their association with distributions from the exponential family will be discussed and illustrated next.

Let $F_Y(y)$ be a member of the exponential family. Its probability function is expressed by:

$$f_Y(y;\theta,\phi) = \exp[\phi\{y\theta - a(\theta)\} + b(y,\phi)]$$

where θ is the canonical parameter and ϕ is the scale parameter. It is well known that $\mu_Y = a'(\theta)$ and $\sigma^2 = a''(\theta)/\phi$ (Bickel and Doksum, 1977). The ability to express the expectation in terms of the canonical parameter is of practical interest. One consequence is that a link

function may be chosen on this basis. An example presented in detail may clarify this idea. Let y be a binomial random variable. Its probability function in usual notation is: $f_Y(y;\pi) = {}_nC_y \cdot \pi^y \cdot (1-\pi)^{n-y}$. This may be rewritten in exponential family form as follows:

$$\begin{aligned} f_Y(y;\pi) &= \exp[\ln[{}_nC_y \cdot \pi^y \cdot (1-\pi)^{n-y}]] \\ &= \exp[\ln({}_nC_y) + \ln(\pi^y) + \ln[(1-\pi)^{n-y}]] \\ &= \exp[\ln({}_nC_y) + y \cdot \ln(\pi) + (n - y) \cdot \ln(1-\pi)] \\ &= \exp[y \cdot [\ln(\pi)-\ln(1-\pi)] + n \cdot \ln(1-\pi) + \ln({}_nC_y)] \\ &= \exp[y \cdot \ln[\pi/(1-\pi)] + n \cdot \ln(1-\pi) + \ln({}_nC_y)] \ . \end{aligned}$$

Inspection of the first term implies that $\ln[\pi/(1-\pi)]$ equals θ, the canonical parameter. To express the second term as $a(\theta)$, π must first be solved in terms of θ. Use of algebra shows π to equal $e^\theta/(1 + e^\theta)$. This is the logistic response function. Substituting these expressions into the above derivation yields

$$\begin{aligned} f_Y(y;\theta,\phi) &= \exp[y\theta + n \cdot \ln[1 - e^\theta/(1+e^\theta)] + \ln({}_nC_y)] \\ &= \exp[y\theta + n \cdot \ln[1/(1+e^\theta)] + \ln({}_nC_y)] \\ &= \exp[y\theta - n \cdot \ln(1+e^\theta) + \ln({}_nC_y)] \\ &= \exp[1 \cdot [y\theta - n \cdot \ln(1+e^\theta)] + 1 \cdot \ln({}_nC_y)] \\ &= \exp[\phi[y\theta - a(\theta)] + b(y,\phi)] \ , \end{aligned}$$

which is in the exponential family form. Successive differentiations of $a(\theta)$ yield the expectation and variance of y: $\mu = a'(\theta) = ne^\theta/(1 + e^\theta)$ and $\sigma^2 = a''(\theta)/\phi = ne^\theta/(1 + e^\theta)^2$. We recall that π was earlier shown to equal $e^\theta/(1 + e^\theta)$. Upon substitution and simple algebra we see that $\mu = n\pi$ and $\sigma^2 = n\pi(1-\pi)$, which are the usual forms of these statistical functions.

Next, determination of the canonical link function is performed. We start by expressing the mean in terms of the canonical parameter: $\mu = n\pi = f(\theta) = ne^{\theta}/(1 + e^{\theta})$. Solving for the canonical parameter shows θ equals $\ln[\mu/(n - \mu)]$. The canonical link function is derived upon equating the linear predictor η to θ:

$$h(\mu) = \ln[\mu/(n - \mu)] = \eta = \underline{x}^{T}\underline{\beta}.$$

The desirability of choosing the link function in terms of the canonical parameter is because the mean is modeled directly as a function of the concomitant information.

Maximum Likelihood Estimation

Although the general model is attractive in form, solving for the estimates of $\underline{\beta}$ is not at all straightforward. The ones that will be considered here are called maximum likelihood (ML) estimates. To begin, a likelihood function is defined in the same manner as the probability function except that the emphasis switches to the unknown parameters rather than the response. Using their realizations, the responses may be envisioned as fixed. Because the covariates are also fixed, the only unknown or unestimated entities remaining are the parameters. The idea in maximum likelihood is to determine values for these parameters such that the probability of having realized this specific set of responses is maximized.

Notation denoting the likelihood for a given subject is $l(\underline{\beta}; y_i, \underline{x}_i)$. Because of independence across subjects, the likelihood for the sample is the product of the individual likelihoods:

$$l(\underline{\beta}; \underline{y}, X) = \prod_{i=1}^{k} l(\underline{\beta}; y_i, \underline{x}_i) .$$

Instead of maximizing the sample likelihood it is usually much easier to work with the log-likelihood, denoted $L(\underline{\beta};\underline{y},X)$. Because the logarithmic transformation is monotonic, maximization will be attained with the same solution. Once again an illustration will be helpful.

Suppose the response can attain either 0 or 1. One possible distribution is the Bernoulli, which has the probability function:

$$f_Y(y) = \pi^y \cdot (1-\pi)^{1-y} = \exp[y \cdot \ln[(\pi/(1-\pi)] + \ln(1-\pi)] \ .$$

We recall that π may be expressed in terms of the canonical parameter information via the logistic response function. The likelihood for the sample then becomes expressible as:

$$l(\underline{\beta};\underline{y},X) = \prod_{i=1}^{k} \exp[y_i \cdot \ln[\pi_i/(1-\pi_i)] + \ln(1-\pi_i)]$$

$$= \exp[\sum_{i=1}^{k} \{y_i\theta_i - \ln(1 + \exp(\theta_i))\}]$$

$$= \exp[\sum_{i=1}^{k} \{y_i \cdot (\underline{x}_i^T\underline{\beta}) - \ln(1 + \exp(\underline{x}_i^T\underline{\beta}))\}] \ .$$

Given this form the log-likelihood is easily attainable. Upon performing the logarithmic transformation, the log-likelihood for the sample becomes:

$$L(\underline{\beta};\underline{y},X) = \sum_{i=1}^{k} \{y_i \cdot (\underline{x}_i^T\underline{\beta}) - \ln[1 + \exp(\underline{x}_i^T\underline{\beta})]] \ .$$

The maximization step is performed by differentiating with respect to $\underline{\beta}$ and setting the result equal to $\underline{0}$:

$$L'(\underline{\beta};\underline{y},X) = \frac{\partial}{\partial \underline{\beta}} \left[\sum_{i=1}^{k} \{y_i \cdot (\underline{x}_i^T\underline{\beta}) - \ln[1 + \exp(\underline{x}_i^T\underline{\beta})]] \right]$$

$$= \sum_{i=1}^{k} \underline{x}_i^T [y_i - \exp(\underline{x}_i^T \underline{\beta})/[1 + \exp(\underline{x}_i^T \underline{\beta})]]$$

$$= \sum_{i=1}^{k} \underline{x}_i^T (y_i - \pi_i) = X^T(\underline{y} - \underline{\pi}) = \underline{0} .$$

If the order of $\underline{\beta}$ equals p, then the above constitutes a homogenous system of p nonlinear equations in p unknowns. One notes that these bear a resemblance to Fisher's score equations. Due to their nonlinearity, a solution is generally obtained using iterative techniques. Methods are discussed by McCullagh and Nelder (1983) and Seber and Wild (1989), and a specific technique is thoroughly explored in Chapter 3.

It has been shown that univariate responses distributed as members of the exponential family have properties that are highly desirable within the context of linear modeling. However it should be apparent that not all responses may have distributions from within this family. Moreover, it is conceivable that analytic forms for the probability functions do not exist. In the event this last statement is true, it will be impossible to construct the likelihood function. A method proposed to address this situation is considered next.

Quasi-Likelihood Estimation

Quasi-likelihood functions (Wedderburn, 1974) require only the specification of the relation between the mean and variance of the responses. Suppose $\underline{y}$ is a vector of independent responses. Let y_i be an element of $\underline{y}$ and assume its variance σ_i^2 can be expressed as some known function, $V(\mu_i)$, of the mean. The quasi-likelihood $q(y_i, \mu_i)$ is defined by the partial differential relation:

$$\frac{\partial\, q(y_i;\mu_i(\underline{\beta}))}{\partial\, \mu_i} = \frac{y_i - \mu_i}{V(\mu_i)} .$$

Recalling that the linear predictor is also a function of the mean, the system above can be reparameterized in terms of the unknown coefficients $\underline{\beta}$. Under hypothesis, the variance is modeled through the mean only. Solution of the quasi-likelihood of the sample proceeds in the manner prescribed for maximizing the log-likelihood:

$$\frac{\partial\, q(\underline{y},\underline{\mu})}{\partial\, \underline{\beta}} = \sum_{i=1}^{k} \left[\frac{\partial\, \mu_i}{\partial\, \underline{\beta}} \cdot \frac{\partial\, q(y_i,\mu_i)}{\partial\, \mu_i} \right] = \sum_{i=1}^{k} \left[\frac{\partial\, \mu_i}{\partial\, \underline{\beta}} \cdot \frac{(y_i - \mu_i)}{V(\mu_i)} \right] = \underline{0} .$$

The solution of these p simultaneous nonlinear equations yields the quasi-likelihood estimates. In fact, the quasi-likelihood function is identical to the log-likelihood when distributions are from the one-parameter exponential family. Extensions of this technique were investigated by McCullagh (1983) and Nelder and Pregibon (1987).

Thus far results have been shown whenever the response vector is considered to be composed of independent univariate outcomes. As mentioned earlier, the dearth of multivariate distributions allowing correlation structure presents a major obstacle. Maximum likelihood methods are usually not possible due to the lack of explicit representations. Neither are those utilizing quasi-likelihoods because it is not possible to postulate a mean-variance relation without explicitly incorporating correlation. However, if a correlational form can be assumed beforehand, an estimation method similar to quasi-likelihood should be applicable. This line of reasoning led to the development of a current technology in the modeling of longitudinal

responses -- generalized estimating equations (GEEs). GEEs can be envisioned as being the multivariate analogue of the quasi-likelihood functions, as will be seen in the following development.

Generalized Estimating Equations

It has been noted several times that incorporation of the correlation structure between dependent responses is a primary requirement for the proper analysis of longitudinal data. Consider the $n_i \times 1$ response vector $\underline{y}_i$ and its associated mean vector $\underline{\mu}_i$ and covariance matrix Σ_i. We begin by mentioning that any covariance matrix can be factored into $S_i^{\frac{1}{2}} \cdot R_i \cdot S_i^{\frac{1}{2}}$ (Seber, 1984), where R_i is the correlation matrix, and S_i is a diagonal matrix consisting of the variances of individual responses for the ith subject. This decomposition permits individual examination of these two components.

First, we adopt the quasi-likelihood concept whereby the variance of each response is modeled through some known function of the mean. For members of the exponential family, the jth diagonal element of S_i is expressible by $\sigma^2_{ijj} = a''(\theta_{ij})/\phi$. Thus S_i is readily constructed. This is not to be the case for the correlation matrix, however, as R_i is postulated to be completely arbitrary. Due to its symmetry, it can contain no more than $n_i \cdot (n_i - 1)$ distinct elements. In view of this, let $\underline{\alpha}$ be a vector of correlation coefficients that completely specifies R_i across all subjects. We formalize this by specifying the correlation matrix for the ith subject to be represented in functional notation as $R_i(\underline{\alpha})$. Through utilization of these notations , the covariance matrix is expressed by: $\Sigma_i = V_i/\phi = A_i^{\frac{1}{2}} \cdot R_i(\underline{\alpha}) \cdot A_i^{\frac{1}{2}}/\phi$.

We recall that maximization of the quasi-likelihood equations produced a score-like system of simultaneous equations. If the scalars

y_i are replaced by dependent response vectors $\underline{Y}_i$, the vectors of concomitant information $\underline{x}_i$ replaced by the matrix X_i, and the variance $V(\mu_i)$ replaced by Σ_i (with matrix inversion substituting for division), the resulting system of simultaneous nonlinear equations, analogous to the quasi-likelihood functions, are the GEEs (Liang and Zeger, 1986):

$$S_k(\underline{\beta}, \underline{\alpha}) = \sum_{i=1}^{k} \frac{\partial \underline{\mu}_i^T}{\partial \underline{\beta}} \cdot \Sigma_i^{-1} \cdot [\underline{Y}_i - \underline{\mu}_i]$$

$$= \phi \cdot \sum_{i=1}^{k} \frac{\partial \underline{\mu}_i^T}{\partial \underline{\beta}} \cdot V_i^{-1} \cdot [\underline{Y}_i - \underline{\mu}_i]$$

$$= \sum_{i=1}^{k} \frac{\partial \underline{\mu}_i^T}{\partial \underline{\beta}} \cdot V_i^{-1} \cdot [\underline{Y}_i - \underline{\mu}_i] = \underline{0} .$$

The first, second, and last factors comprising the GEEs are of dimension $p \times n_i$, $n_i \times n_i$, and $n_i \times 1$, respectively.

The GEEs are functions not only of $\underline{\beta}$ but also $\underline{\alpha}$, the unknown correlation coefficients constituting each R_i. Upon fixing the choice of R_i it is seen that the expected value of the GEEs equals the zero vector. This suggests any estimate of $\underline{\beta}$ is consistent regardless of the postulated structure of R_i, a point proved by Liang and Zeger (1986). However, the computed estimate and its variance are dependent on the choice actually used.

The GEE approach is gaining acceptance among practitioners in several areas. Zeger, Liang, and Albert (1988) highlighted modeling approaches with respect to differences between population-averaged and subject-specific parameter estimates. Qaqish (1990) examined its role in serial binary response regression models with special attention

extended to support extra-binomial variation (EBV). Lipsitz (1991) applied the methodology to categorical data analysis. Carr and Chi (1992) demonstrated its application to repeated measures ANOVA by producing closed-form solutions for balanced models with complete data and identity link.

Because this process has already established itself as a useful tool in the statistical realm, the GEE methodology is continuing to be the subject of extensive research. Crowder and Hand (1990) noted it was possible to fabricate the derivation through partial differentiation of quadratic forms even though the original formulation is based on analogy to quasi-likelihood equations. It is from this perspective that we begin the research phase.

CHAPTER III

Maximum Quasi-Likelihood Generalized Estimating Equations

Let $\underline{Y}_1, \underline{Y}_2, \ldots, \underline{Y}_k$ be independent random response vectors and $X_1, X_2, \ldots, X_k$ be concomitant data matrices. For every $\underline{Y}_i$ we postulate the existence of expectation vector $\underline{\mu}_i$ and covariance matrix Σ_i. We define the Mahalanobis distance D^2 by the quadratic form:

$$D^2 = \sum_{i=1}^{k} D_i^2 = \sum_{i=1}^{k} (\underline{Y}_i - \underline{\mu}_i)^T \Sigma_i^{-1} (\underline{Y}_i - \underline{\mu}_i) .$$

Also, each y_{ij} is postulated to be distributed from a member of the exponential family: $f_Y(y_{ij}; \theta_{ij}, \phi) = \exp[\phi\{y_{ij}\theta_{ij} - a(\theta_{ij})\} + b(y_{ij}, \phi)]$, where θ_{ij} and ϕ are the canonical and scale parameters, respectively.

In the regression context y_{ij} is related to concomitant data $\underline{x}_{ij}$ through a p×1 structural parameter vector $\underline{\beta}$. The GLM concept formalizes this association by equating the linear predictor $\underline{x}_{ij}^T\underline{\beta}$ to a known link function $h(\mu_{ij})$. Typically $h(\mu_{ij})$ is selected to equal the canonical link which implies $\theta_{ij} = h(\mu_{ij}) = \underline{x}_{ij}^T\underline{\beta}$. Under hypothesis the mean and variance of y_{ij} are functions of the canonical parameter: $\mu_{ij} = a'(\theta_{ij})$ and $\sigma_{ijj}^2 = a''(\theta_{ij})/\phi$. By virtue of the linking process both are also observed to be explicit functions of $\underline{\beta}$.

Derivation of the MQL GEEs

Inference on $\underline{\beta}$ is the goal of our investigation, but we must first establish an optimality property on which to base an estimator. Barnett

(1976) promoted ML-based rather than the usual minimum-distance (MD), or generalized least squares, estimators. Furthermore, Crowder and Hand (1990) commented that the original GEE methodology does not minimize D^2 because that technique does not directly use information contained within the covariance matrices. In view of these remarks we choose to seek an estimator which is obtainable through actual minimization of the quadratic form. This estimator $\hat{\underline{\beta}}$ of $\underline{\beta}$ is defined as a solution to the system of equations:

$$\frac{\partial}{\partial \underline{\beta}} D^2 = \frac{\partial}{\partial \underline{\beta}} \left[\sum_{i=1}^{k} (\underline{Y}_i - \underline{\mu}_i)^T \Sigma_i^{-1} (\underline{Y}_i - \underline{\mu}_i) \right] = \underline{0} .$$

It was observed in Chapter 2 that a covariance matrix Σ_i factors into the product $S_i^{\frac{1}{2}} \cdot R_i \cdot S_i^{\frac{1}{2}}$, where $S_i^{\frac{1}{2}} = \text{diag}([\sigma_{i11}, \sigma_{i22}, \ldots, \sigma_{inn}])$ and R_i is the correlation matrix. By hypothesis $S_i^{\frac{1}{2}}$ is also equal to $\text{diag}([\{a''(\theta_{i1})/\phi\}^{\frac{1}{2}}, \{a''(\theta_{i2})/\phi\}^{\frac{1}{2}}, \ldots, \{a''(\theta_{in})/\phi\}^{\frac{1}{2}}]) = \phi^{-\frac{1}{2}} A_i^{\frac{1}{2}}$, where $A_i^{\frac{1}{2}} = \text{diag}([\{a''(\theta_{i1})\}^{\frac{1}{2}}, \{a''(\theta_{i2})\}^{\frac{1}{2}}, \ldots, \{a''(\theta_{in})\}^{\frac{1}{2}}])$. Additionally, we also postulate a common correlation matrix R to exist across subjects. Together these hypotheses permit Σ_i to decompose into $(\phi^{-\frac{1}{2}} A_i^{\frac{1}{2}}) R (\phi^{-\frac{1}{2}} A_i^{\frac{1}{2}}) = \phi^{-1}(A_i^{\frac{1}{2}} R A_i^{\frac{1}{2}}) = \phi^{-1} V_i$. Upon substitution, the Mahalanobis distance becomes:

$$D^2 = \sum_{i=1}^{k} (\underline{Y}_i - \underline{\mu}_i)^T \Sigma_i^{-1} (\underline{Y}_i - \underline{\mu}_i)$$

$$= \sum_{i=1}^{k} (\underline{Y}_i - \underline{\mu}_i)^T (\phi^{-1} V_i)^{-1} (\underline{Y}_i - \underline{\mu}_i)$$

$$= \phi \cdot \sum_{i=1}^{k} (\underline{Y}_i - \underline{\mu}_i)^T V_i^{-1} (\underline{Y}_i - \underline{\mu}_i)$$

$$= \phi \cdot Q .$$

The factorization illustrates that minimization of the Mahalanobis distance is based solely on Q because ϕ is a constant scale parameter. Cognizant of this, we begin by expanding Q and then differentiating:

$$\frac{\partial}{\partial \underline{\beta}} Q = \frac{\partial}{\partial \underline{\beta}} \left[\sum_{i=1}^{k} (\underline{Y}_i - \underline{\mu}_i)^T V_i^{-1} (\underline{Y}_i - \underline{\mu}_i) \right]$$

$$= \frac{\partial}{\partial \underline{\beta}} \left[\sum_{i=1}^{k} (\underline{Y}_i^T V_i^{-1} \underline{Y}_i - \underline{Y}_i^T V_i^{-1} \underline{\mu}_i - \underline{\mu}_i^T V_i^{-1} \underline{Y}_i + \underline{\mu}_i^T V_i^{-1} \underline{\mu}_i) \right]$$

$$= \frac{\partial}{\partial \underline{\beta}} \left[\sum_{i=1}^{k} (\underline{Y}_i^T V_i^{-1} \underline{Y}_i - 2\underline{Y}_i^T V_i^{-1} \underline{\mu}_i + \underline{\mu}_i^T V_i^{-1} \underline{\mu}_i) \right]$$

$$= \sum_{i=1}^{k} \left[\frac{\partial}{\partial \underline{\beta}} (\underline{Y}_i^T V_i^{-1} \underline{Y}_i - 2\underline{Y}_i^T V_i^{-1} \underline{\mu}_i + \underline{\mu}_i^T V_i^{-1} \underline{\mu}_i) \right]$$

$$= \sum_{i=1}^{k} \left[\frac{\partial}{\partial \underline{\beta}} (\underline{Y}_i^T V_i^{-1} \underline{Y}_i) - 2 \frac{\partial}{\partial \underline{\beta}} (\underline{Y}_i^T V_i^{-1} \underline{\mu}_i) + \frac{\partial}{\partial \underline{\beta}} (\underline{\mu}_i^T V_i^{-1} \underline{\mu}_i) \right] . \qquad (3.1)$$

The three indicated partial derivatives require evaluation. The original GEE formulation considers V_i known and functionally constant so differentiation is straightforward (Graybill, 1976). However, the current methodology relaxes this premise. Appendix A contains lemmas that assist differentiations of various matrix-vector product combinations. Starting with the first term in Equation 3.1,

$$\frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T V_i^{-1} \underline{Y}_i] = \frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T (A_i^{\frac{1}{2}} R^{-1} A_i^{\frac{1}{2}}) \underline{Y}_i]$$

$$= \frac{\partial}{\partial \underline{\beta}} [(\underline{Y}_i^T A_i^{\frac{1}{2}}) \cdot (R^{-1} A_i^{\frac{1}{2}} \underline{Y}_i)]$$

$$= \left[\frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} \underline{Y}_i + \left[\frac{\partial}{\partial \underline{\beta}} [R^{-1} A_i^{-\frac{1}{2}} \underline{Y}_i]^T \right] \cdot [\underline{Y}_i^T A_i^{-\frac{1}{2}}]^T \quad \text{(by Lemma 3)}$$

$$= \left[\frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} \underline{Y}_i + \left[\frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T A_i^{-\frac{1}{2}} R^{-1}] \right] \cdot A_i^{-\frac{1}{2}} \underline{Y}_i$$

$$= \left[\frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} \underline{Y}_i + \left[\frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} \underline{Y}_i \quad \text{(by Lemma 4)}$$

$$= 2 \left[\frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} \underline{Y}_i \; . \qquad (3.2)$$

Next,

$$\frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T V_i^{-1} \underline{\mu}_i] = \frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T (A_i^{-\frac{1}{2}} R^{-1} A_i^{-\frac{1}{2}}) \underline{\mu}_i]$$

$$= \frac{\partial}{\partial \underline{\beta}} [(\underline{Y}_i^T A_i^{-\frac{1}{2}}) \cdot (R^{-1} A_i^{-\frac{1}{2}} \underline{\mu}_i)]$$

$$= \left[\frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} \underline{\mu}_i + \left[\frac{\partial}{\partial \underline{\beta}} [R^{-1} A_i^{-\frac{1}{2}} \underline{\mu}_i]^T \right] \cdot [\underline{Y}_i^T A_i^{-\frac{1}{2}}]^T \quad \text{(by Lemma 3)}$$

$$= \left[\frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} \underline{\mu}_i + \left[\frac{\partial}{\partial \underline{\beta}} [\underline{\mu}_i^T A_i^{-\frac{1}{2}} R^{-1}] \right] \cdot A_i^{-\frac{1}{2}} \underline{Y}_i$$

$$= \left[\frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} \underline{\mu}_i + \left[\frac{\partial}{\partial \underline{\beta}} [\underline{\mu}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} \underline{Y}_i \; . \qquad (3.3)$$

Lastly,

$$\frac{\partial}{\partial \underline{\beta}} [\underline{\mu}_i^T V_i^{-1} \underline{\mu}_i] = \frac{\partial}{\partial \underline{\beta}} [\underline{\mu}_i^T (A_i^{-\frac{1}{2}} R^{-1} A_i^{-\frac{1}{2}}) \underline{\mu}_i]$$

$$= \frac{\partial}{\partial \underline{\beta}}[(\underline{\mu}_i^T A_i^{-\frac{1}{2}}) \cdot (R^{-1} A_i^{-\frac{1}{2}} \underline{\mu}_i)]$$

$$= \left[\frac{\partial}{\partial \underline{\beta}}[\underline{\mu}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} \underline{\mu}_i + \left[\frac{\partial}{\partial \underline{\beta}}[R^{-1} A_i^{-\frac{1}{2}} \underline{\mu}_i]^T \right] \cdot [\underline{\mu}_i^T A_i^{-\frac{1}{2}}]^T \quad \text{(by Lemma 3)}$$

$$= \left[\frac{\partial}{\partial \underline{\beta}}[\underline{\mu}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} \underline{\mu}_i + \left[\frac{\partial}{\partial \underline{\beta}}[\underline{\mu}_i^T A_i^{-\frac{1}{2}} R^{-1}] \right] \cdot A_i^{-\frac{1}{2}} \underline{\mu}_i$$

$$= \left[\frac{\partial}{\partial \underline{\beta}}[\underline{\mu}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} \underline{\mu}_i + \left[\frac{\partial}{\partial \underline{\beta}}[\underline{\mu}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} \underline{\mu}_i \quad \text{(by Lemma 4)}$$

$$= 2 \left[\frac{\partial}{\partial \underline{\beta}}[\underline{\mu}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} \underline{\mu}_i \; . \qquad (3.4)$$

Upon substitution of Equations 3.2, 3.3, and 3.4 into Equation 3.1, we obtain:

$$\frac{\partial}{\partial \underline{\beta}} Q = \sum_{i=1}^{k} \left[\frac{\partial}{\partial \underline{\beta}}(\underline{Y}_i^T V_i^{-1} \underline{Y}_i) - 2 \frac{\partial}{\partial \underline{\beta}}(\underline{Y}_i^T V_i^{-1} \underline{\mu}_i) + \frac{\partial}{\partial \underline{\beta}}(\underline{\mu}_i^T V_i^{-1} \underline{\mu}_i) \right]$$

$$= \sum_{i=1}^{k} \Bigg[2 \left[\frac{\partial}{\partial \underline{\beta}}[\underline{Y}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} \underline{Y}_i$$

$$- 2 \left[\frac{\partial}{\partial \underline{\beta}}[\underline{Y}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} \underline{\mu}_i$$

$$- 2 \left[\frac{\partial}{\partial \underline{\beta}}[\underline{\mu}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} \underline{Y}_i$$

$$+ 2 \left[\frac{\partial}{\partial \underline{\beta}}[\underline{\mu}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} \underline{\mu}_i \Bigg]$$

$$= \sum_{i=1}^{k} \Bigg[2 \left[\frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} (\underline{Y}_i - \underline{\mu}_i)$$

$$- 2 \left[\frac{\partial}{\partial \underline{\beta}} [\underline{\mu}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} (\underline{Y}_i - \underline{\mu}_i) \Bigg]$$

$$= -2 \sum_{i=1}^{k} \Bigg[\left[\frac{\partial}{\partial \underline{\beta}} [\underline{\mu}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} (\underline{Y}_i - \underline{\mu}_i)$$

$$- \left[\frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} (\underline{Y}_i - \underline{\mu}_i) \Bigg]$$

$$= -2 \sum_{i=1}^{k} \left[\left[\frac{\partial}{\partial \underline{\beta}} [\underline{\mu}_i^T A_i^{-\frac{1}{2}}] - \frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} (\underline{Y}_i - \underline{\mu}_i) \right] .$$

Division by (-2) transforms the minimization problem to one of maximization. As a consequence of this act, we designate the following system of equations evaluated at $\underline{\beta} = \hat{\underline{\beta}}$ as the maximum quasi-likelihood (MQL) GEEs:

$$\sum_{i=1}^{k} \left[\frac{\partial}{\partial \underline{\beta}} [\underline{\mu}_i^T A_i^{-\frac{1}{2}}] - \frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} A_i^{-\frac{1}{2}} (\underline{Y}_i - \underline{\mu}_i) = \underline{0} . \qquad (3.5)$$

As Equation 3.5 stands the MQL GEEs are not in a directly useful format. We continue the expansion by performing the two indicated partial differentiations:

$$\left[\frac{\partial}{\partial \underline{\beta}}[\underline{\mu}_i^T A_i^{-\frac{1}{2}}]\right] = \left[\frac{\partial}{\partial \underline{\beta}}\underline{\mu}_i^T\right]\cdot A_i^{-\frac{1}{2}} + \left[\frac{\partial}{\partial \underline{\beta}}\text{vec}(A_i^{-\frac{1}{2}})\right]\cdot\text{diag}(\underline{\mu}_i) \quad \text{(by Lemma 2)}$$

and $$\left[\frac{\partial}{\partial \underline{\beta}}[\underline{Y}_i^T A_i^{-\frac{1}{2}}]\right] = \left[\frac{\partial}{\partial \underline{\beta}}\text{vec}(A_i^{-\frac{1}{2}})\right]\cdot\text{diag}(\underline{Y}_i) \; . \quad \text{(by Lemma 1)}$$

Substitution of these into Equation 3.5 followed by appropriate rearrangement and collection of terms yields:

$$\sum_{i=1}^{k}\left[\left[\frac{\partial}{\partial \underline{\beta}}\underline{\mu}_i^T\right]\cdot A_i^{-\frac{1}{2}} - \left[\frac{\partial}{\partial \underline{\beta}}\text{vec}(A_i^{-\frac{1}{2}})\right]\cdot\text{diag}(\underline{Y}_i - \underline{\mu}_i)\right]\cdot R^{-1}A_i^{-\frac{1}{2}}(\underline{Y}_i - \underline{\mu}_i) \; . \tag{3.6}$$

There exists a close similarity between Equation 3.6 and the original GEE formulation. In fact, if the covariance matrices are not functions of $\underline{\beta}$ (as is the case for normally distributed responses), it reduces identically to that shown in Zeger and Liang (1986):

$$\sum_{i=1}^{k}\left[\left[\frac{\partial}{\partial \underline{\beta}}\underline{\mu}_i^T\right]\cdot A_i^{-\frac{1}{2}} - \left[\frac{\partial}{\partial \underline{\beta}}\text{vec}(A_i^{-\frac{1}{2}})\right]\cdot\text{diag}(\underline{Y}_i - \underline{\mu}_i)\right]\cdot R^{-1}A_i^{-\frac{1}{2}}(\underline{Y}_i - \underline{\mu}_i)$$

$$= \sum_{i=1}^{k}\left[\left[\frac{\partial}{\partial \underline{\beta}}\underline{\mu}_i^T\right]\cdot A_i^{-\frac{1}{2}} - O\cdot\text{diag}(\underline{Y}_i - \underline{\mu}_i)\right]\cdot R^{-1}A_i^{-\frac{1}{2}}(\underline{Y}_i - \underline{\mu}_i)$$

$$= \sum_{i=1}^{k}\left[\frac{\partial}{\partial \underline{\beta}}\underline{\mu}_i^T\right]\cdot A_i^{-\frac{1}{2}}R^{-1}A_i^{-\frac{1}{2}}(\underline{Y}_i - \underline{\mu}_i)$$

$$= \sum_{i=1}^{k}\left[\frac{\partial}{\partial \underline{\beta}}\underline{\mu}_i^T\right]\cdot V_i^{-1}(\underline{Y}_i - \underline{\mu}_i) \; .$$

This was anticipated, for it establishes an equivalence between the original and MQL GEEs methods for normally distributed response data.

The culminating step requires evaluation of the two terms containing partial derivatives in Equation 3.6. The first term involves

$$\underline{\mu}_i^T = [\, \mu_{i1}, \mu_{i2}, \ldots, \mu_{in} \,]$$

$$= [\, a'(\theta_{i1}), a'(\theta_{i2}), \ldots, a'(\theta_{in}) \,] .$$

We recall $\underline{\theta}_i$ equals $X_i\underline{\beta}$ through the canonical link. Hence, by Lemma 5,

$$\frac{\partial}{\partial \underline{\beta}} \underline{\mu}_i^T = X_i^T \cdot \mathrm{diag}([a''(\theta_{i1}), a''(\theta_{i2}), \ldots, a''(\theta_{in})]) = X_i^T A_i .$$

The first term of Equation 3.6 is thus reduced to:

$$\left[\frac{\partial}{\partial \underline{\beta}} \underline{\mu}_i^T \right] \cdot A_i^{-\frac{1}{2}} = (X_i^T A_i) \cdot A_i^{-\frac{1}{2}} = X_i^T A_i^{\frac{1}{2}} . \qquad (3.7)$$

The second term involves

$$\mathrm{vec}(A_i^{-\frac{1}{2}}) = [\, \{a''(\theta_{i1})\}^{-\frac{1}{2}}, \{a''(\theta_{i2})\}^{-\frac{1}{2}}, \ldots, \{a''(\theta_{in})\}^{-\frac{1}{2}} \,] .$$

Through use of Lemma 5 (and the chain rule for scalars) we obtain:

$$\frac{\partial}{\partial \underline{\beta}} \mathrm{vec}(A_i^{-\frac{1}{2}}) = -X_i^T \cdot \mathrm{diag}([c_{i1}, c_{i2}, \ldots, c_{in}]),$$

where $c_{ij} = \frac{1}{2}[a''(\theta_{ij})]^{-\frac{3}{2}} a'''(\theta_{ij})$. The second term of Equation 3.6 refines to:

$$\left[\frac{\partial}{\partial \underline{\beta}} \mathrm{vec}[A_i^{-\frac{1}{2}}] \right] \cdot \mathrm{diag}(\underline{y}_i - \underline{\mu}_i)$$

$$= -X_i^T \cdot \mathrm{diag}([c_{i1}, c_{i2}, \ldots, c_{in}]) \cdot \mathrm{diag}(\underline{y}_i - \underline{\mu}_i) = -X_i^T W_i \qquad (3.8)$$

where $w_{ijj} = \frac{1}{2}[[a''(\theta_{ij})]^{-\frac{3}{2}} a'''(\theta_{ij})](y_{ij} - \mu_{ij})$.

All indicated differentiations are now completed. Substitution of Equations 3.7 and 3.8 into Equation 3.6 establishes the MQL GEEs in the final representation we denote $g(\underline{\beta})$:

$$g(\underline{\beta}) = \sum_{i=1}^{k} [X_i^T A_i^{\frac{1}{2}} - (-X_i^T W_i)] R^{-1} A_i^{-\frac{1}{2}} (\underline{y}_i - \underline{\mu}_i)$$

$$= \sum_{i=1}^{k} X_i^T (A_i^{\frac{1}{2}} + W_i) R^{-1} A_i^{-\frac{1}{2}} (\underline{y}_i - \underline{\mu}_i). \qquad (3.9)$$

Estimation of β

We next develop a method which generates the estimate $\hat{\underline{\beta}}$ of $\underline{\beta}$ through solution of the MQL GEEs. Following current convention (Seber and Wild, 1989) we refer to $g(\underline{\beta})$ (Equation 3.9) as the gradient vector of Q. Because g is a vector function of $\underline{\beta}$, we assume it can be expressed by a multivariable Taylor series (Olmstead, 1961) expanded about some approximation $\underline{\beta}^*$. This expansion is written:

$$g(\underline{\beta}) = g(\underline{\beta}^*) + \{ (\beta_0 - \beta_0^*) \cdot \frac{\partial}{\partial \beta_0} g(\underline{\beta}^*) + \ldots + (\beta_q - \beta_q^*) \cdot \frac{\partial}{\partial \beta_q} g(\underline{\beta}^*) \}$$

$$+ \{ \text{(higher order terms)} \} = \underline{0} .$$

Obviously the closer $\underline{\beta}^*$ approaches $\underline{\beta}$, the remaining higher order terms become negligible. This motivates the consideration of an iterative approach for estimation.

The technique we adopt is the Newton-Raphson method (Scarborough, 1966). For a given estimate $\underline{\beta}^t$, an adjustment $\underline{\delta}^t$ is generated. The next estimate becomes $\underline{\beta}^{t+1} = \underline{\beta}^t + \underline{\delta}^t$. The iterative process continues until $\underline{\delta}^t$ becomes vanishingly small (stopping rules are discussed by Seber and Wild, 1989). We now formalize this procedure for the MQL

GEEs. Using a variant of the Taylor series expansion shown above, we define the Newton-Raphson formula for the (t+1) iteration by:

$$g(\underline{\beta}^t) + \{ (\beta_0{}^{t+1} - \beta_0{}^t)\cdot\frac{\partial}{\partial \beta_0} g(\underline{\beta}^t) + \ldots + (\beta_q{}^{t+1} - \beta_q{}^t)\cdot\frac{\partial}{\partial \beta_q} g(\underline{\beta}^t) \} \approx \underline{0} .$$

Recalling the relation between $\underline{\beta}^t$ and $\underline{\beta}^{t+1}$ shown above, this formula may be rewritten as:

$$g(\underline{\beta}^t) + \{ \delta_0{}^t\cdot\frac{\partial}{\partial \beta_0} g(\underline{\beta}^t) + \delta_1{}^t\cdot\frac{\partial}{\partial \beta_1} g(\underline{\beta}^t) + \ldots + \delta_q{}^t\cdot\frac{\partial}{\partial \beta_q} g(\underline{\beta}^t) \} \approx \underline{0} .$$

Without loss of generality, replacement of the approximate equality poses no difficulty. The formula, upon rearrangement, becomes:

$$\delta_0{}^t\cdot\frac{\partial}{\partial \beta_0} g(\underline{\beta}^t) + \delta_1{}^t\cdot\frac{\partial}{\partial \beta_1} g(\underline{\beta}^t) + \ldots + \delta_q{}^t\cdot\frac{\partial}{\partial \beta_q} g(\underline{\beta}^t) = -g(\underline{\beta}^t) . \quad (3.10)$$

If there exists no confusion we drop both the superscript denoting iteration and the explicit functional notation of g. This simplifies the appearance of Equation 3.10 to:

$$\delta_0\cdot\frac{\partial}{\partial \beta_0} g + \delta_1\cdot\frac{\partial}{\partial \beta_1} g + \ldots + \delta_q\cdot\frac{\partial}{\partial \beta_q} g$$

$$= \left[\frac{\partial}{\partial \beta_0} g \quad \frac{\partial}{\partial \beta_1} g \quad \ldots \quad \frac{\partial}{\partial \beta_q} g \right]\cdot\underline{\delta} = \left[\frac{\partial}{\partial \underline{\beta}^T} g \right]\cdot\underline{\delta} = H\underline{\delta} = -\underline{g} . \quad (3.11)$$

The matrix H is termed the Hessian. Because g is the partial of Q with respect to $\underline{\beta}$, H is the matrix of its second partial derivatives. Assuming that the order of differentiation is irrelevant, we confirm the fact H is symmetric.

$$H = \begin{bmatrix} \frac{\partial}{\partial \beta_0} g & \frac{\partial}{\partial \beta_1} g & \cdots & \frac{\partial}{\partial \beta_q} g \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0}\left[\frac{\partial}{\partial \underline{\beta}} Q\right] & \frac{\partial}{\partial \beta_1}\left[\frac{\partial}{\partial \underline{\beta}} Q\right] & \cdots & \frac{\partial}{\partial \beta_q}\left[\frac{\partial}{\partial \underline{\beta}} Q\right] \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0}\left[\frac{\partial}{\partial \beta_0} Q\right] & \frac{\partial}{\partial \beta_1}\left[\frac{\partial}{\partial \beta_0} Q\right] & \cdots & \frac{\partial}{\partial \beta_q}\left[\frac{\partial}{\partial \beta_0} Q\right] \\ \frac{\partial}{\partial \beta_0}\left[\frac{\partial}{\partial \beta_1} Q\right] & \frac{\partial}{\partial \beta_1}\left[\frac{\partial}{\partial \beta_1} Q\right] & \cdots & \frac{\partial}{\partial \beta_q}\left[\frac{\partial}{\partial \beta_1} Q\right] \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial}{\partial \beta_0}\left[\frac{\partial}{\partial \beta_q} Q\right] & \frac{\partial}{\partial \beta_1}\left[\frac{\partial}{\partial \beta_q} Q\right] & \cdots & \frac{\partial}{\partial \beta_q}\left[\frac{\partial}{\partial \beta_q} Q\right] \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0}\left[\frac{\partial}{\partial \beta_0} Q\right] & \frac{\partial}{\partial \beta_0}\left[\frac{\partial}{\partial \beta_1} Q\right] & \cdots & \frac{\partial}{\partial \beta_0}\left[\frac{\partial}{\partial \beta_q} Q\right] \\ \frac{\partial}{\partial \beta_1}\left[\frac{\partial}{\partial \beta_0} Q\right] & \frac{\partial}{\partial \beta_1}\left[\frac{\partial}{\partial \beta_1} Q\right] & \cdots & \frac{\partial}{\partial \beta_1}\left[\frac{\partial}{\partial \beta_q} Q\right] \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial}{\partial \beta_q}\left[\frac{\partial}{\partial \beta_0} Q\right] & \frac{\partial}{\partial \beta_q}\left[\frac{\partial}{\partial \beta_1} Q\right] & \cdots & \frac{\partial}{\partial \beta_q}\left[\frac{\partial}{\partial \beta_q} Q\right] \end{bmatrix}$$

$$= \left[\frac{\partial}{\partial \underline{\beta}} \left[\frac{\partial}{\partial \beta_0} Q \right] \quad \frac{\partial}{\partial \underline{\beta}} \left[\frac{\partial}{\partial \beta_1} Q \right] \quad \cdots \quad \frac{\partial}{\partial \underline{\beta}} \left[\frac{\partial}{\partial \beta_q} Q \right] \right]$$

$$= \frac{\partial}{\partial \underline{\beta}} \left[\frac{\partial}{\partial \beta_0} Q \quad \frac{\partial}{\partial \beta_1} Q \quad \cdots \quad \frac{\partial}{\partial \beta_q} Q \right] = \frac{\partial}{\partial \underline{\beta}} g^T .$$

Hence
$$H\underline{\delta} = \left[\frac{\partial}{\partial \underline{\beta}} g^T \right] \cdot \underline{\delta} = -g . \qquad (3.12)$$

The next task faced is that of evaluating the Hessian explicitly. From Equation 3.9 we obtain the completed expression for g. Application of a second partial differentiation reveals the Hessian to equal:

$$H = \frac{\partial}{\partial \underline{\beta}} g^T = \frac{\partial}{\partial \underline{\beta}} \left[\sum_{i=1}^{k} [X_i^T (A_i^{\frac{1}{2}} + W_i) R^{-1} A_i^{-\frac{1}{2}} (\underline{Y}_i - \underline{\mu}_i)]^T \right]$$

$$= \frac{\partial}{\partial \underline{\beta}} \left[\sum_{i=1}^{k} (\underline{Y}_i - \underline{\mu}_i)^T A_i^{-\frac{1}{2}} R^{-1} (A_i^{\frac{1}{2}} + W_i) X_i \right]$$

$$= \sum_{i=1}^{k} \left[\frac{\partial}{\partial \underline{\beta}} (\underline{Y}_i - \underline{\mu}_i)^T A_i^{-\frac{1}{2}} R^{-1} (A_i^{\frac{1}{2}} + W_i) X_i \right]$$

$$= \sum_{i=1}^{k} \left[\frac{\partial}{\partial \underline{\beta}} [(\underline{Y}_i - \underline{\mu}_i)^T A_i^{-\frac{1}{2}} R^{-1} (A_i^{\frac{1}{2}} + W_i)] \cdot X_i \right]$$

$$= \sum_{i=1}^{k} \left[\frac{\partial}{\partial \underline{\beta}} [(\underline{Y}_i - \underline{\mu}_i)^T A_i^{-\frac{1}{2}} R^{-1} (A_i^{\frac{1}{2}} + W_i)] \right] \cdot X_i \quad \text{(by Lemma 4)}$$

$$= \sum_{i=1}^{k} \left[\frac{\partial}{\partial \underline{\beta}} [(\underline{Y}_i - \underline{\mu}_i)^T A_i^{-\frac{1}{2}} R^{-1}] \cdot (A_i^{\frac{1}{2}} + W_i) \right] \cdot X_i$$

$$= \sum_{i=1}^{k} \left[\left[\frac{\partial}{\partial \underline{\beta}} [(\underline{Y}_i - \underline{\mu}_i)^T A_i^{-\frac{1}{2}} R^{-1}] \right] \cdot (A_i^{\frac{1}{2}} + W_i) \right.$$

$$\left. + \left[\frac{\partial}{\partial \underline{\beta}} \text{vec}(A_i^{\frac{1}{2}} + W_i) \right] \cdot \text{diag}[(\underline{Y}_i - \underline{\mu}_i)^T A_i^{-\frac{1}{2}} R^{-1}] \right] \cdot X_i \quad \text{(by Lemma 2)}$$

$$= \sum_{i=1}^{k} \left[\left[\frac{\partial}{\partial \underline{\beta}} [(\underline{Y}_i - \underline{\mu}_i)^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} (A_i^{\frac{1}{2}} + W_i) \right.$$

$$\left. + \left[\frac{\partial}{\partial \underline{\beta}} \text{vec}(A_i^{\frac{1}{2}} + W_i) \right] \cdot \text{diag}[(\underline{Y}_i - \underline{\mu}_i)^T A_i^{-\frac{1}{2}} R^{-1}] \right] \cdot X_i \quad \text{(by Lemma 4)}$$

$$= \sum_{i=1}^{k} \left[\left[\frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T A_i^{-\frac{1}{2}} - m_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} (A_i^{\frac{1}{2}} + W_i) \right.$$

$$\left. + \left[\frac{\partial}{\partial \underline{\beta}} \text{vec}(A_i^{\frac{1}{2}} + W_i) \right] \cdot \text{diag}[(\underline{Y}_i - \underline{\mu}_i)^T A_i^{-\frac{1}{2}} R^{-1}] \right] \cdot X_i$$

$$= \sum_{i=1}^{k} \left[\left[\frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T A_i^{-\frac{1}{2}}] - \frac{\partial}{\partial \underline{\beta}} [m_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1} (A_i^{\frac{1}{2}} + W_i) \right.$$

$$\left. + \left[\frac{\partial}{\partial \underline{\beta}} \text{vec}(A_i^{\frac{1}{2}} + W_i) \right] \cdot \text{diag}[(\underline{Y}_i - \underline{\mu}_i)^T A_i^{-\frac{1}{2}} R^{-1}] \right] \cdot X_i$$

$$= \sum_{i=1}^{k} \left[\left[\frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T A_i^{-\frac{1}{2}}] - \frac{\partial}{\partial \underline{\beta}} [\mu_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1}(A_i^{\frac{1}{2}} + W_i) \cdot X_i \right.$$

$$\left. + \left[\frac{\partial}{\partial \underline{\beta}} \text{vec}(A_i^{\frac{1}{2}} + W_i) \right] \cdot \text{diag}[(\underline{Y}_i - \underline{\mu}_i)^T A_i^{-\frac{1}{2}} R^{-1}] \cdot X_i \right] . \quad (3.13)$$

Continued evaluation of the Hessian is performed on each of its two terms separately.

The first term of Equation 3.13 contains (within the intermediate brackets) the negative of an identical factor found in Equation 3.5. The evaluation of that partial derivative (via Equations 3.7 and 3.8) was shown to equal (from Equation 3.9) $X_i^T(A_i^{\frac{1}{2}} + W_i)$. Use of this result inplies the term reduces to:

$$\left[\frac{\partial}{\partial \underline{\beta}} [\underline{Y}_i^T A_i^{-\frac{1}{2}}] - \frac{\partial}{\partial \underline{\beta}} [\mu_i^T A_i^{-\frac{1}{2}}] \right] \cdot R^{-1}(A_i^{\frac{1}{2}} + W_i) \cdot X_i$$

$$= -X_i^T(A_i^{\frac{1}{2}} + W_i)R^{-1}(A_i^{\frac{1}{2}} + W_i)X_i . \quad (3.14)$$

The second term of Equation 3.13 must be evaluated directly. Let d_{ij} represent the jth element of $\text{vec}(A_i^{\frac{1}{2}} + W_i)$. The jth diagonal elements of $A_i^{\frac{1}{2}}$ and W_i were shown earlier to equal $[a''(\theta_{ij})]^{\frac{1}{2}}$ and $\frac{1}{2}[[a''(\theta_{ij})]^{-\frac{1}{2}}a'''(\theta_{ij})](y_{ij} - \mu_{ij})$, respectively. Hence d_{ij} is their sum and subsequently is a function of θ_{ij}. Through use of Lemma 5 (and the chain rule for scalars) we obtain:

$$\frac{\partial}{\partial \underline{\beta}} \text{vec}(A_i^{\frac{1}{2}} + W_i) = X_i^T \cdot \text{diag}([u_{i1}, u_{i2}, \ldots, u_{in}]) = X_i^T U_i \quad (3.15)$$

where:

$$u_{ijj} = \frac{\partial}{\partial \theta_{ij}} [[a''(\theta_{ij})]^{\frac{1}{2}} + \tfrac{1}{2}[[a''(\theta_{ij})]^{-\frac{3}{2}} a'''(\theta_{ij})](y_{ij} - \mu_{ij})]$$

$$= \tfrac{1}{2}[[a''(\theta_{ij})]^{-\frac{3}{2}} a'''(\theta_{ij}) - \tfrac{3}{2}[a''(\theta_{ij})]^{-\frac{5}{2}}[a'''(\theta_{ij})]^2](y_{ij} - \mu_{ij}) .$$

Substitution of Equations 3.14 and 3.15 into Equation 3.13 culminates the evaluation of the Hessian into a completed form:

$$H(\underline{\beta}) = \sum_{i=1}^{k} [-X_i^T (A_i^{\frac{1}{2}} + W_i) R^{-1} (A_i^{\frac{1}{2}} + W_i) X_i + X_i^T U_i \cdot \mathrm{diag}([\underline{y}_i - \underline{\mu}_i]^T A_i^{-\frac{1}{2}} R^{-1}) X_i]$$

$$= \sum_{i=1}^{k} X_i^T [U_i \cdot \mathrm{diag}([\underline{y}_i - \underline{\mu}_i]^T A_i^{-\frac{1}{2}} R^{-1}) - (A_i^{\frac{1}{2}} + W_i) R^{-1} (A_i^{\frac{1}{2}} + W_i)] X_i . \quad (3.16)$$

Equations 3.9 and 3.16 provide the basis of the framework necessary to generate the estimate $\hat{\underline{\beta}}$. The process begins with the choice of an initial estimate (e.g. $\underline{0}$). We recall from Equation 3.11 that $H\underline{\delta}$ equals $-\underline{g}$. Under usual conditions H is nonsingular and consequently invertible. This implies the existence of $\underline{\delta}$, the update increment. Iteration continues some t number of times until $\underline{\delta}^t$ becomes sufficiently small ($\approx \underline{0}$), at which point we declare the current value of $\underline{\beta}^{t+1}$ to equal $\hat{\underline{\beta}}$, the MQL estimate of $\underline{\beta}$. The gradient and Hessian evaluated at $\hat{\underline{\beta}}$ are designated accordingly:

$$g(\hat{\underline{\beta}}) = \sum_{i=1}^{k} X_i^T (\hat{A}_i^{\frac{1}{2}} + \hat{W}_i) R^{-1} \hat{A}_i^{-\frac{1}{2}} (\underline{y}_i - \hat{\underline{\mu}}_i) = \underline{0} \quad (3.17)$$

and

$$H(\hat{\underline{\beta}}) = \sum_{i=1}^{k} X_i^T [\hat{U}_i \cdot \mathrm{diag}([\underline{y}_i - \hat{\underline{\mu}}_i]^T \hat{A}_i^{-\frac{1}{2}} R^{-1}) - (\hat{A}_i^{\frac{1}{2}} + \hat{W}_i) R^{-1} (\hat{A}_i^{\frac{1}{2}} + \hat{W}_i)] X_i . \quad (3.18)$$

Estimation of ϕ

The scale parameter ϕ is necessary for proper variance estimation. Recall that $D^2 = \phi \cdot Q$, from which it follows that $\phi = D^2/Q$. Q is readily approximated via its evaluation at $\hat{\beta}$. We thus solicit a procedure for estimating Mahalanobis distance.

One method, advocated by McCullagh and Nelder (1983), follows from the asymptotic distribution of D^2. We recall D^2 is a quadratic form, which is itself the sum of independent quadratic forms. Let D_i^2 represent this quantity for the ith subject. It is well known (Johnson and Wichern, 1982) that the distribution of D_i^2 is asymptotically $\chi^2(n_i)$, where n_i is the order of the response vector y_i. As a consequence, $E\{D_i^2\}$ is approximately equal to n_i. By independence D^2 also approaches the chi square in law, implying $E\{D^2\} \approx (n_1 + n_2 + \ldots + n_k)$. This leads us to define: $\hat{D}^2 = [E\{D^2\} - p] = [(\Sigma\, n_i) - p]$, where the adjustment p is recommended due to the estimation of β.

Now that estimates of both Q and D^2 are available, the estimate $\hat{\phi}$ of ϕ is defined as:

$$\hat{\phi} = \hat{D}^2/\hat{Q} = [(\sum_{i=1}^{k} n_i) - p]/\hat{Q} \,. \qquad (3.19)$$

Consistency Generalizations of the MQL Estimate

A desirable property of estimates is that known as consistency. An estimate $\hat{\psi}$ is said to be consistent if the asymptotic expectation $E\{f(\hat{\psi} - \psi)\}$ equals zero for some function f. We next query into the MQL estimate for the property of $k^{\frac{1}{2}}$-consistency: $f = k^{\frac{1}{2}}(\hat{\beta} - \beta)$.

Recall that $g(\hat{\beta})$ is the gradient vector of Q evaluated at $\hat{\beta}$. We now consider its multivariable Taylor series expansion about β:

$$g(\hat{\underline{\beta}}) = g(\underline{\beta}) + \frac{\partial}{\partial \hat{\underline{\beta}}}[g(\underline{\beta})]\cdot(\hat{\underline{\beta}} - \underline{\beta}) + (\text{higher order terms})$$

$$= g(\underline{\beta}) + \frac{\partial}{\partial \hat{\underline{\beta}}}[g(\underline{\beta})]\cdot(\hat{\underline{\beta}} - \underline{\beta}) + O_p(1)$$

$$= g(\underline{\beta}) + H\cdot(\hat{\underline{\beta}} - \underline{\beta}) + O_p(1) \ .$$

But $g(\hat{\underline{\beta}})$ equals $\underline{0}$ by Equation 3.17. Thus:

$$\underline{0} = g(\underline{\beta}) + H\cdot(\hat{\underline{\beta}} - \underline{\beta}) + O_p(1) \ .$$

Upon rearrangement of this expression, and using the results of Wedderburn (1976) (as noted by McCullagh and Nelder, 1983), we have:

$$k^{\frac{1}{2}}(\hat{\underline{\beta}} - \underline{\beta}) = -k^{\frac{1}{2}}H^{-1}\cdot g(\underline{\beta}) + O_p(k^{-\frac{1}{2}}) \qquad (3.20)$$

We now investigate the expectation of $k^{\frac{1}{2}}(\hat{\underline{\beta}} - \underline{\beta})$. Before starting, observe that as k increases the remainder term becomes diminishingly small. This implies the remainder term has expectation zero asymptotically. Therefore our attention rests with $-k^{\frac{1}{2}}H^{-1}\cdot g(\underline{\beta})$. In practice the Hessian is replaced by its expectation, denoted $\tilde{H}$. Replacing H with $\tilde{H}$ and dropping the remainder term in Equation 3.20 yields:

$$E\{k^{\frac{1}{2}}(\hat{\underline{\beta}} - \underline{\beta})\} = E\{-k^{\frac{1}{2}}\tilde{H}^{-1}\cdot g(\underline{\beta})\} = -k^{\frac{1}{2}}\tilde{H}^{-1}\cdot E\{g(\underline{\beta})\}$$

$$= -k^{\frac{1}{2}}\tilde{H}^{-1}\cdot E\{\sum_{i=1}^{k} X_i^T(A_i^{\frac{1}{2}} + W_i)R^{-1}A_i^{-\frac{1}{2}}(\underline{y}_i - \underline{\mu}_i) \ \}$$

$$= -k^{\frac{1}{2}}\tilde{H}^{-1}\cdot\sum_{i=1}^{k} E\{X_i^T(A_i^{\frac{1}{2}} + W_i)R^{-1}A_i^{-\frac{1}{2}}(\underline{y}_i - \underline{\mu}_i)\}$$

$$= -k^{\frac{1}{2}}\tilde{H}^{-1}\sum_{i=1}^{k} E\{X_i^T A_i^{\frac{1}{2}} R^{-1} A_i^{-\frac{1}{2}}(Y_i - \mu_i) + X_i^T W_i R^{-1} A_i^{-\frac{1}{2}}(Y_i - \mu_i)\}$$

$$= -k^{\frac{1}{2}}\tilde{H}^{-1}\sum_{i=1}^{k} [E\{X_i^T A_i^{\frac{1}{2}} R^{-1} A_i^{-\frac{1}{2}}(Y_i - \mu_i)\} + E\{X_i^T W_i R^{-1} A_i^{-\frac{1}{2}}(Y_i - \mu_i)\}]$$

$$= -k^{\frac{1}{2}}\tilde{H}^{-1}\sum_{i=1}^{k} [X_i^T A_i^{\frac{1}{2}} R^{-1} A_i^{-\frac{1}{2}}(E\{Y_i\} - \mu_i) + X_i^T E\{W_i R^{-1} A_i^{-\frac{1}{2}}(Y_i - \mu_i)\}]$$

$$= -k^{\frac{1}{2}}\tilde{H}^{-1}\sum_{i=1}^{k} [\underline{0} + X_i^T E\{W_i R^{-1} A_i^{-\frac{1}{2}}(Y_i - \mu_i)\}]$$

$$= -k^{\frac{1}{2}}\tilde{H}^{-1}\sum_{i=1}^{k} X_i^T E\{W_i R^{-1} A_i^{-\frac{1}{2}}(Y_i - \mu_i)\} . \qquad (3.21)$$

The expectation does not proceed through Equation 3.21 directly because W_i involves Y_i. In order to achieve its evaluation it is necessary to first perform the matrix multiplication across $W_i R^{-1} A_i^{-\frac{1}{2}}(Y_i - \mu_i)$. Recall

$$W_i = \begin{bmatrix} w_{i11} & 0 & \cdots & 0 \\ 0 & w_{i22} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & w_{inn} \end{bmatrix} \qquad R^{-1} = \begin{bmatrix} \varphi_{11} & \varphi_{21} & \cdots & \varphi_{n1} \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{n2} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \varphi_{n1} & \varphi_{n2} & \cdots & \varphi_{nn} \end{bmatrix}$$

$$A_i^{-\frac{1}{2}} = \begin{bmatrix} \phi^{-\frac{1}{2}}\sigma_{i11}^{-1} & 0 & \cdots & 0 \\ 0 & \phi^{-\frac{1}{2}}\sigma_{i22}^{-1} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & \phi^{-\frac{1}{2}}\sigma_{inn}^{-1} \end{bmatrix} \qquad (Y_i - \mu_i) = \begin{bmatrix} (y_{i1} - \mu_{i1}) \\ (y_{i2} - \mu_{i2}) \\ \cdot \\ \cdot \\ \cdot \\ (y_{in} - \mu_{in}) \end{bmatrix}$$

by definition. Because no specific structure for R has been assumed, R^{-1} is expressed in a generic form only.

The matrix multiplication is performed starting from the right, then progressing leftward. The first product formed is $A_i^{-\frac{1}{2}}(\underline{y}_i - \underline{\mu}_i)$:

$$A_i^{-\frac{1}{2}}(\underline{y}_i - \underline{\mu}_i) = \begin{bmatrix} \phi^{-\frac{1}{2}}\sigma_{i11}^{-1} & 0 & \cdots & 0 \\ 0 & \phi^{-\frac{1}{2}}\sigma_{i22}^{-1} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & \phi^{-\frac{1}{2}}\sigma_{inn}^{-1} \end{bmatrix} \cdot \begin{bmatrix} (y_{i1} - \mu_{i1}) \\ (y_{i2} - \mu_{i2}) \\ \cdot \\ \cdot \\ \cdot \\ (y_{in} - \mu_{in}) \end{bmatrix}$$

$$= \begin{bmatrix} \phi^{-\frac{1}{2}}\sigma_{i11}^{-1}(y_{i1} - \mu_{i1}) \\ \phi^{-\frac{1}{2}}\sigma_{i22}^{-1}(y_{i2} - \mu_{i2}) \\ \cdot \\ \cdot \\ \cdot \\ \phi^{-\frac{1}{2}}\sigma_{inn}^{-1}(y_{in} - \mu_{in}) \end{bmatrix}.$$

The next product is $R^{-1} \cdot [A_i^{-\frac{1}{2}}(\underline{y}_i - \underline{\mu}_i)] = R^{-1}A_i^{-\frac{1}{2}}(\underline{y}_i - \underline{\mu}_i)$:

$$R^{-1}A_i^{-\frac{1}{2}}(\underline{y}_i - \underline{\mu}_i) = \begin{bmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{n1} \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{n2} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \varphi_{n1} & \varphi_{n2} & \cdots & \varphi_{nn} \end{bmatrix} \cdot \begin{bmatrix} \phi^{-\frac{1}{2}}\sigma_{i11}^{-1}(y_{i1} - \mu_{i1}) \\ \phi^{-\frac{1}{2}}\sigma_{i22}^{-1}(y_{i2} - \mu_{i2}) \\ \cdot \\ \cdot \\ \cdot \\ \phi^{-\frac{1}{2}}\sigma_{inn}^{-1}(y_{in} - \mu_{in}) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{m=1}^{n} \varphi_{1m}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im}) \\ \sum_{m=1}^{n} \varphi_{2m}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im}) \\ \vdots \\ \sum_{m=1}^{n} \varphi_{nm}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im}) \end{bmatrix}.$$

The final product is $W_i \cdot [R^{-1}A_i^{-\frac{1}{2}}(\underline{Y}_i - \underline{\mu}_i)] = W_i R^{-1} A_i^{-\frac{1}{2}}(\underline{Y}_i - \underline{\mu}_i)$:

$$W_i R^{-1} A_i^{-\frac{1}{2}}(\underline{Y}_i - \underline{\mu}_i) = \begin{bmatrix} w_{i11} & 0 & \cdots & 0 \\ 0 & w_{i22} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & w_{inn} \end{bmatrix} \cdot \begin{bmatrix} \sum_{m=1}^{n} \varphi_{1m}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im}) \\ \sum_{m=1}^{n} \varphi_{2m}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im}) \\ \vdots \\ \sum_{m=1}^{n} \varphi_{nm}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im}) \end{bmatrix}$$

$$= \begin{bmatrix} w_{i11} \cdot \sum_{m=1}^{n} \varphi_{1m}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im}) \\ w_{i22} \cdot \sum_{m=1}^{n} \varphi_{2m}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im}) \\ \vdots \\ w_{inn} \cdot \sum_{m=1}^{n} \varphi_{nm}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im}) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{m=1}^{n} w_{i11}\varphi_{1m}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im}) \\ \sum_{m=1}^{n} w_{i22}\varphi_{2m}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im}) \\ \vdots \\ \sum_{m=1}^{n} w_{inn}\varphi_{nm}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im}) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{m=1}^{n} \{\tfrac{1}{2}[[a''(\theta_{i1})]^{-\frac{3}{2}}a'''(\theta_{i1})](y_{i1} - \mu_{i1})\}\varphi_{1m}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im}) \\ \sum_{m=1}^{n} \{\tfrac{1}{2}[[a''(\theta_{i2})]^{-\frac{3}{2}}a'''(\theta_{i2})](y_{i2} - \mu_{i2})\}\varphi_{2m}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im}) \\ \vdots \\ \sum_{m=1}^{n} \{\tfrac{1}{2}[[a''(\theta_{in})]^{-\frac{3}{2}}a'''(\theta_{in})](y_{in} - \mu_{in})\}\varphi_{nm}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im}) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{m=1}^{n} \tfrac{1}{2}[a''(\theta_{i1})]^{-1}a'''(\theta_{i1})[a''(\theta_{i1})]^{-\frac{1}{2}}\varphi_{1m}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im})(y_{i1} - \mu_{i1}) \\ \sum_{m=1}^{n} \tfrac{1}{2}[a''(\theta_{i2})]^{-1}a'''(\theta_{i2})[a''(\theta_{i2})]^{-\frac{1}{2}}\varphi_{2m}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im})(y_{i2} - \mu_{i2}) \\ \vdots \\ \sum_{m=1}^{n} \tfrac{1}{2}[a''(\theta_{in})]^{-1}a'''(\theta_{in})[a''(\theta_{in})]^{-\frac{1}{2}}\varphi_{nm}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im})(y_{in} - \mu_{in}) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{m=1}^{n} \tfrac{1}{2}[a''(\theta_{i1})]^{-1}a'''(\theta_{i1})\phi^{-\frac{1}{2}}\sigma_{i11}^{-1}\varphi_{1m}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im})(y_{i1} - \mu_{i1}) \\ \sum_{m=1}^{n} \tfrac{1}{2}[a''(\theta_{i2})]^{-1}a'''(\theta_{i2})\phi^{-\frac{1}{2}}\sigma_{i22}^{-1}\varphi_{2m}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im})(y_{i2} - \mu_{i2}) \\ \vdots \\ \sum_{m=1}^{n} \tfrac{1}{2}[a''(\theta_{in})]^{-1}a'''(\theta_{in})\phi^{-\frac{1}{2}}\sigma_{inn}^{-1}\varphi_{nm}\phi^{-\frac{1}{2}}\sigma_{imm}^{-1}(y_{im} - \mu_{im})(y_{in} - \mu_{in}) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{m=1}^{n} \tfrac{1}{2}\phi^{-1}[a''(\theta_{i1})]^{-1}a'''(\theta_{i1})\cdot\varphi_{1m}\cdot(y_{im} - \mu_{im})(y_{i1} - \mu_{i1})/(\sigma_{i11}\sigma_{imm}) \\ \sum_{m=1}^{n} \tfrac{1}{2}\phi^{-1}[a''(\theta_{i2})]^{-1}a'''(\theta_{i2})\cdot\varphi_{2m}\cdot(y_{im} - \mu_{im})(y_{i2} - \mu_{i2})/(\sigma_{i22}\sigma_{imm}) \\ \vdots \\ \sum_{m=1}^{n} \tfrac{1}{2}\phi^{-1}[a''(\theta_{in})]^{-1}a'''(\theta_{in})\cdot\varphi_{nm}\cdot(y_{im} - \mu_{im})(y_{in} - \mu_{in})/(\sigma_{inn}\sigma_{imm}) \end{bmatrix}$$

The expectation of $W_i R^{-1} A_i^{-\frac{1}{2}}(\underline{y}_i - \underline{\mu}_i)$ indicated in Equation 3.21 is now evaluable. Using the expansion derived above we find:

$$E\{W_i R^{-1} A_i^{-\frac{1}{2}}(\underline{y}_i - \underline{\mu}_i)\}$$

$$= \begin{bmatrix} \sum_{m=1}^{n} \tfrac{1}{2}\phi^{-1}[a''(\theta_{i1})]^{-1}a'''(\theta_{i1})\cdot\varphi_{1m}\cdot E\{(y_{im} - \mu_{im})(y_{i1} - \mu_{i1})\}/(\sigma_{i11}\sigma_{imm}) \\ \sum_{m=1}^{n} \tfrac{1}{2}\phi^{-1}[a''(\theta_{i2})]^{-1}a'''(\theta_{i2})\cdot\varphi_{2m}\cdot E\{(y_{im} - \mu_{im})(y_{i2} - \mu_{i2})\}/(\sigma_{i22}\sigma_{imm}) \\ \vdots \\ \sum_{m=1}^{n} \tfrac{1}{2}\phi^{-1}[a''(\theta_{in})]^{-1}a'''(\theta_{in})\cdot\varphi_{nm}\cdot E\{(y_{im} - \mu_{im})(y_{in} - \mu_{in})\}/(\sigma_{inn}\sigma_{imm}) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{m=1}^{n} \frac{1}{2}\phi^{-1}\{a''(\theta_{i1})\}^{-1}a'''(\theta_{i1})\cdot\varphi_{1m}\cdot\sigma_{i1m}/(\sigma_{i11}\sigma_{imm}) \\ \sum_{m=1}^{n} \frac{1}{2}\phi^{-1}\{a''(\theta_{i2})\}^{-1}a'''(\theta_{i2})\cdot\varphi_{2m}\cdot\sigma_{i2m}/(\sigma_{i22}\sigma_{imm}) \\ \vdots \\ \sum_{m=1}^{n} \frac{1}{2}\phi^{-1}\{a''(\theta_{in})\}^{-1}a'''(\theta_{in})\cdot\varphi_{nm}\cdot\sigma_{inm}/(\sigma_{inn}\sigma_{imm}) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{m=1}^{n} \frac{1}{2}\phi^{-1}\{a''(\theta_{i1})\}^{-1}a'''(\theta_{i1})\cdot\varphi_{1m}\cdot\rho_{1m} \\ \sum_{m=1}^{n} \frac{1}{2}\phi^{-1}\{a''(\theta_{i2})\}^{-1}a'''(\theta_{i2})\cdot\varphi_{2m}\cdot\rho_{2m} \\ \vdots \\ \sum_{m=1}^{n} \frac{1}{2}\phi^{-1}\{a''(\theta_{in})\}^{-1}a'''(\theta_{in})\cdot\varphi_{nm}\cdot\rho_{nm} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{2}\phi^{-1}\{a''(\theta_{i1})\}^{-1}a'''(\theta_{i1}) \cdot \sum_{m=1}^{n}\varphi_{1m}\rho_{1m} \\ \frac{1}{2}\phi^{-1}\{a''(\theta_{i2})\}^{-1}a'''(\theta_{i2}) \cdot \sum_{m=1}^{n}\varphi_{2m}\rho_{2m} \\ \vdots \\ \frac{1}{2}\phi^{-1}\{a''(\theta_{in})\}^{-1}a'''(\theta_{in}) \cdot \sum_{m=1}^{n}\varphi_{nm}\rho_{nm} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{2}\phi^{-1}\{a''(\theta_{i1})\}^{-1}a'''(\theta_{i1})\cdot 1 \\ \frac{1}{2}\phi^{-1}\{a''(\theta_{i2})\}^{-1}a'''(\theta_{i2})\cdot 1 \\ \vdots \\ \frac{1}{2}\phi^{-1}\{a''(\theta_{in})\}^{-1}a'''(\theta_{in})\cdot 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}\phi^{-1}\cdot a'''(\theta_{i1})/a''(\theta_{i1}) \\ \frac{1}{2}\phi^{-1}\cdot a'''(\theta_{i2})/a''(\theta_{i2}) \\ \vdots \\ \frac{1}{2}\phi^{-1}\cdot a'''(\theta_{in})/a''(\theta_{in}) \end{bmatrix} . \quad (3.22)$$

This last result concludes that the expectation of $W_i R^{-1} A_i^{-\frac{1}{2}}(Y_i - \mu_i)$ is not, in general, identically the zero vector. Hence

$$\lim_{k\to\infty} E\{k^{\frac{1}{2}}(\hat{\beta} - \beta)\} = \lim_{k\to\infty} -k^{\frac{1}{2}}\tilde{H}^{-1}\cdot\sum_{i=1}^{k} X_i^T E\{W_i R^{-1} A_i^{-\frac{1}{2}}(Y_i - \mu_i)\} \neq \underline{0} . \quad (3.23)$$

The question of whether or not the MQL estimator possesses $k^{\frac{1}{2}}$-consistency depends on the evaluation of $a'''(\theta)$ for specifically postulated response distributions.

In the case that individual responses are distributed from members of the one-parameter exponential family, $a'''(\theta)$ is nonzero. This implies that the MQL estimator in this instance cannot be $k^{\frac{1}{2}}$-consistent. An analogous conclusion was reached by Fedorov (1972) in the more restrictive case of diagonal covariance structures.

On the other hand, $k^{\frac{1}{2}}$-consistency of the estimator is attained for normal response models because $a'''(\theta_{ij})$ is zero (see Appendix B). Because the MQL GEEs were earlier shown to reduce to that of the original GEE formulation under normality, this result affirms the consistency property originally proven by Liang and Zeger (1986).

Estimation of $V\{\hat{\beta}\}$

The negative of the inverse Hessian, divided by the scale, is shown to equal the asymptotic variance-covariance matrix of $\hat{\beta}$ in many

instances where $\hat{\underline{\beta}}$ consistently estimates $\underline{\beta}$ (Bishop, Fienburg, and Holland, 1975). Unfortunately this is not the case in the current situation. Furthermore it may not even be possible to extend a suitable theory. Yet, the Hessian does provide an intuitive measure of precision because its inverse is used for creation of $\underline{\delta}$-increments during the iterative computation of $\hat{\underline{\beta}}$. Upon this rationale, but without a direct theoretical basis, we adopt the aforementioned quantity as the MQL variance-covariance matrix estimator:

$$V\{\hat{\underline{\beta}}\} = -\hat{H}^{-1} / \hat{\phi} , \qquad (3.24)$$

where $\hat{H}$ and $\hat{\phi}$ are defined by Equations 3.18 and 3.19, respectively. In Chapter 6 we endeavor to justify this stance through examinations of empirical evidence.

CHAPTER IV

Correlation Matrix Considerations

Throughout the derivations in Chapter 3 the correlation matrix R is postulated to be constant, but arbitrary. This is because primary inference is directed towards the regression coefficients β. Even so, the correlation matrix directly affects the estimation process due to the presence of its inverse in both g and H (Equations 3.17 and 3.18).

We now investigate the role correlation plays within this framework. Specifically, we will posit conceivable forms of its structure. There are two benefits of doing so. The first is that explicit definition of the correlation matrix may allow for its inverse to be expressed in a simple analytic form, thus foregoing the need of performing extraneous numerical inversions during estimation. The second is that it provides an avenue for estimation of the correlation coefficients. These positions are exercised throughout the remainder of the chapter.

Attributes of Special Correlation Matrices

An arbitrary correlation matrix of dimension n×n is composed of up to $[n\cdot(n-1)]$ individual correlation coefficients. Due to the difficulty of providing an analytic form of the inverse in this general case, the literature advances three structure alternatives: i) independence, ii) exchangable, and iii) auto-regressive. These choices are convenient because the correlation coefficients are functions of the single parameter ρ.

It was noted above that possession of an analytic expression for R^{-1} is appealing for computative purposes. But this is also a fortuitous situation. It is shown in Section 4.2 that first and second derivatives of R^{-1} must also be available for the simultaneous estimation of both β and ρ. To facilitate this need we next derive these matrices for each proposed structure.

Independence structure.

The simplest assumption of form for correlation is that of independence. Its simplicity allows for ease of implementation into the MQL GEEs -- notably, no parameter estimation is actually required because ρ equals zero. This implies R is an identity matrix:

$$R = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

The inverse of this matrix also yields an identity, while consecutive derivatives of R^{-1} generate zero matrices. For these reasons alteration of the model as given in Chapter 3 is unnecessary.

On the other hand, independence exhibits a draw back. Because we intuitively believe this not to be the case with serial observations on the same subject, this structure is not a desirable preference.

Exchangable structure.

Another simple, but more realistic, assumption is that of exchangability. This structure prescribes that any differing pair of

responses for a subject are identically correlated: $R\{y_{ij}, y_{im}\} = \rho$. The representation in matrix form is:

$$R = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \rho & \rho & \cdots & 1 \end{bmatrix} = (1-\rho)\cdot I_n + \rho \cdot \underline{1}_n \underline{1}_n^T .$$

This matrix pattern is seen in Seber (1984) as possessing an inverse:

$$R^{-1} = \begin{bmatrix} \varphi_1 & \varphi_2 & \cdots & \varphi_2 \\ \varphi_2 & \varphi_1 & \cdots & \varphi_2 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \varphi_2 & \varphi_2 & \cdots & \varphi_1 \end{bmatrix}, \qquad (4.1)$$

where: $\varphi_1 = \dfrac{-\rho(n-2) - 1}{\rho^2(n-1) - \rho(n-2) - 1}$ and $\varphi_2 = \dfrac{\rho}{\rho^2(n-1) - \rho(n-2) - 1}$.

Derivatives of R^{-1} with respect to ρ retain the same form as that of Equation 4.1, with elements:

a) $(R^{-1})'$:

$$\varphi_1' = \frac{\rho(n-1)[\rho(n-2) + 2]}{\{(\rho-1)[\rho(n-1) + 1]\}^2}$$

$$\varphi_2' = \frac{-[\rho^2(n-1) + 1]}{\{(\rho-1)[\rho(n-1) + 1]\}^2}$$

b) $(R^{-1})''$:

$$\varphi_1'' = \frac{-2[\rho^3(n-1)^2(n-2) + 3\rho^2(n-1)^2 + (n-1)]}{\{(\rho-1)[\rho(n-1) + 1]\}^3}$$

$$\varphi_2'' = \frac{2[\rho^3(n-1)^2 + 3\rho(n-1) - (n-2)]}{\{(\rho-1)[\rho(n-1) + 1]\}^3} .$$

Auto-regressive structure.

Another conceivable assumption is that of auto-regressivity. This structure prescribes the correlation between pairs of responses on a subject, $R\{y_{ij}, y_{im}\}$, to equal $\rho^{|j-m|}$. Because it is credible that dependence tendency between outcomes may decrease as a function of increased time separation, this structure may be the most realistic proffered. Naturally it also presents the most elaborate configuration. Its representation in matrix form is:

$$R = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-3} & \rho^{n-2} & r^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-4} & \rho^{n-3} & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-5} & \rho^{n-4} & \rho^{n-3} \\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot & \cdot \\ \rho^{n-3} & \rho^{n-4} & \rho^{n-5} & \dots & 1 & \rho & \rho^2 \\ \rho^{n-2} & \rho^{n-3} & \rho^{n-4} & \dots & \rho & 1 & \rho \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & \rho^2 & \rho & 1 \end{bmatrix}$$

It can be shown through row operations (proof not provided) that this matrix pattern possesses an inverse that likewise has the most complex configuration:

$$R^{-1} = \begin{bmatrix} \varphi_1 & \varphi_3 & 0 & \cdots & 0 & 0 & 0 \\ \varphi_3 & \varphi_2 & \varphi_3 & \cdots & 0 & 0 & 0 \\ 0 & \varphi_3 & \varphi_2 & \cdots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & \varphi_2 & \varphi_3 & 0 \\ 0 & 0 & 0 & \cdots & \varphi_3 & \varphi_2 & \varphi_3 \\ 0 & 0 & 0 & \cdots & 0 & \varphi_3 & \varphi_1 \end{bmatrix}, \qquad (4.2)$$

where: $\varphi_1 = \dfrac{-1}{\rho^2 - 1}$, $\varphi_2 = \dfrac{-(1 + \rho^2)}{\rho^2 - 1}$, and $\varphi_3 = \dfrac{\rho}{\rho^2 - 1}$.

Derivatives of R^{-1} contain elements:

a) $(R^{-1})'$:

$$\varphi_1' = \frac{2\rho}{(\rho^2 - 1)^2} \qquad \varphi_2' = \frac{4\rho}{(\rho^2 - 1)^2} \qquad \varphi_3' = \frac{-(\rho^2 + 1)}{(\rho^2 - 1)^2}$$

b) $(R^{-1})''$:

$$\varphi_1'' = \frac{-2(3\rho^2 + 1)}{(\rho^2 - 1)^3} \qquad \varphi_2'' = \frac{-4(3\rho^2 + 1)}{(\rho^2 - 1)^3} \qquad \varphi_3'' = \frac{2\rho(\rho^2 + 3)}{(\rho^2 - 1)^3}.$$

Incorporation of Correlation Information

The attractiveness of either exchangable or auto-regressive structures is that only one additional parameter is necessary to include into the MQL GEEs framework. Because analytic forms for R^{-1} and its derivatives are available, extension of the iterative system (Equation 3.11) to include estimation of ρ follows.

Derivation of the gradient $g(\beta,\rho)$.

The opening discussion in Section 3.1 establishes the goal of the MQL estimator. Now, in addition to $\underline{\beta}$, we desire an estimate of ρ such that when both are simultaneously considered Q is minimized. To begin, we consider the partial differentiation of Q with respect to ρ:

$$\frac{\partial}{\partial \rho} Q = \frac{\partial}{\partial \rho} \left[\sum_{i=1}^{k} (Y_i - \mu_i)^T V_i^{-1} (Y_i - \mu_i) \right]$$

$$= \frac{\partial}{\partial \rho} \left[\sum_{i=1}^{k} (Y_i - \mu_i)^T (A_i^{-\frac{1}{2}} R^{-1} A_i^{-\frac{1}{2}}) (Y_i - \mu_i) \right]$$

$$= \frac{\partial}{\partial \rho} \left[\sum_{i=1}^{k} \{ [(Y_i - \mu_i)^T A_i^{-\frac{1}{2}}] \cdot [R^{-1} A_i^{-\frac{1}{2}} (Y_i - \mu_i)] \} \right]$$

$$= \sum_{i=1}^{k} \left[\frac{\partial}{\partial \rho} \{ [(Y_i - \mu_i)^T A_i^{-\frac{1}{2}}] \cdot [R^{-1} A_i^{-\frac{1}{2}} (Y_i - \mu_i)] \} \right]$$

$$= \sum_{i=1}^{k} \left[[(Y_i - \mu_i)^T A_i^{-\frac{1}{2}}] \cdot \frac{\partial}{\partial \rho} [R^{-1} A_i^{-\frac{1}{2}} (Y_i - \mu_i)] \right] \quad \text{(by Lemma 6)}$$

$$= \sum_{i=1}^{k} \left[[(Y_i - \mu_i)^T A_i^{-\frac{1}{2}}] \cdot \frac{\partial}{\partial \rho} [R^{-1} \cdot A_i^{-\frac{1}{2}} (Y_i - \mu_i)] \right]$$

$$= \sum_{i=1}^{k} (Y_i - \mu_i)^T A_i^{-\frac{1}{2}} \cdot \frac{\partial}{\partial \rho} (R^{-1}) \cdot A_i^{-\frac{1}{2}} (Y_i - \mu_i) \quad \text{(by Lemma 6)}$$

$$= \sum_{i=1}^{k} (Y_i - \mu_i)^T A_i^{-\frac{1}{2}} (R^{-1})' A_i^{-\frac{1}{2}} (Y_i - \mu_i) .$$

One note of caution before continuing. Recall that the expansion of Equation 3.1 leading to the MQL GEEs (Equation 3.5) used a division

by (-2) which effectively changed that problem from one of minimization to maximization. Because the current derivation has required no change of sign, one must be instituted in order to insure conformity with prior results. Upon performing this operation, the gradient scalar $g(\rho)$ of Q is defined as:

$$g(\rho) = -\sum_{i=1}^{k} (\underline{Y}_i - \underline{\mu}_i)^T A_i^{-\frac{1}{2}} (R^{-1})' A_i^{-\frac{1}{2}} (\underline{Y}_i - \underline{\mu}_i) \ . \qquad (4.3)$$

We now have explicit forms for the gradients of Q with respect to both $\underline{\beta}$ and ρ. Concatenation of Equations 3.9 and 4.3 results in the augmented gradient vector $g(\underline{\beta}, \rho)$:

$$g(\underline{\beta}, \rho) = \begin{bmatrix} g(\underline{\beta}) \\ \hline g(\rho) \end{bmatrix} \ . \qquad (4.4)$$

Derivation of the Hessian $H(\underline{\beta}, \rho)$.

The Taylor series expansion technique of Section 3.2 now utilizes the additional parameter ρ. Recall that $H(\underline{\beta})$ is the matrix of partial derivatives of g^T with respect to $\underline{\beta}$. Adaptation of Equation 3.11 for ρ infers that the augmented Hessian equals:

$$H(\underline{\beta}, \rho) = \left[\frac{\partial}{\partial \beta_0} g(\underline{\beta}, \rho) \quad \frac{\partial}{\partial \beta_1} g(\underline{\beta}, \rho) \quad \cdots \quad \frac{\partial}{\partial \beta_q} g(\underline{\beta}, \rho) \quad \frac{\partial}{\partial \rho} g(\underline{\beta}, \rho) \right]$$

$$= \left[\begin{array}{cccc|c} \frac{\partial}{\partial \beta_0} g(\underline{\beta}) & \frac{\partial}{\partial \beta_1} g(\underline{\beta}) & \cdots & \frac{\partial}{\partial \beta_q} g(\underline{\beta}) & \frac{\partial}{\partial \rho} g(\underline{\beta}) \\ \hline \frac{\partial}{\partial \beta_0} g(\rho) & \frac{\partial}{\partial \beta_1} g(\rho) & \cdots & \frac{\partial}{\partial \beta_q} g(\rho) & \frac{\partial}{\partial \rho} g(\rho) \end{array} \right]$$

$$= \left[\begin{array}{c|c} \dfrac{\partial}{\partial \underline{\beta}^T} g(\underline{\beta}) & \dfrac{\partial}{\partial \rho} g(\underline{\beta}) \\ \hline \dfrac{\partial}{\partial \underline{\beta}^T} g(\rho) & \dfrac{\partial}{\partial \rho} g(\rho) \end{array}\right] . \qquad (4.5)$$

Because the Hessian is symmetric, the partitions in the lower left and upper right corners of Equation 4.5 are transpositions. Also, the upper left corner is $H(\underline{\beta})$ while the lower right corner is $H(\rho)$. This allows us to reexpress the augmented Hessian by:

$$H(\underline{\beta}, \rho) = \left[\begin{array}{c|c} H(\underline{\beta}) & \dfrac{\partial}{\partial \rho} g(\underline{\beta}) \\ \hline \dfrac{\partial}{\partial \rho} [g(\underline{\beta})]^T & H(\rho) \end{array}\right] . \qquad (4.6)$$

An explicit expression for $H(\underline{\beta})$ already exists. Applying implicit partial differentiation to $g(\rho)$ (Equation 4.3) shows the Hessian $H(\rho)$ of Q equals:

$$H(\rho) = -\sum_{i=1}^{k} (\underline{y}_i - \underline{\mu}_i)^T A_i^{-\frac{1}{2}} (R^{-1}) " A_i^{-\frac{1}{2}} (\underline{y}_i - \underline{\mu}_i) . \qquad (4.7)$$

Similarly, partial differentiation of $g(\underline{\beta})$ (Equation 3.9) with respect to ρ yields:

$$\frac{\partial}{\partial \rho} g(\underline{\beta}) = \frac{\partial}{\partial \rho} \left[\sum_{i=1}^{k} X_i^T (A_i^{\frac{1}{2}} + W_i) R^{-1} A_i^{-\frac{1}{2}} (\underline{y}_i - \underline{\mu}_i) \right]$$

$$= \sum_{i=1}^{k} X_i^T (A_i^{\frac{1}{2}} + W_i) \cdot \frac{\partial}{\partial \rho} (R^{-1}) \cdot A_i^{-\frac{1}{2}} (\underline{y}_i - \underline{\mu}_i)$$

$$= \sum_{i=1}^{k} X_i^T (A_i^{\frac{1}{2}} + W_i)(R^{-1})' A_i^{-\frac{1}{2}} (Y_i - \mu_i) \ . \qquad (4.8)$$

Substitution of Equations 3.16 , 4.7, and 4.8 into Equation 4.6 results in the completed expression for $H(\underline{\beta}, \rho)$.

Constrained Joint Estimation of $(\underline{\beta}, \rho)$

The iterative system represented by Equation 3.11 has been extended naturally through derivation of the augmented gradient and Hessian (Equations 4.4 and 4.6). However, the range of ρ is restricted: $[-1 \leq \rho \leq 1]$, and the literature remains vague about implementation of constraints in general.

One remedy is programmatic. We assume that an initial estimate of ρ is available. During the iterative process its δ-increment is inspected before updating. The value is adjusted (by half-sizing, for instance) if necessary to meet the requirements of the constraint. This technique has been implemented and seems to work well in practice.

The solution of the MQL GEEs, cognizant of the above constraint, parallels the procedure described at the end of Section 3.2. At convergence we declare $(\hat{\underline{\beta}}, \hat{\rho})$ to be the MQL estimate of $(\underline{\beta}, \rho)$. The gradient and Hessian evaluated at $(\hat{\underline{\beta}}, \hat{\rho})$ are designated:

$$g(\hat{\underline{\beta}}, \hat{\rho}) = \begin{bmatrix} g(\hat{\underline{\beta}}) \\ \hline g(\hat{\rho}) \end{bmatrix} = \underline{0} \qquad (4.9)$$

$$H(\hat{\underline{\beta}}, \hat{\rho}) = \left[\begin{array}{c|c} H(\hat{\underline{\beta}}) & \dfrac{\partial}{\partial \hat{\rho}} g(\hat{\underline{\beta}}) \\ \hline \dfrac{\partial}{\partial \hat{\rho}} [g(\hat{\underline{\beta}})]^T & H(\hat{\rho}) \end{array}\right] \qquad (4.10)$$

where:

$$g(\hat{\beta}) = \sum_{i=1}^{k} X_i^T(\hat{A}_i^{\frac{1}{2}} + \hat{W}_i)\hat{R}^{-1}\hat{A}_i^{\frac{1}{2}}(Y_i - \hat{\mu}_i)$$

$$g(\hat{\rho}) = -\sum_{i=1}^{k}(Y_i - \hat{\mu}_i)^T\hat{A}_i^{\frac{1}{2}}(\hat{R}^{-1})'\hat{A}_i^{\frac{1}{2}}(Y_i - \hat{\mu}_i)$$

$$H(\hat{\beta}) = \sum_{i=1}^{k} X_i^T[\hat{U}_i \cdot \mathrm{diag}([Y_i - \hat{\mu}_i]^T\hat{A}_i^{\frac{1}{2}}\hat{R}^{-1}) - (\hat{A}_i^{\frac{1}{2}} + \hat{W}_i)\hat{R}^{-1}(\hat{A}_i^{\frac{1}{2}} + \hat{W}_i)]X_i$$

$$H(\hat{\rho}) = -\sum_{i=1}^{k}(Y_i - \hat{\mu}_i)^T\hat{A}_i^{\frac{1}{2}}(\hat{R}^{-1})''\hat{A}_i^{\frac{1}{2}}(Y_i - \hat{\mu}_i)$$

$$\frac{\partial}{\partial \hat{\rho}} g(\hat{\beta}) = \sum_{i=1}^{k} X_i^T(\hat{A}_i^{\frac{1}{2}} + \hat{W}_i)(\hat{R}^{-1})'\hat{A}_i^{\frac{1}{2}}(Y_i - \hat{\mu}_i) \ .$$

CHAPTER V

Canonical Link Details

The estimation framework of Equations 3.17 and 3.18, further augmented with correlation structure in Equations 4.9 and 4.10, is established as a generalization for responses from the exponential family: $f_Y(y;\theta,\phi) = \exp[\phi\{y\theta - a(\theta)\} + b(y,\phi)]$. We now examine several members of this family and derive the associated canonical link functions. The purpose is to obtain link-specific expressions for elements of the vector $\underline{\mu}_i$ and diagonal matrices A_i, W_i, and U_i appearing in these equations.

We recall from Equation 3.8 that the general form for the jth diagonal element of W_i is: $w_{ijj} = \frac{1}{2}[\{a''(\theta_{ij})\}^{-\frac{1}{2}}a'''(\theta_{ij})](y_{ij} - \mu_{ij})$. Use of simple algebra allows this to be expressed by:

$$w_{ijj} = \frac{1}{2}\frac{(y_{ij} - \mu_{ij})}{\{a''(\theta_{ij})\}^{\frac{1}{2}}} \cdot \left[\frac{a'''(\theta_{ij})}{a''(\theta_{ij})}\right]. \qquad (5.1)$$

Similarly, Equation 3.15 shows the general form for an element of U_i equals: $u_{ijj} = \frac{1}{2}[\{a''(\theta_{ij})\}^{-\frac{1}{2}}a''''(\theta_{ij}) - \frac{3}{2}\{a''(\theta_{ij})\}^{-\frac{1}{2}}\{a'''(\theta_{ij})\}^2](y_{ij} - \mu_{ij})$. This may be rewritten as:

$$u_{ijj} = \frac{1}{4}\frac{(y_{ij} - \mu_{ij})}{\{a''(\theta_{ij})\}^{\frac{1}{2}}} \cdot \left[2\cdot\frac{a''''(\theta_{ij})}{a''(\theta_{ij})} - 3\cdot\frac{\{a'''(\theta_{ij})\}^2}{\{a''(\theta_{ij})\}^2}\right]. \qquad (5.2)$$

These representations will prove useful in the subsequent derivations. Additionally, all subscripts are suppressed to reduce notational complexity. This should not pose any difficulty.

Identity Link

The normal distribution is a choice for continuous response models when the assumption of homoscedasticity is tenable. It is expressed in exponential family form by:

$$f_Y(y;\theta,\phi) = \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \tfrac{1}{2}[y^2/\sigma^2 + \ln(2\pi\sigma^2)]\right]$$

from which we observe: $\theta = \mu$ and $\phi = (\sigma^2)^{-1}$. Also, $a(\theta) = \mu^2/2 = \theta^2/2$, implying that consecutive derivatives are:

$$a'(\theta) = \theta \quad (= \mu) \qquad a''(\theta) = 1$$

$$a'''(\theta) = 0 \qquad a''''(\theta) = 0 .$$

Using these, expressions for w and u are found to be:

$$w = \frac{(y - \mu)}{\tfrac{1}{2}\{a''(\theta)\}^{\frac{1}{2}}} \cdot \left[\frac{0}{1}\right] = 0$$

and

$$u = \frac{(y - \mu)}{\tfrac{1}{4}\{a''(\theta)\}^{\frac{1}{2}}} \cdot \left[2 \cdot \frac{0}{1} - 3 \cdot \frac{0^2}{1^2}\right] = 0 .$$

The identity link is established by virtue of equality between θ and μ. The range of this link is the entire real line.

Log Link

The Poisson distribution is a choice for count response models. It is expressed in exponential family form by:

$$f_Y(y;\theta,\phi) = \exp\left[\{y\cdot\ln(\lambda) - \lambda\} - \ln(y!)\right]$$

from which we observe: $\theta = \ln(\lambda)$ and $\phi = 1$. Also, $a(\theta) = \lambda = e^{\theta}$, implying that consecutive derivatives are:

$$a'(\theta) = e^{\theta} \ (= \mu) \qquad a''(\theta) = e^{\theta}$$
$$a'''(\theta) = e^{\theta} \qquad a''''(\theta) = e^{\theta}\ .$$

Using these, expressions for w and u are found to be:

$$w = \frac{(y-\mu)}{\{a''(\theta)\}^{\frac{1}{2}}}\cdot\left[\frac{e^{\theta}}{e^{\theta}}\right] = \frac{(y-\mu)}{\{a''(\theta)\}^{\frac{1}{2}}}\cdot[1] = \frac{(y-\mu)}{\{a''(\theta)\}^{\frac{1}{2}}}$$

$$= \frac{(y-e^{\theta})}{e^{\frac{1}{2}\theta}} = \tfrac{1}{2}[ye^{-\frac{1}{2}\theta} - e^{\frac{1}{2}\theta}]$$

and

$$u = \frac{(y-\mu)}{\{a''(\theta)\}^{\frac{1}{2}}}\cdot\left[2\cdot\frac{e^{\theta}}{e^{\theta}} - 3\cdot\frac{e^{2\theta}}{e^{2\theta}}\right] = \frac{(y-m)}{\{a''(\theta)\}^{\frac{1}{2}}}\cdot[-1] = -\frac{(y-\mu)}{\{a''(\theta)\}^{\frac{1}{2}}}$$

$$= \frac{(e^{\theta}-y)}{e^{\frac{1}{2}\theta}} = \tfrac{1}{4}[e^{\frac{1}{2}\theta} - ye^{-\frac{1}{2}\theta}]\ .$$

The link function derivation is: $\mu = e^{\theta}$ implies $\theta = \ln(\mu) = h(\mu)$. The range of the log link is the entire real line.

Logit Link

The binomial distribution is a choice for dichotomous response models. It is expressed in exponential family form by:

$$f_Y(y;\theta,\phi) = \exp\left[\{y\cdot\ln[\pi/(1-\pi)] + n\cdot\ln(1-\pi)\} + \ln({}_nC_y)\right]$$

from which we observe: $\theta = \ln[\pi/(1-\pi)]$ and $\phi = 1$. Also, $a(\theta) = -(n)\cdot\ln(1-\pi) = n\cdot\ln(1 + e^{\theta})$. For single trials (n = 1) per subject this distribution is known as the Bernoulli, and it is in this case interest more often exists. Thus $a(\theta) = \ln(1 + e^{\theta})$, implying that consecutive derivatives are:

$$\begin{aligned} a'(\theta) &= e^{\theta}/(1 + e^{\theta}) \quad (= \mu) \\ a''(\theta) &= e^{\theta}/(1 + e^{\theta})^2 \\ a'''(\theta) &= e^{\theta}(1 - e^{\theta})/(1 + e^{\theta})^3 \\ a''''(\theta) &= e^{\theta}(1 - 4e^{\theta} + e^{2\theta})/(1 + e^{\theta})^4 \ . \end{aligned}$$

Using these, expressions for w and u are found to be:

$$w = \frac{(y - \mu)}{\frac{1}{2}\{a''(\theta)\}^{\frac{1}{2}}}\cdot\left[\frac{e^{\theta}(1 - e^{\theta})/(1 + e^{\theta})^3}{e^{\theta}/(1 + e^{\theta})^2}\right] = \frac{(y - \mu)}{\frac{1}{2}\{a''(\theta)\}^{\frac{1}{2}}}\cdot\frac{(1 - e^{\theta})}{(1 + e^{\theta})}$$

$$= \frac{(y - \mu)}{\frac{1}{2}e^{\frac{1}{2}\theta}/(1 + e^{\theta})}\cdot\frac{(1 - e^{\theta})}{(1 + e^{\theta})} = \frac{(y - \mu)(1 - e^{\theta})}{\frac{1}{2}e^{\frac{1}{2}\theta}} = \tfrac{1}{2}(y - \mu)(e^{-\frac{1}{2}\theta} - e^{\frac{1}{2}\theta})$$

$$= \tfrac{1}{2}(y - [e^{\theta}/(1 + e^{\theta})])(e^{-\frac{1}{2}\theta} - e^{\frac{1}{2}\theta})$$

and

$$u = \frac{(y-\mu)}{\frac{1}{4}[a''(\theta)]^{\frac{1}{2}}} \cdot \left[2 \cdot \frac{e^{\theta}(1-4e^{\theta}+e^{2\theta})/(1+e^{\theta})^4}{e^{\theta}/(1+e^{\theta})^2} - 3 \cdot \frac{[e^{\theta}(1-e^{\theta})/(1+e^{\theta})^3]^2}{[e^{\theta}/(1+e^{\theta})^2]^2} \right]$$

$$= \frac{(y-\mu)}{\frac{1}{4}[a''(\theta)]^{\frac{1}{2}}} \cdot \left[2 \cdot \frac{(1 - 4e^{\theta} + e^{2\theta})}{(1 + e^{\theta})^2} - 3 \cdot \frac{(1 - e^{\theta})^2}{(1 + e^{\theta})^2} \right]$$

$$= \frac{(y-\mu)}{\frac{1}{4}[a''(\theta)]^{\frac{1}{2}}} \cdot \left[\frac{2(1 - 4e^{\theta} + e^{2\theta}) - 3(1 - e^{\theta})^2}{(1 + e^{\theta})^2} \right]$$

$$= \frac{(y-\mu)}{\frac{1}{4}[a''(\theta)]^{\frac{1}{2}}} \cdot \left[\frac{2 - 8e^{\theta} + 2e^{2\theta} - 3 + 6e^{\theta} - e^{2\theta}}{(1 + e^{\theta})^2} \right]$$

$$= \frac{(y-\mu)}{\frac{1}{4}[a''(\theta)]^{\frac{1}{2}}} \cdot \left[\frac{-1 - 2e^{\theta}\, e^{\theta 2}}{(1 + e^{\theta})^2} \right] = \frac{(y-\mu)}{\frac{1}{4}[a''(\theta)]^{\frac{1}{2}}} \cdot \left[\frac{-(1 + e^{\theta})^2}{(1 + e^{\theta})^2} \right]$$

$$= -\frac{(y - [e^{\theta}/(1 + e^{\theta})])}{e^{\frac{1}{2}\theta}/(1 + e^{\theta})} = -\frac{y(1 + e^{\theta}) - e^{\theta}}{e^{\frac{1}{2}\theta}}$$

$$= \frac{1}{4}[e^{\frac{1}{2}\theta}(1 - y) - ye^{-\frac{1}{2}\theta}] \ .$$

The link function derivation is: $\mu = e^{\theta}/(1+e^{\theta})$ implies $\theta = \ln[\mu/(1-\mu)] = h(\mu)$. The range of the logit link is the entire real line.

Inverse Link

The gamma distribution is a choice for continuous response models when variance cannot be assumed constant. It is often observed in experiments that variance increases with the response. Because the

homoscedasticity requirement for usual linear modeling is violated, it may be plausible to postulate constant coefficient of variation (CV) instead. This is a feature of the gamma distribution, which is expressed in exponential family form by:

$$f_Y(y;\theta,\phi) = \exp\left[\, \nu\{y(-\lambda/n) + \ln(\lambda)\} + (\nu-1)\ln(y) - \ln[\Gamma(\nu)]\,\right]$$

from which we observe: $\theta = -\lambda/\nu$ and $\phi = \nu$. Also, $a(\theta) = -\ln(\lambda) = -[\ln(-\theta) + \ln(\nu)]$, implying that consecutive derivatives are:

$$a'(\theta) = -1/\theta \quad (= \mu) \qquad a''(\theta) = 1/\theta^2$$

$$a'''(\theta) = -2/\theta^3 \qquad a''''(\theta) = 6/\theta^4 .$$

Using these, expressions for w and u are found to be:

$$w = \frac{(y-\mu)}{\frac{1}{2}\{a''(\theta)\}^{\frac{1}{2}}} \cdot \left[\frac{-2/\theta^3}{1/\theta^2}\right] = \frac{(y-\mu)}{\frac{1}{2}\{a''(\theta)\}^{\frac{1}{2}}} \cdot [-2/\theta] = \frac{-(y-\mu)}{\{a''(\theta)\}^{\frac{1}{2}}\theta}$$

$$= \frac{-(y - [-1/\theta])}{(1/\theta^2)^{\frac{1}{2}}\theta} = -(y + 1/\theta)$$

and

$$u = \frac{(y-\mu)}{\frac{1}{4}\{a''(\theta)\}^{\frac{1}{2}}} \cdot \left[2 \cdot \frac{6/\theta^4}{1/\theta^2} - 3 \cdot \frac{[-2/\theta^3]^2}{[1/\theta^2]^2}\right]$$

$$= \frac{(y-\mu)}{\frac{1}{4}\{a''(\theta)\}^{\frac{1}{2}}} \cdot \left[\frac{12}{\theta^2} - \frac{12}{\theta^2}\right] = 0 .$$

The link function derivation is: $\mu = -1/\theta$ implies $\theta = -1/\mu = h(\mu)$. Unlike those derived earlier, the inverse link has a constrained range because λ is positive but θ equals $(-\lambda)$. Therefore the range is the negative segment of the real line.

Implementation of Methodology

The previous sections supply link-specific forms for elements of the vectors and matrices comprising the gradient and Hessian (Equations 4.9 and 4.10). These are summarized in Table 1.

TABLE 1

Link-Specific Elements of Vectors and Matrices

Link (Distribution)	μ	A	W	U
Identity (normal)	θ	1	0	0
Log (Poisson)	e^{θ}	e^{θ}	$\frac{1}{2}(ye^{-\frac{1}{2}\theta} - e^{\frac{1}{2}\theta})$	$\frac{1}{4}(e^{\frac{1}{2}\theta} - ye^{-\frac{1}{2}\theta})$
Logit (binomial)	$\frac{e^{\theta}}{1 + e^{\theta}}$	$\frac{e^{\theta}}{(1 + e^{\theta})^2}$	$\frac{1}{2}(y-\mu)(e^{-\frac{1}{2}\theta}-e^{\frac{1}{2}\theta})$	$\frac{1}{4}[e^{\frac{1}{2}\theta}(1-y) - ye^{-\frac{1}{2}\theta}]$
Inverse (gamma)	$-1/\theta$	$1/\theta^2$	$-(y + 1/\theta)$	0

This completes the details necessary for implemention of a flexible algorithm to solve the MQL GEEs. Appendix B contains a programmed implementation of this methodology. Modeling support is currently provided for identity, log, and logit links only; however, all three postulated correlation structures are automatically invoked.

CHAPTER VI

Simulation Studies

The derivations in Chapters 3, 4, and 5 provide means of computing the MQL GEE estimator, but its only statistical property studied thus far is undesirable. The original GEE estimator is well established, plus it possesses several desirable properties (Liang and Zeger, 1986). Except in the case of normal response data practitioners are faced with a choice in methodology.

Criteria on which to base a decision are clearly needed. The one we presently emphasize is assessment of estimator performance. To assist in this evaluation we first compute both MQL and original GEE information-based estimates. These in turn are matched and compared with sampling-based counterparts. Several examples are taken from the literature to motivate this theme. Discussion of Monte Carlo simulation results conclude the chapter.

Jackknife and Bootstrap Procedures

In many situations it may be difficult, if not impossible, to obtain moment-derived representations for the expectation and variance of estimators. Two nonparametric resampling procedures available for approximating these are the jackknife and bootstrap. The jackknife was introduced by Tukey (1958) and subsequently examined by Miller (1974). The bootstrap was conceived by Efron (1979). Both were illustrated by Efron and Gong (1983). These procedures are depicted next.

Consider a resampling scheme in which replicates are generated from the original sample by excluding data for one subject at a time. This technique will produce k distinct replicates of size (k - 1). Let $\hat{\beta}_{(-i)}$ represent the estimate based on the ith replicate. The jackknife estimator $\hat{\beta}_J$ is defined as the arithmetic average of these k combined estimates, and its variance is approximated by:

$$V[\hat{\beta}_J] = \frac{k-1}{k} \sum_{i=1}^{k} (\hat{\beta}_{(-i)} - \hat{\beta}_J)(\hat{\beta}_{(-i)} - \hat{\beta}_J)^T.$$

A different resampling scheme allows creation of replicates of size k formed by drawing from the original sample with replacement. In this scenario $(2^k - 1)$ distinct replicates are possible. Realizing this provides a sizable number of combinations for even modest k we opt to perform resampling at random some large number B of times. Let $\hat{\beta}_{(b)}$ represent the estimate based on the bth replicate. The bootstrap estimator $\hat{\beta}_B$ is defined as the arithmetic average of these B combined estimates, and its variance is approximated by:

$$V[\hat{\beta}_B] = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\beta}_{(b)} - \hat{\beta}_B)(\hat{\beta}_{(b)} - \hat{\beta}_B)^T.$$

Comparison of Estimation Methodologies

We next consider four separate applications of the MQL and original GEE methodologies to independent univariate data:

1) Case 1 -- Binary response (small sample)
2) Case 2 -- Binary response (large sample)
3) Case 3 -- Poisson response (small sample)
4) Case 4 -- Poisson response (large sample).

Each case data is initially fit using both methodologies. By virtue of independence, the latter method degenerates to the usual iteratively reweighted least squares (IRLS). The estimates and standard deviations generated are denoted as being information-based. The alert reader may notice discrepancies between the IRLS information-based standard errors presented here and those reported in the referenced works. This is because the scale parameter ϕ must be incorporated into variance computations and we use Equation 3.19 to produce its estimate. Most commercially available software assume a value of unity for ϕ in both logistic and Poisson regression algorithms.

Analyses continue with the computation of jackknife and (unless noted) bootstrap estimators, the latter of which is based on B = 10,000 replicates. Tabulation of modeling results is provided for visual aid. Standard errors are displayed within parentheses juxtaposed to their parameter estimates. Discussion of comparisons concludes each case.

Binary response -- small sample.

Finney (1947) investigated factors related with vaso-constriction of the fingers (response). Using results obtained from a sample of 39 subjects, he proposed a model relating response to the rate and volume of inspired air. This model is equivalent to the logistic regression reformulation:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \cdot \log(\text{RATE}_i) + \beta_2 \cdot \log(\text{VOLUME}_i) + \varepsilon_i \ .$$

This data is well known, having been referenced in many articles related to logistic regression. Of particular interest is the fact that observations 4 and 18 are known outliers. For this reason, Pregibon (1981) used the data to illustrate regression diagnostic techniques.

More recently, Moulton (1986) applied it to demonstrate variations on the bootstrap estimator for GLM-based models. He noted (p. 26) that certain bootstrap replicates would not lead to finite parameter estimates, and this situation was indeed observed in our investigation. Because exclusion of these replicates would induce selection bias, we elect to bypass bootstrap procedures for this data.

Application of the data to model fitting produces quadratics at convergence of 23.81059 and 34.23381 for MQL and IRLS methods, respectively. As noted earlier, observations 4 and 18 are definite outliers as determined by IRLS ($P < .005$) but are less pronounced when fit via MQL ($P = .029$ and .048, respectively). Table 2 displays results for information- and jackknife-based estimators only.

TABLE 2

Modeling of Binary Response Data Using Small Samples

	MQL		IRLS	
Coef	Information	Jackknife	Information	Jackknife
β_0	-1.3428(0.488)	-1.3508(0.752)	-2.8754(1.288)	-2.9210(3.043)
β_1	2.3866(0.819)	2.3979(0.959)	4.5617(1.792)	4.6217(3.753)
β_2	2.6741(0.756)	2.6869(1.070)	5.1793(1.818)	5.2474(4.124)

Inspection of Table 2 indicates that MQL parameter estimates are roughly on the order of one-half the size of their IRLS counterparts. We return to this matter in Chapter 7. Of special note is the close agreement between the MQL-derived estimators. This contrasts with standard errors for IRLS jackknife parameter estimates being more than twice the size of those obtained from information. It turns out that jackknife replicates 4 and 18 (those excluding each outlier) yield IRLS

parameter estimates considerably different from the other 37, whereas MQL parameter estimates are not apparently affected to the same degree. This may lend the MQL technique to offer some measure of resistence against influence from outlying observations.

Binary response -- large sample.

Hosmer and Lemeshow (1989) provided 200 cases of birth information collected at the Baystate Medical Center (Springfield, MA) during 1986. Low birth weight, defined as body mass less than 2500 g, is a known infant mortality risk factor. On this basis infant weights were dichotomized into one of two categories: low versus normal birth weight (response). Additional maternal data was also included:

1. Age (years)
2. Weight pre-conception (lb)
3. Race (recoded: white or nonwhite)
4. Smoker (yes or no)
5 Prior premature deliveries (count)
6. Hypertensive (yes or no)
7. Uterine irritability (yes or no)
8. Visits to physician during first trimester (count)

The purpose of the study was to determine if factors associated with low birth weight could be identified. Because the response was binary, logistic regression modeling was indicated.

The data was applied to MQL and IRLS fitting techniques. For purposes of illustration let us assume that the full main effects model is appropriate, where the coefficient subscripts are equated to the covariate reference enumeration above. These methods produce quadratics

of 156.03941 and 182.32798, respectively, at convergence. Only one observation is detected as an outlier using MQL whereas seven are implicated by the IRLS counterpart. Table 3 displays results of the modeling.

TABLE 3

Modeling of Binary Response Data Using Large Samples

	MQL		
Coef	Information	Jackknife	Bootstrap
β_0	0.2310(0.7009)	0.2314(0.6578)	0.2833(0.6855)
β_1	-0.0207(0.0232)	-0.0207(0.0198)	-0.0218(0.0203)
β_2	-0.0070(0.0038)	-0.0070(0.0038)	-0.0077(0.0039)
β_3	0.5327(0.2382)	0.5329(0.2261)	0.5628(0.2329)
β_4	0.5489(0.2424)	0.5491(0.2217)	0.5813(0.2330)
β_5	0.2558(0.1976)	0.2561(0.2241)	0.3081(0.2341)
β_6	0.9251(0.4473)	0.9257(0.4153)	1.1074(1.5755)
β_7	0.3998(0.2801)	0.4000(0.2675)	0.4195(0.2744)
β_8	0.0189(0.1067)	0.0189(0.0950)	0.0137(0.0988)
	IRLS		
Coef	Information	Jackknife	Bootstrap
β_0	0.3028(1.1698)	0.3034(1.2849)	0.3913(1.3766)
β_1	-0.0314(0.0371)	-0.0314(0.0374)	-0.0333(0.0392)
β_2	-0.0140(0.0066)	-0.0140(0.0076)	-0.0151(0.0081)
β_3	1.0186(0.4000)	1.0188(0.4231)	1.0541(0.4463)
β_4	0.9934(0.3985)	0.9936(0.4092)	1.0452(0.4425)
β_5	0.5381(0.3480)	0.5386(0.4618)	0.6409(0.4944)
β_6	1.8460(0.7006)	1.8468(0.7573)	2.0898(1.7438)
β_7	0.7514(0.4648)	0.7516(0.5364)	0.7952(0.5626)
β_8	0.0699(0.1717)	0.0698(0.1795)	0.0560(0.1959)

As was the case with the smaller sample, inspection of Table 3 indicates the MQL parameter estimate values to be roughly one-half the size of those produced by IRLS. In contrast to that case, however, is closer agreement between information- and jackknife-based estimates for both techniques. Bootstrap parameter estimates tend to be somewhat larger and more uncertain (especially β_6) across the board.

Poisson response -- small sample.

McCullagh and Nelder (1983) analyzed shipping-induced cargo damage incidences (response) from 34 vessels. The data was supplied by Lloyd's Register of Shipping. Concomitant data included ship type, year of construction, service period, and length of aggregate service (years).

In an attempt to provide some measure of design balance, the data was selected based on categories defined by the first three factors:

Ship type (code) -- A, B, C, D, and E

Year constructed -- 1960-64, 1965-69, 1970-74, and 1975-79

Period of operation -- 1960-74 and 1975-79 .

At first glance it would appear that these would allow for 40 possible combinations; however, some combinations are clearly impossible and it was for this reason only 34 observations were included.

Coding of indicator variables within factor was based on contrasting the first level to the rest. As a result eight design variables were incorporated into the linear predictor. The natural logarithm of aggregate service duration entered into the linear predictor as an offset -- a factor with known parameter coefficient of unity. This was done to encompass cumulative time-dependent hazard effects.

Application of the data to MQL and IRLS methods yield quadratics of 36.39346 and 42.27525, respectively. The same single outlier is identified by both techniques. A summary of results is displayed in Table 4.

TABLE 4

Modeling of Poisson Response Data Using Small Samples

	M Q L		
Coef	Information	Jackknife	Bootstrap
β_0	-6.3722(0.2636)	-6.3708(0.2894)	-6.6831(1.8344)
β_1	-0.5658(0.2145)	-0.5660(0.1575)	-0.4819(0.8997)
β_2	-0.2473(0.3384)	-0.2582(0.8433)	-0.3496(1.0536)
β_3	0.1066(0.3232)	0.0996(0.6435)	-0.5854(3.0511)
β_4	0.4540(0.2763)	0.4552(0.3792)	0.4901(1.0491)
β_5	0.7127(0.1783)	0.7121(0.2605)	0.9611(1.6982)
β_6	0.8073(0.1992)	0.8051(0.2892)	1.0546(1.6827)
β_7	0.4602(0.2684)	0.4577(0.3665)	0.6819(1.7701)
β_8	0.3666(0.1385)	0.3661(0.1897)	0.3463(0.2515)
	I R L S		
Coef	Information	Jackknife	Bootstrap
β_0	-6.4059(0.2828)	-6.4051(0.2727)	-6.7514(1.9232)
β_1	-0.5433(0.2309)	-0.5430(0.1440)	-0.4457(0.9539)
β_2	-0.6874(0.4279)	-0.6881(0.7291)	-0.6358(1.0943)
β_3	-0.0760(0.3779)	-0.0808(0.7430)	-0.7708(3.1742)
β_4	0.3256(0.3067)	0.3292(0.3649)	0.4153(1.1113)
β_5	0.6971(0.1946)	0.6959(0.2545)	0.9528(1.7746)
β_6	0.8184(0.2208)	0.8164(0.2682)	1.0967(1.7610)
β_7	0.4534(0.3032)	0.4502(0.3724)	0.7026(1.8503)
β_8	0.3845(0.1538)	0.3848(0.1822)	0.3726(0.2566)

Inspection of Table 4 seems to indicate closer overall agreement between parameter estimates obtained via MQL and IRLS for Poisson response data than is observed in the binary case. In line with earlier results, information- and jacknife-based parameter estimates obtained within technique remain comparable, with slightly more variability accorded to the jackknife. Bootstrap-based values are typically higher and have standard errors generally several times greater than the other two. This could be attributed to the small sample size.

Poisson response -- large sample.

Zeger (1988) examined 168 monthly poliomyelitis incidence frequencies (response) for the years 1970 through 1983 inclusive. This data was obtained from the U.S. Centers for Disease Control (CDC). The purpose of the analysis was to determine the significance of an observed overall decreasing trend in incidence. Three models were contrasted, one of which was the log-linear model that corresponds to the usual Poisson regression.

No exogenous concomitant information was used, and for this reason models were presumed to be functions of time only. Time units were in months, with centering (t = 0) occurring on January, 1977 (the article stated 1976 but this was an apparent misprint). Factors included were:

Linear component

1. trend ($t \times 10^{-3}$)

Cyclical components

2. $\cos(2\pi t/12)$
3. $\sin(2\pi t/12)$
4. $\cos(2\pi t/6)$
5. $\sin(2\pi t/6)$.

The data was applied to MQL and IRLS fitting methods. Quadratics produced at convergence are 245.59862 and 318.72163, respectively. Five outliers are identified by the MQL technique, whereas six are observed using the model determined by IRLS. Results are listed in Table 5.

TABLE 5

Modeling of Poisson Response Data Using Large Samples

	M Q L		
Coef	Information	Jackknife	Bootstrap
β_0	0.6019(0.0743)	0.6016(0.1032)	0.8876(0.1858)
β_1	-4.1119(1.3574)	-4.1101(2.3523)	-3.9145(1.9796)
β_2	-0.1743(0.0963)	-0.1742(0.1386)	-0.1644(0.1232)
β_3	-0.4197(0.1045)	-0.4197(0.1795)	-0.4203(0.1499)
β_4	0.1273(0.1000)	0.1273(0.1541)	0.1357(0.1299)
β_5	-0.4994(0.0974)	-0.4991(0.1620)	-0.4623(0.1392)
	I R L S		
Coef	Information	Jackknife	Bootstrap
β_0	0.1494(0.1078)	0.1492(0.1051)	0.1198(0.1021)
β_1	-4.7987(1.9678)	-4.7983(2.3077)	-4.7107(2.2239)
β_2	-0.1487(0.1364)	-0.1487(0.1402)	-0.1463(0.1351)
β_3	-0.5319(0.1530)	-0.5319(0.1651)	-0.5338(0.1574)
β_4	0.1691(0.1386)	0.1691(0.1433)	0.1728(0.1374)
β_5	-0.4321(0.1414)	-0.4321(0.1525)	-0.4259(0.1468)

Examination of Table 5 discloses patterns of parameter estimates similar to those seen in the previous Poisson case. On the other hand, standard error estimates deserve closer attention. The information-based MQL values are somewhat smaller than those derived from jackknife and bootstrap procedures, whereas IRLS yields values in closer agreement

across all three. It is not known whether or not this is an artifact of the data or a property of the MQL for Poisson data.

Discussion of estimator comparisons.

Several issues need to be explored in regards to the cases presented. First, the MQL and IRLS fitting methodologies produce disparate parameter estimates. Direct comparison of method performance cannot be made solely on the basis of these values. On the other hand, similar patterns emerge in the estimates generated by both methods -- information- and jackknife-based values tend to be in closer proximity whereas bootstrap-based values are somewhat larger.

The MQL method fits the data "closer" under conditions postulated in Chapter 3, and this is confirmed by the reported quadratics. A possible attribute of this technique, noted only in binary cases, is the perceived regression resistence offered against outliers. This characteristic is highly desirable, especially when models are produced for predictive purposes. Also observed is the implication of fewer responses being tagged as outliers. Because the overall fit is optimal with respect to the quadratic, it seems reasonable that individual fits should benefit accordingly. On the basis of these accounts the MQL method embeds a favorable impression.

Both MQL and IRLS methods signify that sample size is a factor in measurement of precision for parameters. For small samples (Table 4) it is observed that values of bootstrap standard errors are generally several times greater than their information- and jackknife-based counterparts. This discrepancy is ameliorated for larger samples, but more important is the observation that MQL information-based standard

errors are reasonably comparable to those generated by the jackknife and bootstrap in these cases (see Tables 3 and 5).

We recall that the negative inverse of the Hessian, divided by the scale parameter, is used as the asymptotic variance-covariance (Equation 3.24) without theoretical basis. More work is clearly indicated to attain an adequate resolution of the conjecture. Until then, one could argue that general agreement with sampling-based estimates provides evidence to justify its continued use.

Monte Carlo Simulation

The parameter estimates generated in Section 6.2 are based on known samples. Next, we consider procedures in which parameter values are fixed, but responses are generated at random. The collective term for this family of techniques is known as Monte Carlo simulation. Rubinstein (1981) provided a rigorous development of its many aspects.

The particular implementation we use is called the hit-or-miss method. Hosmer and Lemeshow (1989) describe its application in relation to logistic regression. We next adapt their development of a modeling procedure for use in contrasting MQL and IRLS methodologies.

The vital status (response) of 200 intensive care unit (ICU) patients were recorded along with age (years) and chronic renal failure history (CRN) indicator. The logistic model is fit using both MQL and IRLS methods. Parameter estimates obtained are:

$$\text{MQL:} \quad \underline{\beta} = (-1.5238,\ 0.0126,\ 0.5094)^T$$

$$\text{IRLS:} \quad \underline{\beta} = (-3.0299,\ 0.0250,\ 1.0199)^T$$

with respect to the intercept, age, and CRN indicator, respectively. These values are considered fixed during subsequent simulations.

TABLE 6

Monte Carlo Simulations

	I. MQL-Derived Parameters		II. IRLS-Derived Parameters	
	i	ii	i	ii
A. 40 obs/rep				
Coef	MQL	IRLS	MQL	IRLS
β_0	-0.8156(0.720)	-1.6006(1.374)	-1.8131(1.376)	-3.3927(2.152)
β_1	0.0068(0.012)	0.0131(0.023)	0.0159(0.020)	0.0286(0.032)
β_2	-0.4380(8.462)	-0.1697(8.875)	-1.0009(8.443)	-0.4379(8.833)
B. 160 obs/rep				
Coef	MQL	IRLS	MQL	IRLS
β_0	-0.7857(0.291)	-1.5679(0.578)	-1.5877(0.404)	-3.1449(0.767)
β_1	0.0066(0.005)	0.0131(0.009)	0.0134(0.006)	0.0264(0.012)
β_2	0.2180(0.893)	0.4724(1.075)	0.4613(0.894)	0.9626(1.080)
C. 640 obs/rep				
Coef	MQL	IRLS	MQL	IRLS
β_0	-0.7763(0.141)	-1.5320(0.281)	-1.5304(0.201)	-3.0541(0.391)
β_1	0.0064(0.002)	0.0127(0.005)	0.0128(0.003)	0.0254(0.006)
β_2	0.2546(0.146)	0.5096(0.291)	0.5035(0.152)	1.0084(0.302)

Before discussing simulation results, a note on the covariates is in order. The original data set showed that 40/200 (20.0%) patients incurred the terminal event. Of these, 8 (20.0%) had a history of CRN. On the other hand, only 11/160 (6.9%) survivors had a similar history. This indicates the obvious reason for inclusion of CRN into the

predictor. Similarly, there existed a trend for higher mortality in older patients. Thus, age was deemed important.

Overall, 19/200 (9.5%) patients were coded as CRN-positive. Citing this, we generated values based on comparison with the pseudo-random deviate $u \sim U(0,1)$: CRN = 1 if $u < 0.095$; CRN = 0 otherwise. The average age of CRN-negative patients was 56.53 years, whereas 67.33 years was noted for the CRN-positive group. It was decided to generate ages from $N(56.53, 18^2)$, with 10.68 years added to those indicated as CRN-positive.

The simulation study performed is based on 1000 replicates. These vary in sizes of: A) 40, B) 160, and C) 640 observations. Each replicate is formed by combining identical pseudo-randomly generated covariate data with the assumed known parameters obtained from MQL and IRLS methodologies. The resulting linear predictors are in turn used to calculate expectations of response π_{MQL} and π_{IRLS}. Responses are formed randomly based on hit-or-miss:

1) generate $u \sim U(0,1)$

2) assign $y = \begin{cases} 1, & \text{if } \pi > u \\ 0, & \text{otherwise} \end{cases}$

In other words, a terminal event is predicted only if its expected value is greater than a uniform pseudo-random deviate. Because MQL and IRLS methods yield different expectations, only the generated responses vary between otherwise identical sets of replicates.

The results of the simulation study are reported in Table 6. Each replicate is fitted using both MQL and IRLS. Consequently, two sets of parameter estimates (i and ii) are listed under each replicate basis (I

and II). The first notable remark is that larger replicate sizes yield smaller standard errors in every category. Significance is not even attained for replicates of size 40, and this is remindful of the bootstrap estimates noted in Table 4. This indicates that sample sizes may need to be larger in practice, as demonstrated in the results for sizes of 160 and 640. Next, values in columns Iii) and IIii) evince a reproductive ability of IRLS. It is known that this method produces consistent and unbiased estimates, and this is clearly evident from the close agreement between the values produced and their progenitor parameters. This is in obvious contrast to columns Ii) and IIi) which contain values highly different from those on which the replicates are based. We had already suspected the existence of estimation bias by the MQL method given its general inconsistency property. The values displayed here, and in Tables 2 and 3, highlight its level for the binary response case. The general trend is for MQL parameter estimates to be on the order of 50% smaller than expected. We suspect the bias is less pronounce, or even inconsequential, when modeling responses from distributions permitting greater range of outcome (e.g. Poisson). This last remark is only a conjecture and is therefore a subject worthy of further investigation.

CHAPTER VII

Summary and Conclusion

We confirmed that the original GEE formulation is possible through differentiation of the Mahalanobis distance D^2. At that juncture we relaxed the constraint of modeling variances solely through known functions of expectations. Instead, minimization of the quadratic form was chosen as the optimality property of the derived estimator $\hat{\underline{\beta}}$, which we denoted the MQL GEE estimator of $\underline{\beta}$. It was found that this estimator is generally inconsistent, and its variance estimator was derived on an intuitive, rather than theoretical, basis.

Correlation structure incorporation based on three possibilities advanced in the literature was augmented onto the estimation framework. This induced the estimation of a solitary additional parameter. Finally, several distributions from the exponential family were examined in respect to the elements of vectors and matrices comprising the gradient and Hessian.

In spite of the statistical problems encountered, case examinations of performance supported the MQL GEE method as a contender to the original GEE formulation. Especially notable was the perceived regression resistance capability in binary response models and the implication of fewer outliers. Also important was the general agreement observed between the information-based standard errors and sampling-based measures for large sample cases.

Before we continue, the reader may wonder why case studies were performed using univariate response data only. The primary reason is because the data sets used are readily available. For example, the Finney (1947) data is well known and has been the object of considerable attention in logistic regression literature. Because its outliers are highly influential under IRLS (Pregibon, 1981), it was of interest to see the effect they would inflict on the MQL method. The remaining cases were chosen on the basis of sample size and response category.

This research has identified several areas that suggest further attention. First, the observed general agreement of estimator standard errors does not infer confirmation of Equation 3.24 as being legitimate. Next, the regression resistance perception needs clarification due to the importance of this concept for predictive models. Also, the empirical relation between MQL and IRLS estimators in binary response cases needs resolution. It seems more than coincidental that values from MQL are roughly one half those generated by IRLS; however, no connection has been found thus far analytically. Finally, the related topic of bias assessment for the methodology in general requires investigation.

REFERENCES

Allison, P. D. (1984). Event history analysis: Regression for longitudinal event data. Beverly Hills, CA: Sage.

Barnett, W. A. (1976). Maximum likelihood and iterated Aitken estimation of nonlinear systems of equations. Journal of the American Statistical Association, 71, 354-360.

Bickel, P. J. & Doksum, K. A. (1977). Mathematical statistics: Basic ideas and selected topics. Oakland, CA: Holden-Day.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). Discrete multivariate analysis: Theory and practice. Cambridge, MA: MIT Press.

Bonney, G. E. (1987). Logistic regression for dependent binary observations. Biometrics, 43, 951-973.

Carr, G. J. & Chi, E. M. (1992). Analysis of variance for repeated measures data: A generalized estimating equations approach. Statistics in Medicine, 11, 1033-1040.

Connolly, M. A. & Liang, K. Y. (1988). Conditional logistic regression models for correlated binary data. Biometrika, 75, 501-506.

Cook, N. R. (1982). A general linear model approach to longitudinal data analysis. Unpublished doctoral dissertation, Harvard University, Boston.

Cook, N. R. & Ware, J. H. (1983). Design and analysis methods for longitudinal research. Annual Review of Public Health, 4, 1-23.

Cook, N. R., Scherr, P. A., Evans, D. A., Laughlin, L. W., Chapman, W. G., Rosner, B., Kass, E. H., Taylor, J. O., & Hennekens, C. H. (1985). Regression analysis of changes in blood pressure with oral contraceptive use. American Journal of Epidemiology, 121, 530-540.

Crowder, M. J. & Hand, D. J. (1990). Analysis of repeated measures. New York: Chapman and Hall.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. Annals of Statistics, 7, 1-26.

Efron, B. & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician, 37, 36-48.

Fearn, T. (1975). A Bayesian approach to growth curves. Biometrika, 62, 89-100.

Fedorov, V. V. (1972). Theory of optimal experiments. New York: Academic Press.

Finney, D. J. (1947). The estimation from individual records of the relationship between dose and quantal response. Biometrika, 34, 320-334.

Forsythe, G. E., Malcolm, M. A., & Moler, C. B. (1977). Computer Methods for Mathematical Computations. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Geisser, S. (1970). Bayesian analysis of growth curves. Sanhkyá Series A, 32, 53-64.

Glindmeyer, H. W., Diem, J. E., Jones, R. N., & Weill, H. (1982). Noncomparability of longitudinally and cross-sectionally determined annual change in spirometry. American Review of Respiratory Disease, 125, 544-548.

Graybill, F. A. (1976). Theory and application of the linear model. Pacific Grove, CA: Wadsworth & Brooks/Cole.

Grizzle, J. E. & Allen, D. M. (1969). Analysis of growth and dose response curves. Biometrics, 25, 357-381.

Hosmer, D.W. & Lemeshow, S. (1989). Applied logistic regression. New York: John Wiley & Sons.

Johnson, R. A. & Wichern, D. W. (1982). Applied multivariate statistical analysis. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Korn, E. L. & Whittemore, A. S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. Biometrics, 35, 795-802.

Laird, N. M. & Ware, J. H. (1982). Random-effects models for longitudinal data. Biometrics, 38, 963-974.

Lee, E. T. (1980). Statistical methods for survival data analysis. Belmont, CA: Lifetime Learning Publications.

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

Lindley, D. V. & Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). Journal of the Royal Statistical Society Series B, 34, 1-41.

Lipsitz, S. R. (1991). Practical uses of GEEs for repeated categorical responses. Boston: Harvard University Department of Biostatistics and Dana Farber Cancer Institute.

McCullagh, P. (1983). Quasi-likelihood functions. The Annals of Statistics, 11, 59-67.

McCullagh, P. & Nelder, J. A. (1983). Generalized linear models. London: Chapman and Hall.

Miller, R. G. (1974). The jackknife -- a review. Biometrika, 61, 1-15.

Moulton, L. H. (1986). Bootstrapping generalized linear models with application to longitudinal data. Dissertation Abstracts International, 47, 4745B. (University Microfilms No. 87-07,284)

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. Journal of the Royal Statistical Society Series A, 135, 370-384.

Nelder, J. A. & Pregibon, D. (1987). An extended quasi-likelihood function. Biometrika, 74, 221-232.

Olmsted, J. M. H. (1961). Advanced calculus. New York: Appleton-Century-Crofts, Inc.

Potthoff, R. F. & Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika, 51, 313-326.

Pregibon, D. (1981). Logistic regression diagnostics. The Annals of Statistics, 9, 705-724.

Qaqish, B. F. (1990). Multivariate regression models using generalized estimating equations. Dissertation Abstracts International.

Qu, Y., Williams, G. W., Beck, G. J., & Goormastic, M. (1987). A generalized model of logistic regression for clustered data. Communications in Statistics, Theory and Methods, 16, 3447-3476.

Rao, C. R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. Biometrika, 52, 447-458.

Rao, C. R. (1975). Simultaneous estimation of parameters in different linear models and applications to biometric problems. Biometrics, 31, 545-554.

Rosner, B., Hennekens, C. H., Kass, E. H., & Miall, W. E. (1977). Age-specific correlation analysis of longitudinal blood pressure data. American Jounal of Epidemiology, 106, 306-313.

Rubinstein, R. Y. (1981). Simulation and the Monte Carlo method. New York: John Wiley & Sons.

Scarborough, J. B. (1966). Numerical mathematical analysis (6th ed.). Baltimore: The Johns Hopkins Press.

Seber, G. A. F. (1984). Multivariate observations. New York: John Wiley & Sons.

Seber, G. A. F. & Wild, C. J. (1989). Nonlinear regression. New York: John Wiley & Sons.

Stiratelli, R., Laird, N., & Ware, J. H. (1984). Random-effects models for serial observations with binary response. Biometrics, 40, 961-971.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. Annals of Mathematical Statistics, 29, 614. (Abstract)

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. The American Statistician, 39, 95-101.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika, 61, 439-447.

Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. Biometrika, 63, 27-32.

Zeger, S. L. & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. Biometrics, 42, 121-130.

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

Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. Biometrics, 44, 1049-1060.

Zinner, S. H., Levy, P. S., & Kass, E. H. (1971). Familial aggregation of blood pressure in childhood. The New England Journal of Medicine, 284, 401-404.

Zinner, S. H., Martin, L. F., Sacks, F., Rosner, B., & Kass, E. H. (1975). A longitudinal study of blood pressure in children. American Journal of Epidemiology, 100, 437-442.

APPENDIX A

Lemmas

Lemma 1

Let $\underline{w}$ be a constant vector and D be a diagonal matrix such that both are conformable to multiplication. Let the elements of D, d_{jj}, each be functions of the vector $\underline{\beta}$. Then:

$$\frac{\partial}{\partial \underline{\beta}} [\underline{w}^T \cdot D] = \frac{\partial}{\partial \underline{\beta}} [\text{vec}(D)] \cdot \text{diag}(\underline{w}) .$$

Proof:

Let $\underline{w} = \begin{bmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ \cdot \\ w_n \end{bmatrix}$ and $D = \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & d_{nn} \end{bmatrix}$.

Thus $$\underline{w}^T \cdot D = \left[w_1 d_{11}, w_2 d_{22}, \ldots, w_n d_{nn} \right] .$$

Hence $$\frac{\partial}{\partial \underline{\beta}} [\underline{w}^T \cdot D] = \frac{\partial}{\partial \underline{\beta}} \left[w_1 d_{11}, w_2 d_{22}, \ldots, w_n d_{nn} \right]$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0} \left[w_1 d_{11}, w_2 d_{22}, \ldots, w_n d_{nn} \right] \\ \frac{\partial}{\partial \beta_1} \left[w_1 d_{11}, w_2 d_{22}, \ldots, w_n d_{nn} \right] \\ \vdots \\ \frac{\partial}{\partial \beta_q} \left[w_1 d_{11}, w_2 d_{22}, \ldots, w_n d_{nn} \right] \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0}(w_1 d_{11}) & \frac{\partial}{\partial \beta_0}(w_2 d_{22}) & \cdots & \frac{\partial}{\partial \beta_0}(w_n d_{nn}) \\ \frac{\partial}{\partial \beta_1}(w_1 d_{11}) & \frac{\partial}{\partial \beta_1}(w_2 d_{22}) & \cdots & \frac{\partial}{\partial \beta_1}(w_n d_{nn}) \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial}{\partial \beta_q}(w_1 d_{11}) & \frac{\partial}{\partial \beta_q}(w_2 d_{22}) & \cdots & \frac{\partial}{\partial \beta_q}(w_n d_{nn}) \end{bmatrix}$$

$$= \begin{bmatrix} w_1 \cdot \frac{\partial}{\partial \beta_0}(d_{11}) & w_2 \cdot \frac{\partial}{\partial \beta_0}(d_{22}) & \cdots & w_n \cdot \frac{\partial}{\partial \beta_0}(d_{nn}) \\ w_1 \cdot \frac{\partial}{\partial \beta_1}(d_{11}) & w_2 \cdot \frac{\partial}{\partial \beta_1}(d_{22}) & \cdots & w_n \cdot \frac{\partial}{\partial \beta_1}(d_{nn}) \\ \vdots & \vdots & \cdots & \vdots \\ w_1 \cdot \frac{\partial}{\partial \beta_q}(d_{11}) & w_2 \cdot \frac{\partial}{\partial \beta_q}(d_{22}) & \cdots & w_n \cdot \frac{\partial}{\partial \beta_q}(d_{nn}) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0}(d_{11}) & \frac{\partial}{\partial \beta_0}(d_{22}) & \cdots & \frac{\partial}{\partial \beta_0}(d_{nn}) \\ \frac{\partial}{\partial \beta_1}(d_{11}) & \frac{\partial}{\partial \beta_1}(d_{22}) & \cdots & \frac{\partial}{\partial \beta_1}(d_{nn}) \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial}{\partial \beta_q}(d_{11}) & \frac{\partial}{\partial \beta_q}(d_{22}) & \cdots & \frac{\partial}{\partial \beta_q}(d_{nn}) \end{bmatrix} \cdot \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0}\left[d_{11}, d_{22}, \ldots, d_{nn} \right] \\ \frac{\partial}{\partial \beta_1}\left[d_{11}, d_{22}, \ldots, d_{nn} \right] \\ \vdots \\ \frac{\partial}{\partial \beta_q}\left[d_{11}, d_{22}, \ldots, d_{nn} \right] \end{bmatrix} \cdot \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}$$

$$= \frac{\partial}{\partial \underline{\beta}}\left[d_{11}, d_{22}, \ldots, d_{nn} \right] \cdot \mathrm{diag}(\underline{w})$$

$$= \frac{\partial}{\partial \underline{\beta}}\left[\mathrm{vec}(D) \right] \cdot \mathrm{diag}(\underline{w}) \ .$$

Lemma 2

Let $\underline{v}$ be a vector and D be a diagonal matrix such that both are conformable to multiplication. Let the elements of each, v_i and d_{jj}, respectively, be functions of the vector $\underline{\beta}$. Then:

$$\frac{\partial}{\partial \underline{\beta}} [\underline{v}^T \cdot D] = \frac{\partial}{\partial \underline{\beta}} [\underline{v}^T] \cdot D + \frac{\partial}{\partial \underline{\beta}} [\text{vec}(D)] \cdot \text{diag}(\underline{v}) .$$

Proof:

$$\text{Let} \quad \underline{v} = \begin{bmatrix} v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_n \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & d_{nn} \end{bmatrix} .$$

$$\text{Thus} \quad \underline{v}^T \cdot D = \begin{bmatrix} v_1, & v_2, & \ldots, & v_n \end{bmatrix} \cdot \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & d_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} v_1 d_{11}, & v_2 d_{22}, & \ldots, & v_n d_{nn} \end{bmatrix} .$$

$$\text{Hence} \quad \frac{\partial}{\partial \underline{\beta}} [\underline{v}^T \cdot D] = \frac{\partial}{\partial \underline{\beta}} \begin{bmatrix} v_1 d_{11}, & v_2 d_{22}, & \ldots, & v_n d_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0}\left[v_1 d_{11},\ v_2 d_{22},\ \ldots,\ v_n d_{nn} \right] \\ \frac{\partial}{\partial \beta_1}\left[v_1 d_{11},\ v_2 d_{22},\ \ldots,\ v_n d_{nn} \right] \\ \vdots \\ \frac{\partial}{\partial \beta_q}\left[v_1 d_{11},\ v_2 d_{22},\ \ldots,\ v_n d_{nn} \right] \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0}(v_1 d_{11}) & \frac{\partial}{\partial \beta_0}(v_2 d_{22}) & \cdots & \frac{\partial}{\partial \beta_0}(v_n d_{nn}) \\ \frac{\partial}{\partial \beta_1}(v_1 d_{11}) & \frac{\partial}{\partial \beta_1}(v_2 d_{22}) & \cdots & \frac{\partial}{\partial \beta_1}(v_n d_{nn}) \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial}{\partial \beta_q}(v_1 d_{11}) & \frac{\partial}{\partial \beta_q}(v_2 d_{22}) & \cdots & \frac{\partial}{\partial \beta_q}(v_n d_{nn}) \end{bmatrix}$$

$$= \begin{bmatrix} (v_1 \cdot \frac{\partial}{\partial \beta_0} d_{11} + d_{11} \cdot \frac{\partial}{\partial \beta_0} v_1) & (v_2 \cdot \frac{\partial}{\partial \beta_0} d_{22} + d_{22} \cdot \frac{\partial}{\partial \beta_0} v_2) & \cdots & (v_n \cdot \frac{\partial}{\partial \beta_0} d_{nn} + d_{nn} \cdot \frac{\partial}{\partial \beta_0} v_n) \\ (v_1 \cdot \frac{\partial}{\partial \beta_1} d_{11} + d_{11} \cdot \frac{\partial}{\partial \beta_1} v_1) & (v_2 \cdot \frac{\partial}{\partial \beta_1} d_{22} + d_{22} \cdot \frac{\partial}{\partial \beta_1} v_2) & \cdots & (v_n \cdot \frac{\partial}{\partial \beta_1} d_{nn} + d_{nn} \cdot \frac{\partial}{\partial \beta_1} v_n) \\ \vdots & \vdots & \cdots & \vdots \\ (v_1 \cdot \frac{\partial}{\partial \beta_q} d_{11} + d_{11} \cdot \frac{\partial}{\partial \beta_q} v_1) & (v_2 \cdot \frac{\partial}{\partial \beta_q} d_{22} + d_{22} \cdot \frac{\partial}{\partial \beta_q} v_2) & \cdots & (v_n \cdot \frac{\partial}{\partial \beta_q} d_{nn} + d_{nn} \cdot \frac{\partial}{\partial \beta_q} v_n) \end{bmatrix}$$

$$= \begin{bmatrix} d_{11}\cdot\frac{\partial}{\partial\beta_0}v_1 & d_{22}\cdot\frac{\partial}{\partial\beta_0}v_2 & \cdots & d_{nn}\cdot\frac{\partial}{\partial\beta_0}v_n \\ d_{11}\cdot\frac{\partial}{\partial\beta_1}v_1 & d_{22}\cdot\frac{\partial}{\partial\beta_1}v_2 & \cdots & d_{nn}\cdot\frac{\partial}{\partial\beta_1}v_n \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ d_{11}\cdot\frac{\partial}{\partial\beta_q}v_1 & d_{22}\cdot\frac{\partial}{\partial\beta_q}v_2 & \cdots & d_{nn}\cdot\frac{\partial}{\partial\beta_q}v_n \end{bmatrix} + \begin{bmatrix} v_1\cdot\frac{\partial}{\partial\beta_0}d_{11} & v_2\cdot\frac{\partial}{\partial\beta_0}d_{22} & \cdots & v_n\cdot\frac{\partial}{\partial\beta_0}d_{nn} \\ v_1\cdot\frac{\partial}{\partial\beta_1}d_{11} & v_2\cdot\frac{\partial}{\partial\beta_1}d_{22} & \cdots & v_n\cdot\frac{\partial}{\partial\beta_1}d_{nn} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ v_1\cdot\frac{\partial}{\partial\beta_q}d_{11} & v_2\cdot\frac{\partial}{\partial\beta_q}d_{22} & \cdots & v_n\cdot\frac{\partial}{\partial\beta_q}d_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial\beta_0}v_1 & \frac{\partial}{\partial\beta_0}v_2 & \cdots & \frac{\partial}{\partial\beta_0}v_n \\ \frac{\partial}{\partial\beta_1}v_1 & \frac{\partial}{\partial\beta_1}v_2 & \cdots & \frac{\partial}{\partial\beta_1}v_n \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \frac{\partial}{\partial\beta_q}v_1 & \frac{\partial}{\partial\beta_q}v_2 & \cdots & \frac{\partial}{\partial\beta_q}v_n \end{bmatrix} \cdot \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & d_{nn} \end{bmatrix}$$

$$+ \begin{bmatrix} \frac{\partial}{\partial\beta_0}d_{11} & \frac{\partial}{\partial\beta_0}d_{22} & \cdots & \frac{\partial}{\partial\beta_0}d_{nn} \\ \frac{\partial}{\partial\beta_1}d_{11} & \frac{\partial}{\partial\beta_1}d_{22} & \cdots & \frac{\partial}{\partial\beta_1}d_{nn} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \frac{\partial}{\partial\beta_q}d_{11} & \frac{\partial}{\partial\beta_q}d_{22} & \cdots & \frac{\partial}{\partial\beta_q}d_{nn} \end{bmatrix} \cdot \begin{bmatrix} v_1 & 0 & \cdots & 0 \\ 0 & v_2 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & v_n \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0}\left[\, v_1, v_2, \ldots, v_n \,\right] \\ \frac{\partial}{\partial \beta_1}\left[\, v_1, v_2, \ldots, v_n \,\right] \\ \vdots \\ \frac{\partial}{\partial \beta_q}\left[\, v_1, v_2, \ldots, v_n \,\right] \end{bmatrix} \cdot \begin{bmatrix} d_{11} & 0 & \ldots & 0 \\ 0 & d_{22} & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & d_{nn} \end{bmatrix}$$

$$+ \begin{bmatrix} \frac{\partial}{\partial \beta_0}\left[\, d_{11}, d_{22}, \ldots, d_{nn} \,\right] \\ \frac{\partial}{\partial \beta_1}\left[\, d_{11}, d_{22}, \ldots, d_{nn} \,\right] \\ \vdots \\ \frac{\partial}{\partial \beta_q}\left[\, d_{11}, d_{22}, \ldots, d_{nn} \,\right] \end{bmatrix} \cdot \begin{bmatrix} v_1 & 0 & \ldots & 0 \\ 0 & v_2 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & v_n \end{bmatrix}$$

$$= \frac{\partial}{\partial \underline{\beta}}\left[\, v_1, v_2, \ldots, v_n \,\right] \cdot D + \frac{\partial}{\partial \underline{\beta}}\left[\, d_{11}, d_{22}, \ldots, d_{nn} \,\right] \cdot \mathrm{diag}(\underline{v})$$

$$= \frac{\partial}{\partial \underline{\beta}}\left[\underline{v}^T\right] \cdot D + \frac{\partial}{\partial \underline{\beta}}\left[\mathrm{vec}(D)\right] \cdot \mathrm{diag}(\underline{v}) \; .$$

Lemma 3

Let $\underline{v}$ and $\underline{w}$ be vectors such that both are conformable to multiplication. Let the elements of each, v_i and w_j, respectively, be functions of the vector $\underline{\beta}$. Then:

$$\frac{\partial}{\partial \underline{\beta}} [\underline{v}^T \cdot \underline{w}] = \frac{\partial}{\partial \underline{\beta}} [\underline{v}^T] \cdot \underline{w} + \frac{\partial}{\partial \underline{\beta}} [\underline{w}^T] \cdot \underline{v} .$$

Proof:

Let $\underline{v} = \begin{bmatrix} v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_n \end{bmatrix}$ and $\underline{w} = \begin{bmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ \cdot \\ w_n \end{bmatrix}$.

Thus $\underline{v}^T \cdot \underline{w} = (v_1w_1 + v_2w_2 + \ldots + v_nw_n)$.

Hence $$\frac{\partial}{\partial \underline{\beta}} [\underline{v}^T \cdot \underline{w}] = \frac{\partial}{\partial \underline{\beta}} (v_1w_1 + v_2w_2 + \ldots + v_nw_n)$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0} (v_1w_1 + v_2w_2 + \ldots + v_nw_n) \\ \frac{\partial}{\partial \beta_1} (v_1w_1 + v_2w_2 + \ldots + v_nw_n) \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial}{\partial \beta_q} (v_1w_1 + v_2w_2 + \ldots + v_nw_n) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0}(v_1 w_1) + \frac{\partial}{\partial \beta_0}(v_2 w_2) + \ldots + \frac{\partial}{\partial \beta_0}(v_n w_n) \\ \frac{\partial}{\partial \beta_1}(v_1 w_1) + \frac{\partial}{\partial \beta_1}(v_2 w_2) + \ldots + \frac{\partial}{\partial \beta_1}(v_n w_n) \\ \vdots \\ \frac{\partial}{\partial \beta_q}(v_1 w_1) + \frac{\partial}{\partial \beta_q}(v_2 w_2) + \ldots + \frac{\partial}{\partial \beta_q}(v_n w_n) \end{bmatrix}$$

$$= \begin{bmatrix} w_1 \cdot \frac{\partial}{\partial \beta_0} v_1 + w_2 \cdot \frac{\partial}{\partial \beta_0} v_2 + \ldots + w_n \cdot \frac{\partial}{\partial \beta_0} v_n \\ w_1 \cdot \frac{\partial}{\partial \beta_1} v_1 + w_2 \cdot \frac{\partial}{\partial \beta_1} v_2 + \ldots + w_n \cdot \frac{\partial}{\partial \beta_1} v_n \\ \vdots \\ w_1 \cdot \frac{\partial}{\partial \beta_q} v_1 + w_2 \cdot \frac{\partial}{\partial \beta_q} v_2 + \ldots + w_n \cdot \frac{\partial}{\partial \beta_q} v_n \end{bmatrix}$$

$$+ \begin{bmatrix} v_1 \cdot \frac{\partial}{\partial \beta_0} w_1 + v_2 \cdot \frac{\partial}{\partial \beta_0} w_2 + \ldots + v_n \cdot \frac{\partial}{\partial \beta_0} w_n \\ v_1 \cdot \frac{\partial}{\partial \beta_1} w_1 + v_2 \cdot \frac{\partial}{\partial \beta_1} w_2 + \ldots + v_n \cdot \frac{\partial}{\partial \beta_1} w_n \\ \vdots \\ v_1 \cdot \frac{\partial}{\partial \beta_q} w_1 + v_2 \cdot \frac{\partial}{\partial \beta_q} w_2 + \ldots + v_n \cdot \frac{\partial}{\partial \beta_q} w_n \end{bmatrix}$$

$$
= \begin{bmatrix}
\frac{\partial v_1}{\partial \beta_0} & \frac{\partial v_2}{\partial \beta_0} & \cdots & \frac{\partial v_n}{\partial \beta_0} \\
\frac{\partial v_1}{\partial \beta_1} & \frac{\partial v_2}{\partial \beta_1} & \cdots & \frac{\partial v_n}{\partial \beta_1} \\
\cdot & \cdot & \cdots & \cdot \\
\cdot & \cdot & \cdots & \cdot \\
\cdot & \cdot & \cdots & \cdot \\
\frac{\partial v_1}{\partial \beta_q} & \frac{\partial v_2}{\partial \beta_q} & \cdots & \frac{\partial v_n}{\partial \beta_q}
\end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ \cdot \\ w_n \end{bmatrix}
$$

$$
+ \begin{bmatrix}
\frac{\partial w_1}{\partial \beta_0} & \frac{\partial w_2}{\partial \beta_0} & \cdots & \frac{\partial w_n}{\partial \beta_0} \\
\frac{\partial w_1}{\partial \beta_1} & \frac{\partial w_2}{\partial \beta_1} & \cdots & \frac{\partial w_n}{\partial \beta_1} \\
\cdot & \cdot & \cdots & \cdot \\
\cdot & \cdot & \cdots & \cdot \\
\cdot & \cdot & \cdots & \cdot \\
\frac{\partial w_1}{\partial \beta_q} & \frac{\partial w_2}{\partial \beta_q} & \cdots & \frac{\partial w_n}{\partial \beta_q}
\end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_n \end{bmatrix}
$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0}\left[v_1, v_2, \ldots, v_n \right] \\ \frac{\partial}{\partial \beta_1}\left[v_1, v_2, \ldots, v_n \right] \\ \vdots \\ \frac{\partial}{\partial \beta_q}\left[v_1, v_2, \ldots, v_n \right] \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

$$+ \begin{bmatrix} \frac{\partial}{\partial \beta_0}\left[w_1, w_2, \ldots, w_n \right] \\ \frac{\partial}{\partial \beta_1}\left[w_1, w_2, \ldots, w_n \right] \\ \vdots \\ \frac{\partial}{\partial \beta_q}\left[w_1, w_2, \ldots, w_n \right] \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

$$= \frac{\partial}{\partial \underline{\beta}}\left[v_1, v_2, \ldots, v_n \right] \cdot \underline{w}$$

$$+ \frac{\partial}{\partial \underline{\beta}}\left[w_1, w_2, \ldots, w_n \right] \cdot \underline{v}$$

$$= \frac{\partial}{\partial \underline{\beta}}\left[\underline{v}^T\right] \cdot \underline{w} + \frac{\partial}{\partial \underline{\beta}}\left[\underline{w}^T\right] \cdot \underline{v} .$$

Lemma 4

Let $\underline{v}$ be a vector and C be a constant matrix such that both are conformable to multiplication. Let the elements of $\underline{v}$, v_i, each be functions of the vector $\underline{\beta}$. Then:

$$\frac{\partial}{\partial \underline{\beta}} [\underline{v}^T \cdot C] = \frac{\partial}{\partial \underline{\beta}} [\underline{v}^T] \cdot C .$$

Proof:

$$\text{Let} \quad \underline{v} = \begin{bmatrix} v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_n \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ c_{n1} & c_{n2} & \cdots & c_{nm} \end{bmatrix} .$$

$$\text{Thus} \quad \underline{v}^T \cdot C = \begin{bmatrix} v_1, & v_2, & \ldots, & v_n \end{bmatrix} \cdot \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ c_{n1} & c_{n2} & \cdots & c_{nm} \end{bmatrix}$$

$$= \left[\left(\sum_{j=1}^{n} v_j c_{j1}\right), \left(\sum_{j=1}^{n} v_j c_{j2}\right), \ldots, \left(\sum_{j=1}^{n} v_j c_{jm}\right) \right] .$$

$$\text{Hence} \quad \frac{\partial}{\partial \underline{\beta}} [\underline{v}^T \cdot C] = \frac{\partial}{\partial \underline{\beta}} \left[\left(\sum_{j=1}^{n} v_j c_{j1}\right), \left(\sum_{j=1}^{n} v_j c_{j2}\right), \ldots, \left(\sum_{j=1}^{n} v_j c_{jm}\right) \right]$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0} \left[\left(\sum_{j=1}^{n} v_j c_{j1}\right), \left(\sum_{j=1}^{n} v_j c_{j2}\right), \ldots, \left(\sum_{j=1}^{n} v_j c_{jm}\right) \right] \\ \frac{\partial}{\partial \beta_1} \left[\left(\sum_{j=1}^{n} v_j c_{j1}\right), \left(\sum_{j=1}^{n} v_j c_{j2}\right), \ldots, \left(\sum_{j=1}^{n} v_j c_{jm}\right) \right] \\ \vdots \\ \frac{\partial}{\partial \beta_q} \left[\left(\sum_{j=1}^{n} v_j c_{j1}\right), \left(\sum_{j=1}^{n} v_j c_{j2}\right), \ldots, \left(\sum_{j=1}^{n} v_j c_{jm}\right) \right] \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0} \left(\sum_{j=1}^{n} v_j c_{j1}\right) & \frac{\partial}{\partial \beta_0} \left(\sum_{j=1}^{n} v_j c_{j2}\right) & \cdots & \frac{\partial}{\partial \beta_0} \left(\sum_{j=1}^{n} v_j c_{jm}\right) \\ \frac{\partial}{\partial \beta_1} \left(\sum_{j=1}^{n} v_j c_{j1}\right) & \frac{\partial}{\partial \beta_1} \left(\sum_{j=1}^{n} v_j c_{j2}\right) & \cdots & \frac{\partial}{\partial \beta_1} \left(\sum_{j=1}^{n} v_j c_{jm}\right) \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial}{\partial \beta_q} \left(\sum_{j=1}^{n} v_j c_{j1}\right) & \frac{\partial}{\partial \beta_q} \left(\sum_{j=1}^{n} v_j c_{j2}\right) & \cdots & \frac{\partial}{\partial \beta_q} \left(\sum_{j=1}^{n} v_j c_{jm}\right) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{j=1}^{n} \left[\frac{\partial}{\partial \beta_0}(v_j c_{j1})\right] & \sum_{j=1}^{n} \left[\frac{\partial}{\partial \beta_0}(v_j c_{j2})\right] & \cdots & \sum_{j=1}^{n} \left[\frac{\partial}{\partial \beta_0}(v_j c_{jm})\right] \\ \sum_{j=1}^{n} \left[\frac{\partial}{\partial \beta_1}(v_j c_{j1})\right] & \sum_{j=1}^{n} \left[\frac{\partial}{\partial \beta_1}(v_j c_{j2})\right] & \cdots & \sum_{j=1}^{n} \left[\frac{\partial}{\partial \beta_1}(v_j c_{jm})\right] \\ \vdots & \vdots & \cdots & \vdots \\ \sum_{j=1}^{n} \left[\frac{\partial}{\partial \beta_q}(v_j c_{j1})\right] & \sum_{j=1}^{n} \left[\frac{\partial}{\partial \beta_q}(v_j c_{j2})\right] & \cdots & \sum_{j=1}^{n} \left[\frac{\partial}{\partial \beta_q}(v_j c_{jm})\right] \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{j=1}^{n} \left[\frac{\partial}{\partial \beta_0} v_j\right] \cdot c_{j1} & \sum_{j=1}^{n} \left[\frac{\partial}{\partial \beta_0} v_j\right] \cdot c_{j2} & \cdots & \sum_{j=1}^{n} \left[\frac{\partial}{\partial \beta_0} v_j\right] \cdot c_{jm} \\ \sum_{j=1}^{n} \left[\frac{\partial}{\partial \beta_1} v_j\right] \cdot c_{j1} & \sum_{j=1}^{n} \left[\frac{\partial}{\partial \beta_1} v_j\right] \cdot c_{j2} & \cdots & \sum_{j=1}^{n} \left[\frac{\partial}{\partial \beta_1} v_j\right] \cdot c_{jm} \\ \vdots & \vdots & \cdots & \vdots \\ \sum_{j=1}^{n} \left[\frac{\partial}{\partial \beta_q} v_j\right] \cdot c_{j1} & \sum_{j=1}^{n} \left[\frac{\partial}{\partial \beta_q} v_j\right] \cdot c_{j2} & \cdots & \sum_{j=1}^{n} \left[\frac{\partial}{\partial \beta_q} v_j\right] \cdot c_{jm} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0} \left[v_1, v_2, \ldots, v_n \right] \\ \frac{\partial}{\partial \beta_1} \left[v_1, v_2, \ldots, v_n \right] \\ \vdots \\ \frac{\partial}{\partial \beta_q} \left[v_1, v_2, \ldots, v_n \right] \end{bmatrix} \cdot \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nm} \end{bmatrix}$$

$$\frac{\partial}{\partial \underline{\beta}} \left[v_1, v_2, \ldots, v_n \right] \cdot C$$

$$= \frac{\partial}{\partial \underline{\beta}} \left[\underline{v}^T\right] \cdot C \, .$$

Lemma 5

Let $\underline{v}$ be a vector such that its elements $v_1, v_2, \ldots, v_n$ are functions of $\theta_1, \theta_2, \ldots, \theta_n$, respectively. Also, let X be a constant matrix and $\underline{\beta}$ be a vector such that both are conformable to multiplication. Furthermore, let $\underline{\theta} = X\underline{\beta}$. Then:

$$\frac{\partial}{\partial \underline{\beta}} \underline{v}^T = X^T \cdot \mathrm{diag}([v_1'(\theta_1), v_2'(\theta_2), \ldots, v_n'(\theta_n)]) .$$

Proof:

$$\text{Let } \underline{v} = \begin{bmatrix} v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_n \end{bmatrix}, \quad \underline{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \cdot \\ \cdot \\ \cdot \\ \theta_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1q} \\ 1 & x_{21} & \cdots & x_{2q} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 1 & x_{n1} & \cdots & x_{nq} \end{bmatrix} \text{ and } \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_q \end{bmatrix}$$

where $\underline{\theta} = X\underline{\beta}$.

Hence
$$\frac{\partial}{\partial \underline{\beta}} \underline{v}^T = \frac{\partial}{\partial \underline{\beta}} \left[v_1, v_2, \ldots, v_n \right]$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0} \left[v_1, v_2, \ldots, v_n \right] \\ \frac{\partial}{\partial \beta_1} \left[v_1, v_2, \ldots, v_n \right] \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial}{\partial \beta_q} \left[v_1, v_2, \ldots, v_n \right] \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0} v_1 & \frac{\partial}{\partial \beta_0} v_2 & \cdots & \frac{\partial}{\partial \beta_0} v_n \\ \frac{\partial}{\partial \beta_1} v_1 & \frac{\partial}{\partial \beta_1} v_2 & \cdots & \frac{\partial}{\partial \beta_1} v_n \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \frac{\partial}{\partial \beta_q} v_1 & \frac{\partial}{\partial \beta_q} v_2 & \cdots & \frac{\partial}{\partial \beta_q} v_n \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_0} \theta_1 \cdot \frac{\partial}{\partial \theta_1} v_1 & \frac{\partial}{\partial \beta_0} \theta_2 \cdot \frac{\partial}{\partial \theta_2} v_2 & \cdots & \frac{\partial}{\partial \beta_0} \theta_n \cdot \frac{\partial}{\partial \theta_n} v_n \\ \frac{\partial}{\partial \beta_1} \theta_1 \cdot \frac{\partial}{\partial \theta_1} v_1 & \frac{\partial}{\partial \beta_1} \theta_2 \cdot \frac{\partial}{\partial \theta_2} v_2 & \cdots & \frac{\partial}{\partial \beta_1} \theta_n \cdot \frac{\partial}{\partial \theta_n} v_n \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \frac{\partial}{\partial \beta_q} \theta_1 \cdot \frac{\partial}{\partial \theta_1} v_1 & \frac{\partial}{\partial \beta_q} \theta_2 \cdot \frac{\partial}{\partial \theta_2} v_2 & \cdots & \frac{\partial}{\partial \beta_q} \theta_n \cdot \frac{\partial}{\partial \theta_n} v_n \end{bmatrix}$$

$$= \begin{bmatrix} 1 \cdot \frac{\partial}{\partial \theta_1} v_1 & 1 \cdot \frac{\partial}{\partial \theta_2} v_2 & \cdots & 1 \cdot \frac{\partial}{\partial \theta_n} v_n \\ x_{11} \cdot \frac{\partial}{\partial \theta_1} v_1 & x_{21} \cdot \frac{\partial}{\partial \theta_2} v_2 & \cdots & x_{n1} \cdot \frac{\partial}{\partial \theta_n} v_n \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ x_{1q} \cdot \frac{\partial}{\partial \theta_1} v_1 & x_{2q} \cdot \frac{\partial}{\partial \theta_2} v_2 & \cdots & x_{nq} \cdot \frac{\partial}{\partial \theta_n} v_n \end{bmatrix}$$

$$
= \begin{bmatrix}
1 \cdot v_1'(\theta_1) & 1 \cdot v_2'(\theta_2) & \cdots & 1 \cdot v_n'(\theta_n) \\
x_{11} \cdot v_1'(\theta_1) & x_{21} \cdot v_2'(\theta_2) & \cdots & x_{n1} \cdot v_n'(\theta_n) \\
\vdots & \vdots & \cdots & \vdots \\
x_{1q} \cdot v_1'(\theta_1) & x_{2q} \cdot v_2'(\theta_2) & \cdots & x_{nq} \cdot v_n'(\theta_n)
\end{bmatrix}
$$

$$
= \begin{bmatrix}
1 & 1 & \cdots & 1 \\
x_{11} & x_{21} & \cdots & x_{n1} \\
\vdots & \vdots & \cdots & \vdots \\
x_{1q} & x_{2q} & \cdots & x_{nq}
\end{bmatrix}
\cdot
\begin{bmatrix}
v_1'(\theta_1) & 0 & \cdots & 0 \\
0 & v_2'(\theta_2) & \cdots & 0 \\
\vdots & \vdots & \cdots & 0 \\
0 & 0 & \cdots & v_n'(\theta_n)
\end{bmatrix}
$$

$$
= X^T \cdot \mathrm{diag}([v_1'(\theta_1), v_2'(\theta_2), \ldots, v_n'(\theta_n)]) \; .
$$

Lemma 6

Let $\underline{w}$ be a constant vector and P be a matrix such that both are conformable to multiplication. Let the elements of P, p_{ij}, each be functions of the scalar ρ. Then:

$$\frac{\partial}{\partial \rho}\,[\underline{w}^T \cdot P] \;=\; \underline{w}^T \cdot \frac{\partial}{\partial \rho}\,[P]\ .$$

Proof:

$$\text{Let}\qquad \underline{w} = \begin{bmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ \cdot \\ w_n \end{bmatrix} \qquad \text{and} \qquad P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ p_{n1} & p_{n2} & \cdots & p_{nm} \end{bmatrix}.$$

$$\text{Thus}\qquad \underline{w}^T \cdot P = \begin{bmatrix} w_1, & w_2, & \ldots, & w_n \end{bmatrix} \cdot \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ p_{n1} & p_{n2} & \cdots & p_{nm} \end{bmatrix}$$

$$= \left[\left(\sum_{j=1}^{n} w_j p_{j1}\right), \left(\sum_{j=1}^{n} w_j p_{j2}\right), \ldots, \left(\sum_{j=1}^{n} w_j p_{jm}\right) \right].$$

$$\text{Hence}\qquad \frac{\partial}{\partial \rho}\,[\underline{w}^T \cdot P] = \frac{\partial}{\partial \rho} \left[\left(\sum_{j=1}^{n} w_j p_{j1}\right), \left(\sum_{j=1}^{n} w_j p_{j2}\right), \ldots, \left(\sum_{j=1}^{n} w_j p_{jm}\right) \right]$$

$$= \left[\frac{\partial}{\partial \rho} \left(\sum_{j=1}^{n} w_j p_{j1} \right) \quad \frac{\partial}{\partial \rho} \left(\sum_{j=1}^{n} w_j p_{j2} \right) \quad \cdots \quad \frac{\partial}{\partial \rho} \left(\sum_{j=1}^{n} w_j p_{jm} \right) \right]$$

$$= \left[\sum_{j=1}^{n} \left[\frac{\partial}{\partial \rho} (w_j p_{j1}) \right] \quad \sum_{j=1}^{n} \left[\frac{\partial}{\partial \rho} (w_j p_{j2}) \right] \quad \cdots \quad \sum_{j=1}^{n} \left[\frac{\partial}{\partial \rho} (w_j p_{jm}) \right] \right]$$

$$= \left[\sum_{j=1}^{n} \left[w_j \cdot \frac{\partial}{\partial \rho} p_{j1} \right] \quad \sum_{j=1}^{n} \left[w_j \cdot \frac{\partial}{\partial \rho} p_{j2} \right] \quad \cdots \quad \sum_{j=1}^{n} \left[w_j \cdot \frac{\partial}{\partial \rho} p_{jm} \right] \right]$$

$$= \left[w_1, w_2, \ldots, w_n \right] \cdot \begin{bmatrix} \frac{\partial}{\partial \rho} p_{11} & \frac{\partial}{\partial \rho} p_{12} & \cdots & \frac{\partial}{\partial \rho} p_{1m} \\ \frac{\partial}{\partial \rho} p_{21} & \frac{\partial}{\partial \rho} p_{22} & \cdots & \frac{\partial}{\partial \rho} p_{2m} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \frac{\partial}{\partial \rho} p_{n1} & \frac{\partial}{\partial \rho} p_{n2} & \cdots & \frac{\partial}{\partial \rho} p_{nm} \end{bmatrix}$$

$$= \underline{w}^T \cdot \frac{\partial}{\partial \rho} [P] \, .$$

APPENDIX B

Program Implementation

The MQL GEE framework defined by Equations 4.9 and 4.10 is implemented as a Fortran-77 computer program compilable using the Microsoft Fortran Version 5.1 software. The program follows the American National Standards Institute (ANSI) standard very closely, and as such should be downward compatible (with minor modification) to most commercially available Fortran-IV compilers.

In addition to the usual implicit function calls, three external subroutines are required during the object linking process. The first two, DECOMP and SOLVE, are used for matrix inversion. The description and source code for these are available in Forsythe, Malcolm, and Moler (1977). The last, PROBCHI2, computes approximate probability points for the chi square distribution. This subroutine is based on source code available in Lee (1980).

User-required modification is necessary for: (1) choosing a link function, (2) defining the number of beta parameters, (3) equating physical files to logical units, and (4) describing the input data format. These (and related) issues are detailed within the program listing under the heading "Special Programming Modification."

The program source code, which follows, is available on diskette. It is offered without charge for academic and noncommercial purposes. As such, no warranty is expressed nor implied.

```
C pgm: GEE_MQLE.FOR   --   Maximum Quasi-Likelihood Estimation
C                          of BETA and RHO (exchangable & AR-1)
C MQL Generalized Estimating Equations for correlated responses
C
C ================================================================
C General Program Information
C
C   Implementation : Microsoft Fortran-77 Version 5.1 for IBM-PCs
C                    Double-Precision Floating-Point
C   Capability:
C     GLM link choices: 1=Logit, 2= Identity, 3=Log
C     Correlation structures:  Independence, AR-1, and Exchangable --
C                                 each automatically invoked
C   External Subroutines:
C     1. DECOMP   -- decomposes a square matrix by Gaussian
C                    elimination, and reports its condition
C                    SOURCE: Forsythe, Malcolm, & Moler (1977)
C     2. SOLVE    -- solves a linear system using a DECOMPed matrix
C                    (used for matrix inversion in this program)
C                    SOURCE: Forsythe, Malcolm, & Moler (1977)
C     3. PROBCHI2 -- probability points of the chi-square distribution
C                    SOURCE: modified from Lee (1980)
C
C   Comments:
C     This program processes data on a subject basis, and as such is
C     quite compact.  This allows very large data sets to be analyzed,
C     but sacrifices speed.  If your computer allows software disk-
C     caching, physical I/O operations should decrease and thus
C     improve overall throughput.  Use of a math-coprocessor (on PCs)
C     is HIGHLY recommended for similar reasons.
C
C   Caveats:
C     1. Needs rough estimate of RHO (wild guesses seem to work OK)
C     2. Requires complete data per observation, but "should"
C        accommodate entire missing observations (although this has
C        not been thoroughly wrung-out).
C
C ================================================================
C Global Environment Settings
C
        IMPLICIT DOUBLE PRECISION (A-H,O-Z)
        IMPLICIT INTEGER (I-N)
C
        PARAMETER  (KMAX = 10)
        PARAMETER  (LMAX = 10)
        PARAMETER  (MMAX = 11)
C
C KMAX is the max number of observations per subject
C LMAX is the max order of the beta parameter vector in the linear
C      predictor ( i.e. # of betas )
C MMAX is the max size of the SCORE (gradient), HESSIAN, and DELTA;
C      typically: atleast 1 bigger than LMAX to accommodate RHO param
C
C ================================================================
```

```
C Input-Output Specifications by Logical Unit
C
C INPUT
C  04: control records defining the number of observation records
C        in unit-05 for each subject (see stmt #0110 for details)
C  05: observation records (see next section on required mods)
C
C OUTPUT
C  06: report file -- summary stats and parameter info
C  07: detail file -- lengthy details during iterative process
C
C INPUT/OUTPUT
C  *: keyboard -- used to enter RHO initial estimate
C
C ====================================================================
C Special Programming Modification
C
C Within the program, the user MUST:
C   1.  modify LINKTYPE (stmt #0001) for the link desired
C   2.  modify LDIM (stmt #0002) to equal the exact order of beta
C   3.  modify OPEN stmts (#0011 - #0014) to equate logical units
C         with physical file locations( or devices)
C   4.  modify READ and FORMAT stmts (#0120 - #0121) to match the
C         incoming data from logical unit-05
C
C Optionally, the following may be changed:
C   5.  modify MAXITER (stmt #0003) for the maximum iteration count
C   6.  modify SIGNIF (stmt #0004) which is the score (gradient)
C         equations' "zero" level
C
C --------------------------------------------------------------------
C
C ********************************************************************
C     B  E  G  I  N:   D A T A   D E F I N I T I O N S
C ********************************************************************
C
C --------------------------------------------------------------------
C -- Work-space for the subroutines DECOMP & SOLVE
C
      DIMENSION IPVT(50)
      DIMENSION WORK(50)
C
C --------------------------------------------------------------------
C -- Matrix and Vector Definitions
C
      DIMENSION X(KMAX,LMAX),Y(KMAX),R(KMAX,KMAX),RINV(KMAX,KMAX),
     + RINVDER1(KMAX,KMAX),RINVDER2(KMAX,KMAX),
     + E(KMAX),S(KMAX),V(KMAX),AHALF(KMAX),T(KMAX),U(KMAX),
     + SCORE(MMAX),HESSIAN(MMAX,MMAX),HESSCOPY(MMAX,MMAX),DELTA(MMAX),
     + COVAR(MMAX,MMAX),CORREL(MMAX,MMAX),
     + HESSCROS(LMAX),SUBJSCOR(LMAX),IBETASUB(LMAX),
     + SCORBETA(LMAX),HESSBETA(LMAX,LMAX),BETA(LMAX),
     + TMP1(LMAX,KMAX),TMP2(LMAX,KMAX),TMP3(LMAX,KMAX),
     + TMP5(KMAX),TMP6(KMAX,KMAX),TMP7(LMAX,KMAX),TMP8(LMAX)
```

```
C
C ---------------------------------------------------------------
C
C -- Boolean Variable Definitions
C
C  SUCCESS -- convergence indicator
      LOGICAL SUCCESS
C  LNKLOGIT, LNKIDENT, or LNKLOGAR -- chosen link function
      LOGICAL LNKLOGIT,LNKIDENT,LNKLOGAR
C  INDEPEND, EXCHANGE, or AUTOREG1 -- current correlation structure
      LOGICAL INDEPEND,EXCHANGE,AUTOREG1
C
C ---------------------------------------------------------------
C --  Labels used for printed reports
C
      CHARACTER*20 LBLLINK(3)
      CHARACTER*16 LBLCORR(3)
C
         DATA LBLLINK(1)/'LOGIT (binary)      '/
         DATA LBLLINK(2)/'IDENTITY (normal)   '/
         DATA LBLLINK(3)/'LOGARITHM (Poisson) '/
C
         DATA LBLCORR(1)/'Independent     '/
         DATA LBLCORR(2)/'Autoregressive  '/
         DATA LBLCORR(3)/'Exchangable     '/
C
C ***************************************************************
C        B  E  G  I  N:   P R O C E D U R A L   S E C T I O N
C ***************************************************************
C ===============================================================
C -- choose the LINK for the desired type of responses
0001   LINKTYPE = 1
C -- define the order of the beta parameter vector (1 + # covariates)
0002   LDIM = 7
C -- declare the maximum number of iterations to attempt
0003   MAXITER = 39
C -- declare the convergence criteria "zero" size
0004   SIGNIF = 1.0D-4
C ===============================================================
      LNKLOGIT = .FALSE.
      LNKIDENT = .FALSE.
      LNKLOGAR = .FALSE.
      IF ( LINKTYPE .EQ. 1 ) LNKLOGIT = .TRUE.
      IF ( LINKTYPE .EQ. 2 ) LNKIDENT = .TRUE.
      IF ( LINKTYPE .EQ. 3 ) LNKLOGAR = .TRUE.
C
      ZERO = 0.0D0
      ONE  = 1.0D0
      TWO  = 2.0D0
      THREE= 3.0D0
      FOUR = 4.0D0
C
C -- equate data files to their logical units
C
```

```
0011    OPEN(4,FILE='GEE04.DTA',STATUS='OLD')
0012    OPEN(5,FILE='GEE05.DTA',STATUS='OLD')
0013    OPEN(6,FILE='REPORT06.PRT',STATUS='NEW')
0014    OPEN(7,FILE='DETAIL07.PRT',STATUS='NEW')
C ---------------------------------------------------------------------
C
        WRITE(6,6000) LBLLINK(LINKTYPE)
        WRITE(7,6000) LBLLINK(LINKTYPE)
6000    FORMAT(1H1,/,'* * MQL GEEs  for ',A20,' -- pgm GEE_MQLE * *')
C
        WRITE(*,6001)
6001    FORMAT(1X,'Enter the RHO to be used initially for AR-1 & EXCH:')
        READ(*,*) RHOINPUT
        WRITE(6,6002) RHOINPUT
        WRITE(7,6002) RHOINPUT
6002    FORMAT(1X,/,'* * * Initial est of RHO =',F6.3,' * * *')
C
C -- Init loop counter for various correlation structures to use
C
        LOOPCORR = 0
C
C = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
C
C -- Start of current correlation structure to be used
C
C = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
0050    LOOPCORR = LOOPCORR + 1
        IF ( LOOPCORR .GT. 3 ) GO TO 9999
C -- get start time of this regression
        CALL GETTIM(IHOURBEG,IMINBEG,ISECBEG,I100BEG)
C -- reset Boolean switches for this correlation structure
        INDEPEND = .FALSE.
        AUTOREG1 = .FALSE.
        EXCHANGE = .FALSE.
        IF (LOOPCORR .EQ. 1) INDEPEND = .TRUE.
        IF (LOOPCORR .EQ. 2) AUTOREG1 = .TRUE.
        IF (LOOPCORR .EQ. 3) EXCHANGE = .TRUE.
C -- set the total number of parameters to estimate -- MDIM
        MDIM = LDIM
        IF( AUTOREG1 .OR. EXCHANGE ) MDIM = LDIM + 1
        KDIM = KMAX
C
        WRITE(*,6051) LOOPCORR,LBLCORR(LOOPCORR)
        WRITE(6,6051) LOOPCORR,LBLCORR(LOOPCORR)
        WRITE(7,6051) LOOPCORR,LBLCORR(LOOPCORR)
6051    FORMAT(1X,//,' LOOP-',I2,' -- Correlation structure = ',A16)
        IF ( .NOT. INDEPEND ) GO TO 0060
C
C -- Init the 'working' correlation & inverse for INDEPEND
C
       DO 0059 I=1,KDIM
          DO 0055 J=1,KDIM
            R(I,J) = ZERO
            RINV(I,J) = ZERO
```

```
0055        CONTINUE
            R(I,I) = ONE
            RINV(I,I) = ONE
0059     CONTINUE
C
C -- Init estimates of BETA, its subscript labels, RHO, & Φ
C
0060     CONTINUE
         DO 0069 I=1,LDIM
           BETA(I) = ZERO
           IBETASUB(I) = I - 1
0069     CONTINUE
         RHO = ZERO
         IF ( EXCHANGE .OR. AUTOREG1 ) RHO = RHOINPUT
         SCALE = ONE
C
C -- reset iteration counter to zero
C
         ITER = 0
         SUCCESS = .FALSE.
C - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
C ******* Begin iteration for current correlation structure ********
C - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
0100     CONTINUE
         ITER = ITER + 1
         IF ( ITER .GT. MAXITER ) GO TO 9998
         WRITE(7,6101) ITER,RHO,(BETA(J),J=1,LDIM)
6101     FORMAT(1X,/,' ITERATION #',I3,' - RHO estimate is:',D12.6,/,
     + '   BETA estimates are:',/,10(D12.6,1X))
C
C -- re-set the subject-info (04), subject-data (05) files, & EVERYTHING
         REWIND 04
         REWIND 05
         N = 0
C -- init Quadratic distance (Q² = D²/Φ) measure and outlier count
         DISTSQRD = ZERO
         IOUTLIER = 0
C
C -- init score and hessian accumulators for BETA, cross-terms, & RHO
         DO 0104 I=1,LDIM
           SCORBETA(I) = ZERO
           HESSCROS(I) = ZERO
           DO 0103 J=1,LDIM
             HESSBETA(I,J) = ZERO
0103       CONTINUE
0104     CONTINUE
         SCORERHO = ZERO
         HESSRHO = ZERO
C ---------------------------------------------------------------------
C - Begin subject processing
C ---------------------------------------------------------------------
0110     CONTINUE
         READ(4,4001,END=0200) ISUBJECT,IOBSCNT
4001     FORMAT(I6,2X,I2)
```

```
        KDIM = IOBSCNT
        ISUBJN = 0
C
        IF( INDEPEND )   GO TO 0120
C
C -- Set up R, inv(R), [inv(R)]', and [inv(R)]" for AR-1 and EXCH
C    because INDEPEND has already been done
C
C -- special code is required to do either AUTOREG1 or EXCHANGE
        IF ( EXCHANGE ) GO TO 0116
C ---------------------------------------------------------------
C -- AUTOREG1 correlation
        P = RHO
        DENOM = (P**2 - ONE)
        A = -ONE/DENOM
        B = -(ONE + P**2)/DENOM
        C = P/DENOM
        ADER1 = TWO*P/(DENOM**2)
        BDER1 = TWO*ADER1
        CDER1 = -(ONE + P**2)/(DENOM**2)
        ADER2 = -TWO*(THREE*(P**2) + ONE)/(DENOM**3)
        BDER2 = TWO*ADER2
        CDER2 = TWO*P*(P**2 + THREE)/(DENOM**3)
        DO 0114 I=1,KDIM
          DO 0113 J=1,KDIM
            IF ( IABS(I-J) .EQ. 1 ) GO TO 0111
            RINV(I,J) = ZERO
            RINVDER1(I,J) = ZERO
            RINVDER2(I,J) = ZERO
            GO TO 0113
0111        RINV(I,J) = C
            RINVDER1(I,J) = CDER1
            RINVDER2(I,J) = CDER2
0113      CONTINUE
          RINV(I,I) = B
          RINVDER1(I,I) = BDER1
          RINVDER2(I,I) = BDER2
0114    RINV(1,1) = A
        RINV(KDIM,KDIM) = A
        RINVDER1(1,1) = ADER1
        RINVDER1(KDIM,KDIM) = ADER1
        RINVDER2(1,1) = ADER2
        RINVDER2(KDIM,KDIM) = ADER2
        GO TO 0120
C ---------------------------------------------------------------
C -- EXCHANGE correlation
0116    Q = DFLOAT(KDIM)
        P = RHO
        DENOM = (P-ONE)*(P*(Q-ONE)+ONE)
        TOP1ON = -(P*(Q-TWO)+ONE)
        TOP1OFF= P
        TOP2ON = P*(Q-ONE)*(P*(Q-TWO)+TWO)
        TOP2OFF= -((P**2)*(Q-ONE)+ONE)
        TOP3ON = -TWO*(Q-ONE)*
```

```
     +     ( (P**3)*(Q-ONE)*(Q-TWO) + THREE*(P**2)*(Q-ONE)+ONE )
       TOP3OFF= TWO*
     +     ( (P**3)*((Q-ONE)**2) + THREE*P*(Q-ONE) - (Q-TWO) )
       C    = TOP1ON/DENOM
       D    = TOP1OFF/DENOM
       CDER1 = TOP2ON/(DENOM**2)
       DDER1 = TOP2OFF/(DENOM**2)
       CDER2 = TOP3ON/(DENOM**3)
       DDER2 = TOP3OFF/(DENOM**3)
       DO 0119 I=1,KDIM
         DO 0118 J=1,KDIM
           R(I,J) = RHO
           RINV(I,J) = D
           RINVDER1(I,J) =  DDER1
           RINVDER2(I,J) =  DDER2
0118     CONTINUE
         R(I,I) = ONE
         RINV(I,I) = C
         RINVDER1(I,I) =  CDER1
         RINVDER2(I,I) =  CDER2
0119   CONTINUE
CCCCCC GO TO 0120
C -------------------------------------------------------------
0120   CONTINUE
C
C -- Read in all observations for this subject (specified in Unit-04)
C
       DO 0121 I=1,KDIM
C        The following code MUST be user-modified!!!
C <*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>
C
         READ(5,5001) RESPONSE,(X(I,J),J=2,LDIM)
5001     FORMAT(F1.0,F5.2,F4.2,F4.2,F4.2,F6.3,F6.3,5X)
         Y(I) = RESPONSE
         X(I,1) = ONE
C
C           ( Additions and modifications to covariates)
C                    ( should go in here)
C
C <*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>
C
         ISUBJN = ISUBJN + 1
0121   CONTINUE
       N = N + ISUBJN
C
C -- Compute the linear predictor ETA for this subject
C -- Then equate it to the chosen canonical link
C
       DO 0126 I=1,KDIM
         ETA = ZERO
         DO 0122 J=1,LDIM
           ETA = ETA + X(I,J)*BETA(J)
0122     CONTINUE
C <*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>
```

```
C            ( Add any fixed "OFFSET" to ETA here )
C <*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>*<*>
C
          IF (LNKLOGIT) GO TO 0123
          IF (LNKIDENT) GO TO 0124
C ---- LOGARITHMIC LINK CODE (Poisson data) IS HERE ----
          REXP = DEXP(ETA)
          E(I) = REXP
          V(I) = REXP
          AHALF(I) = DSQRT( V(I) )
          S(I) = Y(I) - E(I)
          T(I) = ( AHALF(I) + Y(I)/AHALF(I) )/TWO
          U(I) = -(S(I)/AHALF(I))/FOUR
          GO TO 0125
C ---- LOGIT LINK CODE (binary data) IS HERE ----
0123      REXP = DEXP(ETA)
          RLOGIST = REXP/(ONE + REXP)
          E(I) = RLOGIST
          V(I) = RLOGIST*(ONE - RLOGIST)
          AHALF(I) = DSQRT( V(I) )
          S(I) = Y(I) - E(I)
          T(I) = ( (ONE-Y(I))*DSQRT(REXP) + Y(I)/DSQRT(REXP) )/TWO
          U(I) = -(S(I)/AHALF(I))/FOUR
          GO TO 0125
C ---- IDENTITY LINK CODE (Gaussian data) IS HERE ----
0124      E(I) = ETA
          V(I) = ONE
          AHALF(I) = ONE
          S(I) = Y(I) - E(I)
          T(I) = ONE
          U(I) = ZERO
          GO TO 0125
C ---- (next link to go here)
0125      CONTINUE
0126    CONTINUE
C
C -- Compute SCORE and HESSIAN for current subject
C
C -- Compute SCORBETA: [X' * (A^½+W) * inv(R) * inv(A^(½))] * S
C
C -- 1. compute TMP1 = X'*T  (where T = A^½ + W)
        DO 0139 I=1,LDIM
          DO 0135 J=1,KDIM
            TMP1(I,J) = X(J,I)*T(J)
0135      CONTINUE
0139    CONTINUE
C
C -- 2. compute TMP2 = TMP1*inv(R)
        DO 0149 I=1,LDIM
          DO 0147 J=1,KDIM
            SUM = ZERO
            DO 0145 K=1,KDIM
              SUM = SUM + TMP1(I,K)*RINV(K,J)
0145        CONTINUE
```

```
            TMP2(I,J) = SUM
0147       CONTINUE
0149     CONTINUE
C
C -- 3. compute TMP3 = TMP2 * inv(A^(½))
         DO 0159 I=1,LDIM
           DO 0155 J=1,KDIM
             TMP3(I,J) = TMP2(I,J)/AHALF(J)
0155       CONTINUE
0159     CONTINUE
C
C -- 4. compute SCORBETA = TMP3 * S
         DO 0169 I=1,LDIM
           SUM = ZERO
           DO 0165 J=1,KDIM
             SUM = SUM + TMP3(I,J)*S(J)
0165       CONTINUE
           SUBJSCOR(I) = SUM
           SCORBETA(I) = SCORBETA(I) + SUM
0169     CONTINUE
C
C -- Compute HESSBETA = -[ X' * ( BIG expression=TMP6 ) * X]
C
C -- 5. compute TMP5 (vector) = U * matdiag[S*A^(-½)*inv(R)]
         DO 0173 I=1,KDIM
           SUM = ZERO
           DO 0172 J=1,KDIM
             SUM = SUM + S(J)/AHALF(J)*RINV(J,I)
0172       CONTINUE
           TMP5(I) = SUM*U(I)
0173     CONTINUE
C -- & compute TMP6 = TMP5(matrix) - T*inv(R)*T
         DO 0175 I=1,KDIM
           DO 0174 J=1,KDIM
             UTERM = ZERO
             IF ( I .EQ. J ) UTERM = TMP5(I)
             TMP6(I,J) = UTERM - T(I)*RINV(I,J)*T(J)
0174       CONTINUE
0175     CONTINUE
C -- 6. compute HESSBETA = X'*TMP6*X
         DO 0184 I=1,LDIM
           DO 0183 J=1,KDIM
             SUM = ZERO
             DO 0182 K=1,KDIM
               SUM = SUM + X(K,I)*TMP6(K,J)
0182         CONTINUE
             TMP7(I,J) = SUM
0183       CONTINUE
0184     CONTINUE
C
         DO 0189 I=1,LDIM
           DO 0188 J=1,I
             SUM = ZERO
             DO 0187 K=1,KDIM
```

```
              SUM = SUM + TMP7(I,K)*X(K,J)
0187          CONTINUE
              HESSBETA(I,J) = HESSBETA(I,J) + SUM
              HESSBETA(J,I) = HESSBETA(I,J)
0188        CONTINUE
0189      CONTINUE
          IF ( INDEPEND )   GO TO 0194
C -- compute HESSIAN and SCORE for cross-terms and RHO
          DO 0191 I=1,KDIM
            SUM1 = ZERO
            SUM2 = ZERO
            DO 0190 J=1,KDIM
              SUM1 = SUM1 + S(J)/AHALF(J)*RINVDER1(J,I)
              SUM2 = SUM2 + S(J)/AHALF(J)*RINVDER2(J,I)
0190        CONTINUE
            TMP8(I) = SUM1*T(I)
            SCORERHO = SCORERHO + SUM1*S(I)/AHALF(I)
            HESSRHO = HESSRHO + SUM2*S(I)/AHALF(I)
0191      CONTINUE
          DO 0193 I=1,LDIM
            SUM = ZERO
            DO 0192 J=1,KDIM
              SUM = SUM + TMP8(J)*X(J,I)
0192        CONTINUE
            HESSCROS(I) = HESSCROS(I) + SUM
0193      CONTINUE
C
C -- 8. compute this subject's Q² and D² distance
C           Q² = [S'·inv(V)·S] = [S'·inv(A^½)·inv(R)·inv(A^½)·S]
0194      CONTINUE
          DISTSUBJ = ZERO
          DO 0198 I=1,KDIM
            SUM = ZERO
            DO 0197 J=1,KDIM
              SUM = SUM + (S(J)/AHALF(J))*RINV(I,J)
0197        CONTINUE
            DISTSUBJ = DISTSUBJ + ( SUM*(S(I)/AHALF(I)) )
0198      CONTINUE
          DISTSQRD = DISTSQRD + DISTSUBJ
          IF ( .NOT. SUCCESS ) GO TO 0199
C
C -- convergence assured, now print some diagnostics (if outlier)
C
          EXPTSUBJ = DISTSUBJ*SCALE
          PVALUE = ONE - PROBCHI2(EXPTSUBJ,KDIM)
          IF ( PVALUE .GT. 0.05D0 ) GO TO 199
          IOUTLIER = IOUTLIER + 1
C --
CCCCC     GO TO 0199 (make this a true GO TO to bypass printing)
C --
          WRITE(7,6197) ISUBJECT,EXPTSUBJ,PVALUE,KDIM
6197      FORMAT(1X,'Subj#',I4,' Mahalanobis D²=',D15.8,', P=',F5.3,
     +    ' on ',I2,' DF')
          WRITE(7,6198) (Y(I),I=1,KDIM)
```

```
6198   FORMAT(1X,' Observed ',7(F13.6,1X))
       WRITE(7,6199) (E(I),I=1,KDIM)
6199   FORMAT(1X,' Expected ',7(F13.6,1X))
C
C -- go back to process next subject
C
0199   CONTINUE
       GO TO 0110
C ---------------------------------------------------------------
C -- finished processing of all subjects for current iteration
C ---------------------------------------------------------------
0200   CONTINUE
       RN = DFLOAT(N)
       RMDIM = DFLOAT(MDIM)
C -- compute the expected Mahalanobis distance D²
       EXPTSQRD = (RN - RMDIM)
C -- compute the scale parameter Φ = D²/Q²
       SCALE = EXPTSQRD/DISTSQRD
C
       IDFDIST = N - MDIM
       PVALUE = ONE - PROBCHI2(DISTSQRD,IDFDIST)
       IF (SUCCESS) WRITE(6,6200) EXPTSQRD,DISTSQRD,IDFDIST,PVALUE,
     +                            SCALE
       WRITE(7,6200) EXPTSQRD,DISTSQRD,IDFDIST,PVALUE,
     +               SCALE
6200   FORMAT(1X,'   Mahalanobis D²:',D18.11,/,
     +        1X,'   Quadratic   Q²:',D18.11,' on ',I5,' DF, P=',F6.4,
     + /,'      Computed Scale      = ',D14.8,2X)
       WRITE(7,6201) SCORERHO
6201   FORMAT(1X,' * Score (RHO):',D12.6)
       WRITE(7,6202) (SCORBETA(I),I=1,LDIM)
6202   FORMAT(1X,' Score (BETAs):',10(D12.6,1X))
C
C Augment scores and hessian for BETA, RHO, and cross-terms
C -- move BETA first
       DO 0207 I=1,LDIM
         SCORE(I) = SCORBETA(I)
         DO 0205 J=1,LDIM
           HESSIAN(I,J) = HESSBETA(I,J)
0205     CONTINUE
0207   CONTINUE
       IF ( INDEPEND ) GO TO 0210
C -- move RHO next
       DO 0209 J=1,LDIM
         HESSIAN(J,MDIM) = HESSCROS(J)
         HESSIAN(MDIM,J) = HESSCROS(J)
0209   CONTINUE
C -- because we are maximizing the -(Least Squares), CHANGE the sign
       HESSIAN(MDIM,MDIM) = -HESSRHO
       SCORE(MDIM) = -SCORERHO
C ---------------------------------------------------------------
0210   CONTINUE
C ---------------------------------------------------------------
C
```

```
C -- move -SCORE TO DELTA,  this to solve  HESSIAN*DELTA = -SCORE
      DO 0219 I=1,MDIM
        DELTA(I) = -SCORE(I)
        DO 0215 J=1,MDIM
          HESSCOPY(I,J) = HESSIAN(I,J)
0215    CONTINUE
0219  CONTINUE
C
C -- Solve for DELTA and prepare to update BETA, etc
C
      CALL DECOMP(MMAX,MDIM,HESSIAN,COND,IPVT,WORK)
      IF( COND .NE. COND+ONE ) GO TO 0230
0225  WRITE(7,6225)
6225  FORMAT(1HO,'*** Ill cond -HESSIAN matrix @ 0225 ***',/)
      GO TO 9998
0230  CALL SOLVE(MMAX,MDIM,MMAX,1,HESSIAN,DELTA,IPVT)
      WRITE(7,6230) (DELTA(I),I=1,MDIM)
6230  FORMAT(1X,'   Correction to BETA (&RHO): DELTA',/,
     + 2X,10(D12.6,1X))
C
C -- If already converged in previous pass, let's wrap it up
C
      IF ( SUCCESS ) GO TO 0301
C
C -- Otherwise, check for convergence (via diminishing SCORE)
C
      ICONVERG = ZERO
      DO 0249 I=1,MDIM
        ABSSCORE = DABS(SCORE(I))
        IF ( ABSSCORE .GT. SIGNIF ) ICONVERG = ICONVERG + 1
0249  CONTINUE
C
      IF ( ICONVERG .EQ. 0 ) GO TO 0300
C
C -- Update BETA (and possibly RHO) and go another iteration
0260  CONTINUE
      DO 0269 I=1,LDIM
        BETA(I) = BETA(I) + DELTA(I)
0269  CONTINUE
      IF ( INDEPEND ) GO TO 0279
C -- special code to constraint RHO, but help accelerate it also
      DELTARHO = DELTA(MDIM)
      RHOINCRE = DELTARHO
      RHOWORK = RHOINCRE+RHO
      IF ( DABS(RHOWORK) .LT. ONE ) GO TO 0275
      RHOINCRE = DELTARHO/TWO
      RHOWORK = RHOINCRE+RHO
      IF ( DABS(RHOWORK) .LT. ONE ) GO TO 0275
      RHOINCRE = DELTARHO/FOUR
      RHOWORK = RHOINCRE+RHO
      IF ( DABS(RHOWORK) .LT. ONE ) GO TO 0275
      GO TO 0279
C -- but try to accelerate it if too small of update
0275  IF( RHOINCRE .LE. 1.0D-4 ) RHOWORK = RHOINCRE*1.5D0 + RHO
```

```
      RHO = RHOWORK
C - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
C       ******** Finish current iteration ********
C - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
0279   GO TO 0100
C
C ###################################################################
C       # # #   Convergence on Score Equations met   # # #
C ###################################################################
0300   CONTINUE
C
C -- Set the switch, & go back one more time to really get good results
C
       SUCCESS = .TRUE.
       GO TO 0260
C
C -- Now, report the final results for this correlation structure
C
0301   CONTINUE
       CALL GETTIM(IHOUREND,IMINEND,ISECEND,I100END)
       WRITE(6,6302) IHOURBEG,IMINBEG,ISECBEG,IHOUREND,IMINEND,ISECEND
6302   FORMAT(1X,'Start ',I2,':',I2,':',I2,' - End ',I2,':',I2,':',I2)
       WRITE(7,6310) SIGNIF,ITER
       WRITE(6,6310) SIGNIF,ITER
6310   FORMAT(1X,/,'Convergence at',F10.7,' lvl in ',I3,' iterations')
       WRITE(7,6311) IOUTLIER
       WRITE(6,6311) IOUTLIER
6311   FORMAT(1X,'   Number of OUTLIERs = ',I6)
C
C -- retrieve HESSIAN (via HESSCOPY), and compute COVAR = inv(-HESSIAN)
C
       DO 0319 I=1,MDIM
         DO 0315 J=1,MDIM
           COVAR(I,J) = ZERO
           HESSIAN(I,J) = -HESSCOPY(I,J)
0315     CONTINUE
         COVAR(I,I) = ONE
0319   CONTINUE
       CALL DECOMP(MMAX,MDIM,HESSIAN,COND,IPVT,WORK)
       IF ( COND .NE. COND+ONE ) GO TO 0330
0325   WRITE(7,6325)
6325   FORMAT(1H0,'*** Ill cond -HESSIAN matrix @ 0325 ***',/)
       GO TO 9999
0330   CALL SOLVE(MMAX,MDIM,MMAX,MDIM,HESSIAN,COVAR,IPVT)
       WRITE(7,6331)
6331   FORMAT(1X,/,'Asymptotic VAR/COVARIANCE matrix of BETA (&RHO)')
       DO 0339 I=1,MDIM
C -- Divide inv(-H) by Φ to yield COV(β)
         DO 0335 J=1,MDIM
           COVAR(I,J) = COVAR(I,J)/SCALE
0335     CONTINUE
         WRITE(7,6333) (COVAR(I,J),J=1,MDIM)
6333   FORMAT(1X,10(D12.6,1X))
0339   CONTINUE
```

```
      WRITE(7,6341)
6341  FORMAT(1X,/,'       and CORRELATION structure')
      DO 0349 I=1,MDIM
        STDERI = DSQRT(COVAR(I,I))
        DO 0345 J=1,MDIM
          STDERJ = DSQRT(COVAR(J,J))
          CORREL(I,J) = COVAR(I,J)/(STDERI*STDERJ)
0345    CONTINUE
        WRITE(7,6333) (CORREL(I,J),J=1,MDIM)
0349  CONTINUE
C -- report the scale paramter Φ
      WRITE(7,6370) SCALE
6370  FORMAT(1X,/,'Scale parameter Φ =',D14.6,/)
C --
C -- print parameter estimate information
C
0400  CONTINUE
      IF ( AUTOREG1 .OR. EXCHANGE ) WRITE(6,6402) RHO
6402  FORMAT(1X,'    * * * * * * RHO (est) = ',F12.8,' * * * * * *')
      WRITE(6,6404)
6404  FORMAT(1X,/,16X,'Analysis of M.Q.L. Estimates',//,
     + 5X,'Parameter  Estimate    Std Err    Chi-sqr          P',/)
      DO 0415 J=1,LDIM
        ISUB = IBETASUB(J)
        PARM = BETA(J)
        STDERR = DSQRT(COVAR(J,J))
        CHISQ = (PARM/STDERR)**2
        PVALUE = ONE - PROBCHI2(CHISQ,1)
        WRITE(6,6410) ISUB,PARM,STDERR,CHISQ,PVALUE
6410    FORMAT(1X,7X,'ß',I1,4X,F10.4,2X,F9.4,1X,F13.4,2X,F7.4)
0415  CONTINUE
C = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
C -- Finish of current correlation structure
C = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
      GO TO 0050
C
C ***************************************************************
C ***************************************************************
C
C -- Convergence not attained, so abort the program
C
9998  CONTINUE
      WRITE(7,6998) ITER
6998  FORMAT(1H0,'! NO CONVERGENCE AFTER ',I3,' ITERATIONS -- ABORT!')
C
C ---------------------------------------------------------------
C
C -- EXIT PROGRAM
C
9999  CONTINUE
      ENDFILE 06
      ENDFILE 07
      REWIND 06
      REWIND 07
```

```
CLOSE (4)
CLOSE (5)
CLOSE (6)
CLOSE (7)
STOP
END
```

GRADUATE SCHOOL
UNIVERSITY OF ALABAMA AT BIRMINGHAM
DISSERTATION APPROVAL FORM

Name of Candidate John David Bass

Major Subject Biostatistics

Title of Dissertation A Contribution to Longitudinal Data Analysis: Maximum Quasi-Likelihood Generalized Estimating Equations

Dissertation Committee:

_____, Chairman

_____

_____

_____

_____

Director of Graduate Program _____

Dean, UAB Graduate School _____

Date 6/8/93