9
Pseudo-Likelihood

9.1 Introduction

Full marginal maximum likelihood, as discussed in Chapters 6 and 7, can become prohibitive in terms of computation when measurement sequences are of moderate to large length. This is one of the reasons why generalized estimating equations (GEE, Chapter 8) have become so popular. One way to view the genesis of GEE is by modifying the score equations to simpler estimating equations, thereby preserving consistency and asymptotic normality, upon using an appropriately corrected variance-covariance matrix. Alternatively, the (log-)likelihood itself can be simplified to a more manageable form. This is, broadly speaking, the idea behind *pseudo-likelihood* (PL). For example, when a joint density is of the Bahadur (Section 7.2), probit (Section 7.6), or Dale (Section 7.7) form, calculating the higher-order probabilities needed to evaluate the score vector and Hessian matrix can be prohibitive while, at the same time, interest can be confined to a small number of lower-order moments. The idea is then to replace the single joint density by, for example, all univariate densities, or all pairwise densities over the set of all possible pairs within a sequence of repeated measures. As a simple illustration, a three-way density

$$L_i = f_i(y_{i1}, y_{i2}, y_{i3}|\boldsymbol{\theta}_i) \tag{9.1}$$

would be replaced by the product

$$L_i^* = f_i(y_{i1}, y_{i2}\boldsymbol{\theta}_i^*) \cdot f_i(y_{i1}, y_{i3}\boldsymbol{\theta}_i^*) \cdot f_i(y_{i2}, y_{i3}|\boldsymbol{\theta}_i^*). \tag{9.2}$$

Such a change is computationally advantageous, asymptotics can be rescued, and modeling (9.2) is equally simple, if not simpler, than modeling (9.1), as the parameter vector $\boldsymbol{\theta}_i^*$ in (9.2) typically is a subvector of $\boldsymbol{\theta}_i$ in (9.1).

Section 9.2 introduces pseudo-likelihood in a formal way, and such that it can be of use, not only here in marginal applications, but also for conditional (Chapter 12) and subject-specific (Chapters 21 and 25) applications. Appropriate test statistics are given in Section 9.3. The specific situation of PL for marginal models is the topic of Section 9.4, and a comparison between marginal PL and GEE is presented in Section 9.5. The methodology is illustrated using the NTP data (Section 9.6).

9.2 Pseudo-Likelihood: Definition and Asymptotic Properties

To formally introduce pseudo-likelihood, we will use the convenient general definition given by Arnold and Strauss (1991). See also Geys, Molenberghs, and Ryan (1999) and Aerts *et al* (2002). Without loss of generality we can assume that the vector $\boldsymbol{Y}_i$ of binary outcomes for subject i $(i = 1, \ldots, N)$ has constant dimension n. The extension to variable lengths n_i for $\boldsymbol{Y}_i$ is straightforward.

9.2.1 Definition

Define S as the set of all $2^n - 1$ vectors of length n, consisting solely of zeros and ones, with each vector having at least one non-zero entry. Denote by $\boldsymbol{y}_i^{(s)}$ the subvector of $\boldsymbol{y}_i$ corresponding to the components of s that are non-zero. The associated joint density is $f_s(\boldsymbol{y}_i^{(s)}|\boldsymbol{\theta}_i)$. To define a pseudo-likelihood function, one chooses a set $\delta = \{\delta_s | s \in S\}$ of real numbers, with at least one non-zero component. The log of the pseudo-likelihood is then defined as

$$p\ell = \sum_{i=1}^{N} \sum_{s \in S} \delta_s \ln f_s(\boldsymbol{y}_i^{(s)}|\boldsymbol{\theta}_i). \tag{9.3}$$

Adequate regularity conditions have to be assumed to ensure that (9.3) can be maximized by solving the pseudo-likelihood (score) equations, the latter obtained by differentiation of the logarithm of PL and setting the derivative equal to zero.

The classical log-likelihood function is found by setting $\delta_s = 1$ if s is the vector consisting solely of ones, and 0 otherwise.

9.2.2 Consistency and Asymptotic Normality

Before stating the main asymptotic properties of the PL estimators, we first list the required regularity conditions on the density functions $f_s(\boldsymbol{y}^{(s)}|\boldsymbol{\theta})$.

A0 The densities $f_s(\boldsymbol{y}^{(s)}|\boldsymbol{\theta})$ are distinct for different values of the parameter $\boldsymbol{\theta}$.

A1 The densities $f_s(\boldsymbol{y}^{(s)}|\boldsymbol{\theta})$ have common support, which does not depend on $\boldsymbol{\theta}$.

A2 The parameter space Ω contains an open region ω of which the true parameter value $\boldsymbol{\theta}_0$ is an interior point.

A3 ω is such that for all s, and almost all $\boldsymbol{y}^{(s)}$ in the support of $\boldsymbol{Y}^{(s)}$, the densities admit all third derivatives

$$\frac{\partial^3 f_s(\boldsymbol{y}^{(s)}|\boldsymbol{\theta})}{\partial\theta_{k_1}\partial\theta_{k_2}\partial\theta_{k_3}}.$$

A4 The first and second logarithmic derivatives of f_s satisfy

$$E_{\boldsymbol{\theta}}\left(\frac{\partial \ln f_s(\boldsymbol{y}^{(s)}|\boldsymbol{\theta})}{\partial\theta_k}\right) = 0, \qquad k = 1,\ldots,p,$$

and

$$0 < E_{\boldsymbol{\theta}}\left(\frac{-\partial^2 \ln f_s(\boldsymbol{y}^{(s)}|\boldsymbol{\theta})}{\partial\theta_{k_1}\partial\theta_{k_2}}\right) < \infty, \qquad k_1, k_2 = 1,\ldots,p.$$

A5 The matrix I_0, to be defined in (9.5), is positive definite.

A6 There exist functions $M_{k_1k_2k_3}$ such that

$$\sum_{s\in S}\delta_s E_{\boldsymbol{\theta}}\left|\frac{\partial^3 \ln f_s(\boldsymbol{y}^{(s)}|\boldsymbol{\theta})}{\partial\theta_{k_1}\partial\theta_{k_2}\partial\theta_{k_3}}\right| < M_{k_1k_2k_3}(\boldsymbol{y})$$

for all $\boldsymbol{y}$ in the support of f and for all $\boldsymbol{\theta} \in \omega$ and $m_{k_1k_2k_3} = E_{\boldsymbol{\theta}_0}[M_{k_1k_2k_3}(Y)] < \infty$.

Theorem 9.1, proven by Arnold and Strauss (1991), guarantees the existence of at least one solution to the pseudo-likelihood equations, which is consistent and asymptotically normal. Without loss of generality, we can assume $\boldsymbol{\theta}$ is constant. Replacing it by $\boldsymbol{\theta}_i$ and modeling it as a function of covariates is straightforward.

Theorem 9.1 (Consistency and Asymptotic Normality) *Assume that* $(\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_N)$ *are i.i.d. with common density that depends on* $\boldsymbol{\theta}_0$. *Then under regularity conditions (A1)–(A6):*

1. *The pseudo-likelihood estimator* $\tilde{\boldsymbol{\theta}}_N$*, defined as the maximizer of (9.3), converges in probability to* $\boldsymbol{\theta}_0$*.*

2. $\sqrt{N}(\tilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)$ *converges in distribution to*

$$N_p[\mathbf{0}, I_0(\boldsymbol{\theta}_0)^{-1} I_1(\boldsymbol{\theta}_0) I_0(\boldsymbol{\theta}_0)^{-1}], \tag{9.4}$$

with $I_0(\boldsymbol{\theta})$ *defined by*

$$I_{0,k_1k_2}(\boldsymbol{\theta}) = -\sum_{s\in S} \delta_s E_{\boldsymbol{\theta}} \left(\frac{\partial^2 \ln f_s(\boldsymbol{y}^{(s)}|\boldsymbol{\theta})}{\partial\theta_{k_1}\partial\theta_{k_2}} \right) \tag{9.5}$$

and $I_1(\boldsymbol{\theta})$ *by*

$$I_{2,k_1k_2}(\boldsymbol{\theta}) = \sum_{s,t\in S} \delta_s\delta_t E_{\boldsymbol{\theta}} \left(\frac{\partial \ln f_s(\boldsymbol{y}^{(s)}|\boldsymbol{\theta})}{\partial\theta_{k_1}} \frac{\partial \ln f_t(\boldsymbol{y}^{(t)}|\boldsymbol{\theta})}{\partial\theta_{k_2}} \right). \tag{9.6}$$

Similar in spirit to generalized estimating equations (Chapter 8), the asymptotic normality result provides an easy way to consistently estimate the asymptotic covariance matrix. Indeed, the matrix I_0 is found from evaluating the second derivative of the log PL function at the PL estimate. The expectation in I_1 can be replaced by the cross-products of the observed scores. We will refer to I_0^{-1} as the model based variance estimator (which should not be used as it overestimates the precision), to I_1 as the empirical correction, and to $I_0^{-1} I_1 I_0^{-1}$ as the empirically corrected variance estimator. In the context of generalized estimating equations, this is also known as the sandwich estimator.

As discussed by Arnold and Strauss (1991), and exactly the same as with GEE, the Cramèr-Rao inequality implies that $I_0^{-1} I_1 I_0^{-1}$ is greater than the inverse of I (the Fisher information matrix for the maximum likelihood case), in the sense that $I_0^{-1} I_1 I_0^{-1} - I^{-1}$ is positive semi-definite. Strict inequality holds if the PL estimator fails to be a function of a minimal sufficient statistic. Therefore, a PL estimator is always less efficient than the corresponding ML estimator. Note that, for maximum likelihood, the full density f would be used, rather than the pseudo-likelihood contributions.

9.3 Pseudo-Likelihood Inference

The close connection of PL to likelihood is an attractive feature. It enabled Geys, Molenberghs, and Ryan (1999) to construct pseudo-likelihood ratio test statistics that have easy-to-compute expressions and intuitively appealing limiting distributions. In contrast, likelihood ratio test statistics for GEE (Rotnitzky and Jewell 1990) are slightly more complicated.

In practice, one will often want to perform a flexible model selection. Therefore, one needs extensions of the Wald, score, or likelihood ratio test statistics to the pseudo-likelihood framework. Rotnitzky and Jewell (1990) examined the asymptotic distributions of generalized Wald and score tests, as well as likelihood ratio tests, for regression coefficients obtained by generalized estimating equations for a class of marginal generalized linear models for correlated data. Using similar ideas, we derive different test statistics, as well as their asymptotic distributions for the pseudo-likelihood framework. Liang and Self (1996) have considered a test statistic, for one specific type of pseudo-likelihood function, which is similar in form to one of the tests we will present below.

Suppose we are interested in testing the null hypothesis $H_0 : \boldsymbol{\gamma} = \boldsymbol{\gamma}_0$, where $\boldsymbol{\gamma}$ is an r-dimensional subvector of the vector of regression parameters $\boldsymbol{\theta}$ and write $\boldsymbol{\theta}$ as $(\boldsymbol{\gamma}', \boldsymbol{\delta}')'$. Then, several test statistics can be used.

9.3.1 Wald Statistic

Because of the asymptotic normality of the PL estimator $\tilde{\boldsymbol{\theta}}_N$,

$$W^* = N(\tilde{\gamma}_N - \gamma_0)' \Sigma_{\gamma\gamma}^{-1} (\tilde{\gamma}_N - \gamma_0)$$

has an asymptotic χ_r^2 distribution under the null hypothesis, where $\Sigma_{\gamma\gamma}$ denotes the $r \times r$ submatrix of $\Sigma = I_0^{-1} I_1 I_0^{-1}$. In practice, the matrix Σ can be replaced by a consistent estimator, obtained by substituting the PL estimator $\tilde{\theta}_N$. Although the Wald test is in general simple to apply, it is well-known to be sensitive to changes in parameterization. The Wald test statistic is therefore particularly unattractive for conditionally specified models, as marginal effects are likely to depend in a complex way on the model parameters (Diggle, Heagerty, Liang, and Zeger 2002).

9.3.2 Pseudo-Score Statistics

As an alternative to the Wald statistic, one can propose the *pseudo-score statistic*. A score test has the advantage that it can be obtained by fitting the null model only. Furthermore, it is invariant to reparameterization. Let us define

$$S^*(e.c.) = \frac{1}{N} U_\gamma[\gamma_0, \tilde{\delta}(\gamma_0)]' I_0^{\gamma\gamma} \Sigma_{\gamma\gamma}^{-1} I_0^{\gamma\gamma} U_\gamma[\gamma_0, \tilde{\delta}(\gamma_0)],$$

where 'e.c.' denotes empirically corrected and $\tilde{\delta}(\gamma_0)$ denotes the maximum pseudo-likelihood estimator in the subspace where $\gamma = \gamma_0$, $I_0^{\gamma\gamma}$ is the $r \times r$ submatrix of the inverse of I_0, and $I_0^{\gamma\gamma} \Sigma_{\gamma\gamma}^{-1} I_0^{\gamma\gamma}$ is evaluated under H_0. Geys, Molenberghs, and Ryan (1999) showed that this pseudo-score statistic is asymptotically χ_r^2 distributed under H_0. As discussed by Rotnitzky and Jewell (1990) in the context of generalized estimating equations, such a

score statistic may suffer from computational stability problems. A model based test that may be computationally simpler is:

$$S^*(m.b.) = \frac{1}{N} U_\gamma[\gamma_0, \tilde{\delta}(\gamma_0)]' I_0^{\gamma\gamma} U_\gamma[\gamma_0, \tilde{\delta}(\gamma_0)].$$

However, its asymptotic distribution under H_0 is complicated and given by $\sum_{j=1}^{r} \lambda_j \chi^2_{1(j)}$ where the $\chi^2_{1(j)}$ are independently distributed as χ^2_1 variables and $\lambda_1 \geq \ldots \geq \lambda_r$ are the eigenvalues of $(I_0^{\gamma\gamma})^{-1}\Sigma_{\gamma\gamma}$, evaluated under H_0. The score statistic $S^*(m.b.)$ can be adjusted such that it has an approximate χ^2_r distribution, which is much easier to evaluate. Several types of adjustments have been proposed in the literature (Rao and Scott 1987, Roberts, Rao, and Kumar 1987). Similar to Rotnitzky and Jewell (1990), Geys, Molenberghs, and Ryan (1999) proposed an adjusted pseudo-score statistic

$$S^*_a(m.b.) = S^*(m.b.)/\overline{\lambda},$$

where $\overline{\lambda}$ is the arithmetic mean of the eigenvalues λ_j. Note that no distinction can be made between $S^*(e.c.)$ and $S^*_a(m.b.)$ for $r = 1$. Moreover, in the likelihood-based case, all eigenvalues reduce to one and thus all three statistics coincide with the model based likelihood score statistic.

9.3.3 Pseudo-Likelihood Ratio Statistics

Another alternative is provided by the pseudo-likelihood ratio test statistic, which requires comparison of full and reduced model:

$$G^{*2} = 2\left[p\ell(\tilde{\theta}_N) - p\ell(\gamma_0, \tilde{\delta}(\gamma_0))\right].$$

Geys, Molenberghs, and Ryan (1999) showed that the asymptotic distribution of G^{*2} can also be written as a weighted sum $\sum_{j=1}^{r} \lambda_j \chi^2_{1(j)}$, where the $\chi^2_{1(j)}$ are independently distributed as χ^2_1 variables and $\lambda_1 \geq \ldots \geq \lambda_r$ are the eigenvalues of $(I_0^{\gamma\gamma})^{-1}\Sigma_{\gamma\gamma}$. Alternatively, the adjusted pseudo-likelihood ratio test statistic, defined by

$$G^{*2}_a = G^{*2}/\overline{\lambda},$$

is approximately χ^2_r distributed. Their proof shows that G^{*2} can be rewritten as an approximation to a Wald statistic. The covariance structure of the Wald statistic can be calculated under the null hypothesis, but also under the alternative hypothesis. Both versions of the Wald tests are asymptotically equivalent under H_0 (Rao 1973, p. 418). Therefore, it can be argued that the adjustments in G^{*2}_a can also be evaluated under the null as well as under the alternative hypothesis. These adjusted statistics will then be denoted by $G^{*2}_a(H_0)$ and $G^{*2}_a(H_1)$, respectively. In analogy with the Wald test statistic, we expect $G^{*2}_a(H_1)$ to have high power. A similar

reasoning suggests that the score test $S_a^*(m.b.)$ might closely correspond to $G_a^{*2}(H_0)$, as both depend strongly on the fitted null model. Analogous results were obtained by Rotnitzky and Jewell (1990). Aerts *et al* (2002) reported on extensive simulations to compare the behavior of the various test statistics.

The asymptotic distribution of the pseudo-likelihood based test statistics are weighted sums of independent χ_1^2 variables where the weights are unknown eigenvalues. In Aerts and Claeskens (1999) it is shown theoretically that the parametric bootstrap leads to a consistent estimator for the null distribution of the pseudo-likelihood ratio test statistic. The bootstrap approach does not need any additional estimation of unknown eigenvalues and automatically corrects for the incomplete specification of the joint distribution in the pseudo-likelihood. Similar results hold for the robust Wald and robust score test. The simulation study of Aerts and Claeskens (1999) indicates that the χ^2 tests often suffer from inflated type I error probabilities, which are nicely corrected by the bootstrap. This is especially the case for the Wald statistic, whereas the asymptotic χ^2 distribution of the robust score statistic test is performing quite well. The parametric bootstrap is expected to break down if the likelihood of the data is grossly misspecified. Aerts *et al* (2002, Chapter 11) present a more robust semiparametric bootstrap, based on resampling the score and differentiated score values.

9.4 Marginal Pseudo-Likelihood

A marginally specified odds ratio model (Molenberghs and Lesaffre 1994, 1999, Glonek and McCullagh 1995, Lang and Agresti 1994, see also Section 7.7) becomes prohibitive in computational terms when the number of replications within a unit gets moderate to large. In such a situation, both GEE and PL are viable alternatives. The connection between GEE based on odds ratios (Section 8.6) and the corresponding PL is strong and will be developed in Section 9.5. Marginal PL methodology has been proposed, among others, by le Cessie and van Houwelingen (1994) and Geys, Molenberghs, and Lipsitz (1998).

9.4.1 Definition of Marginal Pseudo-Likelihood

Again, assume there are $i = 1, \ldots, N$ units with $j = 1, \ldots, n_i$ measurements per unit. We will start with a general form and then focus on clustered binary data, where the outcomes Y_{ij} are replaced by a summary statistic $Z_i = \sum_{j=1}^{n_i} Y_{ij}$, the total number of successes within the ith cluster.

9.4.1.1 First Form

le Cessie and van Houwelingen (1994) replace the true contribution of a vector of correlated binary data to the full likelihood, written as $f(y_{i1}, \ldots, y_{in_i})$, by the product of all pairwise contributions $f(y_{ij_1}, y_{ij_2})$ $(1 \leq j_1 < j_2 \leq n_i)$, to obtain a *pseudo-likelihood function*. Grouping the outcomes for subject i into a vector $\boldsymbol{Y}_i$, the contribution of the ith cluster to the log pseudo-likelihood is

$$p\ell_i = \sum_{1 \leq j_1 < j_2 \leq n_i} \ln f(y_{ij_1}, y_{ij_2}), \tag{9.7}$$

if it contains more than one observation. Otherwise $p\ell_i = f(y_{i1})$. In what follows, we restrict our attention to clusters of size larger than 1. Units of size 1 contribute to the marginal parameters only. This specific version of pseudo-likelihood is often referred to as pairwise likelihood.

Using a bivariate Plackett distribution (Plackett 1965, Section 7.7.1), the joint probabilities $f(y_{ij_1}, y_{ij_2})$, denoted by $\mu_{ij_1j_2}$, can be specified using (7.40), with the pairwise odds ratio as in (7.39). The contributions of the form $f(y_{ij_1}, y_{ij_2})$ can then be combined into a pseudo-likelihood function $p\ell$ (9.7), which can be maximized as if it where a genuine bivariate log-likelihood. The asymptotic variance-covariance matrix of the parameter estimates then follows from (9.4).

9.4.1.2 Under Exchangeability

For binary data and taking the exchangeability assumption into account, the log pseudo-likelihood contribution $p\ell_i$ can be formulated as:

$$p\ell_i = \binom{z_i}{2} \ln \mu^*_{i11} + \binom{n_i - z_i}{2} \ln \mu^*_{i00} + z_i(n_i - z_i) \ln \mu^*_{i10}. \tag{9.8}$$

In this formulation, μ^*_{i11} and μ^*_{i00} denote the bivariate probabilities of observing two *successes* or two *failures*, respectively, and μ^*_{i10} is the probability for the first component being 1 and the second being 0. Under exchangeability, this is identical to the probability μ^*_{i01} for the first being 0 and the second being 1. If we consider the following reparameterization:

$$\begin{aligned} \mu_{i11} &= \mu^*_{i11}, \\ \mu_{i10} &= \mu^*_{i11} + \mu^*_{i10} = \mu_{01}, \\ \mu_{i00} &= \mu^*_{i11} + \mu^*_{i10} + \mu^*_{i01} + \mu^*_{i00} = 1, \end{aligned}$$

then this one-to-one reparameterization maps the three, common within-cluster, two-way marginal probabilities $(\mu^*_{i11}, \mu^*_{i10}, \mu^*_{i00})$ to two one-way marginal probabilities (which under exchangeability are both equal to μ_{i10}) and one two-way probability $\mu_{i11} = \mu^*_{i11}$. Hence, equation (9.8) can be reformulated as:

$$p\ell_i = \binom{z_i}{2} \ln \mu_{i11} + \binom{n_i - z_i}{2} \ln(1 - 2\mu_{i10} + \mu_{i11})$$

$$+z_i(n_i - z_i)\ln(\mu_{i10} - \mu_{i11}), \tag{9.9}$$

and the pairwise odds ratio ψ_{ijk} reduces to:

$$\psi_i = \frac{\mu_{i11}(1 - 2\mu_{i10} + \mu_{i11})}{(\mu_{i10} - \mu_{i11})^2}.$$

To enable model specification, we assume a composite link function $\boldsymbol{\eta}_i = (\eta_{i1}, \eta_{i2})'$ with a mean and an association component:

$$\begin{aligned} \eta_{i1} &= \ln(\mu_{i10}) - \ln(1 - \mu_{i10}), \\ \eta_{i2} &= \ln(\psi_i) = \ln(\mu_{i11}) + \ln(1 - 2\mu_{i10} + \mu_{i11}) - 2\ln(\mu_{i10} - \mu_{i11}). \end{aligned}$$

From these links, the univariate and pairwise probabilities are easily derived (Plackett 1965), leading to a specific version of (7.40):

$$\mu_{i10} = \frac{\exp(\eta_{i1})}{1 + \exp(\eta_{i1})}$$

and

$$\mu_{i11} = \begin{cases} \frac{1+2\mu_{i10}(\psi_i-1)-S_i}{2(\psi_i-1),} & \text{if } \psi_i \neq 1 \\ \mu_{i10}^2 & \text{if } \psi_i = 1, \end{cases}$$

with

$$S_i = \sqrt{[1 + 2\mu_{i10}(\psi_i - 1)]^2 + 4\psi_i(1 - \psi_i)\mu_{i10}^2}.$$

Finally, we can assume a linear model $\boldsymbol{\eta}_i = X_i\boldsymbol{\theta}$, with X_i a known design matrix and $\boldsymbol{\theta}$ a vector of unknown regression parameters. The maximum pseudo-likelihood estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is then defined as the solution to the pseudo-score equations $\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{0}$. Using the chain rule, $\boldsymbol{U}(\boldsymbol{\theta})$ can be written as:

$$\boldsymbol{U}(\boldsymbol{\theta}) = \sum_{i=1}^{N} X_i'(T_i^{-1})' \frac{\partial p\ell_i}{\partial \boldsymbol{\mu}_i} \tag{9.10}$$

with $\boldsymbol{\mu}_i = (\mu_{i10}, \mu_{i11})'$ and $T_i = \partial\boldsymbol{\eta}_i/\partial\boldsymbol{\mu}_i$. Newton-Raphson starts with a vector of initial estimates $\boldsymbol{\theta}^{(0)}$ and updates the current value of the parameter vector $\boldsymbol{\theta}^{(s)}$ by

$$\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)} + W(\boldsymbol{\theta}^{(s)})^{-1}\boldsymbol{U}(\boldsymbol{\theta}^{(s)}).$$

Here, $W(\boldsymbol{\theta})$ is the matrix of the second derivatives of the log pseudo-likelihood with respect to the regression parameters $\boldsymbol{\theta}$:

$$W(\boldsymbol{\theta}) = \sum_{i=1}^{N} X_i' \left[\boldsymbol{F}_i + (T_i^{-1})' \frac{\partial^2 p\ell_i}{\partial\boldsymbol{\mu}_i \partial\boldsymbol{\mu}_i'} (T_i^{-1}) \right] X_i,$$

and $\boldsymbol{F}_i$ is defined by (McCullagh 1987, p. 5; Molenberghs and Lesaffre 1999, see also Section 7.12.2):

$$(F_i)_{pq} = \sum_s \sum_{a,b,c} \frac{\partial^2 \eta_{ia}}{\partial \mu_{ib} \partial \mu_{ic}} \frac{\partial \mu_{is}}{\partial \eta_{ia}} \frac{\partial \mu_{ib}}{\partial \eta_{ip}} \frac{\partial \mu_{ic}}{\partial \eta_{iq}} \frac{\partial p\ell_i}{\partial \mu_{is}}.$$

The Fisher scoring algorithm is obtained by replacing the matrix $W(\boldsymbol{\theta})$ by its expected value:

$$E[W(\boldsymbol{\theta})] = \sum_{i=1}^{N} X_i' (T_i^{-1})' A_i (T_i^{-1}) X_i,$$

with A_i the expected value of the matrix of second derivatives of the log pseudo-likelihood $p\ell_i$ with respect to $\boldsymbol{\mu}_i$.

The sandwich estimator (9.4) can now be written as:

$$W(\widehat{\boldsymbol{\theta}})^{-1} \left[\sum_{i=1}^{N} \boldsymbol{U}_i(\widehat{\boldsymbol{\theta}}) \boldsymbol{U}_i(\widehat{\boldsymbol{\theta}})' \right] W(\widehat{\boldsymbol{\theta}})^{-1}.$$

9.4.1.3 Second Form

A non-equivalent specification of the pseudo-likelihood contribution (9.7) is:

$$p\ell_i^* = p\ell_i/(n_i - 1).$$

The factor $1/(n_i - 1)$ corrects for the feature that each response Y_{ij} occurs $n_i - 1$ times in the ith contribution to the PL, and it ensures that the PL reduces to full likelihood under independence, as then (9.9) simplifies to:

$$p\ell_i = (n_i - 1) \left[z_i \ln(\mu_{i10}) + (n_i - z_i) \ln(1 - \mu_{i10}) \right].$$

We can replace $p\ell_i$ by $p\ell_i^*$. However, if $(n_i - 1)$ is considered random it is not obvious that the expected value of $U_i(\boldsymbol{\theta})/(n_i - 1)$ equals zero. To ensure that the solution to the new pseudo-score equation is consistent, we have to assume that n_i is independent of the outcomes given the covariates for the ith unit. When all n_i are equal, the PL estimator $\boldsymbol{\theta}$ and its variance-covariance matrix remain the same, no matter whether we use $p\ell_i$ or $p\ell_i^*$ in the definition of the log pseudo-likelihood.

9.4.2 A Generalized Linear Model Representation

To obtain the pseudo-likelihood function described in Section 9.4.1, we replaced the true contribution $f(y_{i1}, \ldots, y_{in_i})$ of the ith unit to the full likelihood by the product of all pairwise contributions $f(y_{ij_1}, y_{ij_2})$ with

$1 \leq j_1 < j_2 \leq n_i$. This implies that a particular response y_{ij} occurs $n_i - 1$ times in $p\ell_i$. Therefore, it is useful to construct for each response y_{ij}, $n_i - 1$ replicated $y_{ij_1}^{(j_2)}$ with $j_2 \neq j_1$. The dummy response $y_{ij_1}^{(j_2)}$ is to be interpreted as the particular replicate of y_{ij} that is paired with the replicate $y_{ij_2}^{(j_1)}$ of y_{ij_2} in the pseudo-likelihood function. Using this at first sight odd but convenient device, we are able to rewrite the gradient of the log pseudo-likelihood $p\ell$ in an appealing generalized linear model type representation. With notation introduced in the previous section, the gradient can now be written as

$$\boldsymbol{U}(\boldsymbol{\theta}) = \sum_{i=1}^{N} X_i'(T_i^{-1})' V_i^{-1}(\boldsymbol{Z}_i - \boldsymbol{\mu}_i),$$

or, using the second representation $p\ell_i^*$, as

$$\boldsymbol{U}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{1}{n_i - 1} X_i'(T_i^{-1})' V_i^{-1}(\boldsymbol{Z}_i - \boldsymbol{\mu}_i),$$

where we now define

$$\boldsymbol{Z}_i = \begin{pmatrix} \sum_{j_1=1}^{n_i} \sum_{j_2 \neq j_1} Y_{ij_1}^{(j_2)} \\ \frac{1}{2} \sum_{j_1=1}^{n_i} \sum_{j_2 \neq j_1} Y_{ij_1}^{(j_2)} Y_{ij_2}^{(j_1)} \end{pmatrix}, \quad \mu_i = \begin{pmatrix} n_i(n_i - 1)\mu_{i10} \\ \binom{n_i}{2}\mu_{i11} \end{pmatrix},$$

and V_i is the covariance matrix of $\boldsymbol{Z}_i$. Geys, Molenberghs, and Lipsitz (1998) have shown that the elements of V_i take appealing expressions and are easy to implement. One only needs to evaluate first- and second-order probabilities. Under independence, the variances reduce to well-known quantities. To obtain a suitable PL estimator for $\boldsymbol{\theta}$, we can use the Fisher-scoring algorithm where the matrix A_i in the previous section is now replaced by the inverse of V_i. The asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}$ is estimated in a similar fashion as before.

9.5 Comparison with Generalized Estimating Equations

In the previous sections, we described one alternative estimating procedure for full maximum likelihood estimation in the framework of a marginally specified odds ratio model, which is easier and much less time consuming. Several questions arise such as to how the different methods compare in terms of efficiency and in terms of computing time and what the mathematical differences and similarities are. At first glance, there is a fundamental difference. A pseudo-likelihood function is constructed by modifying a joint density. Parameters are estimated by setting the first derivatives of

this function equal to zero. On the contrary, generalized estimating equations follow from specification of the first few moments and by adopting assumptions about the higher order moments. We will explore similarities and differences in some detail.

In Section 9.4.2, we have rewritten the PL score equations as contrasts of observed and fitted frequencies, establishing some agreement between PL and GEE2. Both procedures lead to similar estimating equations. The most important difference is in the evaluation of the matrix $V_i = \text{Cov}(Z_i)$. This only involves first- and second-order probabilities for the pseudo-likelihood procedure. In this respect, PL resembles GEE1. In contrast, GEE2 also requires evaluation of third- and fourth-order probabilities. This makes the GEE2 score equations harder to evaluate and also more time consuming.

Both pseudo-likelihood and generalized estimating equations yield consistent and asymptotically normally distributed estimators, provided an empirically corrected variance estimator is used and provided the model is correctly specified. This variance estimator is similar for both procedures, the main difference being the evaluation of V_i.

If we define the log of the pseudo-likelihood contribution for clusters with size larger than one as $p\ell_i^* = p\ell_i/(n_i - 1)$, the first component of the PL vector contribution $\boldsymbol{S}_i = \boldsymbol{Z}_i - \boldsymbol{\mu}_i$ equals that of GEE2. On the contrary, the association component differs by a factor of $1/(n_i - 1)$. Yet, if we would define the log pseudo-likelihood as $p\ell = \sum_{i=1}^{N} p\ell_i$, then the second components would be equal, while the first components would differ by a factor of $n_i - 1$. Therefore, in studies where the main interest lies in the marginal mean parameters, one would prefer $p\ell^*$ over $p\ell$. However, if primary interest focuses on the estimation of the association parameters, we advocate the use of $p\ell$ instead. GEE1 in that case should be avoided, as its goal is limited to estimation of the mean model parameters, whereas GEE2 is computationally more complex.

Aerts *et al* (2002) compared PL, GEE1, and GEE2 in terms of asymptotic and small sample relative efficiency. It was found that the behavior of PL is generally highly acceptable. In particular, the behavior of PL was very similar to GEE2, while in terms of computational complexity it is closer to GEE1 than to GEE2. Liang, Zeger, and Qaqish (1992) suggested GEE1, GEE2, and PL may be less efficient when the number of repeated measures per unit are unequal.

9.6 Analysis of NTP Data

We apply the PL and first- and second-order GEE estimating procedures to data from the DEHP and DYME studies, described in Section 2.7 and analyzed, using the Bahadur model, in Section 7.2.3 and, using a number of GEE methods (GEE1, GEE2, and ALR), in Section 8.9. The model

TABLE 9.1. *NTP Data. Parameter estimates (empirically corrected standard errors) for pseudo-likelihood (PL), GEE1, and GEE2 with exchangeable odds ratio, fitted to the collapsed outcome in the DEHP and DYME studies.* β_0 *and* β_d *are the marginal intercept and dose effect, respectively;* α *is the log odds ratio;* ψ *is the odds ratio.*

Study	β_0	β_d	α	ψ
	Newton-Raphson PL Estimates			
DEHP	-3.98(0.30)	5.57(0.61)	1.10(0.27)	3.00(0.81)
DYME	-5.73(0.46)	8.71(0.94)	1.42(0.31)	4.14(1.28)
	Fisher scoring PL Estimates			
DEHP	-3.98(0.30)	5.57(0.61)	1.11(0.27)	3.03(0.82)
DYME	-5.73(0.47)	8.71(0.95)	1.42(0.35)	4.14(1.45)
	GEE2 Estimates			
DEHP	-3.69(0.25)	5.06(0.51)	0.97(0.23)	2.64(0.61)
DYME	-5.86(0.42)	8.96(0.87)	1.36(0.34)	3.90(1.32)
	GEE1 Estimates			
DEHP	-4.02(0.31)	5.79(0.62)	0.41(0.34)	1.51(0.51)
DYME	-5.89(0.42)	8.99(0.87)	1.46(0.75)	4.31(3.23)

used in the earlier analyses is retained, using intercept (β_0) and dose (β_d) parameters. The log odds ratio ψ_i is modeled as $\ln \psi_i = \alpha$, in agreement with, for example, Table 8.5. Table 9.1 shows that the parameter estimates, obtained by either the pseudo-likelihood or the generalized estimating equations approach, are comparable. Note that the GEE1 and GEE2 parameter estimates differ somewhat from the ones obtained in Tables 8.2–8.5, as here the odds ratio is used to measure association, whereas we used the correlation coefficient in Tables 8.2–8.4. Table 8.5 used the odds ratio as well, but there ALR was used as estimation method. Because main interest is focused on the dose effect, we used $p\ell^*$ rather than $p\ell$. Dose effects and association parameters are, again, significant throughout, except for the GEE1 association estimates. For this procedure, β_a is not significant for the DEHP study and marginally significant for the DYME study. The GEE1 standard errors for β_a are much larger than for their PL and GEE2 counterparts. The GEE2 standard errors are the smallest among the different estimating approaches, which is in agreement with findings in previous sections. Furthermore, it is observed that the standard errors of the Newton-Raphson PL algorithm are generally slightly smaller than those obtained using Fisher scoring, which is in line with other empirical findings. On the other hand, the Newton-Raphson procedure is computationally slightly more complex in this case. The time gain of Fisher scoring, however, is negligible. PL based on the classical representation of Section 9.4.1 only needs 11% of the computation time needed for GEE2. For the GLM, based representation

of Section 9.4.2, this becomes 7%. The corresponding figure for GEE1 is 2.5%.