Determining Linkage and Mode of Inheritance: Mod Scores and Other Methods

Kung-Yee Liang, Paul J. Rathouz, and Terri H. Beaty

Departments of Biostatistics (K.-Y.L., P.J.R.) and Epidemiology (T.H.B.), Johns Hopkins University, School of Hygiene and Public Health, Baltimore, Maryland

The lod score method remains a popular approach for detecting linkage and estimating the recombination fraction \( \theta \) between a marker locus and a trait locus. However, its implementation requires knowledge of all parameters of the genetic mechanism, including the number of loci involved and the genotype specific penetrance, which could depend on factors such as age. When some of the penetrance parameters \( \phi \) are unknown, several methods are available, and have been reviewed by Hodge and Elston [1994] Genet Epidemiol 11:329–342]. These include the “mod score” (lod score maximized over \( \theta \) under a wrong value of \( \phi \)) and “mod score” (lod score maximized over both \( \theta \) and \( \phi \)) methods for inference on \( \theta \). It has further been proposed that the mod score also be used for estimating \( \phi \). In this paper, we review and assess the adequacy of these two methods for inferences on both \( \phi \) and \( \theta \). In particular, all of the methods can be seen as variations on likelihood inference, using the information in the conditional likelihood for the marker data, given the trait data. The loss of efficiency of the mod is compared to that of the full likelihood, which utilizes all information available in the trait data. We also propose an alternative, based on the pseudo-likelihood, where \( \phi \) is estimated via the trait information and plugged into the conditional likelihood. This method is compared to the mod score method, and the advantages and disadvantages of each are elucidated. In particular, it is seen that the pseudo-likelihood method can be more efficient than the mod score method if the ascertainment scheme can be modeled. As examples, both a random sample and a multiplex ascertainment scheme are considered. In addition, the pseudo-likelihood method leads to likelihood ratio tests for detecting linkage with a simple, known asymptotic reference distribution, a feature not shared by the mod score. Finally, we discuss the advantages of using the pseudo-likelihood method over the full likelihood method, both of which are valid when the ascertainment scheme is known.

© 1996 Wiley-Liss, Inc.

Received for publication April 16, 1996. Revision accepted July 17, 1996.

Address reprint requests to Kung-Yee Liang, Department of Biostatistics, Johns Hopkins University, School of Hygiene and Public Health, 615 North Wolfe Street, Baltimore, MD 21205.

© 1996 Wiley-Liss, Inc.
INTRODUCTION

The lod score method has been and remains a popular approach for detecting linkage and estimating the recombination fraction between a marker locus and a trait locus. The term “lod score” stands for the ratio of the likelihoods in the logarithm scale of base 10, evaluated at some recombination fraction value $\theta \leq 0.5$ vs. that when $\theta = 0.5$, the latter indicating no linkage between the locus for trait ($T$) and that for some marker ($M$). In this paper, $T$ and $M$ represent, respectively, the phenotype, i.e., trait, and the marker genotype of the subject. Implicit in the calculation of lod scores is that trait parameters, $\phi$, which include, e.g., penetrance and gene frequency at the trait locus, are known without error. However, these parameters are often not known or are known only partially. Elston [1989; see also Greenberg, 1989] makes a distinction in parameters describing the “mode of inheritance,” categorizing them as either “genetic mechanism parameters” or “penetrance parameters.”

The number of loci involved, e.g., is a genetic mechanism parameter, while the mode of inheritance (dominant or recessive) or age specific penetrance function are penetrance parameters, since they describe the relationship between the observed trait and its genotype. In this paper, we will be concentrating on unknown penetrance parameters, $\phi$, and assume that all genetic mechanism parameters are known, except for $\theta$.

Recently, the so-called “mod score” method has received a good deal of attention in the literature [e.g., Risch, 1984; Clerget-Darpoux et al., 1986; Greenberg, 1989]. This method amounts to the maximization of lod scores with respect to both $\theta$ and $\phi$. Thus it achieves the dual goals of 1) detecting linkage and 2) determining the appropriate genetic model. It is important to point out that this method is different, conceptually, from the conventional maximum likelihood method in that it is the likelihood ratio rather than the likelihood itself that is to be maximized. Recently, Hodge and Elston [1994] examined the validity of and provided guidelines for the use of the mod score method. They also discussed the comparison of mod scores, lod scores, and the so-called “wrod scores,” which are lod scores where $\phi$ may be incorrectly specified. The purposes of this paper are twofold. First, we will reinforce and clarify some of the main points made by Hodge and Elston [1994]. Specifically, they note that we use the mod score at the price of a loss of some information. In what follows, we elucidate the nature of that information loss. Second, we propose as an alternative a pseudo-likelihood method for linkage detection and determination of the genetic model as characterized by $\phi$.

In a modern linkage analysis setting, we may effectively have many values of $\theta$, since we are often performing an analysis involving many markers. Furthermore, unlike many inferential problems, our desire for precision in estimation of $\theta$ is highly related to the value of $\theta$ itself. We are inclined to require more accuracy for smaller $\theta$, but are willing to tolerate larger margins of error when $\theta$ is close to 0.5. On the other hand, our requirements for precise inferences on trait parameters $\phi$ do not vary with the value of $\theta$. This unusual feature of the inferential problem with respect to $\theta$ in linkage analysis plays a role in the methods we ultimately choose to use.
DEFINITIONS AND SOME BASIC PROPERTIES

Throughout, we will follow the notations used in Hodge and Elston [1994] as closely as possible. As such, let $L(\theta, \phi)$ denote the likelihood function indexed by $\theta$ and $\phi$. Here, as noted earlier, $\theta$ represents the recombination fraction between a trait locus, and a marker locus and $\phi$ represents unknown penetrance parameters for the trait locus. For simplicity, we assume $\theta$ is a scalar. For the moment, then, we are focusing on a simple two-point analysis involving one marker locus and one trait locus; however, we consider the generalization of this to the multiple marker situation at various points in the paper. Furthermore, we note that $\theta \in [0, 1/2] = \Theta$. One important implication, from the statistical point of view, is that the $\theta$ value of 1/2, which represents no linkage between the marker and the trait loci, is on the boundary of the parameter space $\Theta$.

Lods and Wrods

The lod score, considered a function of $\theta$, is defined as

$$Z(\theta) = \log_{10} \frac{L(\theta, \phi_0)}{L(1/2, \phi_0)}, \quad \theta \in \Theta$$

where $\phi_0$ is the true $\phi$ value. Likewise, the wrod score is defined as

$$W(\theta) = \log_{10} \frac{L(\theta, \phi^*)}{L(1/2, \phi^*)}, \quad \theta \in \Theta$$

where $\phi^*$ is an arbitrary $\phi$ value, which may be incorrect, i.e., $\phi^* \neq \phi_0$. While many linkage packages (e.g., LINKAGE, HOMOG) report lod scores for some selected $\theta$ values, it is a common practice that one draws inference for testing $H_0: \theta = 1/2$ vs. $H_1: \theta < 1/2$ based on a single number. To this end, one can define

$$Z_{\text{max}} = \max_{\theta \in \Theta} Z(\theta)$$
$$W_{\text{max}} = \max_{\theta \in \Theta} W(\theta)$$

and $\hat{\theta}_Z$ and $\hat{\theta}_W$, the $\theta$ values which maximize $Z(\theta)$ and $W(\theta)$, respectively. Equivalently, because the denominator terms in $Z(\theta)$ and $W(\theta)$ do not contain $\theta$, $\hat{\theta}_Z$ and $\hat{\theta}_W$ also maximize $L(\theta, \phi_0)$ and $L(\theta, \phi^*)$. Consequently, we note the close relationship between the conventional likelihood ratio test statistics for testing $H_0$ and Eq. (3) as

$$LR_Z = 2 \log_{10} \frac{L(\hat{\theta}_Z, \phi_0)}{L(1/2, \phi_0)} = \kappa Z_{\text{max}}$$
$$LR_W = 2 \log_{10} \frac{L(\hat{\theta}_W, \phi^*)}{L(1/2, \phi^*)} = \kappa W_{\text{max}}$$

where $\kappa = 2 \log_{10}(10)$. It has been pointed out that asymptotically, $LR_W$ has the same distribution under $H_0$ as $LR_Z$ even if the model is misspecified, i.e., $\phi^* \neq \phi_0$.
[Williamson and Amos. 1990]. Here "asymptotic" refers to the situation where the number of informative meioses increases without bound [Hodge and Elston, 1994]. The common asymptotic distribution is, however, a 50:50 mixture of $\chi^2$ and zero, rather than $\chi^2$, due to the boundary value of $\theta = 1/2$ in $\Theta$ [Self and Liang, 1987]. The implication is that the critical value for either $LR_2$ or $LR_4$ with, e.g., a type I error of 0.05 is 2.71, the 90th percentile of $\chi^2$. MacLean et al. [1993] show that this mixture distribution appears to hold even under fairly small samples, such as 25 informative meioses. As these authors point out, the effect of misspecifying $\phi$ is a loss of power to detect linkage when it is present. Finally, as noted in Hodge and Elston [1994], valid inferences about $\theta$, when $\phi$ is misspecified, can only be drawn when $\theta = 1/2$, so that the word score is only useful for detecting linkage, and not for estimating $\theta$ when linkage is present.

**Mods: Is This a Good Idea?**

Instead of treating trait parameters $\phi$ as known, the mod score method searches for estimates for $\theta$ and $\phi$ by maximizing lod scores, $\log_{10}[L(\theta, \phi)/L(1/2, \phi)]$, over both $\theta$ and $\phi$ simultaneously. That is, one finds $\hat{\theta}_M$ and $\hat{\phi}_M$ such that

$$M_{\text{max}} = \max_{\theta, \phi} \log_{10} \frac{L(\theta, \phi)}{L(1/2, \phi)} = \log_{10} \frac{L(\hat{\theta}_M, \hat{\phi}_M)}{L(1/2, \hat{\phi}_M)}.$$  \hspace{1cm} (5)

As Hodge and Elston [1994] have indicated, generally speaking, it is wrong to maximize the likelihood ratio for the purpose of deriving estimates for two sets of parameters, $\theta$ and $\phi$ in this case. However, when applying the mod score method in the context of linkage analysis, this turns out to be a valid procedure. This is due to the fact that 1) the likelihood function can be expressed as

$$L(\theta, \phi) \geq L(M \mid T; \theta, \phi) L(T; \phi)$$

in which the probability (density) function for the trait locus depends solely on $\phi$, and 2) $L(M \mid T; 1/2, \phi) = L(M \mid T; 1/2)$, i.e., the conditional distribution of $M$ given $T$ contains no information about trait parameters $\phi$ under the null hypothesis of no linkage. Consequently, maximizing lod scores over $\theta$ and $\phi$ is equivalent to maximizing a genuine conditional likelihood, namely, $L(M \mid T; \theta, \phi)$ [e.g., Elston, 1989; Clerget-Darpoux et al., 1992; Clerget-Darpoux and Bonaiti-Pellié, 1992], i.e.,

$$\log_{10} \frac{L(\theta, \phi)}{L(1/2, \phi)} = \log_{10} \frac{L(M \mid T; \theta, \phi)}{L(M \mid T; 1/2)} \geq \log_{10} L(M \mid T; \theta, \phi).$$  \hspace{1cm} (6)

Thus, some, although not all, of the standard properties possessed by likelihood methods may be carried over to the mod score method in linkage analysis. There are, however, some undesirable properties of the mod score method which are the focus of the next section.

It is worth reiterating one of the main strengths in using the mod score method as stated in Hodge and Elston [1994]. Assuming the sample was ascertained through a part of the trait data, $T$, inferences based on the likelihood conditional on all trait
data, i.e., based on \( L(M \mid T; \theta, \phi) \), would lead to valid inferences for \( \phi \) and hence \( \theta \) as well. This method is especially useful when the mode of ascertainment is either unclear or too complex to be modeled properly. On the other hand, it comes at a price of potentially losing information for \( \phi \), which can in turn affect inferences on \( \theta \) as well. In the next section we investigate the nature and consequences of this information loss.

**SOME STATISTICAL ISSUES CONCERNING MODS**

In this section we discuss in more detail some statistical issues faced if one adopts the mod score method. For illustrative purposes, we consider a simple situation in which the trait is governed by a single autosomal dominant locus with two alleles, \( D \) and \( d \), and the markers have two codominant alleles, \( A \) and \( B \). Table I gives the observed frequencies and theoretical joint probabilities for each of four possible values of \((T, M)\) for offsprings from the phase-known double-backcross mating, \( DA/dB \times dB/dB \): see also Tables II and III of Hodge and Elston [1994]. Our model assumes a random sample of offspring from phase-known parents in the double-backcross matings. Note that, for a given offspring, \( \Pr(M = AB \mid T; \theta, \phi) \) in this case is equal to \( 1 - \theta \) if the offspring is affected and equal to 

\[
\frac{1 - (1 - \theta)\phi}{2 - \phi}
\]

if unaffected. Here \( \phi \) reflects the incomplete penetrance for the trait. Note that both conditional probabilities are indeed independent of \( \phi \) when \( \theta = 0.5 \). In Table I, \( n \) is the total number of informative meioses. We will examine the issues both asymptotically, i.e., as \( n \) increases without bound, and in finite samples based on two hypothetical example data sets: see Table II. The major difference between these two examples is their estimated \( \theta \) values: the first example is characterized by a high \( \theta \) value (i.e., loose linkage), whereas the second example is more tightly linked. Both examples could be characterized by a penetrance of 0.4 (i.e., 20% offspring are affected, whereas the probability of inheriting a \( D \) allele is 50%).

Hodge and Elston [1994], Clerget-Darpoux et al. [1986], and Elston [1989] indicate that the mod score method is a valid inference procedure. They conclude this by

**TABLE I. Observed Frequencies (Upper Entry) and Probability Distribution (Lower Entry) of Offspring Phenotypes From the Phase-Known Double-Backcross Mating \( DA/dB \times dB/dB \)**

<table>
<thead>
<tr>
<th>Marker</th>
<th>Trait</th>
<th>AB</th>
<th>BB</th>
<th>Marginal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Affected</td>
<td>( n_{11} )</td>
<td>( n_{12} )</td>
<td>( n_{1+} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \theta \phi )</td>
<td>( \theta \phi )</td>
<td>( \theta \phi )</td>
</tr>
<tr>
<td></td>
<td>Unaffected</td>
<td>( n_{21} )</td>
<td>( n_{22} )</td>
<td>( n_{2+} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( 1 - \theta \phi )</td>
<td>( 1 - \theta \phi )</td>
<td>( 1 - \theta \phi )</td>
</tr>
<tr>
<td></td>
<td>Marginal</td>
<td>( n_{+1} )</td>
<td>( n_{+2} )</td>
<td>( n )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \frac{1}{2} )</td>
<td>( \frac{1}{2} )</td>
<td>1</td>
</tr>
</tbody>
</table>
TABLE II. Two Hypothetical Examples

<table>
<thead>
<tr>
<th>Trait</th>
<th>Example 1 Marker</th>
<th>Example 2 Marker</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AB</td>
<td>BB</td>
</tr>
<tr>
<td>AFFECTED</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>UNAFFECTED</td>
<td>36</td>
<td>44</td>
</tr>
<tr>
<td>MARGINAL</td>
<td>48</td>
<td>52</td>
</tr>
</tbody>
</table>

Demonstrating that it is a proper conditional likelihood. There are two issues that are alluded to in that work, for which we would like to explore further the implications for inference. First, as pointed out earlier, the null value of the parameter \( \theta \) exists on the boundary of the parameter space, \( \Theta \). Thus, although mod scores give rise to valid LR statistics, their distribution under no linkage is not necessarily \( \chi^2 \). Second, and perhaps more importantly, at the value of \( \theta = 0.5 \), the conditional distribution implicit in the mod function contains no information on any genetic parameters.

**Estimation of \( \phi \)**

Hodge and Elston [1994] correctly point out that inference for \( \phi \) based on the conditional likelihood \( L(M \mid T; \theta, \phi) \), i.e., based on the mod score method, could be very inefficient, especially when \( \theta \) is close to 0.5. This is because most of the information about \( \phi \) is likely to be carried by \( L(T; \phi) \), and the information contained in \( L(M \mid T; \theta, \phi) \) diminishes as \( \theta \) approaches 0.5, at which point no knowledge about \( \phi \) is conveyed by the conditional likelihood.

The ratio of the asymptotic variance of \( \hat{\phi} \), the maximum likelihood estimate of \( \phi \) based on the full likelihood \( L(\theta, \phi) \), to that of \( \hat{\phi}_M \) defined in Eq. (5) provides an index of relative efficiency of the mod score compared to the full likelihood approach. A ratio of, e.g., 0.2 would imply that the variance of \( \hat{\phi}_M \) is five times as high as that of \( \hat{\phi} \). Equivalently, we could say that the efficiency for estimating \( \phi \) or the information on \( \phi \), using the mod score relative to the full likelihood is 20%. Tidious, but straightforward calculations show that the relative efficiency for \( \phi \) is given by

\[
\frac{\text{Var}(\hat{\phi})}{\text{Var}(\hat{\phi}_M)} = \frac{\phi(2\theta - 1)^2}{2 - \phi}
\]

This is plotted as a function of \( \phi \) for several values of \( \theta \) in Figure 1. The plots suggest that \( \hat{\phi}_M \) is less efficient relative to the "gold standard." Specifically, the efficiency decreases as \( \theta \) approaches 0.5, and the penetrance decreases. Indeed, the ratio is equal to zero when \( \theta = 0.5 \), reflecting the lack of information on \( \phi \) carried by \( L(M \mid T) \) when \( \theta = 0.5 \). On the other hand, when \( \theta \) is small and the penetrance is fairly high, much of the information about \( \phi \) is retained. Another concern is that information loss for \( \phi \) is dependent on the true value of \( \phi \), which may well be unknown.

The above observations are consistent with results from the two hypothetical examples. In example 1 where \( \theta \) is estimated at 0.4, the estimated variance (0.694)^2 of \( \phi_M \) is approximately 75 times as large as that of \( \phi \) ((0.08)^2); see Table III. The
Fig 1. Asymptotic relative efficiency comparing the mod score method to the full likelihood method for estimation of the penetrance, \( \phi \), as a function of \( \phi \) and the recombination fraction, \( \theta \).

Discrepancy in estimated variance is less dramatic (i.e., \((0.223)^2 \) vs. \((0.071)^2 \)) in example 2, where \( \theta \) is estimated at about 0.1.

**Estimation of \( \theta \)**

Contrary to the claim that \( \hat{\theta}_M \), the estimate of \( \theta \) which maximizes the conditional likelihood \( L(M \mid T; \theta, \phi) \), is fully efficient [Hodge and Elston, 1994], this estimate
TABLE III. Estimates of $\theta$ and $\phi$ and Their Estimated Standard Errors From the Two Hypothetical Examples

<table>
<thead>
<tr>
<th>Example 1</th>
<th>Estimate</th>
<th>Standard error</th>
<th>Correlation between $\hat{\theta}$ and $\hat{\phi}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mod score</td>
<td>$\bar{\theta}_m$</td>
<td>0.400</td>
<td>0.109</td>
</tr>
<tr>
<td></td>
<td>$\bar{\phi}_m$</td>
<td>0.665</td>
<td>0.694</td>
</tr>
<tr>
<td></td>
<td>$\hat{\theta}$</td>
<td>0.380</td>
<td>0.097</td>
</tr>
<tr>
<td></td>
<td>$\hat{\phi}$</td>
<td>0.405</td>
<td>0.080</td>
</tr>
<tr>
<td>Pseudo-likelihood</td>
<td>$\hat{\theta}_p$</td>
<td>0.380</td>
<td>0.097</td>
</tr>
<tr>
<td></td>
<td>$\hat{\phi}_p$</td>
<td>0.400</td>
<td>0.080</td>
</tr>
<tr>
<td>Example 2</td>
<td>$\hat{\theta}_M$</td>
<td>0.100</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td>$\hat{\phi}_M$</td>
<td>0.220</td>
<td>0.223</td>
</tr>
<tr>
<td></td>
<td>$\hat{\theta}$</td>
<td>0.115</td>
<td>0.073</td>
</tr>
<tr>
<td></td>
<td>$\hat{\phi}$</td>
<td>0.375</td>
<td>0.071</td>
</tr>
<tr>
<td>Pseudo-likelihood</td>
<td>$\hat{\theta}_p$</td>
<td>0.120</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>$\hat{\phi}_p$</td>
<td>0.400</td>
<td>0.080</td>
</tr>
</tbody>
</table>

may be somewhat inefficient relative to $\hat{\theta}$, the maximum likelihood estimate based on $L(\theta, \phi)$. This is mainly because $\hat{\theta}_M$ is potentially highly correlated with $\hat{\phi}_M$, which is very inefficient relative to $\hat{\theta}$. Again, the ratio of the asymptotic variance of $\hat{\theta}$ vs. that of $\hat{\theta}_M$ provides a measure of efficiency. This ratio, plotted in Figure 2 as a function of $\theta$ for various values of $\phi$, is

\[
\frac{\text{Var}(\hat{\theta})}{\text{Var}(\hat{\theta}_M)} = 2\phi\theta^2 - 2\phi \theta + 1.
\]

Judging from the magnitude of the ratio, the degree of efficiency loss for $\theta$ estimation by using the mod score method is less severe compared to the $\phi$ estimation; see Figure 1. As expected, the loss of efficiency increases as $\theta$ approaches 0.5 and the penetrance decreases; when $\theta$ is small, very little information on $\theta$ is lost. The highest information loss is incurred for the case of complete penetrance, when at least 75% of the information on $\theta$ is retained for $\theta < 0.14$. Even when $\theta$ is close to 0.5, the information in the mod score is always at least half of that in the full likelihood. Taking into account that our interest in precise estimates for $\theta$ is greatest for small $\theta$, using the mod score for estimation of $\theta$ appears to be a practical alternative to a full likelihood approach.

Returning to the two hypothetical examples shown in Table II, we find comparable efficiency for estimating $\theta$ for the mod and the full likelihood. The variance estimates for the mod and the full likelihood approaches are $(0.109)^2$ and $(0.097)^2$ for example 1. In example 2, they are $(0.067)^2$ and $(0.073)^2$; see Table III. The variance estimate under the full likelihood approach is slightly larger because it is calculated under different parameter estimates than that for the mod score approach. In the first example, we note that the estimated correlation between $\hat{\theta}_M$ and $\hat{\phi}_M$ is high (0.70), whereas in the second example, it is only 0.15.
Fig. 2. Asymptotic relative efficiency comparing the mod score method to the full likelihood method for estimation of the recombination fraction, $\theta$, as a function of $\theta$ and the penetrance, $\phi$.

It is important to note that the above comparison between mod scores and the full likelihood method is meaningful only if the mode of ascertainment is known to the investigators. If one does not know the mode of ascertainment, then the full likelihood method would lead to inconsistent estimation for both $\theta$ and $\phi$ in which case the comparison becomes less meaningful [e.g., Ewens and Shute, 1986; Hodge, 1988].
Hypothesis Testing of $H_0 : \theta = 0.5$

To detect linkage, the likelihood ratio test using mod scores has the form

$$LR_M = 2 \log e \frac{L(M \mid T; \hat{\theta}_M, \hat{\delta}_M)}{L(M \mid T; 1/2)} = \kappa M_{\text{max}}.$$  (7)

See Eqs. (5) and (6). However, $LR_M$ does not converge under the null hypothesis of no linkage to an asymptotic $\chi^2$ distribution as stated in Hodge and Elston [1994: p 338]. Indeed, it is not certain to what distribution $LR_M$ does converge, as indicated by MacClean et al. [1993]. The complication is mainly due to the fact that the conditional likelihood contains no information about $\phi$ when $\theta = 1/2$, a non-standard situation which prevents the conventional asymptotic results from being applied [e.g., Davies, 1977]. This uncertainty further complicates the issue of multiple tests, as discussed in Hodge and Elston [1994]. Here, multiple tests refer to the notion that one maximizes lod scores over multiple models. We will return to this issue in the next section.

To examine the behavior of $LR_M$ under no linkage, we simulated 5,000 independent replicates of a multinomial distribution of size $n = 100$ whose probability distribution is governed by that in Table I with $(\theta, \phi)$ equal to (0.5, 0.5) and (0.5, 0.8), respectively. Figure 3 gives the empirical distribution functions for $LR_M$ with $\phi = 0.5$ and 0.8, respectively. It is evident from these plots that the distribution of $LR_M$ under $H_0$ is very peculiar, and is quite different from either $\chi^2_1$ or a 50:50 mixture of $\chi^2_1$ and $\chi^2_0$. This behavior, it turns out, is due to the facts that there is no information on $\phi$ at the null value of $\theta = 1/2$. It appears, however, that this distribution is well approximated by $\chi^2_1$ in the upper tail, at least for the specific situation considered in Table I.

ALTERNATIVE METHOD: PSEUDO-LIKELIHOOD

The three methods discussed so far (lods, wrods, and mods) represent completely different likelihood approaches. On the one end, both lod scores and wrod scores use, in principle, the full likelihood function assuming that trait parameters $\phi$ are completely known. For inferences on $\theta$, however, this is equivalent to the conditional likelihood due to the special form of the likelihood function. On the other end, the mod score method begins with the conditional likelihood and tries to learn about both $\theta$ and $\phi$ simultaneously using only this part of the likelihood function. Recognizing that there is a dichotomy as to whether there is linkage or not regarding the true state of nature, Hodge and Elston [1994] recommended that one should use wrod scores when $\theta = 0.5$ and use mod scores when $\theta < 0.5$. The pseudo-likelihood (PL) method discussed below serves as a compromise to the wrod and the mod score approaches in that it uses the marginal trait data to estimate $\phi$ rather than either assuming a known $\phi$ (wrod) or estimating $\phi$ simultaneously with $\theta$ using the marker data $M \mid T$ (mod). It provides a unified approach in that it avoids the dichotomy mentioned above, and addresses concerns raised earlier for the mod score method. Note that this compromise may be achieved if the mode of ascertainment is known.
Fig. 3. Simulated cumulative distribution functions (CDF) of likelihood ratio (LR) statistics under the hypothesis of no linkage. Simulation models are (top) \( \phi = 0.8 \) and (bottom) \( \phi = 0.5 \). Upper dotted line (-----): CDF of a random variable that is a 50:50 mixture of \( \chi^2 \) and zero. Lower dotted line (-----): CDF of \( \chi^2 \) random variable (-----). CDF of LR based on mod score (LR_M). (-----): CDF of LR based on pseudo-likelihood (LR_P). (-----): CDF of LR based on lod or wrod with \( \phi^* = 0.5 \) (LR_P or LR_L). (-----): CDF of LR based on lod or wrod with \( \phi^* = 0.8 \) (LR_L or LR_P). The distributions of the LR statistics based on the lod, wrod, and the pseudo-likelihood are very close as predicted by the asymptotic theory.
Background

The pseudo-likelihood method consists of a two-step procedure. Step one requires finding a well-behaved estimate of \( \phi \), called \( \hat{\phi}_P \). It is well behaved in the sense that \( \hat{\phi}_P \) is asymptotically normally distributed. In step two, one replaces \( \phi \) by \( \hat{\phi}_P \) in \( L(\theta, \phi) \) and makes inferences for \( \theta \) on

\[
L_P(\theta) = L(\theta, \hat{\phi}_P)
\]

the pseudo-likelihood for \( \theta \). Let \( \hat{\theta}_P \) be the estimate of \( \theta \) which maximizes the PL. For a more detailed discussion on the motivation, applications and the asymptotic behavior of the PL method, see Gong and Samaniego [1981] and Liang and Self [1996].

Estimation of \( \phi \)

A key question in using the PL method is how to find \( \hat{\phi}_P \). In the context of linkage analysis, one approach is to estimate \( \phi \) by performing a separate segregation analysis on a different data set prior to linkage analysis. Another approach is to estimate \( \phi \) based on the marginal likelihood \( L(T; \phi) \), effectively performing a segregation analysis on the observed trait data. For the \( 2 \times 2 \) example, this latter approach leads to \( \hat{\phi}_P = \min\{2(n_{11} + n_{21})/n, 1\} \). A commonly used ascertainment scheme (especially in psychiatric research) is to sample multiplex sibships with two or more affected individuals, where the sibship, not the individual, is the sampling unit. A simple modification then leads to \( \hat{\phi}_P = \min\{2(n_{11} + n_{21} - k)/(n - k), 1\} \), where \( k \) is the number of sibships ascertained times the required minimum number of affected individuals per sampled sibship. It is intuitively clear that multiplex ascertainment results in a less efficient estimate of \( \phi \) compared to random sampling of sibship of this mating type. In the \( 2 \times 2 \) example, the relative efficiency may be measured by \( \lambda = (n - k)/n \), the ratio of the number of informative individuals that are available for estimating \( \phi \), the penetrance, to the total number of offspring. Assuming an average sibship size of four offspring, a \( \lambda \) value of 0.5 means that each ascertained sibship contains at least two affected individuals, compared to simply a random sample with, on average, four relatives per sibship. Note that estimation of \( \theta \) and \( \phi \) based on \( L(\theta, \phi) \) incorporating ascertainment would be considerably more complicated, even in this simple multiplex sampling scheme.

The PL method enjoys several properties not shared by the mod score method. First, \( \hat{\theta}_P \) and \( \hat{\phi}_P \) are asymptotically uncorrelated, whereas \( \hat{\theta}_M \) and \( \hat{\phi}_M \), as seen before, may be highly correlated. Second, while \( \hat{\phi}_P \) may not be fully efficient, the informativeness of \( \hat{\phi}_P \) about \( \phi \) is the same regardless of whether \( \theta = 0.5 \) or not. Recall that the relative information about \( \phi \) contained in \( L(M \mid T; \theta, \phi) \) decreases and approaches zero as \( \theta \) approaches 0.5, and for small penetrances, where multiplex sampling schemes are most commonly used.

Figure 4 provides plots of \( \text{var}(\hat{\phi}_P)/\text{var}(\hat{\phi}_M) \) vs. penetrance (\( \phi \)) for some selected \( \theta \) and \( \lambda \) values. In many cases, the estimate of \( \phi \) based on \( L(T; \phi) \), \( \hat{\phi}_P \), is
more efficient relative to \( \hat{\delta}_M \). Exceptions occur with large penetrance (\( \geq 0.6 \)), smaller \( \lambda \) (e.g., 0.4), and small \( \theta \) (e.g., \( \leq 0.1 \)). The efficiency patterns are similar to those shown in Figure 1. Specifically, the ratio of the variance decreases as \( \theta \) increases and \( \phi \) decreases, the pattern being the same for different \( \lambda \) values. From these plots, it can be seen that, especially in cases of significantly reduced penetrance, weak linkage, or high \( \lambda \) values, the pseudo-likelihood outperforms the mod, perhaps by a factor of four or more. In cases of low \( \lambda \) values, it is intuitive that the information in the marginal likelihood is reduced, so the mod may be more efficient. When linkage is tight and penetrance is near complete, the mod outperforms the pseudo-likelihood primarily because the data in the unaffected subjects are informative.

**Estimation of \( \theta \)**

Generally speaking, \( \hat{\theta}_P \), the maximum likelihood estimate based on the pseudo-likelihood, is often more efficient than \( \hat{\theta}_M \), the estimate based on the mod score method. This is evident in Figure 5, which shows the plots of \( \text{var}(\hat{\theta}_P)/\text{var}(\hat{\theta}_M) \) vs. \( \theta \) for selected \( \phi \) and \( \lambda \) values. Exceptions occur again only in the special cases of very high penetrance and tight linkage; here \( \hat{\phi}_M \) outperforms \( \hat{\phi}_P \). For inferences on \( \theta \), the pseudo-likelihood is almost always as efficient as the mod score, and when it is not, the information loss is generally less than 10%. Thus, for estimation of \( \theta \), the pseudo-likelihood method is comparable to or better than the mod score method across wide range values of \( \theta \) and \( \phi \). Table III gives \( \hat{\theta}_P \), \( \hat{\phi}_P \), and their standard error estimates for hypothetical examples 1 and 2. In both examples, the results are very similar to those obtained from the full likelihood approach, i.e., \( \hat{\theta} \) and \( \hat{\phi} \).

Again, the comparison between the PL approach and mod scores is meaningful if the mode of ascertainment is known, for the same reasons discussed earlier.

**Hypothesis Testing of \( H_0 : \theta = 0.5 \)**

To test the hypothesis of no linkage, one can use the pseudo-likelihood and calculate

\[
LR_P = 2 \log \frac{L_P(\hat{\theta}_P)}{L_P(0.5)} = 2 \log \frac{L(M \mid T; \hat{\theta}_P, \hat{\phi}_P)}{L(M \mid T; 0.5)}.
\]

(8)

Note the similarity between Eqs. (8) and (7). However, unlike the mod score method, one can make the following claim for \( LR_P \) [Liang and Self, 1996].

**Proposition.** Assuming that \( \hat{\phi}_P \) converges asymptotically to \( \bar{\phi} \), a constant which may be different than the true \( \phi \) value, \( \phi_0 \), then \( LR_P \) converges, under the null hypothesis, asymptotically to a distribution which is a 50:50 mixture of \( \chi^2_1 \) and zero.

The result suggests that the likelihood ratio test statistics based on lod scores, wrod scores, or the pseudo-likelihood all have the same null distribution asymptotically. This is not surprising as conceptually, the first two approaches may be viewed...
as special cases of the pseudo-likelihood. Recall that \( \phi \) is either correctly specified (lod scores) or incorrectly specified (wrod score). This knowledge of \( \phi \) is presumably obtained a priori, i.e., before linkage analysis was conducted. One such scenario is that a segregation analysis was carried out prior to linkage analysis, either on the same data set or on a different one. In other words, either \( \phi_0 \) or \( \phi^* \) may be considered \( \hat{\phi}_P \) with very little uncertainty due to the presumed fact that the sample size in segregation analysis is much larger than that in the subsequent linkage analysis. The PL method is flexible in that one is allowed to estimate \( \phi \) externally through segregation analysis or internally through maximization of \( L(T; \phi) \). It is also conceivable to combine external information with that in \( L(T; \phi) \) to obtain an estimate of \( \phi \). The key element is that \( \hat{\phi}_P \) be statistically independent of \( L(M \mid T; \theta, \phi) \), the latter being the kernel for the four likelihood ratio test statistics (lods, wrods, mods, and pseudo-likelihood). This observation also explains how the concern of multiple testing raised earlier for the mod score method does not arise here, since the maximization over genetic models is done statistically independently of the LR procedure which is based on \( L(M \mid T) \).

Returning to the simulation study described earlier, we note as shown in Figure 3 that the distribution of the likelihood ratio test statistic based on \( L(\theta, \hat{\phi}_P) \) (the pseudo-likelihood) is in excellent agreement with the 50:50 mixture of \( \chi^2 \) and zero. Similarly, this mixture of \( \chi^2 \) random variables is also the asymptotic distribution of \( L(\theta, \phi_0) \), the lod score, and \( L(\theta, \phi^*) \), the wrod score.

The key assumption in the proposition that \( \hat{\phi}_P \) converges to some \( \bar{\phi} \), but not necessarily to the true \( \phi \) value deserves more detailed discussion. One situation in which \( \bar{\phi} \) might not be equal to \( \phi_0 \) is when one mistakenly assumes a random sample while indeed, the sample was ascertained through a multiplex scheme. In the \( 2 \times 2 \) example where \( \phi \) is the penetrance, \( \hat{\phi}_P = 2 \min\{n_1, n_2, n_1 + n_2\} \) would converge to \( \phi = (1 - c)\phi_0 + c \), instead of to \( \phi_0 \). Here, \( c \) is the limiting value of \( k/n \), the proportion of the total sample size who are affected by requirement of the ascertainment scheme. Another situation where \( \bar{\phi} \neq \phi_0 \) could occur is when the genetic mechanism was incorrectly specified. For example, the true disease mechanism might involve two loci, but a single-locus model was mistakenly assumed. In this example, however, \( \hat{\phi}_P \) will converge to something that involves the true but unspecified parameters, and it is difficult to work out analytically what \( \bar{\phi} \) is.

It is worth emphasizing that the genetic mechanism is allowed to be misspecified only if interest is on detecting linkage, i.e., testing the hypothesis \( \theta = 0.5 \). As pointed out by Hodge and Elston [1994], one needs to assume that the genetic mechanism is the true one, in order to make valid inference for \( \theta \leq 0.5 \) and especially for \( \phi \).

**DISCUSSION**

In this paper we examined the use of the mod score method for linkage analysis. As suggested by Elston [1989], we have attempted to carry out our study using analytic (vs. empirical) methods to the extent possible. There are two main advantages of the mod score approach. First, it is able to both detect linkage and estimate appropriate genetic models. Second, it is an ascertainment "assumption-free" method [Ewens and Shute, 1986; Hodge, 1988]. Ewens and Shute [1986] and Hodge [1988] discuss this method in general terms of conditioning on the part of the data that
contain information on ascertainment. In the case of the mod score, no assumption
of ascertainment model is needed to make valid inferences about linkage (θ), and
to some extent on trait parameters such as penetrance (ϕ), because ascertainment is
through the affected status T on which we condition. Even in the worst case, for linkage
(θ) inferences, at most half of the information on ϕ is lost. On the other hand,
the use of this method is potentially hampered by the following undesirable prop-
eties. First, this method may result in non-trivial loss of efficiency, especially for
inference on ϕ. Under very weak or no linkage, inferences about ϕ based on the
mod may vary considerably from inferences which we may draw from the marginal
information, T. Also, the information loss depends on the true value of ϕ itself. Sec-
ond, the asymptotic distribution of M_{max} or LR_{θ} for detecting H_0 is unknown even
under H_0 : θ = 0.5. As a result, one cannot determine in a straightforward manner
the statistical significance for detecting linkage when using M_{max}. Third, one needs
to deal with the multiple test issue whose resolution is further complicated by the
second concern mentioned above. Fourth, since one never knows in advance the true
state of nature and noting that L(M | T; θ, ϕ) carries no information about ϕ when
θ = 0.5, execution of the mod score method must be a sequential one. That is, one
needs to test for linkage first and to maximize L(M | T; θ, ϕ) only if the hypothesis
of no linkage were rejected, and fairly tight linkage is supported by the data. This
procedure creates an additional complication in that one needs to determine a criteria
for rejection of H_0, a difficult task given that distribution of LR_{θ} cannot be specified.
Finally, if one is performing an analysis with multiple markers, it is possible that the
estimates of ϕ obtained by using the conditional likelihood for each marker could
be quite different, and it is not clear how one would combine this information across
the markers in order to make inferences on ϕ.

As an alternative that more thoroughly exploits the information available to us
in linkage studies, we propose the use of the pseudo-likelihood method. This method
maintains the dual objectives of detecting linkage and determining genetic models.
In many practical situations, this method provides more efficient estimates for θ and
ϕ than the mod score method. Furthermore, this method leads to a testing procedure for H_0 : θ = 0.5 with a well-defined asymptotic reference distribution under the
null hypothesis. The concern of performing multiple tests is also alleviated since the
maximization over genetic models can be carried out in a fashion which does not sta-
tistically affect the inference for detecting linkage. Furthermore, unlike the mod score
method, we can achieve the dual goals simultaneously rather than sequentially. Fi-
nally, with the PL method, in a multipoint analysis, ϕ needs be calculated only once,
and the same value will apply to all markers. Consequently, this method is readily
available by using the existing statistical software for segregation analysis to derive
ϕ, and standard linkage software to test and estimate linkage by plugging in ϕ.
For a summary and comparison of the mod score and PL methods, see Table IV.

In advocating use of the pseudo-likelihood, it must be understood that we are
placing some confidence in the marginal information contained in the phenotypes,
T. A reasonable question is then, why not use the full likelihood “gold standard”
as discussed by Hodge and Elston [1994]? The answer is that the PL approach off-
fers several advantages and few disadvantages over the full likelihood approach. The
pseudo-likelihood is almost as efficient as the full likelihood, since the marginal data,
T, contains most of the information on ϕ and the conditional data, M | T, contains
TABLE IV. Summary and Comparison of the Mod Score and the Pseudo-Likelihood (PL) Methods

<table>
<thead>
<tr>
<th></th>
<th>Mod</th>
<th>PL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ascertainment model</td>
<td>None required</td>
<td>Must specify</td>
</tr>
<tr>
<td>Distribution under $H_0$</td>
<td>Unknown</td>
<td>$\frac{1}{2} \hat{d} + \frac{1}{2} X_1^2$</td>
</tr>
<tr>
<td>$\phi$ efficiency</td>
<td>Fair*</td>
<td>Good</td>
</tr>
<tr>
<td>$\theta$ efficiency: $\theta \sim \frac{1}{2}$</td>
<td>Good</td>
<td>Better</td>
</tr>
<tr>
<td>$\theta$ efficiency: $\theta \sim 0$</td>
<td>Very good</td>
<td>Very good</td>
</tr>
<tr>
<td>Multiple markers</td>
<td>New $\phi$ estimates</td>
<td>Same $\phi$ estimates</td>
</tr>
<tr>
<td>Software</td>
<td>?</td>
<td>Existing</td>
</tr>
</tbody>
</table>

*Efficiency depends on $\theta$ and $\phi$.

most of the information on $\theta$. The pseudo-likelihood offers the following two features not available in the full likelihood approach. First, in the PL approach, it is straightforward to incorporate data from segregation studies with the linkage data to obtain an estimate of $\phi$. Second, the pseudo-likelihood has the correct type I error rate, regardless of how poorly $\phi$ is estimated. Therefore, just as with the wrod score method, we are protected against an excess of false positives even if our ascertainment model is incorrect, but we maintain close to full efficiency when our ascertainment model is correct. Also, in the analysis of multiple markers, the full likelihood approach would require new $\phi$ estimates for each marker, whereas these remain the same in the PL approach. Finally, as a practical matter, the PL approach allows us to use segregation analysis software to estimate $\phi$, and, plugging in the estimate, linkage analysis software to estimate $\theta$.

In the situation where the mode of ascertainment is clearly understood, we recommend that the pseudo-likelihood be used instead of the mod score method. While the latter method may be the choice when one does not know precisely how the data were ascertained, the potential gain in efficiency in using the PL approach with a reasonably accurate ascertainment model—if this is possible—highlights the need to avoid situations of unknown method of ascertainment. Indeed, Greenberg [1986] has shown that wrongly assumed mode of ascertainment can have major effects on one's conclusions drawn from segregation analysis. We suggest two possible strategies. One such strategy is to perform linkage studies which are well planned in the design stage to make information in the marginal trait status data available for inferences on trait parameters. For example, MacClean et al. [1993] refer to a “high-density” sampling scheme, in which families are sampled such that, on average, half of the sampled subjects are affected. Depending on how the sample is drawn, information contained in the ascertainment of the sample used for linkage is relevant to trait parameters, so that well-defined sampling protocols could produce data that could be used for estimation of $\phi$ with $\hat{\phi}$. A second strategy, which may be necessary in studies of complex or very rare diseases, is to plan segregation studies that will provide efficient estimates of trait parameters, and to use these estimates via the pseudo-likelihood to test for linkage. In either of these approaches, it will ultimately be necessary to confront the ascertainment issue in order to be able to make correct and efficient estimates for genetic parameters.
ACKNOWLEDGMENTS

This work was supported in part by grant GM49909 from the National Institutes of Health. We thank Ms. Yen-Feng Chiu for her computing help in producing the figures.

REFERENCES


