Case-Deletion Diagnostics for Maximum Likelihood Multipoint Quantitative Trait Locus Linkage Analysis

Maria C.B. Mendoza^{a, c} Trudy L. Burns^b Michael P. Jones^a

Departments of ^aBiostatistics and ^bEpidemiology, The University of Iowa, Iowa City, Iowa, ^cFDA, National Center for Toxicological Research, Division of Personalized Nutrition and Medicine, Jefferson, Ark., USA

Key Words

Case-deletion diagnostics • Empirical influence function • Influential observation • Outlier

Abstract

Objectives: Case-deletion diagnostic methods are tools that allow identification of influential observations that may affect parameter estimates and model fitting conclusions. The goal of this paper was to develop two case-deletion diagnostics, the exact case deletion (ECD) and the empirical influence function (EIF), for detecting outliers that can affect results of sib-pair maximum likelihood quantitative trait locus (QTL) linkage analysis. **Methods:** Subroutines to compute the ECD and EIF were incorporated into the maximum likelihood QTL variance estimation components of the linkage analysis program MAPMAKER/SIBS. Performance of the diagnostics was compared in simulation studies that evaluated the proportion of outliers correctly identified (sensitivity), and the proportion of non-outliers correctly identified (specificity). **Results:** Simulations involving nuclear family data sets with one outlier showed EIF sensitivities approximated ECD sensitivities well for outlier-affected parameters. Sensitivities were high, indicating the outlier was identified a high proportion of the time. Simulations also showed the

enormous computational time advantage of the EIF. Diagnostics applied to body mass index in nuclear families detected observations influential on the lod score and model parameter estimates. **Conclusions:** The EIF is a practical diagnostic tool that has the advantages of high sensitivity and quick computation.

Copyright © 2009 S. Karger AG, Basel

Introduction

Case-deletion diagnostics are well developed in the context of least squares regression models and have widespread use [1–3]. However, there has been little application of case-deletion diagnostics to the analysis of family-based data in the context of maximum likelihood estimation. Wang et al. [4] developed the exact case-deletion, the empirical influence function, the one-step approximation, and the approximate one-step for Class A Regressive Models in segregation analysis. de Andrade et al. [5] discussed diagnostic methods that were applied to quantitative trait polygenic and major gene models in the context of variance components models. They looked for individual outliers that caused departures from normality, and family outliers that may have caused lack of nor-

malities or influenced linkage results. Until now, case-deletion diagnostics have not been developed for linkage analysis. The purpose of this paper is twofold: the first is to develop two case-deletion diagnostic tools, the exact case-deletion (ECD) and the empirical influence function (EIF), for outlier detection in sib-pair quantitative trait locus (QTL) linkage analysis, and the second is to implement these two diagnostic methods in the maximum likelihood QTL variance estimation components of the linkage analysis program MAPMAKER/SIBS, to show how they can be practically applied in the detection of outliers for further investigation.

For sib-pair data, if a QTL exists and is linked to a marker locus, a negative association is expected between the sib-pair absolute trait difference and the proportion of alleles the sib pair shares identical by descent (ibd) at the marker locus. So if there is linkage, the greater the proportion of alleles shared ibd the smaller the expected sib-pair trait difference, and the smaller the proportion of alleles shared ibd the greater the expected sib-pair trait difference. Because of this relationship between ibd allele sharing and the trait difference, the variances of the phenotype differences among the classes of ibd allele sharing (0, 1, or 2) exhibit the constrained relationship $\sigma_0^2 \geq \sigma_1^2 \geq \sigma_2^2$. The maximum likelihood methods in MAPMAKER/SIBS estimate these three variance parameters subject to this constraint, and use them to compute the lod score. If one assumes no dominance variance, $\sigma_1^2 = (\sigma_0^2 + \sigma_2^2)/2$.

Outliers, influential observations or observations that are discrepant on the dependent variable (the phenotype difference), can have a disproportionately large effect on the variance parameter estimates, even to the extent of reversing the negative association between the trait difference and the proportion of alleles shared ibd expected with linkage. For example, an outlier sib pair with a large trait difference and 0 alleles ibd can strengthen the negative association. An outlier sib pair with a large trait difference and 2 alleles ibd can weaken the negative association. In this paper, it is these types of outliers that the ECD and EIF were developed to detect. In this context, outliers can be detected by examining the changes in the variance parameter estimates themselves or the changes in the lod score, due to deleting the outlier. By analyzing the observed distribution of these changes, critical values can be established so that if exceeded, influential observations can be flagged for further investigation.

Methods

Kruglyak and Lander [6] developed QTL maximum likelihood methods for sib-pair linkage analysis and incorporated them into the MAPMAKER/SIBS program. Linkage analysis can be conducted using two models which define the relationship of the sib-pair trait difference and the variance of the difference when sib pair j shares k alleles ibd. The first model is a more general model, in which the three variance parameters, σ_0^2 , σ_1^2 , and σ_2^2 are estimated subject to the constraint $\sigma_0^2 \geq \sigma_1^2 \geq \sigma_2^2$. In this paper, this model is referred to as the 3-parameter model. The second model assumes no dominance variance, that is there is no variance due to dominant effects of alleles. In this model, $\sigma_0^2 \geq \sigma_2^2$ and $\sigma_1^2 = (\sigma_0^2 + \sigma_2^2)/2$; two parameters are estimated. In this paper, this model is referred to as the 2-parameter model. The parameters in both models are estimated by maximizing equation (1), which is the likelihood when the number of alleles a sib pair shares ibd is ambiguously known.

$$L = e^C = \prod_{j=1}^n \sum_{k=0}^2 z_{jk} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp \left[\frac{-D_j^2}{2\sigma_k^2} \right] \quad (1)$$

For each sib pair j , the sum is taken over all possible ibd allele sharing configurations k ($k = 0, 1, \text{ or } 2$), D_j is the sib-pair trait difference, σ_k^2 is the variance of D_j for each class, k , of ibd allele sharing, and z_{jk} is the probability that sib pair j shares k alleles ibd.

ECD and EIF Case-Deletion Diagnostics

The exact case-deletion (ECD) and empirical influence function (EIF) can be applied to the QTL maximum likelihood methods developed by Kruglyak and Lander [6]. The gold standard for delete-one diagnostics is the ECD method, in which one observation at a time is deleted from the data set [1–3]. This method can be described as follows: (1) compute the parameter estimates when all the observations are in the data set; (2) delete the first observation; (3) compute the parameter estimates based on the remaining observations; (4) calculate the difference in the parameter estimates computed in steps 1 and 3; (5) replace the first observation in the data set, and (6) repeat steps 2 through 5 for every subsequent observation in the data set.

Let $\hat{\theta}$ be the maximum likelihood estimate (mle) of a generic parameter based on the complete data set. After each deletion, the mle without the i -th observation $\hat{\theta}_{(-i)}$ is computed and the difference in the mles with and without the i -th observation, $\hat{\theta}_{(-i)} - \hat{\theta}$, is assessed. The mles $\hat{\theta}$ and $\hat{\theta}_{(-i)}$ are both vectors of one or more parameters. A large change in the mle of any of the parameter estimates indicates that the deleted observation is an influential point. The ECD method is very time consuming and computationally expensive since $n + 1$ mles must be computed, where n is the number of observations [3]. Hence, a less time consuming method to approximate the ECD method is preferred.

The EIF, a local influence diagnostic introduced by Hampel [7, 8], is easy to compute. It is used to determine the influence of a single observation on parameter estimates and functions of parameter estimates [7]. Local influence refers to perturbations of the data that are infinitesimally small [2]. For maximum likelihood estimation, the EIF is computed as

$$EIF(x_j) = I^{-1}(\hat{\theta}) \dot{\mathcal{L}}_j^*(\hat{\theta}, x_j),$$

where $I^{-1}(\hat{\theta})$ is the inverse information based on all parameter estimates, x_j is the trait value for the j -th observation, and $\dot{\mathcal{L}}_j^*(\hat{\theta}, x_j)$ is the contribution to the score vector from the j -th observation [8, 9].

Change in the Variances

For both the 2- and 3-parameter models, the EIF can be applied to both the alternative hypothesis of linkage and the null hypothesis of no linkage. For the 2-parameter model under the alternative hypothesis of linkage, the EIF takes the following form:

$$\begin{aligned} EIF_{(-j)} &= I^{-1}(\hat{\sigma}_0^2, \hat{\sigma}_2^2) \dot{\mathcal{L}}_j(\hat{\sigma}_0^2, \hat{\sigma}_2^2) \\ &= \left(\Delta_{(-j)}^{EIF} \hat{\sigma}_0^2, \Delta_{(-j)}^{EIF} \hat{\sigma}_2^2 \right)', \end{aligned} \quad (2)$$

where $I^{-1}(\hat{\sigma}_0^2, \hat{\sigma}_2^2)$ is the 2×2 information matrix and $\dot{\mathcal{L}}_j(\hat{\sigma}_0^2, \hat{\sigma}_2^2)$ is the contribution to the score from the j -th sib pair evaluated at the mles. The term $\Delta_{(-j)}^{EIF} \hat{\sigma}_k^2$ is the change in the mle of σ_k^2 approximated by the EIF method when the j -th sib pair is omitted. The EIF estimates changes in the variances (equation (2)) which, when added to the corresponding mles based on all sib pairs, yields EIF-based variance estimates

$$\hat{\sigma}_{k(-j)}^{2,EIF} = \hat{\sigma}_k^2 + \Delta_{(-j)}^{EIF} \hat{\sigma}_k^2, k = 0, 2 \text{ and } \hat{\sigma}_1^{2,EIF} = \frac{\hat{\sigma}_0^{2,EIF} + \hat{\sigma}_2^{2,EIF}}{2}.$$

Under the null hypothesis of no linkage there is only one parameter σ_{null}^2 . When the j -th sib pair is omitted, the EIF is

$$\begin{aligned} EIF_{(-j)} &= I^{-1}(\hat{\sigma}_{null}^2) \dot{\mathcal{L}}_j(\hat{\sigma}_{null}^2) = \Delta_{(-j)}^{EIF} \hat{\sigma}_{null}^2, \\ \hat{\sigma}_{null(-j)}^{2,EIF} &= \hat{\sigma}_{null}^2 + \Delta_{(-j)}^{EIF} \hat{\sigma}_{null}^2. \end{aligned}$$

In the 2-parameter model, if convergence in the EM algorithm results in $\hat{\sigma}_0^2 < \hat{\sigma}_2^2$, then σ_0^2 is set equal to σ_2^2 because of the constraint and the EM algorithm is repeated to produce a single variance estimate. The constrained variance estimates are used to compute constrained lod scores. In addition to constrained variance estimates, unconstrained variance estimates can be also computed. These use the values of the variances that are obtained when the EM algorithm converges the first time whether or not the constraint is met; hence, either $\hat{\sigma}_0^2 < \hat{\sigma}_2^2$ or $\hat{\sigma}_0^2 > \hat{\sigma}_2^2$. Unconstrained parameter estimates are used to compute unconstrained lod scores.

The EIF is computed similarly for the 3-parameter model except σ_1^2 is also estimated, the inverse information is a 3×3 matrix, and the score vector contains three elements. The EIF is a labor-saving approximation to the ECD when there is no constraint, which is not the case in the 3-parameter model with a constraint. For example, suppose that constrained maximization results in $\hat{\sigma}_0^2 = \hat{\sigma}_1^2 = \hat{\sigma}_2^2$ when all observations are used. If removal of the i -th sib pair results in $\hat{\sigma}_{0(-i)}^2 \geq \hat{\sigma}_{1(-i)}^2 = \hat{\sigma}_{2(-i)}^2$, the ECD approach simply computes the difference between these estimators and the original estimators that are all equal. The EIF formula is based on the 1×1 information matrix from the constrained mle based on all observations and hence is no longer applicable when removal of a sib pair removes the need for the constraint. Modification of the EIF to approximate the ECD is complex and loses any computational advantage. Hence, there is no time-efficient EIF formula for approximating the ECD.

Change in the Lod Score

The EIF approximation to the change in lod score when the j -th sib pair is omitted can also be calculated. The EIF-based lod score is the base 10 logarithm of the likelihood ratio, in which the numerator likelihood is evaluated at the EIF-based variance estimates under the alternative hypothesis of linkage, and the denominator likelihood is evaluated at the EIF-based null variance estimate under the hypothesis of no linkage. For both numerator and denominator, the j -th sib pair is deleted from the likelihood computations, yielding

$$LOD_{(-j)}^{EIF} = \log_{10} \frac{L_{(-j)}(\hat{\sigma}_{0(-j)}^{2,EIF}, \hat{\sigma}_{1(-j)}^{2,EIF}, \hat{\sigma}_{2(-j)}^{2,EIF})}{L_{(-j)}(\hat{\sigma}_{null(-j)}^{2,EIF})}.$$

Multiple Sib-Pair Families

In computing the EIF when there is more than one sib pair in a family, the contributions to the score are summed over all sib pairs in that family, and all sib pairs in that family are deleted from the likelihood under both the alternative and null hypotheses.

Simulation of Pedigrees

Data sets were simulated using the program SLINK [10, 11]. Each simulation consisted of 500 data sets. Each data set consisted of 100 2-sibling nuclear families. For each individual, three loci were simulated. Two were marker loci, each with four equi-frequent alleles. Because the 2-parameter model is a no dominance variance model, the third locus was a biallelic QTL simulated under an additive model with alleles 'a' and 'A', and with genotype-specific phenotype means of 23.0, 28.5, and 34.0 for the genotypes aa, Aa, and AA, respectively. The frequency of allele 'a' was $p = 0.871$ and that of allele 'A' was $1 - p = 0.129$. This was the allele frequency that, when using the genotype-specific means of 23.0, 28.5, and 34.0, yielded a phenotype such that the QTL represented 30% of the total phenotypic variability based on the variance of BMI, 22.7, from the sibling phenotypes [12]. The QTL was either linked to the marker loci in a simulation of 500 data sets or unlinked to the marker loci in a simulation of 500 data sets. When the QTL and marker loci were linked, the QTL was centered between the two markers, located 5 cM from each marker for a total genetic map distance of 10 cM.

Construction of Outlier Sib Pair

The outlier was a single sib pair in each data set of 100 families. The presence of the outlier had the effect of either enhancing or diminishing evidence for linkage between the QTL and marker loci. These effects were created by giving the sib pair a large trait difference, which was generated from a $N(15, 1)$ distribution. Mean sib-pair trait differences with values between 14 and 16 yielded the best results with respect to generating an outlier that was detected some of the time, but not detected all of the time or never detected at all. Therefore, a distribution with mean 15 and a standard deviation of 1 was chosen to generate the outlier sib-pair trait differences. This was considered reasonable since sib-pair trait differences in the Muscatine Study example discussed below were as high as 19. If the QTL and marker loci were unlinked to each other, then an outlier with a large sib-pair trait difference and high proportion of alleles shared ibd would have the effect of maintaining no linkage whereas an outlier with a large

sib-pair trait difference and a small proportion of alleles shared ibd would have the effect of creating linkage. If the QTL and marker loci were linked, then an outlier with a large sib-pair trait difference and high proportion of alleles shared ibd would have the effect of diminishing linkage whereas an outlier with a large sib-pair trait difference and a small proportion of alleles shared ibd would have the effect of enhancing the linkage signal. Creation of alleles shared ibd in the outlier sib pair at each marker locus was accomplished by randomly choosing the offspring genotypes from a list of all possible marker allele combinations conditional on the parental genotypes which were assumed to be double heterozygotes, and the number of alleles ibd the sib pair was to share, either 0 or 2. Data sets were also simulated with two outlier sib pairs under two scenarios with ibd allele sharing created in the same fashion as when there was one outlier sib pair in a data set. Under the first scenario, both outlier sib pairs shared 0 alleles ibd in each data set of 100 families. Under the second scenario, one outlier sib pair shared 0 alleles ibd and the other sib pair shared 2 alleles ibd in each data set of 100 families.

Statistics for Sensitivity and Specificity Analyses

To evaluate how well the EIF approximated the ECD, sensitivity and specificity analyses were conducted using both diagnostics. Sensitivity was defined as the proportion of outlier sib pair(s) that were correctly identified as outliers. Specificity was defined as the proportion of non-outlier sib pairs that were correctly identified as such. Three criterion statistics were constructed for sensitivity and specificity analyses. The first two statistics, S1 and S2, are standardized changes in the variances

$$S1_{(-j)k}^m = \frac{\left| \Delta_{(-j)}^m \hat{\sigma}_k^2 \right|}{\frac{\widehat{se}(\hat{\sigma}_k^2)}{\sqrt{\text{no. of sib pairs}}}},$$

$$S2_{(-j)k}^m = \frac{\left| \Delta_{(-j)}^m \hat{\sigma}_k^2 \right|}{SD_k(m)}$$

where $k = 0, 1, \text{ or } 2$; $m = \text{EIF or ECD}$; $\Delta_{(-j)}^m \hat{\sigma}_k^2$ is the estimated change in $\hat{\sigma}_k^2$ from dropping sib pair j using method m ; $\widehat{se}(\hat{\sigma}_k^2)$ is the standard error based on all sib pairs in the data set; and $SD_k(m)$ is the standard deviation of the $\{\Delta_{(-j)}^m \hat{\sigma}_k^2\}$. Regarding the standardization of the S1 criterion, the variance of $\Delta_{(-j)}^m \hat{\sigma}_k^2$ when deleting one subject contains the variance for the j -th pair divided by n^2 plus a variance (summed over the other pairs) divided by a term to the order of n^4 . The dominant term is $\sigma^2/n^2 = se^2/n$. The change in the variance statistic S2, is standardized by the sample.

The third statistic, S3, is the absolute change in the lod score with and without the j -th sib pair.

$$S3 = | \text{lodscore}_{(-j)} - \text{lodscore} |.$$

Sensitivity was computed as the proportion of the 500 simulated data sets for which the value of a criterion statistic for an outlier sib pair(s), in a data set of 100 families, was greater than or equal to a particular critical value. To compute specificity, the proportion of the non-outlier sib pairs, again in a data set of 100 families, for which the value of the criterion statistic was less than a particular critical value (a correctly classified non-outlier sib pair) was found for each data set. The mean proportion over the 500 data sets was the specificity. Critical values were chosen based

upon the observed distribution of the corresponding criterion statistic. Emphasis was placed on choosing critical values at regular intervals such that the sensitivity was high at low values of the statistic and low at high values of the statistic. For statistics S1 and S2 the chosen critical values were 1.50, 1.75, 2.00, 2.25, 2.50, and 2.75. Results are reported for 1.50, 2.00, and 2.50. For statistic S3 the chosen critical values were 0.05, 0.10, 0.15, 0.20, 0.25, and 0.30 lod score units. Results are reported for 0.05, 0.15, and 0.25.

Results

Simulation Study

Table 1 shows the estimated sensitivities for the standardized change in the variances for the 2-parameter model based on statistic S2 when the QTL and marker loci were unlinked. The outlier sib pair shared 0 alleles ibd in the first simulation (a), 1 allele ibd in the second simulation (b), and 2 alleles ibd in the third simulation (c). For both the constrained and unconstrained parameters, the EIF approximated the ECD very well for the outlier-affected parameter (σ_0^2 when the outlier sib pair shared 0 alleles ibd, σ_1^2 when the outlier sib pair shared 1 allele ibd, and σ_2^2 when the outlier sib pair shared 2 alleles ibd). This was not always the case for the outlier-unaffected parameter. At every critical value, the sensitivity was highest for the outlier-affected parameter indicating that statistic S2, within the context of the 2-parameter model, identified the outlier a high proportion of time. It also yielded high sensitivities for the outlier-unaffected parameters, showing the inability of this statistic under the 2-parameter model to specifically identify, the true nature of the ibd allele sharing in the outlier sib pair. Comparing these results to those of the 3-parameter model in table 2, the unconstrained EIF sensitivities approximated the unconstrained ECD sensitivities very well for all three parameters. In addition, based on the 3-parameter model, statistic S2 identified the outlier a high proportion of time and had the ability to distinguish the true nature of ibd allele sharing in the outlier sib pair, as seen from the high sensitivities for the unconstrained outlier-affected parameter, and the low, if not zero, sensitivities for the unconstrained outlier-unaffected parameters. The statistic S1 sensitivities (not shown) were very similar to those of statistic S2. Table 3 shows the sensitivities for the absolute value of the change in the lod score, statistic S3. The EIF sensitivities approximated the ECD sensitivities very well at all critical values under both the 2- and 3-parameter models. These sensitivities, as one would expect, were unable to distinguish the nature of ibd allele sharing in the outlier sib pair. Specifici-

Table 1. Estimated sensitivities for the change in the variances, based on statistic S2, standardized by the sample, using the 2-parameter model when the QTL and marker loci are unlinked

CV	Constrained						Unconstrained					
	ECD			EIF			ECD			EIF		
	σ_0^2	σ_1^2	σ_2^2	σ_0^2	σ_1^2	σ_2^2	σ_0^2	σ_1^2	σ_2^2	σ_0^2	σ_1^2	σ_2^2
(a) Outlier sib pair shares 0 alleles ibd												
1.50	93.6	89.6	43.0	94.6	88.8	41.4	97.0	87.4	49.4	97.2	86.8	41.8
2.00	81.6	60.4	24.2	81.6	58.6	20.4	92.6	60.2	31.4	92.6	58.6	23.2
2.50	62.4	30.0	10.2	63.4	28.8	9.2	84.6	32.8	20.2	85.2	31.6	12.0
(b) Outlier sib pair shares 1 allele ibd												
1.50	71.2	93.4	42.4	76.2	93.8	44.8	33.2	94.2	23.2	38.4	94.6	29.4
2.00	44.4	71.2	25.6	48.4	72.0	26.8	13.8	71.4	8.6	16.8	72.8	10.6
2.50	19.8	41.6	13.2	21.8	42.6	13.8	3.4	42.8	2.6	5.0	44.6	2.8
(c) Outlier sib pair shares 2 alleles ibd												
1.50	85.0	91.0	95.8	81.0	91.8	96.0	57.6	86.0	98.2	47.2	86.6	97.6
2.00	64.0	68.0	84.8	57.2	68.8	84.8	37.0	60.6	94.0	26.2	60.0	94.2
2.50	40.4	42.8	71.2	34.4	42.2	70.4	22.0	35.6	87.2	14.4	34.6	87.0

CV = Critical value; 500 data sets in each simulation (a), (b), and (c); evaluated at 0 cM; QTL simulated under an additive model; outlier sib-pair trait difference was from a N(15,1) distribution.

Table 2. Estimated sensitivities for the change in the variances, based on statistic S2, standardized by the sample, using the 3-parameter model when the QTL and marker loci are unlinked

CV	Constrained			Unconstrained					
	ECD			ECD			EIF		
	σ_0^2	σ_1^2	σ_2^2	σ_0^2	σ_1^2	σ_2^2	σ_0^2	σ_1^2	σ_2^2
(a) Outlier sib pair shares 0 alleles ibd									
1.50	95.6	40.4	13.6	98.4	7.0	0.4	98.4	3.4	0.0
2.00	88.0	29.2	8.4	96.8	3.8	0.2	97.0	1.4	0.0
2.50	74.6	16.6	4.4	93.4	1.8	0.2	93.0	0.8	0.0
(b) Outlier sib pair shares 1 allele ibd									
1.50	58.6	94.8	33.4	1.6	97.4	3.4	1.0	97.8	1.4
2.00	46.0	82.2	24.4	0.4	92.2	1.2	0.2	92.2	0.4
2.50	30.6	63.0	14.8	0.2	83.0	0.8	0.0	82.8	0.0
(c) Outlier sib pair shares 2 alleles ibd									
1.50	32.6	58.2	96.4	0.0	5.8	99.0	0.0	2.0	99.0
2.00	23.4	45.8	88.0	0.0	2.4	97.6	0.0	0.8	97.6
2.50	13.0	31.6	76.0	0.0	1.6	93.0	0.0	0.4	92.8

CV = Critical value; 500 data sets in each simulation (a), (b), and (c); evaluated at 0 cM; QTL simulated under an additive model; outlier sib-pair trait difference was from a N(15,1) distribution.

ties (results not shown) increased with decreasing sensitivities and were for the most part very high (≥ 90) for statistics S1, S2, and S3. Very few times did the specificities decrease to the high 80s.

Table 4 shows the sensitivities for the S2 standardized change in the variances when there were two outliers in the data set. The EIF sensitivities estimated the ECD sensitivities well when both outliers shared 0 alleles ibd (a).

Table 3. Estimated sensitivities for the change in the lod score when the QTL and marker loci are unlinked

CV	2-Parameter model				3-Parameter model		
	constrained		unconstrained		constrained	unconstrained	
	ECD	EIF	ECD	EIF	ECD	ECD	EIF
(a) Outlier sib pair shares 0 alleles ibd							
0.05	47.6	47.8	74.2	74.4	55.6	77.8	78.0
0.15	15.6	15.8	29.0	28.8	19.0	38.0	37.2
0.25	2.6	2.6	10.6	10.6	4.6	13.6	13.2
(b) Outlier sib pair shares 1 allele ibd							
0.05	0.2	0.2	0.8	0.8	30.0	65.0	65.2
0.15	0.0	0.0	0.0	0.0	3.0	20.4	20.4
0.25	0.0	0.0	0.0	0.0	0.0	4.4	4.4
(c) Outlier sib pair shares 2 alleles ibd							
0.05	35.4	35.4	72.6	72.6	47.4	76.6	76.8
0.15	18.8	18.8	29.8	29.8	22.4	39.8	39.4
0.25	9.8	9.2	11.8	11.0	13.0	17.6	17.6

CV = Critical value; 500 data sets in each simulation (a), (b), and (c); evaluated at 0 cM; QTL simulated under an additive model; outlier sib-pair trait difference was from a N(15,1) distribution.

Table 4. Estimated sensitivities for the standardized change in the variances, based on S2, standardized by the sample, using the 2-parameter model when the QTL and marker loci are unlinked

CV		First outlier				Second outlier				At least one outlier			
		constrained		unconstrained		constrained		unconstrained		constrained		unconstrained	
		ECD	EIF	ECD	EIF	ECD	EIF	ECD	EIF	ECD	EIF	ECD	EIF
(a) Both outlier sib pairs share 0 alleles ibd													
σ_0^2	1.50	92.4	93.0	95.2	95.8	93.6	93.6	95.6	96.0	95.8	96.2	96.8	97.4
	2.00	80.0	80.0	88.6	88.4	80.0	80.6	88.2	88.8	88.0	88.2	93.2	93.4
	2.50	61.6	61.4	75.6	74.8	61.8	62.0	77.2	77.0	73.8	74.2	85.4	85.0
(b) First outlier sib pair shares 0 alleles ibd, the second outlier sib pair shares 2 alleles ibd													
σ_0^2	1.50	94.8	95.0	97.8	98.6	76.6	69.8	52.6	44.4	96.8	97.0	98.4	99.0
	2.00	77.8	79.6	93.4	94.0	55.6	48.2	34.8	26.2	88.0	88.2	96.0	95.8
	2.50	57.6	57.6	87.2	87.2	35.0	27.8	24.2	14.8	69.8	67.6	91.6	90.4
σ_2^2	1.50	43.4	42.6	46.8	39.4	95.4	95.6	98.0	98.0	97.6	97.4	99.2	99.2
	2.00	25.2	24.6	29.6	20.8	87.8	88.0	96.6	96.6	92.2	92.4	98.2	98.2
	2.50	10.4	10.8	19.8	10.6	75.0	75.0	90.8	91.6	78.6	78.8	94.2	93.0

CV = critical value; 500 data sets in each simulation (a) and (b); evaluated at 0 cM; QTL simulated under an additive model; outlier sib-pair trait difference is from a N(15,1) distribution.

When the number of alleles shared ibd was different for the two outlier sib pairs, the σ_0^2 EIF sensitivities approximated the σ_0^2 ECD sensitivities well for the outlier sib pair that shared 0 alleles ibd, the σ_2^2 EIF sensitivities approximated the σ_2^2 ECD sensitivities well for the outlier sib pair

that shared 2 alleles ibd, and the EIF sensitivities approximated the ECD sensitivities well when there was at least one outlier (b). Table 5 shows the sensitivities for the change in the lod scores when there were two outliers in a data set. The EIF sensitivities estimated the ECD sensi-

Table 5. Estimated sensitivities for the change in the lod score using the 2-parameter model when there are two outliers in each data set

CV	First outlier				Second outlier				At least one outlier			
	constrained		unconstrained		constrained		unconstrained		constrained		unconstrained	
	ECD	EIF	ECD	EIF	ECD	EIF	ECD	EIF	ECD	EIF	ECD	EIF
(a) Both outlier sib pairs share 0 alleles ibd												
0.05	60.4	60.4	75.0	75.0	60.0	60.0	74.0	74.0	61.0	61.0	76.4	76.4
0.15	21.8	21.8	27.2	27.2	22.4	22.4	28.4	28.4	26.4	26.4	33.2	33.2
0.25	2.6	2.6	4.8	4.8	3.2	3.4	5.8	6.0	4.0	4.2	6.6	6.8
(b) First outlier sib pair shares 0 alleles ibd, the second outlier sib pair shares 2 alleles ibd												
0.05	27.8	27.8	67.0	67.0	49.6	49.6	73.2	73.2	49.6	49.6	88.0	88.0
0.15	4.2	4.4	25.0	25.2	27.0	26.6	30.8	30.4	27.0	26.6	47.8	47.4
0.25	0.4	0.4	10.6	10.6	16.2	16.0	16.4	16.2	16.2	16.0	26.4	26.2

CV = critical value; 500 data sets in each simulation (a) and (b); evaluated at 0 cM; QTL simulated under an additive model; outlier sib-pair trait difference is from a N(15,1) distribution.

Table 6. Descriptive statistics (mean $\pm$ standard deviation, with range in parentheses) for age, BMI, and standardized residuals for age- and gender-adjusted $\ln(\text{BMI})$ for 233 individuals in the genotyped Muscatine Study nuclear families

	Age	BMI	Adjusted $\ln(\text{BMI})$
Fathers (n = 50)	42.8 $\pm$ 5.7 (32–68)	29.09 $\pm$ 4.7 (20.5–38.4)	0.1 $\pm$ 1.0 (–2.0 to 2.0)
Mothers (n = 50)	40.6 $\pm$ 4.09 (32–49)	28.3 $\pm$ 6.4 (17.4–43.9)	0.2 $\pm$ 1.1 (–2.0 to 2.4)
Female offspring (n = 69)	14.2 $\pm$ 4.5 (5–31)	22.5 $\pm$ 5.6 (13.4–38.8)	0.2 $\pm$ 1.2 (–2.3 to 3.2)
Male offspring (n = 64)	14.1 $\pm$ 3.1 (6–21)	21.8 $\pm$ 5.1 (14.7–43.4)	0.0 $\pm$ 1.0 (–1.6 to 3.5)

tivities very well. The sensitivities when the QTL and the markers were linked (results not shown) were quite similar to the sensitivities when the QTL and the markers were unlinked.

For both the 2- and 3-parameter models, the EIF was more time efficient than the ECD in computing case-deletion diagnostics on 500 datasets, each containing 100 two-sibling nuclear families. For the 2-parameter model, it took 6 h, 0 min, and 5 s to compute all the ECDs whereas it only took 11 min and 53 s to compute all the EIFs on an HP Series 9000 Model B1000 Unix computer. For the 3-parameter model the ECDs took 7 h, 44 min,

and 42 s (56 s per data set) whereas the EIFs took only 24 min and 3 s to compute (less than 3 seconds per data set).

Application to Muscatine Study Nuclear Families

School-age children in Muscatine, Iowa were recruited and examined (anthropomorphic measurements and fasting blood samples were obtained) along with their siblings and their parents. There were a total of 381 nuclear families in the data set, 50 of which were fully genotyped using an approximate 10 cM screening set of microsatellite markers. Only the 50 genotyped families, consisting of 233 family members, were used in this analysis. There were two families with five children, four families with four children, nineteen families with three children, and the remaining twenty-five families had two children. Unlike the simulations in which there was only one sib pair per family, there were dependent sib pairs in this analysis. The total number of sib pairs was 126. Table 6 describes the age, body mass index (BMI, weight (kg)/height (m^2)), and the standardized residuals of age- and gender-adjusted natural log transformed BMI, for the fathers, mothers, male offspring, and female offspring. The phenotype analyzed was constructed by first taking the natural log transformation of BMI in all 1,610 members of the 381 families, not just the 50 genotyped families. Within each of four groups (mothers, fathers, female siblings, and male siblings), $\ln\text{BMI}$ was regressed on age. The residuals from this regression were divided by the standard deviation, again, within each respective group.

Fig. 1. Empirical influence function versus exact case deletion using the 2-parameter model. Change in the variances and lod score, statistic S2, standardized by the sample. Muscatine Study nuclear family data set; 50 families; chromosome 11; 7 cM.

This yielded age-gender-adjusted standardized ln body mass index (lnBMI).

In the simulations, the analysis unit was the sib pair, as only one sib pair was simulated per family. In the Muscatine Study data set each family had at least one sib pair and the entire sibship or family was used as the analysis unit. In conducting a multipoint linkage analysis for lnBMI, in which all genetic marker information was used to compute the probability distribution of ibd allele sharing at each point throughout the genome, using the 2- and 3-parameter models, there was evidence of linkage at 8.9 cM near marker ATA33B03 on chromosome 11. For the 2-parameter model, a well-defined peak was located at marker ATA33B03 with a lod score of 0.95. For the 3-parameter model, a well-defined peak also occurred at marker ATA33B03 with a lod score of 1.27. An outlier analysis of the Muscatine data set was conducted to answer the following questions: (1) Is there a family or more

than one family that, if deleted, would either diminish or enhance the linkage signal that was seen near marker ATA33B03? and (2) does the EIF provide a good approximation to the ECD in detecting that influential family or families?

Figures 1 and 2 show index plots of the S2 standardized change in the variances and the lod score for the EIF vs. the ECD when the 2- and 3-parameter models were used to analyze lnBMI, respectively. The EIF approximated the ECD very well for each family. Table 7 shows the most influential families detected at the highest critical S2 value, 2.75, for both diagnostic methods using the 2- and 3-parameter models, respectively. Since the constraint was always met, there was no distinction between unconstrained and constrained estimators.

From this analysis, only the results for statistic S2 are shown because this statistic tended to provide a better EIF approximation to the ECD than statistic S1. From the re-

Fig. 2. Empirical influence function versus exact case deletion using the 3-parameter model. Change in the variances and lod score, statistic S2, standardized by the sample. Muscatine Study nuclear family data set; 50 families; chromosome 11; 7 cM.

Table 7. Most highly influential Muscatine Study nuclear families for age- and gender-adjusted lnBMI identified using the change in the variances, based on statistic S2, standardized by the sample

Family	Affected parameter	2-parameter model				3-parameter model			
		$\Delta\sigma^2$		Δ lod score		$\Delta\sigma^2$		Δ lod score	
		ECD	EIF	ECD	EIF	ECD	EIF	ECD	EIF
1	σ_0^2	3.48	3.52	0.19	0.19	3.88	3.95	0.27	0.27
	σ_1^2	2.95	3.00	0.19	0.19	0.45	0.40	0.27	0.27
12	σ_2^2	2.91	2.96	0.26	0.26	3.03	3.00	0.20	0.20
16	σ_2^2	3.11	3.10	0.28	0.28	3.22	3.19	0.21	0.21
36	σ_2^2	4.05	3.96	0.37	0.37	4.07	4.02	0.28	0.28
37	σ_0^2	2.54	2.52	0.24	0.24	3.01	3.00	0.35	0.35
38	σ_0^2	3.31	3.27	0.13	0.13	2.43	2.32	0.06	0.06
	σ_1^2	3.22	3.17	0.13	0.13	3.07	3.19	0.06	0.06

sults using the 2-parameter model, deleting family 36 resulted in the greatest change in any of the variance parameter estimates among the most influential observations. Deletion of this family also resulted in the greatest change in the lod score. From the results using the 3-parameter model, again deleting family 36 resulted in the greatest change in any of the variance parameter estimates among the most influential observations, but not the greatest change in the lod score. Under the 3-parameter model, it was family 37 that showed the greatest change in the lod score of all the outliers identified using statistic S2. There was one sib pair in this family; the siblings shared 0 alleles ibd with probability 0.97 and had a trait difference (3.10) that was more than two standard deviations from the mean sib-pair trait difference of 2.48. The presence of this family in the data set had the effect of increasing σ^2_0 and therefore the lod score. Family 36 also had one sib pair. The probability that this sib pair shared 2 alleles ibd was 1.00 and the trait difference (2.08) was almost two standard deviations less than the mean sib-pair trait difference, but not as large as the sib-pair trait difference for family 37. Families 36 and 37 were masking each other in the data set. There was additional partial masking with two other families in the data set, families 16 and 1. Both families 36 and 16 influenced the estimate of σ^2_2 and the lod score. When only one of these two families was deleted, the lod score increased (from 1.27 to 1.47 when family 16 was deleted and from 1.27 to 1.54 when family 36 was deleted) and σ^2_2 decreased (from 1.10 to 0.94 when family 16 was deleted and from 1.10 to 0.89 when family 36 was deleted), but not as much as when both families were deleted from the data set (1.92 for the lod score and 0.69 for σ^2_2). Families 1 and 37 were influential for σ^2_0 and the lod score. When only one of these two families was deleted, both the lod score and σ^2_0 decreased, but not as much as when both families were deleted from the data set. When family 1 was deleted from the data set, the lod score decreased from 1.27 to 0.99 and σ^2_0 decreased from 2.88 to 2.51. When family 37 was deleted from the data set, the lod score decreased from 1.27 to 0.91 and σ^2_0 decreased from 2.88 to 2.59. When both families 37 and 1 were deleted from the data set, the lod score decreased to 0.64 and σ^2_0 decreased to 2.21.

Discussion

Under the conditions of the simulation study and from the analysis of the Muscatine Study data, it is clear that the diagnostic methods developed in this paper can de-

tect outliers that can impact parameter estimation and therefore the results of a sib-pair linkage analysis. This investigation demonstrates that the unconstrained EIF is a powerful diagnostic tool based on the standardized change in the variances under the 3-parameter model. The EIF has high sensitivity and identifies the ibd allele sharing in the outlier. In addition, the EIF provides a huge time savings advantage over the ECD and yet estimates the ECD very well. The disadvantage of the EIF approach under the 3-parameter model is the lack of a time-efficient version to approximate the constrained ECD. So why would one compute an unconstrained EIF for the 3-parameter model? The unconstrained estimates still provide considerable information, including the absolute magnitude and direction of change in the value of the parameter estimates, the identification of outliers and which parameters, if any, need to be set equal in the EM algorithm for the constraint to be met.

The main advantage under the 2-parameter model is that both constrained and unconstrained EIF diagnostics are computed quickly. For the outlier-affected parameter, the EIF approximates the ECD very well using the standardized change in the variances. However for the other parameters, the EIF approximation is not as good, especially for the 2-parameter model. Furthermore, under the 2-parameter model, the outlier is identified a high proportion of the time, however there is an inability to discern the ibd allele sharing in the outlier. Although the EIF approximates the ECD very well for the change in the lod score, it is not a preferred diagnostic with respect to identifying influential observations. As seen from the Muscatine Study data, partial masking exists when there is more than one outlier in the data set. The degree of masking and the effect on the change in the variances and the change in the lod scores is dependent on the type of allele sharing in the outliers and the sib-pair trait difference.

One might infer that from the advantages of the 3-parameter over the 2-parameter model mentioned above the 3-parameter model is always the model of choice. After all, the 3-parameter model is a more general model, i.e., allows for dominance variance whereas the 2-parameter model is a more restrictive model that assumes no dominance variance. However, it could well be that neither of the two models is an appropriate fit for the data. As seen from the application to the Muscatine Study data, six of the 50 families are identified as influential observations. These influential families represent 12% of the sample. This begs the question of whether or not the appropriate model was chosen to analyze the data. It is quite likely that locus heterogeneity is responsible for trait vari-

ability for a complex phenotype like BMI and that a heterogeneity parameter needs to be built into the likelihood. Under a heterogeneity model, it is quite likely that these six influential families are not necessarily outliers. Although the power of a heterogeneity analysis would be low given the small number of sib pairs in the Muscatine Study data set, in general, with a larger sample size the possibility of locus heterogeneity should be investigated. In addition, the question of whether these diagnostics have more power to detect outliers in smaller or larger sibships should be investigated.

Case-deletion diagnostic methods are useful tools in identifying influential observations and therefore should be part of any standard QTL analysis. Although the ECD is the gold standard, the EIF approximates the ECD very well and has the advantage of quick computation. In our simulation studies, the EIF was 20 to 30 times faster than the ECD, making it a realistic tool for practical use. Once an influential pair is found, the data must be checked for accuracy. Very frequently, the problem is a miscoded data value, which when corrected, removes the influence of that pair. Of course, the analytical results would be incor-

rect without the correction. In general, one desires analytical results that are stable enough not to be seriously altered by one or two pairs. If a few pairs are influential, then one should always report the results from the complete (corrected) data as well as the results from any reduced data set to insure against data-dredging practices. A discussion of the two analyses is always in order. If the number of influential pairs is large, then the entire model should be called into question.

Acknowledgements

We would like to thank the reviewers for contributing their time, carefully reading our manuscript, and making insightful suggestions and comments.

This dissertation research of M.C.B. Mendoza was supported by NIH grant HD029569 which also supported the recruitment and examination of the Muscatine nuclear families.

The views presented in this paper are those of the authors and do not necessarily represent those of the U.S. Food and Drug Administration.

References

- 1 Belsley DA, Kuh E, Welsch RE: Regression Diagnostics Identifying Influential Data and Sources of Collinearity. New York, John Wiley and Sons, 1980.
- 2 Carroll RJ, Ruppert D: Transformation and Weighting in Regression. New York, Chapman and Hall, 1988.
- 3 Cook RD, Weisberg S: Residuals and Influence in Regression. New York, Chapman and Hall, 1982.
- 4 Wang HM, Jones MP, Burns TL: Regression diagnostics for the class a regressive model with quantitative phenotypes. *Genet Epidemiol* 1995;17:174–187.
- 5 de Andrade M, Fridley B, Boerwinkle E, Turner S: Diagnostic tools in linkage analysis for quantitative traits. *Genet Epidemiol* 2003;24:302–308.
- 6 Kruglyak L, Lander ES: Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am J Hum Genet* 1995; 57:439–454.
- 7 Hampel FR: The influence curve and its role in robust estimation. *J Am Stat Assoc* 1974; 69:383–393.
- 8 Weissfeld L, Schneider H: Influence diagnostics for the normal linear model with censored data. *Aus J Stat* 1990;32:11–20.
- 9 Samuels SJ: Survival Analysis from the Viewpoint of Hampel's Theory for Robust Estimation. Chapel Hill, University of North Carolina, 1978.
- 10 Ott J: Computer-simulation methods in human linkage analysis. *Proc Natl Acad Sci* 1989;86:4175–4178.
- 11 Weeks D, Ott J, Lathrop G: SLINK: a general simulation program for linkage analysis. *Am J Hum Genet* 1990;47:A204.
- 12 Moll PP, Burns TL, Lauer RM: The genetic and environmental sources of body variability: the Muscatine ponderosity family study. *Am J Hum Genet* 1991;49:1243–1255.