Influence Analysis on LOD Score Curve
— Threshold determination and comparisons of diagnostics —

Xiaoling Dou　Project Researcher　Transdisciplinary Research Integration Center

1. Introduction

Quantitative trait locus (QTL) analysis is a statistical method for detecting precise location of chromosome regions associated with a particular phenotypic trait. In QTL detection, log odds (LOD) scores are calculated for marker loci as plausibility of the existence of QTLs.

When two peaks located closely on the same chromosome, it is important to determine whether the two peaks were caused by two QTLs or whether they arose because of statistical errors. Particularly, when sample size is small, LOD score curves can be easily influenced by a few individuals, and lead to unstable results. Therefore, influence analysis is important and necessary to identify influential individuals in evaluating the reliability of the QTL analysis results.

2. LOD score and Influence analysis methods

For a given dataset with sample size n, phenotypes y_i $(i = 1, ..., n)$, genotypes $z_i = (z_i^{(1)}, ..., z_i^{(M)})'$, $z_i^{(j)} = \{-1, 0, 1\}$ on M loci and covariates, such as, sex $u_i = \{0, 1\}$, assuming that

$$y_i = \mu + \alpha_j z_i^{(j)} + \beta_j w_i^{(j)} + \nu u_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2), \tag{1}$$

where $w = 1$ $(z = \pm 1)$, -1 $(z = 0)$ is a function of z, α and β indicate the additive effect and the dominance effect, respectively.

LOD scores is defined as

$$\mathrm{LOD}(j) = \frac{1}{\log 10}\left\{\sum_{i=1}^{n} \log f(y_i|z_i^{(j)}, u_i; \widehat{\theta}_j) - \sum_{i=1}^{n} \log g(y_i|*, u_i; \widetilde{\theta})\right\},$$

where f is the normal density function of y_i under model (1) with $\widehat{\theta}_j = (\widehat{\alpha}_j, \widehat{\beta}_j, \widehat{\mu}_j, \widehat{\nu}_j, \widehat{\sigma}_j^2)'$. g is the normal density function of y_i with $\widetilde{\theta}_j = (0, 0, \widetilde{\mu}, \widetilde{\nu}, \widetilde{\sigma}^2)'$. And the empirical influence function of $\mathrm{LOD}(j)$ for individual i is given by

$$\mathrm{EIF}(i; \mathrm{LOD}(j)) = \frac{n}{\log 10}\left\{\ell(j, \widehat{\theta}(j); i) - \ell_0(\widetilde{\theta}; i)\right\} - \mathrm{LOD}(j),$$

where

$$\ell(j, \theta; i) = \log f(y_i|z_i, u_i; j, \theta), \quad \ell_0(\theta; i) = \log g(y_i|*, u_i; \theta).$$

To detect influential individuals affect the shape of LOD score curve, we have proposed some methods, such as the specified projection method and the eigenvector method. The specified projection method is to find

$$\mathrm{EIF}(i; c) = \sum_{j \in J} c_j \mathrm{EIF}(i; j),$$

where c is a vector designed for the shape of the LOD score curve to deal with, and J is the set of interesting loci. The individuals having large $|\mathrm{EIF}(i; c)|$ can be considered as influential candidates.

3. Threshold determination

We propose that a individual should be called influential when its standardized empirical influence function

$$\mathrm{SEIF}(i; c) = \mathrm{EIF}(i; c) / \sqrt{\sum_{i=1}^{n} \mathrm{EIF}^2(i; c)} \tag{2}$$

is larger than the upper 100α percentage point (such as, 5%) of $\max_{i=1,...,n} \mathrm{SEIF}(i; c)$ under the null hypothesis $(\alpha = \beta = 0)$. The distribution of $\max_{i=1,...,n} \mathrm{SEIF}(i; c)$ can be estimated by simulation.

4. Data Analysis and Simulation

Using the specified projection method, we can design a vectot to detect the influential mice which significantly affect the parallel shift, inclination or curvature of the LOD score curve. For its curvature, using vector $c = (1.04, -2.36, 1.31)'$ and the influence matrix $(\mathrm{EIF}(35.4), \mathrm{EIF}(53.8), \mathrm{EIF}(68.4))$, the standardized empirical influence functions $\mathrm{SEIF}(i; c)$ of all the 170 mice are obtained by (2).

We compare the specified projection method with two most often used diagnostics, the standardized residual, Cook's D_i and some other appropriate statistics by simulations.

Suppose that the regression model is

$$y_i = \sum_{j \in M} (\alpha_j z_i^{(j)} + \beta_j w_i^{(j)}) + \mu + \nu u_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2) \tag{3}$$

where M is a set of interesting loci. With a p-vector $b = ((\alpha_j, \beta_j)_{j \in M}, \mu, \nu)'$, the n-vectors $y = (y_1, ..., y_n)'$, $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)'$, $\mathbb{1} = (1, ..., 1)'$, $u = (u_1, ..., u_n)'$ and the $n \times p$ design matrix $X = ((z^{(j)}, w^{(j)})_{j \in M}, \mathbb{1}, u)$, (3) can be rewritten in the matrix form

$$y = Xb + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2 I).$$

Then, the estimates of y, $\hat{y} = \boldsymbol{H}y = X(X'X)^{-1}X'y$, the residuals $e = y - \hat{y}$, and the estimate of σ^2, $s^2 = e'e/(n - p)$ can be calculated. The ith standardized residual and Cook's D_i are defined as

$$r_i = \frac{e_i}{\sqrt{s^2(1 - h_i)}} \qquad \text{and} \qquad D_i = \frac{h_i r_i^2}{p(1 - h_i)}$$

respectively. Where h_i is the ith diagonal element of the hat matrix $\boldsymbol{H}$. Individuals having large absolute values of these statistics are influential observations.

Besides, $\max_{j \in M} |\mathrm{EIF}(i; j)|$, $\max_{j \in M} |r_i^{(j)}|$ and $\max_{j \in M} D_i^{(j)}$ are also suitable indicators for picking out observations cause pronounced changes in the shape of the curve.

1000 simulations are carried out. Each simulated dataset contains 168 normal observations and two special-created influential cases.

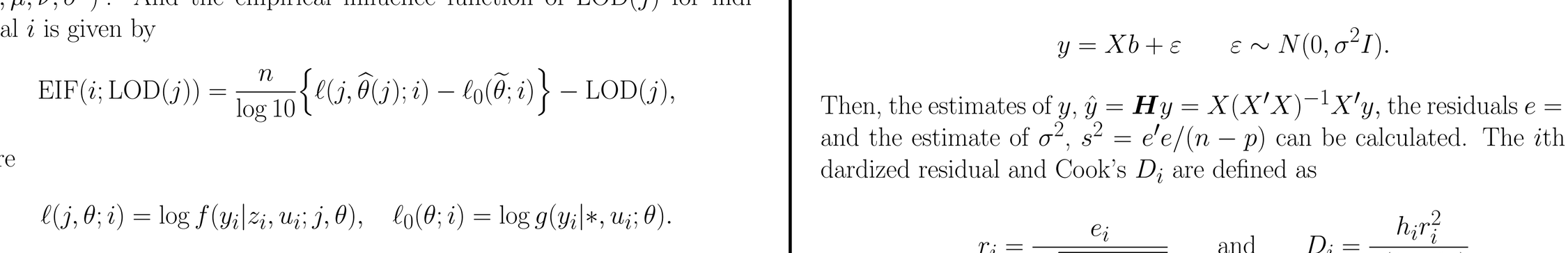

Figure 1: (1) LOD score curve without mouse No. 60. (2) Standardized empirical influence of the 170 mice on curvature of the LOD score curve on chromosome 3. (3) Threshold for detecting influential animals on curvature of the LOD score curve. (4–5)The last two panels show comparisons of the specified projection method with other influence methods by ROC curves. Phenotypes of influential individuals are designed as $N(\mu^*, (3\sigma)^2 I)$ and $\mu^* + (2\sigma)t_3$.