Austin Biometrics and Biostatistics

Open Access

Austin Publishing Group

Research Article

Some Genetic Regression Models for Multiple Quantitative End Points Data

Ao Yuan* and Jaeil Ahn
Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, USA

***Corresponding author:** Ao Yuan, Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington D.C. 20057, USA

Received: May 28, 2015; **Accepted:** January 28, 2016; **Published:** February 08, 2016

Abstract

Multiple endpoints data are common in practice. There are various statistical methods for the analysis of this type of data, however, genetic models for familial observations with multiple endpoints data are relatively few, and the existing methods are basically variations of the Elston-Stewart algorithm. Here we consider several joint statistical models for such data with quantitative measurements with a new algorithm, which is computationally more efficient than the existing method. The proposed method is detailed in some commonly used parametric, semi parametric and nonparametric settings for this type of data. For un-genotyped data, the commonly used models are the mixture and variance components models. We elaborate how these genetic models can be extended for multiple endpoints data with the proposed method

Keywords: Censoring; Endpoints data; Familial structure; Genotype; Missing observation

Introduction

Endpoints data are observed responses from patients of some pre-specified clinical events of interests, such as death, loss of vision, occurrences of certain diseases, or other symptomatic events. In medical research, study participants are often followed for a long time, during which some participants may drop out early, so that random censorship may be present in the data. Such data have missing observations, which may be inhomogeneous across the patients. For example, in one patient we have observations on the lung cancer and kidney disease, and on another we have observations on lung cancer, diabetes and asthma. Analyses of such data largely fall into two categories: hypothesis testing (usually non-model based) and model inference. Here we concentrate on the model inference of such data.

For censored data modeling there are extensive literatures [1-8], just mention a few. For multiple endpoints data analyses, there are various statistical methods [9-13], for example. Wei and Glidden [14] provided an overview for some of the methods in this field.

Family genetic data differ from the ordinary data in that they are collected in familial units, often with varying structures and sizes, and with/without genotyping. These features make the models distinct. The key in the modeling is the familial dependence structure and the implementation of the genetic mechanism, existing methods are basically variations of the Elston-Stewart algorithm, which is a multi-level mixture model, and the computation is often challenging. Genetic models for multiple endpoints data are relatively limited. Here we consider some statistical joint models for such data with quantitative measurements with a new algorithm, which is a one-level mixture model, thus enhance the computation considerably. The parametric method is used when one has some confidence about the model specification. The semi parametric method can be used when there is not enough information about the full parametric model specification. The nonparametric method is used for the robustness of least model assumptions. We elaborate our methods for the parametric, semi parametric and nonparametric cases. The methods we describe below are valid for arbitrary pedigrees; however, in this article we focus on the simpler case of nuclear family for illustration.

In genetic analysis, the data contains genotypes, partially genotypes, or no genotypes. However, even if the data are genotyped, it is still of interest to know whether there are some other unknown gene(s) behind the response functioning. There are reports that with added unknown gene locus, the likelihood Akaike information reduced (e.g., [15], p.1091, [16], which makes sense, as correct parameter(s) added to the model will reduce its AIC), or the segregation analysis guided to some other gene(s) which deserve(s) further investigation. So even for the genotyped data, a segregation model is still of importance. It is also the general model including the genotyped data case. In the following we derive some commonly used regressive models for this case, including the parametric model, semi-parametric proportional hazards model, nonparametric least squares model, variance components model and the competing risks model. Also, hypothesis testing on parameters of interest can be conducted using the likelihood ratio statistics based on the parametric models. Our aim here is to present several new parametric, semi parametric and nonparametric models for this familial data, and thus we focus derivations of the basic forms of these models. Implementations of these models and applications to real data will follow in our future work.

Methods

Suppose there are d responses observed with some clinical events of interests, along with r covariates, for each member in a family. We concentrate on nuclear family structure for simplicity. In practice we only observe a subset of the responses and covariates for each patient in a family. Let $y_i=(y_{if},y_{im},y_{is})(i=1,\ldots,n)$ be the vector of responses for the i-th nuclear family, with corresponding covariates $x_i=(x_{if},x_{im},x_{is})$ $(i=1,\ldots,n)$. Here y_{if} a $d_{if}(<d)$ dimensional vector of responses of the

Citation: Yuan A and Ahn J. Some Genetic Regression Models for Multiple Quantitative End Points Data. Austin Biom and Biostat. 2016; 3(1): 1029.

father in the i-th family, which belongs to a d_{if} dimensional subspace of the d-dimensional space, with a response non-missing indicator vector I_{if} and covariate non-missing indicator vector J_{if}. Similarly, y_{im} denotes a vector of responses of the mother in the i-th family and $y_{is}=(y_{i1},\ldots,y_{ibi})$ is for offspring with each of the y_{ij} s has the same data structure as that for y_{if}. For example, there are three responses to be observed, we have the first and third on the father, then $I_{if}=(1,0,1)$, and $d_{if}=|I_{if}|=2$ is its dimension or cardinality. If there are total of five covariates in the design, and we may only have the first, second, fourth covariates for the father, then $J_{if}=(1,1,0,1,0)$, and $r_{if}=|J_{if}|=3$ is its dimension. We assume random censorship. Let $\delta_{if}=(\delta_{if1},\ldots,\delta_{ifdif})$ be the censoring indicator of y_{if}, i.e. $\delta_{ifj}=1$ if y_{ifj} is uncensored, and $\delta_{ifj}=0$ otherwise. Similar notations are used for the mother. For the off springs, $y_{is}=(y_{i1},\ldots,y_{ki})$ denote the response vector, with y_{ij} be d_{ij} dimensional observation for the j^{th} sib, with response configuration I_{ij} and covariate configuration j_{if}, $(j=1,\ldots,k_i)$. Let $d_i=(d_{if},d_{im},d_{is})$, $I_i=(I_{if},I_{im},I_{is})$, $J_i=(J_{if},J_{im},J_{is})$, $\delta_i=(\delta_{if},\delta_{im},\delta_{is})$ and $\delta_i=(C_{if},C_{im},C_{is})$. The complete data information consists of $Z_i=(y_i,x_i,I_i,J_i,\delta_i)$, $(i=1,\ldots,n)$. Let $\chi(\cdot)$ be the indicator function, i.e. $\chi(g_{ir}=s)=1$ if the genotype of the r^{th} individual is s and zero otherwise. Let π_s be the population proportion of the S-th genotype, $t(s|i,k)$ be the transmission probability of a off string's genotype s given the parents' genotype (i,k), and θ be the collection of all the parameters, including the αs and βs in the mean and the parameters in the within and between individual covariance matrices and the genotype frequencies π_ks and transmission probabilities $t(s|i,k)_s$. With unobserved genotypes, the computation is a serious challenge because of the mixture nature of the model. Let $f(y_r|\theta)$, $F(y_r|\theta)$, and $S(y_r|\theta)=1-F(y_r|\theta)$ be the density function, distribution function and survival function of Y_r, respectively. They may be described by its genotype g_r through the mean function specification. To simplify model specification, we assume random mating so that father and mother can be viewed as independent in most cases. The case of non-random mating or within parent's dependence can be treated similarly with more involved notations.

Note that there is within family dependence but independence among different families. We assume the genotypes of each patient are unobserved; the case of observed genotypes is automatically covered and simpler. Now we describe the methods in some common settings below.

Parametric model

Let the genotypes at the locus of interest be coded as $1,\ldots,k$. We first consider the case of no missing record and censoring. The regressive model assumes $y_{ir}=\mu(g_{ir})+\in_{ir}$, $(i,=1,\ldots,n;\ r=f,m,1,\ldots,n)$,

where $\mu(g_{ir})=\mu_0+\alpha\chi(g_{ir})+\beta x_{ir}$ is the mean phenotypic value, $\alpha=(\alpha_1,\ldots,\alpha_k)'$, $\chi(g_{ir})=(\chi(g_{ir}=1),\ldots,\chi(g_{ir}=k)$, μ_0 is the intercept vector,. The residual error term $\in_{ir}$ is a d dimensional random vector where they are independent across i but dependent across r.

In the case of multivariate familial quantitative response data, under the commonly used Elston-Stewart [15] algorithm or its variants, the likelihood of the observation y_i for the i-th nuclear family is

$$L_i(y_i|\theta)=\sum_k(\pi_k f(y_f|\theta,g_f=k)\sum_j(\pi_j f(y_{im}|\theta,g_{im}=j)K(\theta,k,j)$$
$$\times\prod_{l=1}^{b_i}\sum_r t(r|k,j)f(y_i|\theta,g_f=k,g_{im}=j,g_i=r))),(1)$$

where typically the density is assumed multivariate normal with covariate matrix Σ, $f(\cdot|\theta,k)$ is the density for residual $\in_i$ with genotype g_i in the mean vector specification, $f(y_l|\theta,i,j,r)$ is the density for residual $\in_r$ with genotype g_r and with adjusted mean and variance given by

$$\mu(g_{il}=r)-\Omega\Sigma^{-1}[(y_{if}-\mu(g_f=k))+(y_m-\mu(g_m=j))]\ and\ \Sigma-\Omega\Sigma^{-1}\Omega$$

where $\Sigma=Cov(Y_f,Y_f)$ and $\Omega=Cov(Y_f,Y_m)$; $K(\theta,I,j)$ is a quantity that depends on the parents' genotypes and the mean [18]. It is well known that when the number of genotypes is relatively large, this model is computationally inefficient [19]. Proposed a computational more efficient model. In light of [19], let $G_f=G_m=(\pi_1,\ldots,\pi_L)$, then the joint likelihood for the i-th family is written as

$$L_i(y_i|\theta)=(\sum_k\pi_k f(y_f|\theta,g_f=k))(\sum_j\pi_j f(y_{im}|\theta,g_{im}=j)$$
$$K(\theta)\prod_{l=1}^{b_i}\sum_r T(r)f(y_i|G,\theta,g_i=r),(2)$$

where $K(\theta)=\sum_{i=1}^k\sum_{j=1}^k K(\theta,i,j)$,

$$T(r)=T(r|G_f,G_m)=\sum_{i=1}^k\sum_{s=1}^k t(r|i,s)P(g_f=i)P(g_m=s)=\sum_{i=1}^k\sum_{s=1}^k t(r|i,s)\pi_i\pi_s,$$

In the mean specification $f(y_l|G,\theta,r)$ is the density of residual $\in l$ with genotype $gl=r$ with adjusted mean and variance given by

$$\mu(g_{il}=r)-\Omega_p'\Sigma_p^{-1}(y_p-\mu_p)\quad and\quad \Sigma-\Omega_p'\Sigma_p^{-1}\Omega_p,$$

where $y_p=(y_f,y_m)$,$\mu_p=(\mu_f,\mu_m)$, $\mu_f=\sum_{i=1}^k\pi_i\mu_f(g_f=i)$, similarly for ; μ_m

$$\Omega_p=\begin{pmatrix}\Omega\\ \Omega\end{pmatrix}\quad \Sigma_p=\begin{pmatrix}\Sigma & \Omega\\ \Omega & \Sigma\end{pmatrix}.$$

In comparison, model (1) has three layers of mixing (summation) corresponding to $b_i k^3$ function evaluations that grow exponentially with the number of genotypes. On the other hand, model (2) has only one layer of mixing in three factors each, or $(b_i+2)k$ function evaluations that are linearly proportional the number of genotypes. The reduction of computation will be more significant for multiple loci case.

Here we extend this model in the case of censoring and partial observation. In this case, the mean is modeled as

$$\mu(g_{ir})=I_{ir}\odot(\mu_0+\alpha\chi(g_{ir})+\beta)J_{ir}\odot x_{ir},-0.6cm\quad(3)$$

where the operation $I_{ir}\odot\mu_0$ means the projection of μ_0 onto the subspace corresponding to the nonzero elements of I_{ir}, similarly for $I_{ir}\odot\alpha\chi(g_{ir})$ and $J_{ir}\odot x_{ir}$. The corresponding error is now $I_{ir}\odot\varepsilon_{ir}$.

Recall that in the case of 1-dimensional observation without genetic implementation, the likelihood for an observation y_i with a censoring indicator δ_i is

$$L_i(y_i|\theta)=f(y_i|\theta)^{\delta_i}S(y_i|\theta)^{1-\delta_i}.$$

To extend this to our situation, for any dimension indicator I_i, and any d-variable function $v(\cdot)$, let $I_i\odot v(\cdot)$ be the marginal version of $v(\cdot)$ with respect to the non-zero entry of I_i and $\delta_i I_i\odot v(\cdot)=\delta_i\odot(I_i\odot v(\cdot))$. Let $1-\delta_i$ be the indicator with the same length of δ_i but with 0 and 1 reversed. The full likelihood is

$$L(z|\theta)=\prod_{i=1}^n(\sum_{j=1}^k\pi_j\delta_{if}I_{if}\odot f(y_{if}|\theta,j))(\sum_{j=1}^k\pi_j(1-\delta_{if})I_{if}\odot S(y_{if}|\theta,j))$$
$$\times(\sum_{j=1}^k\pi_j\delta_{im}I_{im}\odot f(y_{im}|\theta,j))(\sum_{j=1}^k\pi_j(1-\delta_{im})I_{im}\odot S(y_{im}|\theta,j))$$
$$\times\prod_{j=1}^{b_i}(\sum_{r=1}^k T(r)\delta_{ij}I_{ij}\odot f(y_j|G,\theta,r))(\sum_{r=1}^k T(r)(1-\delta_{ij})I_{ij}\odot S(y_j|G,\theta,r)).(4)$$

Here extra caution should be taken since the observation vector

from each individual may vary in dimensions and sub-spaces. For a d-dimensional vector v, let $\delta_{ij}I_{ij}\odot v$ be its margin with respect to $\delta_{ij}I_{ij}$; and for a d-dimensional matrix A, $\delta_{ij}I_{ij}\odot A\odot I_{il}\delta_{il}$ denote the sub-matrix by the rows corresponding to the non-zero entry of $\delta_{ij}I_{ij}$ and columns corresponding to the non-zero entry of $\delta_{il}I_{il}$. In particular, $\delta_{ij}I_{ij}\odot f(y_j\mid G,\theta,r)$ has adjusted mean given by

$$\delta_{ij}I_{ij}\odot\mu(g_{ij}=r)-(\delta_{ij}I_{ij}\odot\Omega'_p\odot I_{ip}\delta_{ip})(\delta_{ip}I_{ip}\odot\Sigma_p^{-1}\odot I_{ip}\delta_{ip})(\delta_{ip}I_{ip}\odot(y_p-\mu_p)),$$

and adjusted variance matrix given by

$$\delta_{ij}I_{ij}\odot\Sigma\odot I_{ij}\delta_{ij}-(\delta_{ij}I_{ij}\odot\Omega'_p\odot I_{ip}\delta_{ip})(\delta_{ip}I_{ip}\odot\Sigma_p^{-1}\odot I_{ip}\delta_{ip})(\delta_{ip}I_{ip}\odot\Omega_p\odot I_{ij}\delta_{ij}),$$

where $I_{ip}=(I_{if},I_{im})$. The corresponding adjustment in $(1-\delta_{ij})I_{ij}\odot S(y_j\mid G,\theta,r)$ is made. The parameter θ is estimated by its MLE $\hat{\theta}$ under (3), along with the restriction $\sum_{j=1}^{k}\pi_j=1$.

Semiparametric model

For censored data, a commonly used semi parametric regression model is Cox's proportional hazards model [20,21]. In the univariate case, let $y_{(3)}<y_{(4)}<\ldots<y_{(n)}$ be the ordered observations of $y_1,\ldots,y_n$ (assume no ties for simplicity), $x_{(i)}$ and $\delta_{(i)}$ be the associated quantities, for $y_{(i)}$, of the xi's and δi's. Let $R_{(i)}$ be the i-th risk set, the set of all individuals who are still under study at the 'time' just prior to $y_{(i)}$, U be the set of all uncensored individuals, and

$$\lambda(y\mid x,\theta)=\frac{f(y\mid x,\theta)}{1-F(y\mid x,\theta)}$$

be the hazard function. The proportional hazards model has a form of $\lambda(y/x,\theta)=h(\beta'x)\lambda_0(y)$-0.2cm for some known positive function $h(\cdot)$, and unspecified baseline hazard rate $\lambda_0(\cdot)$, which implies that the distribution belongs to the Lehmann family [4] $1\text{-}F=(1\text{-}F_0)^{\gamma}$ for some $F_0(\cdot)$ and $\gamma>0$. Under these assumptions, the conditional likelihood (partial likelihood [20,21]; marginal rank likelihood, [4]) is

$$L_c(y\mid\theta)=\prod_{i\in U}\frac{h(\beta'x_{(i)})}{\sum_{j\in R_{(i)}}h(\beta'x_{(j)})},$$

where the estimate of θ is the MLE $\hat{\theta}$ under $L_c(y|\theta)$. The optimality property of $\hat{\theta}$ is studied extensively. In the case of multivariate observations, various extensions of this method have focused on each marginal distribution and Markov chain Monte Carlo on the margins [22]. Proposed a multivariate extension of the proportional hazards model, or frailty model, which is equivalent to an exponential specification of the joint survival function [23]. Proposed a class of multivariate failure time distributions, including a multivariate version of Cox's proportional hazards model, in which the within family dependence is modeled by a common latent variable with a known parametric distribution given that all the family members are independent. Then the joint distribution is obtained by taking expectation of the conditional one. All these frailty models assume that there is a shared common dependent latent variable. This assumption basically requires that the distribution be interchangeable among the involved individuals. This is reasonable for some familial data but not generally true. Other existing multivariate proportional hazards models [24-26] are similar in nature. Here we model the within family dependence in a manifest way to be desirable for our genetic analysis. We adopt a successive conditional version of the proportional hazards model where we assume a special semi parametric form of the survival function in order to evaluate the conditioning in closed form easily. More specifically, in our multivariate proportional hazards model, we assume $h(\cdot)$ and $\lambda_0(\cdot)$ are functions of d-variates each. Let $y_{i,(3)},\ldots,y_{i,(ni)}$ be the ordered observations on the i-th variable $(i=1,\ldots,d)$, define $x_{i,(j)},\delta_{i,(j)},R_{i,(j)}$, and U_i accordingly. Note there are structures in $h(\cdot)$ through the dependent effects among the covariates. Recall $\beta'x$ is a d-vector, let

$$h(\beta'x)=e^{-\frac{1}{2}(\beta'x)'\Omega\beta'x},$$

where Ω is the within individual covariance matrix. Then $h(\cdot)$ behaves as a d-variate normal density, and its marginal and conditional versions are well defined and in closed forms, although it is not a proper density function. We need the successive 'conditioning' form of $h(\cdot)$ to apply the proportional hazards method. Specifically, let w_{ij} be the j-th diagonal element of Ω where Ω_j be the upper-left j-dimensional sub-matrix of it, a_j be the first j elements in the j-th column; $[\beta'x]_j$ be the first j components of $\beta'x$,$h_{j+1|j}(\cdot)$, be the conditional version of covariates $[\beta'x]_{j+1}$ given $[\beta'x]_j$. Then $h_{j+1|j}(\beta'x)$ is a univariate normal kernel with mean $-a_j'\Omega_j^{-1}(\beta'x)_j$ and variance $\omega_{jj}-a_j'\Omega_j^{-1}a$, and

$$h(\beta'x)=h_{1|0}([\beta'x]_1)\prod_{j=1}^{d-1}h_{j+1|j}([\beta'x]_{j+1}).(5)$$

Thus, without mixing over gene, for singleton multivariate observations, the joint conditional likelihood is

$$L_c(z\mid\theta)=\prod_{i\in U_1}\frac{h([\beta'x_{(i)}]_1)}{\sum_{l\in R_{1(i)}}h([\beta'x_{(l)}]_1)}\prod_{j=1}^{d-1}\prod_{i\in U_j}\frac{h_{j+1|j}([\beta'x_{(i)}]_j)}{\sum_{l\in R_{j,(i)}}h_{j+1|j}([\beta'x_{(l)}]_j)}.$$

Now for the case of nuclear family, inspired by (4), we assume $h(\cdot)$ has the form

$$h(\mu_i)=\sum_{j=1}^{k}\pi_j h_{I_f}(\mu(g_j,I_f))\sum_{j=1}^{k}\pi_j h_{I_m}(\mu(g_j,I_m))\prod_{l=1}^{b_i}\sum_{j=1}^{k}T(j)h_{I_j}(\mu(g_j,I_j)),(6)$$

Treat $h(\cdot)$ as a 'density'. Recall $\mu_i=(\mu_{if},\mu_{im},\mu_{i1},\ldots,\mu_{ibi})$. The conditioning $[\mu_i]_{j+1}|[\mu_i]_j$ can be applied component-wise, i.e.

$$[\mu_i]_{j+1}|[\mu_i]_j=([\mu_{if}]_{j+1}|[\mu_{if}]_j,\ [\mu_{im}]_{j+1}|[\mu_{im}]_j,\ [\mu_{i1}]_{j+1}|[\mu_{i1}]_j,\ldots,\ [\mu_{ibi}]_{j+1}|[\mu_{ibi}]_j)$$

Now we have

$$h_{j+1|j}([\mu_i]_{j+1})=(\sum_{j=1}^{k}\pi_j h_{I_f|0}([\mu(g_j,I_f)]_1)\prod_{j=1}^{|I_f|-1}h_{I_f,j+1|j}([\mu(g_j,I_f)]_j)$$
$$\times(\sum_{j=1}^{k}\pi_j h_{I_m|0}([\mu(g_j,I_m)]_1)\prod_{j=1}^{|I_m|-1}h_{I_f,j+1|j}([\mu(g_j,I_m)]_j)$$
$$\times\prod_{l=1}^{b_i}(\sum_{j=1}^{k}T(j)h_{I_l|0}([\mu(g_j,I_i)]_1)\prod_{j=1}^{|I_l|-1}h_{I_l,j+1|j}([\mu(g_j,I_i)]_j)).(7)$$

In (7), $[\mu(g_j,I_{ir})]1$, means the first component of $\mu(g_j,I_{ir})$ in I_{ir}, and $|I_{ir}|$ denote its cardinality $(r=f,m,1,\ldots,b_i)$. Now, the conditional likelihood is

$$L_c(z\mid\theta)=\prod_{i\in U_1}\frac{h([\mu_{(i)}]_1)}{\sum_{l\in R_{1(i)}}h([\mu_{(i)}]_1)}\prod_{j=1}^{d-1}\prod_{i\in U_j}\frac{h_{j+1|j}([\mu_{(i)}]_j)}{\sum_{l\in R_{j,(i)}}h_{j+1|j}([\mu_{(i)}]_j)},(8)$$

where $h_{j+1|j}([\mu_{(i)}]j$ is given by (7). The MLE $\hat{\theta}$ of θ is obtained under (8).

Nonparametric model

For univariate censored data, [27,28] considered a class of estimators, including the weighted least squares estimators, for censored data. Here the weights are determined by the ordered statistics of the observations and the associated censoring indicators, and are derived from the empirical survival function, i.e., the Kaplan-Meier product limit estimator [29-32]. Formulated the multivariate Kaplan-Meier estimator. Using the product integral, the mathematical

expressions are quite involved. So instead of choosing the weights according to the multivariate Kaplan-Meier estimator, we use the nonparametric locally weighted least squares method, also called locally linear regression smoothers [33,34]. Let Y and X be the d and J-dimensional random vectors corresponding to the full observation and the covariates for an individual. Let $\mu(x)=E(Y|X=x)$ denote the regression function. In the univariate observation case, the locally linear estimator $\hat{\mu}(x)$ of $\mu(x)$ is first to find $\hat{a}$ and $\hat{b}$ to minimize

$$\sum_{i=1}^{n}(y_i - a - b(x-x_i)^2 K(\frac{x-x_i}{h_n})$$

Where $K(\cdot)$ is a kernel function, h_n is the bandwidth, and $\hat{\mu}(x)=\hat{a}$. In our case, keep the notations in section 1. We choose the kernel to be the J-dimensional standard normal density $\varphi(\cdot)$, and $(\cdot)$ to be its distribution function. To simplify the expression of the likelihood, let

$$\tilde{y}_{ifr} = y_{if} - I_{if} \odot \mu(r), \quad \tilde{x}_{if} = J_{if} \odot (x - x_{if}), \quad \bar{x}_{if} = (1-J_{if}) \odot (x - x_{if}),$$

and similarly for $\tilde{y}_{imr}$, $\tilde{y}_{ijr}$, $\tilde{x}_{im}$, $\bar{x}_{im}$, $\tilde{x}_{ij}$, and $\bar{x}_{ij}$ $(j=1,..,b_i)$. To estimate $\hat{\mu}(x)$, inspired from the univariate locally linear estimator and (4), we first find $(\hat{\mu}_0(x), \hat{\alpha}, \hat{\beta}, \hat{\pi})$ to minimize

$$\begin{aligned}
&\sum_{i=1}^{n}(\sum_{r=1}^{k}\pi_r^2 \tilde{y}'_{ifr}(I_{if}\odot\Omega\odot I_{if})\tilde{y}_{ifr}J_{if}\odot\varphi(\frac{\tilde{x}_{if}}{h_n})(1-(1-J_{if})\odot\Phi(\frac{\bar{x}_{if}}{h_n})\\
&+\sum_{r=1}^{k}\pi_r^2 \tilde{y}'_{imr}(I_{im}\odot\Omega\odot I_{im})\tilde{y}_{imr}J_{im}\odot\varphi(\frac{\tilde{x}_{im}}{h_n})(1-(1-J_{im})\odot\Phi(\frac{\bar{x}_{im}}{h_n})\\
&+\sum_{j=1}^{b_i}\sum_{r=1}^{k}T^2(r)\tilde{y}'_{ijr}(I_{ij}\odot\Omega\odot I_{ij})y_{ijr}J_{ij}\odot\varphi(\frac{\tilde{x}_{ij}}{h_n}|\tilde{x}_p)(1-(1-J_{ij})\odot\Phi(\frac{\bar{x}_{ij}}{h_n}|\bar{x}_p))
\end{aligned} \quad (9)$$

Where Ω is the within individual variance matrix, $\phi(\cdot|\tilde{x}_p)$ and $\Phi(\cdot|\bar{x}_p)$ are the adjusted quantities as those in (4). And $I_{ir}\odot\Omega\odot I_{ir}$ is the sub-matrix of Ω with rows and columns corresponding to the non-zero elements of I_{im}.

We estimate

We estimate Ω by $\hat{\Omega}=(\hat{\omega}_{\mathfrak{s}})$ with

$$\hat{\omega}_{\mathfrak{s}} = \frac{1}{n_{\mathfrak{s}}-1}\sum_{l=1}^{n_{\mathfrak{s}}}(z_{\mathfrak{l}}-\bar{z}_l)(z_{\mathfrak{t}}-\bar{z}_s) \quad i,j=1,...,d,$$

where n_{rs} is the total number of individuals with non-missing (r,s)-th components, z_rs are the rearrangement of the r-th component of $y_{it}-\sum_{j=1}^{k}\pi_k I_{it}\odot\mu(j)$ $(t=f,m,1,\ldots)$ for which the (r,s)-th components are non-missing.

Let $(\hat{\mu}(r,x), \hat{\pi})$ be the minimize of (9), where the full $\hat{\mu}(r,x)$ depends on the genotype r and the point value x. It has the intercept term $\hat{\mu}_0(x)$ (recall (3)), and $\hat{\mu}(x)$ is approximated by setting $\hat{\mu}(x)=\hat{\mu}_0(x)$. Direct computation of $(\hat{\mu}(r,x), \hat{\pi})$ in (7) is not easy, instead we use an iterative procedure as in the following steps.

Select starting values $\pi^{(0)}$ for π. With this $\pi^{(0)}$, compute $\Omega^{(0)}$, and $T^{(0)}(r)$s. Let $\eta=(\mu_0,\alpha,\beta)$ be the full representation of the regression parameters, $X_{ir}(r=f,m,1,\ldots,n_r)$ be the corresponding design matrix for the r-th individual in the i-th family. In iterations 1-m do the following

(i) Fix $\pi^{(i)}$, $\Omega^{(i)}$ and $T^{(i)}(r)$ s minimize (7) with respect to η to get

$$\begin{aligned}
\eta^{(i)} &= (\sum_{i=1}^{n}\sum_{r=1}^{k}\sum_{l}\varpi_{lr}^{(i)2}X_{il}X'_{il}J_{il}\odot\varphi(\frac{\tilde{x}_{il}}{h_n})(1-(1-J_{il})\odot\Phi(\frac{\bar{x}_{il}}{h_n})))^{-1}\\
&\times\sum_{i=1}^{n}\sum_{r=1}^{k}\sum_{l}\varpi_{lr}^{(i)2}y_{il}X'_{il}J_{il}\odot\varphi(\frac{\tilde{x}_{il}}{h_n})(1-(1-J_{il})\odot\Phi(\frac{\bar{x}_{il}}{h_n})),
\end{aligned} \quad (10)$$

where $\varpi_{\mathfrak{r}}^{(i)}=\pi_r^{(i)}$ for l=f,m, and $\varpi_{\mathfrak{r}}^{(i)}=T^{(i)}(r)$ for $l=1,\ldots,n_i$. (ii) Fix $^{(i)}$, minimize (7) with respect to π, with the constraint $\sum_{j=1}^{k}\pi_j=1$, to get

$$\pi_r^{(i+1)} = \frac{\sum_{i=1}^{n}\sum_{l=f,m}\tilde{y}_{ilr}(I_{il}\odot\Omega\odot I_{il})\tilde{y}_{ilr}J_{il}\odot\varphi(\frac{\tilde{x}_{il}}{h_n})(1-(1-J_{il})\odot\Phi(\frac{\bar{x}_{il}}{h_n}))}{\sum_{i=1}^{n}\sum_{l=f,m}\sum_{r=1}^{k}\tilde{y}_{ilr}(I_{il}\odot\Omega\odot I_{il})\tilde{y}_{ilr}J_{il}\odot\varphi(\frac{\tilde{x}_{il}}{h_n})(1-(1-J_{il})\odot\Phi(\frac{\bar{x}_{il}}{h_n}))}, \quad (11)$$

(r=1,…,k) (11)

and update $\Omega^{(i+1)}$, and $T^{(i+1)}(r)$ with $^{(i+1)}$.For some pre-specified $\epsilon>0$, when the relative errors

$$\frac{|(\mu^{(m)}(r,x)\ \pi^{(m)})-(\mu^{(m-1)}(r,x)\ \pi^{(m-1)})|}{|(\mu^{(m-1)}(r,x)\ \pi^{(m-1)})|}\le\varepsilon$$

we stop the process at the last step m, and take $(\hat{\mu}(r,x)\ \hat{\pi})=(\mu^{(m)}(r,x)\ \pi^{(m)})$.

For arbitrary kernel and reasonably chosen band width h_n, various asymptotic results are established in case of standard non-mixture data. We conjecture that similar results will hold under some regularity conditions.

Lastly, the bandwidth determines the smoothness of the estimate. Interesting research that addresses the crucial problem of bandwidth selection can be found in [35]. There are considerable literatures for automatic methods that attempt to minimize a lack-of-fit criterion such as an integrated squared error. But most of the methods provide an optimal h_n determined by some unknown quantities. For simplicity, let $k=|J_{ij}|$ be the dimension of the observed covariate of the j-th $(j=f,m,1,\ldots,n_i)$ individual in the i-th family, for the corresponding kernel, we choose $h_n=Cn^{-1/(k+1)}$, for some constant $C>0$, and C can be selected through numerical trial.

Variance components model

As an alternative to the mixture models considered above, the Variance Components (VC) model [36,37] has received much attention recently due to very efficient in computation as well as relatively robustness to model misspecification [38-46].

Let y_i be the trait vector of the i-th individual in the family, in case without censoring and missing records, the commonly used VC model describing the trait value is

$$y_i=\mu+g_i+G_i+\eta x_i+e_i,$$

Where μ is the overall mean, g_i is the unobserved random vector of major gene effects at the trait locus with alleles A and B, G_i is the unobserved polygenic effects vector, the η_j's are effects associated with the covariates x_{ij}'s, and e_i is the residual random error vector. The usual assumption is that g_i,G_i and e_i are uncorrelated and $E(g_i)=E(G_i)=E(e_i)=0$. When missing records are present, the model is modified as

$$y_i = I_i\odot(\mu+g_i+G_i+\eta)J_i\odot x_i + I_i\odot e_i. \quad (12)$$

In this model, the parameters of interests are specified in the family variance matrix, thus computation can be carried out efficiently without the multiple mixing. Let y_k,π_k and Ω_k be the observation, its mean and variance matrix of the k-th family. We can define I_k, δ_k and J_k accordingly. The commonly used model for quantitative traits is the multivariate normal distribution, thus the total likelihood is

$$L(z|\theta)=\sum_{k=1}^{K}\delta_k I_k\odot\varphi(y_k-\mu_k|\Omega_k)(1-\delta_k)I_k\odot\Phi y_k-\mu_k|\Omega_k).$$

Here ϕ is the distribution function of the normal distribution with mean 0 and variance Ω.

The key lies in the specification of the variance matrices Ω_ks, which we illustrate in the following settings.

In the simplest case of Hardy-Weinberg equilibrium among locus alleles without linkage to marker, and without censoring and missing records, the covariance matrix between individuals *i* and *j* of a given family can be found, for example, in [38]. Modified to our case, it is

$$Cov(Y_{ki},Y_{kj})=\begin{cases} I_i\odot(\sigma_a^2+\sigma_d^2+\sigma_G^2+\sigma_e^2)\odot I_j & if\ \ i=j \\ 2\Phi_{ij}I_i\odot(\sigma_a^2+\Delta_{7ij}\sigma_d^2+2\Phi_{ij}\sigma_G^2)\odot I_j, & if\ \ i\neq j \end{cases} \quad (13)$$

where σ_a^2 is the additive genetic variance matrix due to the locus, σ_d^2 is the dominant genetic variance matrix, $\Phi_{ij}=\Delta_{7ij}/2+\Delta_{8ij}/4$ is the kinship coefficient between individuals *i* and *j* [47], and $\Delta_{7ij},\Delta_{8ij},\Delta_{9ij}$, etc. are the condensed kinship coefficient of Jacquard [48], between individuals *i* and *j*.

In the more general Hardy-Weinberg disequilibrium case, let *f* be the within population inbreeding coefficient *f* at the trait locus [49-51]. Introduced CV model in this case, which modified in our case is

$$Cov(Y_i,Y_j\mid f)=\begin{cases} I_i\odot((1+\frac{f}{2})\sigma_a^2+(1-f)\sigma_d^2+f\sigma_0^2+\sigma_G^2+\sigma_e^2)\odot I_j, & if\ \ i=j \\ I_i\odot(\Delta_{7ij}\gamma_7(f)+\Delta_{8ij}\gamma_8(f)+2\Phi_{ij}\sigma_G^2)\odot I_j, & if\ \ i\neq j \end{cases} \quad (14)$$

where $\gamma_l(f)$s are matrices determined by σ_a^2, σ_d^2 and *f* etc., see there for details.

In the case of linkage to marker with both Hardy-Weinberg and linkage equilibrium, the covariance in our case can be specified based on that of, for example [40], in the same way as above. In the case of linkage to marker with either one or both Hardy-Weinberg and linkage disequilibrium, the covariance in our case can be specified based on that of [51], in the same way.

Competing risks

Now suppose that the response y_i is the failure time and only the failure for one of the *d* diseases is observed for each individual. For the *i*-th family, the data have the form (y_i,δ_i,j_i,x_i), where $y_i=(y_{if},y_{im},y_{i1},\cdots,y_{ibr})$, similarly for δ_i,j_i and x_i, where j_i is the observed disease type indicator. For example if the observed disease for the father is type 2, then $j_{if}=2$. Given the data (y_i,δ_i,j_i,x_i)s, we like to investigate the objective of interests for each of the *d* disease. This problem is that of the competing risks. Note here the response for each individual is one-dimensional, and hence the corresponding quantities have simple notations. We are interested in the genetic regression analysis for the competing risks. We use a variant of the proportional hazards model. The mean of the *j*-th type, *r*-th member of the *i*-th family is specified as

$$\mu_j(g_{ir})=I_{ir}\odot(\mu_{0j}+\alpha_j\chi(g_{ir})+\beta_i)J_{ir}\odot x_{ir},\quad (r=f,m,1,...,b_i).$$

For a reasonably chosen function $h(\cdot)$, we specify

$$h(\mu_i)=(\sum_{l=1}^{k}\pi_l h(\mu_{j_f}(g_l)))(\sum_{l=1}^{k}\pi_l h(\mu_{j_m}(g_l)))\prod_{l=1}^{b_i}(\sum_{s=1}^{k}T(s)h(\mu_{j_l}(g_s))).$$

More convenient below is to use the notations

$$h_r(\mu_i)=\sum_{l=1}^{k}\pi_l h(\mu_{j_r}(g_l)),\quad (r=f,m)$$

and

$$h_r(\mu_i)=\sum_{s=1}^{k}T(s)h(\mu_{j_r}(g_s)\mid y_p)\quad (r=1,...,b_i)$$

Let $y_{j1}<\ldots<y_{jkj}$ be the k_j failures of type $j(j=1,\ldots,d)$, $R(y_{ji})$,be the risk set at y_{ji}, the partial likelihood is

$$L(y\mid\theta)=\prod_{j=1}^{d}\prod_{i=1}^{k_j}\frac{h_{i_j}(\mu_j)}{\sum_{l\in R(y_j)}h_{i_l}(\mu_j)}.(15)$$

Asymptotic heuristic

For IID data, various asymptotic results can be obtained. The results from the score function, the likelihood ratio statistic, and the MLE are equivalent. These results can be used to establish confidence intervals or hypothesis testing, etc. for θ. Here we are more interested in using the MLE. For general dependence model, usually the treatment is non-standard. But for our model, since the log-likelihood is in the form of several additive pieces, standard method can be used to derive the asymptotic distribution of the MLE. Let $z_i=(y_i,\delta_i)(i=1,\ldots,n)$. For the IID data, it is well known that under mild regularity conditions, the MLE $\hat{\theta}_n$ is strongly consistent and asymptotically distributed normal with mean at the true parameter value θ_0, and variance matrix given by the inverse of the Fisher information. Here the observations are unbalanced, the asymptotic variance is the Fisher information times a weight matrix. To derive it, we need some notations, and mainly concentrate on model (4).

Let N_{ip} be the total number of parents with the *i*-th measurement non-missing $(i=1,\ldots,d)$, N_{is}s be those for the siblings, *N* be the total number of individuals in the study, $\gamma_{Nir}=N_{ir}/N(r=p,s)$. Assume $lim_{N\to\infty}\gamma_{Nir}=\gamma_{ir}>0$ exists $(r=p,s;i=1,\ldots,d)$. Let Y_p and Y_j be general random vectors associated with a parent and sib respectively, and Δ_p and Δ_j be the corresponding random vectors.

For model (4), let

$$H(\theta\mid Y_p,Y_j)=\Delta_p\odot\log(\sum_{l=1}^{k}\pi_l f(Y_p\mid\theta,l))+(1-\Delta_p)\odot\log(\sum_{l=1}^{k}\pi_l SY_p\mid\theta,l))$$
$$+\Delta_j\odot\log(\sum_{l=1}^{k}T(l)f(Y_j\mid G,\theta,l))+(1-\Delta_j)\odot\log(\sum_{l=1}^{k}T(l)S(Y_j\mid G,\theta,l)) \quad (16)$$

here we use the notation Δ_p $v(\cdot)$ to represent the marginal version of $v(\cdot)$ corresponding to the non-zero components of Δ_p. The Fisher information matrix is

$$I(\theta)=-E(\frac{\partial^2 H(\theta\mid Y_p,Y_j)}{\partial\theta'\partial\theta}).(17)$$

The above expectation is more involved than it looks, since that involves summations of all possible combinations of non-zero elements of it with respect to Δ_r, and also the unknown distribution of it. Instead, an empirical version of it has a known form

$$I_N(\theta)=-\frac{1}{N}\frac{\partial^2\log L(y\mid\theta)}{\partial\theta'\partial\theta}|_{\theta=\hat{\theta}_n},(18)$$

where $L(y\backslash\theta)$ is given by (4). At the true data generating parameter θ_0, I_N is strongly consistent for $I(\cdot)$. To obtain the weight matrix, we need to specify the parameter order in θ. We arrange the first *k-1* entry to be $\pi_1,\ldots,\pi_{k-1}$, next we arrange all the regression parameters for the first response variable,..., all the regression parameters for the last response variable, then all the independent parameters in the variance matrix Σ and covariance matrix Ω in the similar order. It is clear that, in the weight matrix *W*, for *(i,j)* corresponding to the first *k-1* components in θ, the weight should be $\gamma_p=\gamma_{1p}+\ldots+\gamma_{dp}$; for *(i,j)* corresponding to the *r*-th and the *l*-th regression parameters, the weight is $\sqrt{(\gamma_{rp}+\gamma_{rs})(\gamma_{lp}+\gamma_{ls})}$; for *(i,j)* corresponding to the *(a,b)*-th and the *(u,v)*-th variance or covariance, the weight is

$[(\gamma_{ap}+\gamma_{as})\ (\gamma_{bp}+\gamma_{bs})\ (\gamma_{up}+\gamma_{us})\ (\gamma_{vp}+\gamma_{vs})]^{1/4}$. Let θ_0 be the true unknown data generating parameter, $\xrightarrow{d}$ stands for convergence in distribution. Then, we have

$$\sqrt{N}(\hat{\theta}_N-\theta_0)\xrightarrow{d}N(0,I^{-1}(\theta_0)\otimes W)\ (19)$$

where $A\otimes B$ stands for the Kronecker product of matrices A and B. Since $I(\cdot)$ involves unknown quantities, equivalently

$$\sqrt{N}(\hat{\theta}_N-\theta_0):N(0,I_N^{-1}(\hat{\theta}_n)\otimes W_N)\ (20)$$

where W_N is W with the γs replaced by the γNs.

The above ideal applies to the other models in this paper, but the results will be more involved, and we only discuss them briefly.

For the proportional hazards model, even for the IID data case, the conditional likelihood looks much different from the full likelihood. Interestingly, the MLE from this model (under the assumption of correct model specification and some regularity conditions) has the same asymptotic distribution as that from the full likelihood [21,52]. For the proportional hazards model, it is noted [4] that the survival function can be written as

$S(y\backslash x,\beta)=-S_0(y)^{h(\beta' x)}$,

and $f(y\backslash x,\beta)=-dS(y\backslash x,\beta)/dy$. So to get the full log-likelihood (14), we need the estimate of $S_0(y)$ for the d-dimensional case [29,30]. Proposed the multi-dimensional generalization of the Kaplan-Meyer nonparametric estimator of $S(y)$, similar technique can be used here for the construction of $S_0(y)$. Then (18) continues to hold in this case. Due to technical involvement, we will not pursue the details here.

For the least squares estimator, since the weight involves the kernel and h_n, the treatment is different from those above, and in the case of full observation generally the asymptotic result is of the form

$$\sqrt{\boldsymbol{h}}_n^d(\hat{\theta}_n-\theta_0-h_n^2C-o_p(h_n^2))\xrightarrow{d}N(0,\Omega)$$

for some constant C and matrix Ω determined by the kernel and the true (unknown) data and censoring distributions [53,54]. In our case of partial observation, the above result holds with Ω replaced by $\Omega\otimes W$.

For the competing risks model, the structure is similar to that of the proportional hazards model. Here the response is one dimensional so that the survival function can be estimated by the Kaplan-Meyer estimator and the weight matrix W is the identity.

Discussion

We have considered several statistical methods, parametric, semi parametric and nonparametric models, for the genetic regression analysis of familial multiple endpoints data, with possible missing records. Here we only considered the case of nuclear families and the parameters are independent of time. The cases of arbitrary pedigrees and/or the time dependent parameter can be treated similarly. The variance components method can also be applied to the proportional hazards model and in the analysis of competing risks. There are some marginal models for the multiple endpoints data, which work well in practice. But we think the joint model is more appropriate when the within responses structure is important in the analysis. Another commonly used method to deal with the missing data is the EM algorithm [55], which can be implemented into the models considered here. But for the multiple endpoints data, the proportion of missing part is usually large; the EM algorithm may not be efficient. When the missing pattern is non-ignorable, more complicated approaches need to be considered to reduce potential biases. Hypothesis testing for parameters of interests can be conducted using the likelihood ratio statistics based on the parametric models. We only derived the basic forms of these models; more features can be implemented to them in particular applications.

References

1. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association. 1958; 53: 457-481.
2. Breslow N. Covariance analysis of censored survival data. Biometrics. 1974; 30: 89-99.
3. Fleming TR, Lin DY. Survival analysis in clinical trials: past developments and future directions. Biometrics. 2000; 56: 971-983.
4. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data, Wiley. 1980.
5. Andersen PK, Gill RD. Cox's regression model for counting processes: A large sample study. Annals of Statistics. 1982; 10: 1100-1120.
6. Arjas E. Survival models and martingale dynamics (with discussion). Scandinavian Journal of Statistics. 1989; 16: 177-225.
7. Flemming TR, Harrington DP. Counting Processes and Survival Analysis. Wiley, New York. 1991.
8. Andersen PK, Borgan O, Gill RD, Keiding N. Statistical Models Based on Counting Processes. Springer-Verlag. 1991
9. Printice R, Williams BJ, Peterson AV. On the regression analysis of multiple failure time data. Biometrika. 1981; 68: 373-379.
10. Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. Journal of the American Statistical Association. 1989; 84: 1065-1073.
11. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. Stat Med. 1992; 11: 167-178.
12. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med. 1996; 125: 605-613.
13. Zhang J, Quan H, Ng J, Stepanavage ME. Some statistical methods for multiple endpoints in clinical trials. Controlled Clinical Trials. 1997; 18: 204-221.
14. Wei LJ, Glidden DV. An overview of statistical methods for multiple failure time data in clinical trials. Stat Med. 1997; 16: 833-839.
15. Remington DL. Effects of genetic and environmental factors on trait network predictions from quantitative trait locus data. Genetics. 2009; 181: 1087-1099.
16. Knüppel S, Rohde K, Meidtner K, Drogan D, Holzhütter HG, Boeing H, et al. Evaluation of 41 candidate gene variants for obesity in the EPIC-Potsdam cohort by multi-locus stepwise regression. PLoS One. 2013; 8: e68941.
17. Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. Hum Hered. 1971; 21: 523-542.
18. Bonney GE. Compound regressive models for family data. Hum Hered. 1992; 42: 28-41.
19. Yuan A, Bonney GE. Two new recursive likelihood calculation methods for genetic analysis. Hum Hered. 2002; 54: 82-98.
20. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society. 1972; 34: 187-202.
21. Cox DR. Partial likelihood. Biometrika. 1975; 62: 269-276.
22. Clayton DG. A Monte Carlo method for Bayesian inference in frailty models. Biometrics. 1991; 47: 467-485.

23. Hougaard P. A class of multivariate failure time distributions. Biometrika. 1986; 73: 671-678.

24. Holt JD, Prentice RL. Survival analysis in twin studies and matched pairs experiments. Biometrika. 1974; 61: 17-30.

25. Oakes D. A model for association in bivariate survival data. Journal of the Royal Statistical Society. 1982; 44: 414-422.

26. Wild CJ. Failure time models with matched data. Biometrika. 1983; 70: 633-641.

27. Stute W. Consistent estimation under random censorship when covariable is present. Journal of Multivariate Analysis. 1993; 45: 89-103.

28. Stute W. Distributional convergence under random censorship when covariables are present.Scandinavian Journal of Statistics. 1996; 23: 462-471.

29. Dabrowska DM. Kaplan Meier estimate on the plane. Annals of Statistics. 1988; 16: 1475-1489.

30. Dabrowska DM. Kaplan Meier estimate on the plane: weak convergence, LIL and the bootstrap. Journal of Multivariate Analysis. 1988; 29: 308-325.

31. Gill R. Multivariate survival analysis. Theory of Probability and Its Applications. 1992a; 37: 18-31.

32. Gill R. Multivariate survival analysis: Part 2. Theory of Probability and Its Applications. 1992b; 37: 284-301.

33. Fan J. Locally linear regression smoothers and their minimax efficiencies. Annals of Statistics. 1993; 21: 196-216.

34. Ruppert D, Wand MP. Multivariate locally weighted least squares regression. Annals of Statistics. 1994; 22: 1346-1370.

35. Fan J, Gijbels I. Variable bandwidth and local linear regression smoothers. Annals of Statistics. 1992; 20: 2008-2036.

36. Fisher RA. The correlation between relatives on the supposition of Mendel inheritance. Trans Roy Soc. Edinb. 1918; 52: 399-433.

37. Harris DL. Genotypic Covariances Between Inbred Relatives. Genetics. 1964; 50: 1319-1348.

38. Lange K, Boehnke M. Extensions to pedigree analysis IV. Covariance components models for multivariate traits. Am J Med Genet. 1983; 14: 513-524.

39. Goldgar DE. Multipoint analysis of human quantitative genetic variation. Am J Hum Genet. 1990; 47: 957-967.

40. Amos CI. Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet. 1994; 54: 535-543.

41. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet. 1998; 62: 1198-1211.

42. Fulker DW, Cherny SS, Sham PC, Hewitt JK. Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet. 1999; 64: 259-267.

43. Amos CI, Gu X, Chen J, Davis BR. Least squares estimation of variance components for linkage. Genet Epidemiol. 2000; 19: S1-7.

44. Blangero J, Williams JT, Almasy L. Robust LOD scores for variance component-based linkage analysis. Genet Epidemiol. 2000; 19: S8-14.

45. Sham PC, Purcell S. Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. Am J Hum Genet. 2001; 68: 1527-1532.

46. de Andrade M, Guéguen R, Visvikis S, Sass C, Siest G, Amos CI. Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis. Genet Epidemiol. 2002; 22: 221-232.

47. Lange K. Statistics for Biology and Health. Springer. 1997.

48. Jacquard A. The genetic structure of populations. 1974.

49. Wright S. The genetical structure of populations. Ann Eugen. 1951; 15: 323-354.

50. Cockerham CC. Variance of gene frequencies. Evolution. 1969; 23: 72-84.

51. Yuan A, Chen G, Yang Q, Rotimi C, Bonney G. Variance components model with disequilibria. Eur J Hum Genet. 2006; 14: 941-952.

52. Tsiatis A. A large sample study of the estimate for the integrated hazard function in Cox's regression model for survival data. Technical Report No. 562. 1978; 172-180.

53. Fan J, Hu I, Truong Y. Robust nonparametric function estimation. Scandinavian Journal of Statistics. 1994; 21: 433-446.

54. Cai Z. Weighted local linear approach to censored nonparametric regression. Akritas MG, Politis DM. In: Recent Advances and Trends in Nonparametric Statistics. 2003; 217-231.

55. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. 1977; 39: 1-38.

Austin Biom and Biostat - Volume 3 Issue 1 - 2016
Submit your Manuscript | www.austinpublishinggroup.com

Citation: Yuan A and Ahn J. Some Genetic Regression Models for Multiple Quantitative End Points Data. Austin Biom and Biostat. 2016; 3(1): 1029.