FRAILTY MODELS WITH APPLICATIONS IN LINKAGE ANALYSIS

Benilton de Sá Carvalho[1], **Hildete P. Pinheiro**[2], and **Mariza de Andrade**[3]

[1]Department of Biostatistics, University of Johns Hopkins, USA
E-mail: bcarvalh@jhsph.edu

[2] Departamento de Estatística, Universidade Estadual de Campinas, Caixa Postal 6065, CEP 13083-970, Campinas, SP, Brazil.
E-mail:hildete@ime.unicamp.br

[3] Division of Biostatistics, Mayo Clinic Rochester, MN, USA
E-mail: mandrade@mayo.edu

Abstract

The paper presents the Frailty Model using the hazard function in a logistic form. The model developed here is a survival analysis method which aggregates linkage analysis role. The model structure is based on a Cox model extension. The hazard function has a parametric form, the logistic one (Mackenzie, 1996). The censure is defined using the current age and the age at onset , thus if one has the age at onset greater than the current age, then a censured observation is characterized. The additive frailties are constructed from a gamma densities and it incorporates genetic and environmental contributions on the trait of interest. The use of these techniques in a unified way has been shown to be efficient, once the age at onset is usually collected in genetic mapping studies, moreover it is shown to be associated to complex diseases. Then, this is a useful model, which is able to be applied to Cox model situations and genetic mapping cases. The model is adjusted by maximization of retrospective likelihood (Whittemore, 1996), using an iterative algorithm based on the Kuhn-Tucker equations. The proposed model is ascertained by the analysis of simulated data created using G.A.S.P. package and by comparisons with the results from the joint analysis with SAS and GeneHunter softwares.

1 Introduction

Usually in survival analysis it is assumed that there is independence among the observations. However, sometimes we are interested in study age-at-onset of a certain disease in individuals of the same family. In this case it is expected that the behavior of the observed age at onset among related individuals showed some similarity that are not observed between unrelated individuals. The common practice to analyze age at onset for unrelated individuals is the application of the Cox models (Cox, 1972). Recently, frailty models has been used as one solution to deal with the lack of independence observed age at onset among related individuals (Chang & Hsiung, 1995; Yau& McGilchrist, 1998). Among the frailty models, the gamma distribution is the most used to model the frailty variable (Clayton & Cuzick, 1985, Nielsen et al., 1992).

Frailty models are random effects models designed to work with censored survival data, in which the differences between homogeneous groups are modelled by adding a non-observable factor in the hazard function.

The main purpose of this article is to present a hazard function in the logistic form (Mackenzie, 1996) in frailty models and compare it with other models using frailties such as the Cox models using Martingale residuals with applications in linkage analysis. These models are applied to linkage analysis, as seen in Li and Zhong (2002).

2 Methods

In this section we will describe the different phases of the survival models using frailty. First, we will describe the classical gamma frailty model. Then, we will extend the gamma frailty model using the hazard function in the logistic form (Mackenzie, 1996). The main advantage of using a hazard function in the logistic form is that it is very flexible, in the sense that it can have different forms. The logistic hazard is a function of time and the logit of the baseline hazard is linear in time.

2.1 The Gamma Frailty

The gamma frailty models were first introduced by Clayton (1978) and Vaupel et al. (1979). They assumed models without covariates, where the frailty variable affects the hazard function in a multiplicative way. Later, Clayton and Cuzick (1985) extended the model proposed in Clayton (1978) to include covariates. The model has the form:

$$\lambda_{ij}(t \mid Z_i, \mathbf{X}_{ij}) = Z_i Y_{ij}(t) \alpha_0(t) \text{exp}(\boldsymbol{\beta}' \mathbf{X}_{ij}) \tag{1}$$

where $Y_{ij} = 1$ if the j-th individual of group i is at risk until t^- (time immediate before t) and zero otherwise; $\alpha_0(t)$ is the baseline hazard function (arbitrary); $\boldsymbol{\beta}$ is the vector of regression coefficients; $\mathbf{X}_{ij}$ is the vector of covariates for individual j of group i and $Z_i \sim Gamma(\nu, \eta)$ are independent with $\nu \geq 0$, $\eta \geq 0$ and $\nu = \eta = \theta^{-1}$, i.e., $\text{E}(Z_i) = 1$ and $\text{Var}(Z_i) = \theta$.

Oakes (1982) discusses a reparametrization of the model introduced by Clayton (1978) using bivariate life tables. A few years later, some authors (Self and Prentice, 1986, Nielsen et al., 1992 and Anderson et al., 1993) used the formulation of counting process to study the Gamma frailty models considering the semiparametric Cox Model.

2.2 Frailty Models for Affected Sibpairs

Let Z_j be the unobserved frailty and T_j be the random variable representing the age-at-onset of a particular disease for the j-th sib. $Z_1, \ldots, Z_n$ are correlated due to the genetic segregation and shared frailty, so are $T_1, \ldots, T_n$.

Assuming that T_j given Z_j are independent and based on the model (4), we can see that conditioning on the frailty vector $\mathbf{Z}$, the joint density and survival function can be written as

$$S(t_1, \ldots, t_n | Z_1, \ldots, Z_n) = e^{-\Lambda_1(t_1)Z_1 - \cdots - \Lambda_n(t_n)Z_n},$$

where $\Lambda_j(t_j) = \Lambda_0(t_j)e^{\mathbf{X}_j'\boldsymbol{\beta}}; \quad j = 1, \ldots, n.$

where λ_0 is the baseline hazard function, $\mathbf{X}_j$ is the vector of covariates for the j-th sib and $\boldsymbol{\beta}$ is a vector of parameters, Z_j is the unobserved frailty. $Z_1, \ldots, Z_n$ are correlated due to the genetic segregation and shared frailty, so are $T_1, \ldots, T_n$.

Integrating over $\mathbf{Z}$, we get the following marginal joint survival function, which is demonstrated in the Appendix:

$$\begin{aligned} S(t_1, \ldots, t_n) &= \left\{ \prod_{i=1}^{4} \frac{\eta^{v_d/2}}{\left[\sum_{j=1}^{n} \Lambda_j(t_j)a_{ji} + \eta\right]^{v_d/2}} \right\} \times \\ &\quad \times \left\{ \frac{\eta^{v_p}}{\left[\sum_{j=1}^{n} \Lambda_j(t_j) + \eta\right]^{v_p}} \right\}. \qquad (2) \end{aligned}$$

Usually, the observations are censored and we need both the survival function and combinations of density and survival functions. For a sibship with a affected sibs $(j = 1, \ldots, a)$ and $n - a$ unaffected, the joint survival and density function is

$$\begin{aligned} P(t_1, \delta_1 = 1, \ldots, t_a, \delta_a = 1, t_{a+1}, \delta_{a+1} = 0, \ldots, t_n, \delta_n = 0) = \\ (-1)^a \frac{\partial^a S(t_1, \ldots, t_n)}{\partial t_1 \ldots \partial t_a}. \end{aligned}$$

When all the sibs are unaffected $(a = 0)$, we use the survival function itself, since the density function is used only when there are censored observations.

For a sibship with all sibs affected, the joint density function is:

$$P(t_1, \delta_1 = 1, \ldots, t_n, \delta_n = 1) = (-1)^n \frac{\partial^n S(t_1, \ldots, t_n)}{\partial t_1 \ldots \partial t_n}.$$

For a sibship with two sibs, the joint survival and density function can be determined for a sib pair who shares 0, 1 and 2 alleles identical by descendent at locus d. These joint functions are shown at Table 2.1[1], in which $\Lambda_j^* = \Lambda_j(t_j) + \eta$, $j = 1, 2$ and $\Lambda_{12} = \Lambda_1 + \Lambda_2 + \eta$. We can see that, when $v_d = 0$, the joint survival function does not depend on the number of IBD alleles at locus d, indicating that there is no linkage between disease and locus d.

[1] Demonstrations in the Appendix.

Table 2.1: Joint Survival and Density Function - Bivariate Case

<table>
<tr><th colspan="4">Joint Survival and Density Function</th></tr>
<tr><td colspan="4">$P(t_1, \delta_1 = 0, t_2, \delta_2 = 0) = S(t_1, t_2)$</td></tr>
<tr><td colspan="4">$P(t_1, \delta_1 = 1, t_2, \delta_2 = 0) = C_1(t_1, t_2)\lambda_1(t_1)S(t_1, t_2)$</td></tr>
<tr><td colspan="4">$P(t_1, \delta_1 = 0, t_2, \delta_2 = 1) = C_2(t_1, t_2)\lambda_2(t_2)S(t_1, t_2)$</td></tr>
<tr><td colspan="4">$P(t_1, \delta_1 = 1, t_2, \delta_2 = 1) = \left[C_1(t_1, t_2)C_2(t_1, t_2) + C(t_1, t_2)\right]\lambda_1(t_1)\lambda_2(t_2)S(t_1, t_2)$</td></tr>
<tr><td></td><td>$IBD_d = 0$</td><td>$IBD_d = 1$</td><td>$IBD_d = 2$</td></tr>
<tr><td>$S(t_1, t_2)$</td><td>$\left(\frac{\eta^2}{\Lambda_1^*\Lambda_2^*}\right)^{v_d}\left(\frac{\eta}{\Lambda_{12}}\right)^{v_p}$</td><td>$\left(\frac{\eta^3}{\Lambda_1^*\Lambda_2^*\Lambda_{12}}\right)^{v_d/2}\left(\frac{\eta}{\Lambda_{12}}\right)^{v_p}$</td><td>$\left(\frac{\eta}{\Lambda_{12}}\right)^{v_d+v_p}$</td></tr>
<tr><td>$C_1(t_1, t_2)$</td><td>$\frac{v_d}{\Lambda_1^*} + \frac{v_p}{\Lambda_{12}}$</td><td>$\frac{v_d/2}{\Lambda_1^*} + \frac{v_d/2+v_p}{\Lambda_{12}}$</td><td>$\frac{v_d+v_p}{\Lambda_{12}}$</td></tr>
<tr><td>$C_2(t_1, t_2)$</td><td>$\frac{v_d}{\Lambda_2^*} + \frac{v_p}{\Lambda_{12}}$</td><td>$\frac{v_d/2}{\Lambda_2^*} + \frac{v_d/2+v_p}{\Lambda_{12}}$</td><td>$\frac{v_d+v_p}{\Lambda_{12}}$</td></tr>
<tr><td>$C(t_1, t_2)$</td><td>$\frac{v_p}{\Lambda_{12}^2}$</td><td>$\frac{v_d/2+v_p}{\Lambda_{12}^2}$</td><td>$\frac{v_d+v_p}{\Lambda_{12}^2}$</td></tr>
</table>

Consider a sibship with n sibs and let F be their father and M, their mother. If the parents are independent, there are only four identical by descendent alleles for a given locus. Let be also a region in the chromosome where exists the disease locus. If d is a point inside this region, it is the interest to know when there is a suspect gene at the locus d. Let $(1, 2)$ be the father's chromosome and $(3, 4)$ the mother's one. The sibship allelic inheritance vector at locus d is the vector

$$A_d = (a_1, a_2, \ldots, a_{2j-1}, a_{2j}, \ldots, a_{2n-1}, a_{2n}),$$

where $a_{2j-1} = 1$ or 2 and $a_{2j} = 3$ or 4, ie., the odd indexes represent the father's alleles transmitted and the even ones represent the mother's alleles transmitted to the sibship.

It is important to have the IBD_d (Identical by descendent) definition, which shows how many alleles a sibpair shares at locus d, as in Andrade and Pinheiro (2002). As Figure 2.1 shows, a sibpair can share 0, 1 or 2 alleles.

The parents' genetic frailties due to locus d are defined as

$$\begin{cases} Z_{dF} = U_{d1} + U_{d2} \\ Z_{dM} = U_{d3} + U_{d4}, \end{cases}$$

where the index F means father and M mother, and U_{d1} and U_{d2} are the genetic frailties due to the information contained in the father's chromosomes; U_{d3} and U_{d4} have the analogous interpretation for the mother. It's assumed that the mother and father's frailties are independent.

For a given inheritance vector v_d at locus d for a sibship, the frailty for the j-th sib is defined as

$$Z_{dj} = U_{da_{2j-1}} + U_{da_{2j}}; \;\; j = 1, \ldots, n.$$

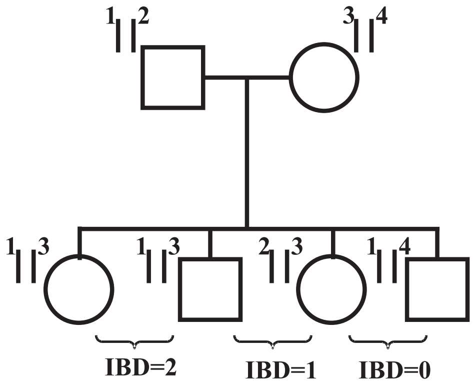

Figure 2.1: Identical by Descendent

We assume U_{d1}, U_{d2}, U_{d3} and U_{d4} are independent and follow $\Gamma(v_d/2, \eta)$, where η is the reciprocal of the scale parameter and v_d is the shape parameter. Then,

$$Z_{dj} \sim \Gamma(v_d, \eta); \quad j = 1, \ldots, n.$$

Influent factors different from genetic contributions due to locus d are considered adding another random frailty term,U_p, to the genetic frailty. Then the genetic frailty for the j-th sib is defined as:

$$Z_j = Z_{dj} + U_p = U_{da_{2j-1}} + U_{da_{2j}} + U_p,$$

where $U_p \sim \Gamma(v_p, \eta)$ over different sibships. Then, $Z_j \sim \Gamma(v_d + v_p, \eta)$. It's easy to check that the means of the frailties are

$$E(Z_1) = E(Z_2) = \cdots = E(Z_n) = \frac{v_p + v_d}{\eta}$$

and the variances are

$$V(Z_1) = V(Z_2) = \cdots = V(Z_n) = \frac{v_p + v_d}{\eta^2}.$$

Then the parameter v_d represents the frailty genetic variance proportion explained by locus d. It's assumed that $v_d + v_p = \eta$, to make the baseline hazard $\lambda_0(t)$ identifiable, which sets $E(Z_j) = 1, \forall j$ and prevents arbitrary scaling in the model. This way, the frailty variance is $1/(v_d + v_p)$.

The sibship frailties can be written in a matrix form,

$$\mathbf{Z} = \mathbf{H}\mathbf{U}, \tag{3}$$

where

$$\begin{aligned}\mathbf{Z} &= (Z_1, Z_2, \ldots, Z_n)' \\ \mathbf{H} &= \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & a_{n4} & 1 \end{pmatrix} \\ \mathbf{U} &= (U_{d1}, U_{d2}, U_{d3}, U_{d4}, U_p)',\end{aligned}$$

with

$$\begin{aligned}&a_{j1} = I(a_{2j-1} = 1); \qquad a_{j2} = I(a_{2j-1} = 2) \\ &a_{j3} = I(a_{2j} = 3); \qquad a_{j4} = I(a_{2j} = 4); \quad j = 1, \ldots, n,\end{aligned}$$

and

$\mathbf{Z}$ is the vector of frailties for the n sibs,

$\mathbf{H}$ is the matrix which its elements indicates the alleles transmitted to the sibs,

$\mathbf{U}$ is the vector of frailties; and

$I(\mathcal{A})$ is the indicator function of $\mathcal{A}$ (ie., $I(\mathcal{A}) = 1$ if the event $\mathcal{A}$ occurs and $I(\mathcal{A}) = 0$, otherwise).

2.2.1 The Additive Genetic Gamma Frailty Model

Let be a sibship with n sibs and T_j the random variable of age at disease onset for the j-th sib. Then, (t_j, δ_j) is the dataset where t_j is the age at onset (if $\delta_j = 1$) or age at censoring (if $\delta_j = 0$). It is assumed that the hazard function of developing the disease for the j-th sib at age t_j is modeled by the proportional hazard model with random efect Z_j,

$$\lambda_j(t|Z_j) = \lambda_0(t) e^{\mathbf{X}'_j \boldsymbol{\beta}} Z_j; \quad j = 1, \ldots, n, \tag{4}$$

where λ_0 is the baseline hazard function, $\mathbf{X}_j$ is the vector of covariates for the j-th sib and $\boldsymbol{\beta}$ is a vector of parameters, Z_j is the unobserved frailty. $Z_1, \ldots, Z_n$ are correlated due to the genetic segregation and shared frailty, so are $T_1, \ldots, T_n$.

2.2.2 The Baseline Logistic Hazard

The logistic hazard frailty model is built using the a baseline hazard of the form

$$\lambda_0^*(t|\alpha, \gamma) = \frac{\exp(t\alpha + \gamma)}{1 + \exp(t\alpha + \gamma)}, \tag{5}$$

resulting in a parametric case of (4). The cumulative hazard function , $\Lambda_j(t_j)$, used at the survival function (2) must be replaced by the expression

$$\begin{aligned}\Lambda_0(t) &= \zeta\int_0^t \frac{e^{u\alpha+\gamma}}{1+e^{u\alpha+\gamma}}\,du\\ &= \ln\left(\frac{1+e^{t\alpha+\gamma}}{1+e^{\gamma}}\right)^{\frac{\zeta}{\alpha}}. \qquad (6)\end{aligned}$$

Thus,

$$\lambda_j(t_j|Z_j) = \lambda_0(t)e^{\mathbf{X}_j'\boldsymbol{\beta}}Z_j; \quad j=1,\ldots,n,$$

where

$$\lambda_0(t|\alpha,\gamma) = \zeta\frac{\exp(t\alpha+\gamma)}{1+\exp(t\alpha+\gamma)},$$

and

$$\begin{aligned}S(t_1,\ldots,t_n) &= \left\{\prod_{i=1}^{4}\frac{\eta^{v_d/2}}{\left[\sum_{j=1}^n \Lambda_j(t_j)a_{ji}+\eta\right]^{v_d/2}}\right\}\times\\ &\times\left\{\frac{\eta^{v_p}}{\left[\sum_{j=1}^n \Lambda_j(t_j)+\eta\right]^{v_p}}\right\}. \qquad (7)\end{aligned}$$

where

$$\begin{aligned}\Lambda_0(t_j) &= \ln\left(\frac{1+e^{t_j\alpha+\gamma}}{1+e^{\gamma}}\right)^{\frac{\zeta}{\alpha}} \text{ and} \qquad (8)\\ \Lambda_j(t_j) &= \Lambda_0(t_j)e^{\mathbf{X}_j'\boldsymbol{\beta}}. \qquad (9)\end{aligned}$$

Since the hazard function represents the instantaneous failure probability, the use of a hazard function of a logistic form is justified, because the logistic function is often used when modelling probabilities dependent on time.

2.3 Hypotheses Testing

The survival model above can be used to develop a likelihood ratio test to be used in linkage analysis. As shown at Table 2.1, when $v_d = 0$, the joint survival and density function does not depend on the number of IBD alleles shared at locus d. Thus, the linkage test between the disease and locus d can be done testing $H_0 : v_d = 0$.

Let the i-th sibship with n_i sibs and $(t_i, \delta_i) = (t_{i1}, \delta_{i1}, \ldots, t_{in_i}, \delta_{in_i})$ the age at onset or at censoring. Consider also a marker M_i for the i-th sibship. The

(M_i, t_i, δ_i) vector can be seen as the retrospective likelihood of M_i given the phenotypes (t_i, δ_i), as in Whittemore (1996).

The retrospective likelihood for the i-th sibship is

$$\begin{aligned} L_i(v_d, v_p, \Lambda_0(t), \boldsymbol{\beta}) &= P(M_i|t_i, \delta_i) \\ &= \frac{\sum_{a_d} P(t_i, \delta_i|\ A_d = a_d)P(A_d = a_d|M_i)}{\sum_{a_d} P(t_i, \delta_i|A_d = a_d)P(A_d = a_d)} P(M_i), \end{aligned}$$

where $P(t_i, \delta_i|A_d = a_d)$ is shown in the Appendix for the bivariate case under the notation $P(t_i, \delta_i|IBD = k)$, $P(A_d = a_d)$ is the prior probability of the inheritance vector A_d and $P(A_d = a_d|M_i)$ can be determined using multipoint methods (Kruglyak et al., 1996). For sib pair data, the retrospective likelihood is:

$$\begin{aligned} &L_i(v_d, v_p, \Lambda_0(t), \boldsymbol{\beta}) = P(M_i|t_i, \delta_i) = \\ = \ &\frac{\sum_{k=0}^{2} P(t_{i1}, \delta_{i1}, t_{i2}, \delta_{i2}|IBD_d = k)P(IBD_d = k|M_i)}{\sum_{k=0}^{2} P(t_{i1}, \delta_{i1}, t_{i2}, \delta_{i2}|IBD_d = k)P(IBD_d = k)} \times P(M_i), \end{aligned}$$

where $P(t_{i1}, \delta_{i1}, t_{i2}, \delta_{i2}|IBD_d = k)$ is given on Table 2.1 and $P(IBD_d = k)$ is the prior probability of a sib pair sharing k IBD alleles.

This likelihood function depends only on the cumulative hazard function and when $v_d = 0$, $L_i(0, v_p, \Lambda_0(t)) = P(M_i)$, thus the likelihood ratio statistic is given by

$$LR_i(v_d, v_p, \boldsymbol{\beta}) = \frac{\sum_{a_d} P(t_i, \delta_i|A_d = a_d)P(A_d = a_d|M_i)}{\sum_{a_d} P(t_i, \delta_i|A_d = a_d)P(A_d = a_d)}.$$

Assuming there are K families, the logarithm of odds (LOD) score, (Olson et al., 1999; Elandt-Johnson, 1971), at locus d is defined as

$$Lod_d = \max_{v_d, v_p, \beta} \sum_{i=1}^{K} \log_{10} LR_i(v_d, v_p, \boldsymbol{\beta}). \qquad (10)$$

The Lod score is used at linkage analysis (Elandt-Johnson, 1971; Ott, 1991), in order to help on the genetic mapping. Based on the likelihood tests theory when the null hypothesis is on the parameter space boundary (Self and Liang, 1987) we have that, under H_0, $2Lod_d \ln(10)$ follows a mixture with equal probability of point mass zero and chi-square with one degree of freedom.

3 Application

3.1 Description of the Simulated Data

In this work, simulated data (created using the Genometric Analysis Simulation Program - G.A.S.P.) were analyzed. The age at onset were created based on a logistic density, shown at equation (11), where $\alpha = 0.1$, $\gamma = 0.1$ and $\zeta = 0.005$, which implies in the density presented at Figure 3.1. The actual ages were simulated using U(60,80).

$$f_0(t|\zeta, \alpha, \gamma) = \lambda_0(t|\zeta, \alpha, \gamma) S_0(t|\zeta, \alpha, \gamma), \tag{11}$$

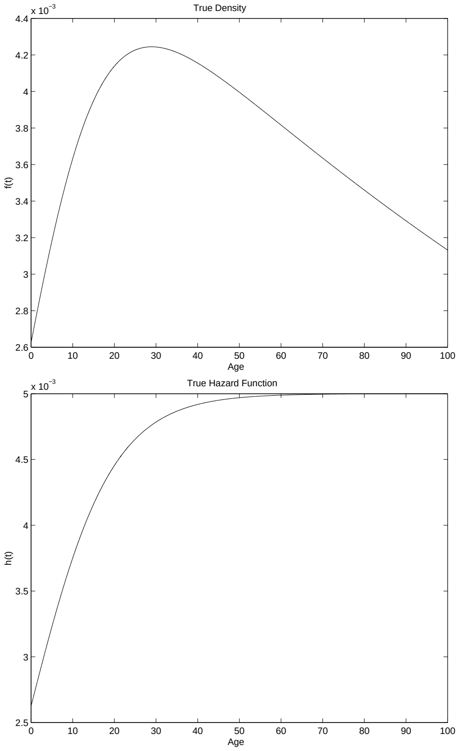

Figure 3.1: Density and Hazard Functions

For the first data set, a binary trait due to a locus with two alleles linked to the markers 1, 2, 3, 4 and 5 was simulated. This binary trait indicates the presence/absence of disease. Then, the age at onset and the actual age were simulated, using the disease information. The second data set was generated in a similar way, but there weren't linkage evidences between the markers and the locus.

Both data sets have one thousand nuclear families with two kids, thus four members per family (father, mother, first son/daugther, second son/daugther).

Confidence intervals for the parameter were built using resampling techniques, such as Bootstrap.

The GenHunter package adjusts the linkage analysis model using a different point of view from the Logistic Hazard Frailty Model. GeneHunter uses the fact of being (or not) affected as response, forgetting all the existing survival analysis structure. Then, we used another method in order to compare the results.

In order to o validate the Logistic Hazard Frailty Model, it was adjusted a Cox model for each of the data sets and then the Martingale Residuals were computed (using SAS). These residuals were used as a quantitative trait to be analyzed by GeneHunter, using non-parametric methods, getting the Z-score, widely used in quantitative traits linkage analysis (QTL). Then, the models are compared and the Logistic Hazard Frailty Model purpose is validated using the Z-score and Lod-score p-values, since $Z^2 \sim \chi^2_1$ and $2Lod_d \ln(10) \sim \frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$, as shown in Kruglyak (1996), Li and Zhong (2002) and Self and Liang (1987).

It is possible to see, as shown in Figures 3.1 and 3.2, that the proposed method did not estimate correctly the parameters, which implied in a significant difference between the true and estimated densities. Then, the bad parameters estimation is reflected on the estimated hazard functions, as in Figures 3.1 and 3.2.

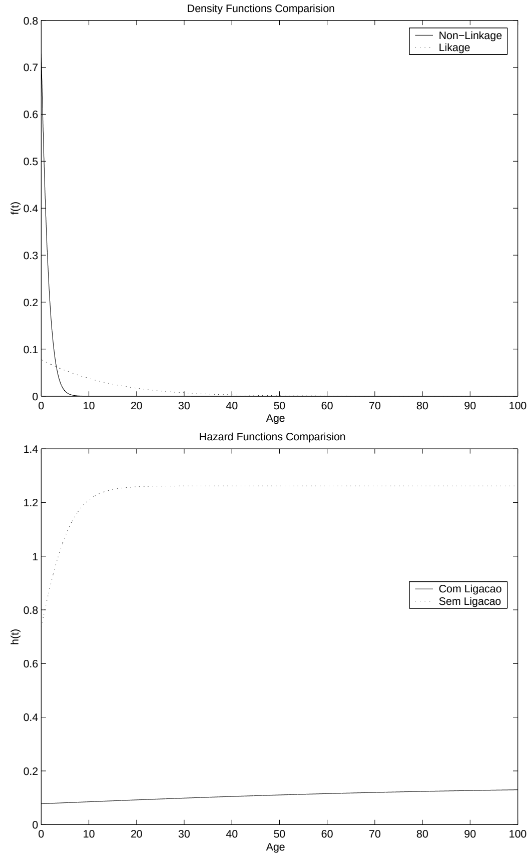

Figure 3.2: Comparisons of Density and Hazard Functions

3.2 Results

3.2.1 Under the assumption of linkage

The purpose of linkage analysis using the Logistic Hazard Frailty Model (LHFM) shows similar results when compared to the Cox Model with Martingale Residuals (CMMR) method, adjusted by SAS/GeneHunter.

The five markers were simulated with strong linkage evidences. This fact was detected by both methods (Logistic Hazard Frailty Model and Cox Model with Martingale Residuals), because the p-values is less than 0.05.

It is possible to verify that the LHFM is more conservative when compared to CMMR method. It is important to say that both methods detected strong linkage evidence at position 7.38 and its neighborhood. Inside the region that contains the locus that causes the disease, the LHFM presents minimum significance values, rejecting the non-linkage hypotheses.

Table 3.2: Comparision between CMMR and LHFM - Linkage Case

Position	P-value		Position	P-value		Position	P-value	
	CMMR	LHFM		CMMR	LHFM		CMMR	LHFM
0.00	10^{-16}	0	7.38	0	0	14.75	10^{-16}	0
1.05	10^{-16}	0	8.43	0	0	15.80	10^{-16}	0
2.11	0	0	9.48	0	0	16.86	10^{-15}	0
3.16	0	0	10.54	0	0	17.91	10^{-15}	0
4.21	0	0	11.59	0	0	18.96	10^{-15}	0
5.27	0	0	12.64	0	0	20.02	10^{-14}	0
6.32	0	0	13.70	10^{-16}	0	21.07	10^{-14}	0

Table 3.3: True Parameters and their Estimates - Linkage Case

	v_d	v_p	α	γ	ζ_0	β
True	0.4000	0.0005	0.1000	0.1000	0.0050	2.0000
Lower Bound 95%	0.3346	0.0008	0.0202	0.1472	0.1311	0.6634
Estimate	0.3958	0.0010	0.0208	0.1663	0.1429	0.7837
Upper Bound 95%	0.4569	0.0011	0.0215	0.1854	0.1547	0.9039

Although there were differences between the true values and the estimates (as shown in Table 3.3), the LHFM was able to detected the same linkage behavior detected by CMMR method, as shown in Figure 3.3, because there was a (perfect) superposition of p-values through all the region analyzed.

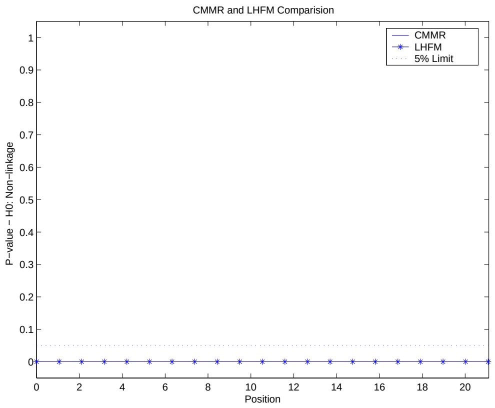

Figure 3.3: P-values for CMMR and LHFM - Linkage Case

As Figure 3.4 shows, it is possible to see the same linkage pattern between CMMR and LHFM, where the latter presents the linkage behavior much stronger than the first method. The region where the maximum is located is the disease's cause location.

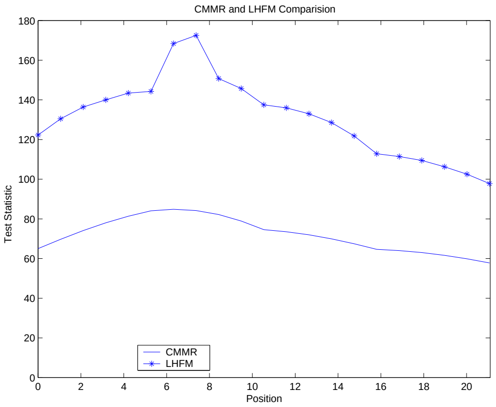

Figure 3.4: Test Statistics for CMMR and LHFM - Linkage Case

The estimates behavior does not match the expected one when compared to the pattern used for the generation of the simulated data on Table 3.3. This data set was generated in such a way that there was linkage between the markers and disease locus and, as said before, this is revealed as long as v_d is greater then 0. It is possible to verify the matching between the true and estimated values for v_d, v_p and γ, but there is a reasenable difference between the true and estimated values for β and ζ_0. This problem can be caused by the estimation procedure, which depends on the start conditions.

3.2.2 Under the assumption of no linkage

For the data set which refers to the five simulated markers without linkage, the LHFM presents good results. Since the dataset was created under the null hypothesis, it is desired to not reject this hypothesis: this fact was observed better when using LHFM (Table 3.4). The LHFM can detect the linkage pattern just as CMMR, Figure 3.5, because the null hypothesis was not rejected anywhere. It is important to say that the p-values associated to the LHFM were always greater than those associated to the CMMR method. Thus the LHFM works better than the CMMR model under the null hypothesis.

Table 3.4: Comparision between CMMR and LHFM - Non-Linkage Case

Position	P-value		Position	P-value		Position	P-value	
	CMMR	LHFM		CMMR	LHFM		CMMR	LHFM
0.00	0.9641	1.00	7.38	0.6191	1.00	14.75	0.6072	1.00
1.05	0.8712	1.00	8.43	0.6735	1.00	15.80	0.6187	1.00
2.11	0.7616	1.00	9.48	0.7587	1.00	16.86	0.5335	1.00
3.16	0.6719	1.00	10.54	0.8433	1.00	17.91	0.4590	1.00
4.21	0.6280	1.00	11.59	0.7582	1.00	18.96	0.4284	1.00
5.27	0.6179	1.00	12.64	0.6731	1.00	20.02	0.4504	1.00
6.32	0.6064	1.00	13.70	0.6195	1.00	21.07	0.5003	1.00

Table 3.5: True Parameters and their Estimates - Non-Linkage Case

	v_d	v_p	α	γ	ζ_0	β
True	0.0050	0.0500	0.1000	0.1000	0.0050	2.0000
Lower Bound 95%	0.0047	0.0422	0.2189	0.2621	0.9793	1.3454
Estimate	0,0061	0.0559	0.2803	0.3356	1.2618	1.5496
Upper Bound 95%	0.0076	0.0696	0.3418	0.4090	1.5463	1.7539

As shown in Figure 3.6, none of the methods detected linkage, since their test statistics are close to zero.

There are differences between the real parameters and their estimates. But the v_d estimate (responsible for the linkage detection) is very close from the true value; the same is valid for v_p, as shown in Table 3.5. Once more, the start condition and process stability are the possible causes of this incorrect parameters estimation.

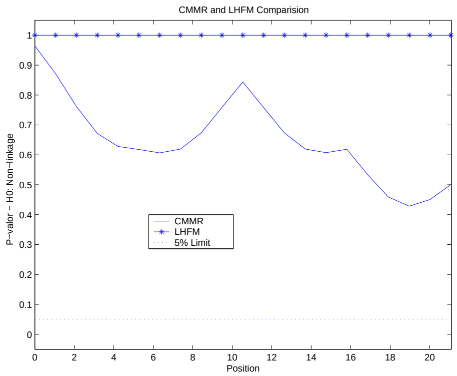

Figure 3.5: P-values for CMMR and LHFM - Non-Linkage Case

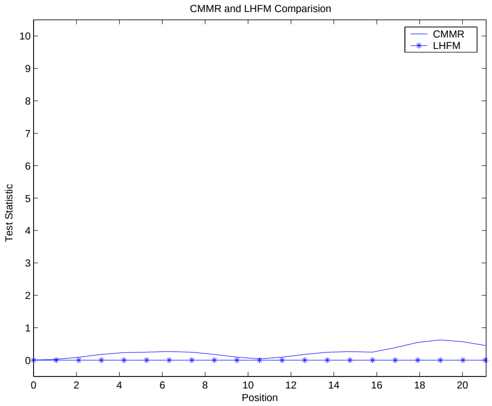

Figure 3.6: Test Statistics for CMMR and LHFM - Non-Linkage Case

Appendix

Proof: Joint Survival and Density Function - Bivariate Case

A.1 $IBD_d = 0$

$$S(t_1, t_2|IBD = 0) \quad = \quad \left(\frac{\eta^2}{\Lambda_1^* \Lambda_2^*}\right)^{v_d} \left(\frac{\eta}{\Lambda_{12}}\right)^{v_p}$$

Assuming non-informative censoring:

$$\begin{aligned}
P(t_1,\delta_1=1,t_2,\delta_2=0|IBD=0) &\propto -\frac{\partial S(t_1,t_2)}{\partial t_1} \\
&= \left(\frac{\eta^2}{\Lambda_1^*\Lambda_2^*}\right)^{v_d}\left(\frac{\eta}{\Lambda_{12}}\right)^{v_p}\lambda_1(t_1)\left(\frac{v_d}{\Lambda_1^*}+\frac{v_p}{\Lambda_{12}}\right) \\
&= \lambda_1(t_1)S(t_1,t_2)\left(\frac{v_d}{\Lambda_1^*}+\frac{v_p}{\Lambda_{12}}\right) \\
&= \lambda_1(t_1)S(t_1,t_2)C_1(t_1,t_2)
\end{aligned}$$

Similarly,

$$P(t_1,\delta_1=0,t_2,\delta_2=1|IBD=0)\propto\lambda_2(t_2)S(t_1,t_2)C_2(t_1,t_2).$$

To compute $P(t_1,\delta_1=1,t_2,\delta_2=1|IBD=0)$,

$$\begin{aligned}
P(t_1,\delta_1=1,t_2,\delta_2=1|IBD=0) &\propto \frac{\partial^2 S(t_1,t_2)}{\partial t_1\partial t_2} \\
&= \left[\left(\frac{v_d}{\Lambda_1^*}+\frac{v_p}{\Lambda_{12}}\right)\left(\frac{v_d}{\Lambda_2^*}+\frac{v_p}{\Lambda_{12}}\right)+\frac{v_p}{\Lambda_{12}^2}\right]\times \\
&\quad \times\lambda_1(t_1)\lambda_2(t_2)\left(\frac{\eta^2}{\Lambda_1^*\Lambda_2^*}\right)^{v_d}\left(\frac{\eta}{\Lambda_{12}}\right)^{v_p} \\
&= [C_1(t_1,t_2)C_2(t_1,t_2)+C(t_1,t_2)]\times \\
&\quad \times\lambda_1(t_1)\lambda_2(t_2)S(t_1,t_2)
\end{aligned}$$

A.2 $IBD_d=1$

$$S(t_1,t_2|IBD=1) = \left(\frac{\eta^3}{\Lambda_1^*\Lambda_2^*\Lambda_{12}}\right)^{v_d/2}\left(\frac{\eta}{\Lambda_{12}}\right)^{v_p}$$

Again, the non-informative censoring hypothesis is used:

$$\begin{aligned}
P(t_1,\delta_1=1,t_2,\delta_2=0|IBD=1) &\propto -\frac{\partial S(t_1,t_2)}{\partial t_1} \\
&= \left[\frac{v_d/2}{\Lambda_1^*}+\frac{v_d/2+v_p}{\Lambda_{12}}\right]\lambda_1(t_1)S(t_1,t_2) \\
&= C_1(t_1,t_2)\lambda_1(t_1)S(t_1,t_2)
\end{aligned}$$

In a analogous way,

$$\begin{aligned}
P(t_1,\delta_1=0,t_2,\delta_2=1|IBD=1) &\propto -\frac{\partial S(t_1,t_2)}{\partial t_2} \\
&= \left[\frac{v_d/2}{\Lambda_2^*}+\frac{v_d/2+v_p}{\Lambda_{12}}\right]\lambda_2(t_2)S(t_1,t_2) \\
&= C_2(t_1,t_2)\lambda_2(t_2)S(t_1,t_2)
\end{aligned}$$

Thus,

$$\begin{aligned}
P(t_1,\delta_1=1,t_2,\delta_2=1|IBD=1) &\propto \frac{\partial^2 S(t_1,t_2)}{\partial t_1 \partial t_2} \\
&= \lambda_1(t_1)\lambda_2(t_2)S(t_1,t_2)\left\{\left[\frac{(v_d/2)}{\Lambda_1^*}+\frac{(v_d/2+v_p)}{\Lambda_{12}}\right]\right. \\
&\quad \left.\times\left[\frac{(v_d/2)}{\Lambda_2^*}+\frac{(v_d/2+v_p)}{\Lambda_{12}}\right]+\frac{(v_d/2+v_p)}{\Lambda_{12}^2}\right\} \\
&= [C_1(t_1,t_2)C_2(t_1,t_2)+C(t_1,t_2)]\times \\
&\quad \times\lambda_1(t_1)\lambda_2(t_2)S(t_1,t_2)
\end{aligned}$$

A.3 $IBD_d = 2$

$$S(t_1,t_2|IBD=2) = \left(\frac{\eta}{\Lambda_{12}}\right)^{v_d+v_p}$$

Assuming non-informative censoring:

$$\begin{aligned}
P(t_1,\delta_1=1,t_2,\delta_2=0|IBD=2) &\propto -\frac{\partial S(t_1,t_2)}{\partial t_1} \\
&= \frac{v_d+v_p}{\Lambda_{12}}\lambda_1(t_1)S(t_1,t_2) \\
&= C_1(t_1,t_2)\lambda_1(t_1)S(t_1,t_2).
\end{aligned}$$

In a analogous way,

$$\begin{aligned}
P(t_1,\delta_1=0,t_2,\delta_2=1|IBD=2) &\propto -\frac{\partial S(t_1,t_2)}{\partial t_2} \\
&= \frac{v_d+v_p}{\Lambda_{12}}\lambda_2(t_2)S(t_1,t_2) \\
&= C_2(t_1,t_2)\lambda_2(t_2)S(t_1,t_2).
\end{aligned}$$

Thus,

$$\begin{aligned}
P(t_1,\delta_1=1,t_2,\delta_2=1|IBD=2) &\propto \frac{\partial^2 S(t_1,t_2)}{\partial t_1 \partial t_2} \\
&= \frac{1}{\Lambda_{12}^2}\left[(v_p+v_d)S(t_1,t_2)(1+v_p+v_d)\lambda_1(t_1)\lambda_2(t_2)\right] \\
&= \left[\left(\frac{v_p+v_d}{\Lambda_{12}}\right)^2+\frac{v_d+v_p}{\Lambda_{12}^2}\right]\lambda_1(t_1)\lambda_2(t_2)S(t_1,t_2) \\
&= [C_1(t_1,t_2)C_2(t_1,t_2)+C(t_1,t_2)]\times \\
&\quad \times\lambda_1(t_1)\lambda_2(t_2)S(t_1,t_2)
\end{aligned}$$

Acknowledgements

This research was funded in part by FAEP 2001/13224-7, FAPESP 00/12212-2 and 01/13224-7.

References

Anderson, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical Model Based on Counting Process.* Springer-Verlag, New York.

Andrade, M. and Pinheiro, H. P. (2002). *Métodos Estatísticos Aplicados em Genética Humana.* 15o. SINAPE - ABE, São Paulo.

Chang, I.S. and Hsiung, C.A. (1995). An efficient estimator for proportional hazards models with frailties. Unpublished manuscript.

Clayton (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**: 141-151.

Clayton, D. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model. *Journal of Royal Statistical Society* **A 148**: 82-117.

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society.* **B 34**: 187-220.

Elandt-Johnson, R.C. (1971). *Probability Models and Statistical Methods in Genetics.* John Wiley & Sons, New York.

Kruglyak, L. and Daly, M. J. and Reeve-Daly, M. P. and Lander, E. S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics.* **58**, 1347-1363.

Li, H. and Zhong, X. (2002). Multivariate survival models induced by genetic frailties, with application to linkage analysis. *Biostatistics* **3,1**, 57-75.

Mackenzie, G. (1996). Regression models for survival data: the generalized time-dependent logistic family. *The Statistician.* **45**, 21-34.

Nielsen, G.G., Gill, R.D., Anderson, P.K. and Sørensen, T.I.A. (1992). A counting process approach to maximum likelihood estimation in frailty models, *Scandinavian Journal of Statistics* **19**: 25-43.

Olson, J. M. and Witte, J. S. and Elston, R. C. (1999). Tutorial in biostatistics genetic mapping of complex traits. *Statistics in Medicine.* **18**, 2961-2981.

Ott, J. (1991). *Analysis of Human Genetic Linkage.* The Johns Hopkins University Press. Baltimore and London.

Self, S. G. and Liang, K. Y. (1982). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of American Statistical Association* **82**, 605-610.

Self, S.G. and Prentice, R.L. (1986). Incorporating random effects into multivariate relative risk regression models, *in Modern Statistical Methods in Chronic Disease Epidemiology*, S.H. Moolgavkar & R.L. Prentice (eds), New York: Wiley, pp.167-177.

Vaupel, J.M. and Manton, K.G. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439-454.

Whittemore, A. S. (1996). Genome scanning for linkage: an overview. *American Journal of Human Genetics.* **59**, 704-716.

Yau, K.K.W. and McGilchrist, C.A. (1998). ML and REML estimation in survival analysis with time dependent correlated frailty. *Statistics in Medicine* **17**: 1201-1213.