PhD Thesis

How Genes Affect Longevity in Heterogeneous Populations: Binomial Frailty Models and Applications

Qihua Tan

Institute of Public Health, Faculty of Health Science

University of Southern Denmark

2000

A dissertation submitted for the degree of Doctor of Philosophy

in

Medicine

in the

Falculty of Health Science

of the

University of Southern Denmark

by

Qihua Tan

Bachelor of Medical Science (Shandong Medical University, China), 1985

Master of Medical Science (Shanxi Medical University, China), 1991

Supervised by:

Professor James W. Vaupel

Professor Anatoli I. Yashin

Professor Kaare Christensen

Committee in charge:

Professor Werner Vach, Chair

Professor Thorkild I.A. Sørensen

Dr. Philip Hougaard

CONTENTS

Chapter 4 Estimating the number of longevity genes

PART III Models for Gene Marker Data on Unrelated Individuals

Chapter 5 The binomial frailty model for gene marker data

PREFECE

In 1995, one winter morning when I was at the Centre for Health and Social Policy (CHS) of Odense University Medical School, I remember Dr. James W.Vaupel coming to me and saying that he and Dr. Anatoli I.Yashin had two projects that they would like me to work on. After a brief discussion, I learned that the two somehow connected projects were (1) to develop a new model to study the influence of observed covariates on the life spans of unrelated individuals and (2) to model the polygenic influence on family correlation of life spans. Now it has turned out that these two projects form the body of the dissertation presented here.

I enjoyed the working environment at CHS very much. Particularly, I enjoyed the frequent scientific communications I had with Dr. Vaupel and Dr. Yashin during my stay there. After extensive work that winter, we were invited to the first longevity conference held by the IPSEN foundation in Paris. There we presented our first results on a family study. The project continued after I came to Dr. Vaupel's new institute in Rostock, Germany, the Max-Planck Institute for Demographic Research (MPIDR) in Oct. 1996. Here at MPIDR, further simulations had been done but most importantly, I came to the stage of model application with genetic data available from both Italy and Denmark. Results on simulation and application had been presented at the 3rd European Research Workshop on Longevity in Ancona, Italy in 1997; the 1st and 2nd GAAC Workshops on the Social and Biological Determinants of Longevity in Davis, California 1997 and Rostock 1998; and the Biometrie 2000 Kolloquium Rostock.

Looking back over the entire research process, I must admit that the work couldn't be done without the scientific inspirations and ideas as well as the great help from my supervisors, Dr. Vaupel, Dr. Yashin and Dr. Kaare Christensen. I very much enjoyed working with them from which I learned not only a great deal, but also appreciated their rigorous attitude toward pursuing scientific research as well as their modest personalities. I can draw on these experiences for the rest of my life. As my

supervisor at University of Southern Denmark (the former Odense University), Dr. Christensen has provided me with direction for my research and assitance in my enrollment in the PhD program at SDU, for which I am very grateful.

My sincere gratitude goes to Dr. G. De Benedictis, Ms. Karen A. Ranberg and Dr. Bertrand Desjardins for their generosity in allowing me to use their data from the Italian centenarian studies, Danish centenarian studies and Quebec genealogy for my thesis. Their research interests as presented in their data have also inspired me and enriched my thesis.

I also want to extend thanks to Dr. Bernard Jeune, Mr. Ivan Iachine, Mr. Axel Skytthe at SDU in Odense, Dr. Kirill Andereev, Dr. Scott Pletcher, Dr. Andreas Wienke and Dr. Heiner Maier at MPIDR in Rostock, for their encouragement and useful discussions. I am especially thankful to Dr. Jutta Gampe at MPIDR for her contribution in improving my thesis through careful reading and providing valuable suggestions. I am grateful to Dr. Karl Brehmer at MPIDR and to Ms. Susan Mazur at University of California Riverside for helping me to edit the text of this thesis and to Miss Annette Erlangsen at MPIDR for translating the summary into Danish.

Finally I thank my wife, Shuxia Li, and my son, Ming Tan, who helped me in many ways. Without their understanding and support, I would never have completed the work. I hope they will learn from now on what a real weekend means.

Rostock, August 2000

Qihua Tan

INTRODUCTION

The determinants of longevity have the drawn attentions of researchers from a variety of disciplines such as sociologists, biologists, gerontologists, psychologists and medical scientists. As a quantitative trait, life span is affected by multi-factors that can be genetic, environmental, medical, etc (Christensen & Vaupel 1996). Social and environmental changes, including advances in medicine, play important roles in reducing death rate or mortality. In terms of mortality change, there had been a large reduction in death rate at younger ages before 1950 and at old ages after 1950 due to improvements in environment and biomedical achievement (Vaupel et al. 1998). As a result, mean life span in developed countries has experienced a remarkable increase due to improved nutrition, better living conditions (environment, social welfare and health care system) and success in treating fatal diseases. With more and more people celebrating their 100-year birthdays, we need to think about why these people survive where others failed and what can help to explain the life span heterogeneity. To answer the question, we look at individual factors like behaviour, socio-economical background and even individual's genetic make-up. Using twin data, a genetic correlation of life spans has been detected in intensive studies conducted in recent decades (Wyskak 1978; Carmelli 1982; Harris et al. 1992; Hayakawa et al. 1992; McGue 1993; Herskind et al. 1996). The heritability of life span was estimated to be 0.23 for males and 0.2 for females from the well-known Danish twin data (Herskind et al. 1996). By applying both shared-frailty models (Vaupel et al. 1991; Yashin & Iachine 1995) and correlated-frailty models (Yashin & Iachine 1995, 1997; Yashin et al. 1999a) to twin data, the coefficients of frailty correlation or the variance of shared frailty instead of correlation of life span can be estimated. Heritability of genetic frailty has been estimated about 0.5 from the Danish twin data (Yashin & Iachine 1995, 1997), which is roughly twice the heritability of life span. Recently, these models have been applied to family data in order to find patterns of life span

correlation among relatives and the age-pattern of life span correlation (Tan et al. submitted). The results indicate that kindred correlation in frailty follows the pattern of additive genetic inheritance, and the roughly estimated heritability of frailty is consistent with the results of major twin studies (Yashin & Iachine 1995, 1997).

Given the genetic influence on life span, a chanllenging task is to find out the genes behind it. However, populations are heterogeneous. Individuals differ along multitudinous dimensions. In any particular study most of these dimensions are unobserved. The frail tend to die first. So these individuals who survive to advanced ages are systematically different from population of individuals at younger ages. Differential survival in heterogeneous population fundamentally complicates effects to assess genetic and non-genetic risk factors that influence survival. In order to make proper inferences regarding the effect of observed genetic covariates on life span, one must bear in mind that life span is a complex trait such that no single gene or attribute can be considered to be an independent predictor of it. This is different from the situation where a single locus is responsible for a distinct dichotomous phenotype regardless of the existing environment or genotypes at other loci (McClearn 1997). This polygenetic feature combined with environmental interference results in individual heterogeneity, meaning that inference upon the effect of one single attribute on life span can't be made properly without consideration of individual variation in the unobserved frailty that contributes to survival. It is imperative to study how the influence from a single gene can be perceived given this complexity and what can be done to minimise the deviation on the conclusion resulted from the unobserved heterogeneity (both genetic and environmental).

Taking advantage of the rapid development in molecular genetics, the study of genes and longevity has intensified during the past few decades (Gerkins et al. 1974; Proust et al. 1982; Takata et al. 1987; Schachter et al.1994; Zhang et al. 1998; Bathum et al. 1998; De Benedictis et al. 1998, 1999; Bladbjerg et al. 1999; Gerdes et al. 2000). Genetic information has been collected from unrelated individuals at younger ages as controls and from older people (usually centenarians) as cases. With a case-control study, the frequencies of a certain gene allele or genotype can be compared to see if any differences exist; a significant frequency increase or decrease from the control to the centenarians leads to the conclusion that the gene allele or genotype is beneficial or detrimental to longevity. Although easy to implement and popular in use, there are

many problems in these case-control studies. As will be discussed later, the major drawbacks originate (1) from the polygenic nature of life span and (2) from the cross-sectional design. The polygenic feature of life span determines that it is continuously distributed with considerable environmental influence. The multi-factorial determination of life span creates individual heterogeneity in their frailty compositions. The cross-sectional design brings up the problem of mortality changes over time since participants in the studies were born in different cohorts. The cohorts involved in the studies have experienced quite different mortality changes since an enormous reduction in the death rate has been achieved during the past century (Vaupel et al. 1998). Under these considerations, the cross-sectional designs engaged in genetic studies of longevity on unrelated individuals (1) do not make full use of the survival information from each individual, (2) ignore the existence of other unobserved biological and environmental factors that also contribute to life span, and (3) ignore the cohort effect in mortality improvement in making inferences on the observed genes or genotypes. All of these problems result in a low rate of efficiency for the cross-sectional approach and sometimes even misleading conclusions (Hayflick 1994). A longitudinal design can help to avoid difficulties from mortality changes, but unfortunately it requires long-term cooperation from both the researchers and the participants, which is time consuming and expensive. Although, with many difficulties, as discussed by Hayflick (1994), the cross-sectional design has turned out to be a feasible way to do a longevity study on unrelated individuals due to its lower expenses and quick outcome. However, as more and more data containing individual genetic information become available from increasingly emerging longevity studies, newer and more powerful models are called for in order to help to find genes with potential contributions to longevity.

The basic purpose of this Ph.D. thesis is to explore how heterogeneity modulates the effect of genetic factors on longevity. A new approach to the analysis of genetic data from cross-sectional longevity studies will be provided. This approach is based on the binomial frailty model, which will be introduced in Part I. Based on the assumption of proportional hazards and a binomial distribution for the gene alleles or genotypes, this model allows the incorporation of polygenic influences as well as heterogeneity in unobserved frailty.

Part I and Part II are mainly devoted to introducing the model, exploring its properties, and investigating its suitability for different applications. This is partly done by theoretical considerations, partly by simulation studies. Readers who want to skip these technicalities and are more interested in the empirical findings may directly proceed to Chapter 6. Relevant sections from earlier chapters are referred to for efficient reading when needed.

The detailed outline is as follows:

In Chapter 1 the model will be defined and some core results will be derived. Further insights from this model will be given in Chapter 2 where, with simulation scenarios, the influence of an individual's genetic make-up on his or her life-span is studied in the presence of a heterogeneous background. Genotype frequency trajectories by age and the phenomenon of risk compensation will also be explored.

Part II is devoted to the analysis of related individuals: In Chapter 3 the binomial frailty model will be extended to study life span correlation and the correlation of individual frailty among first-degree relatives. Estimating the number of longevity genes from Danish twin data, Quebec genealogy data, and data on European noble families will be tackled in Chapter 4.

Part III focuses on applications of the binomial frailty model to gene marker data on unrelated individuals. In Chapter 5 the model is adapted to this situation and an estimating procedure is proposed. To study the effectiveness of this approach as well as its sensitivity to different data characteristics, a simulation study has been conducted. Problems related to cross-sectional data will be discussed, and finally modeling of heterogeneity, interactions, and controlling for sampling bias and confounding factors will be considered.

In Chapter 6 the model is applied in search of candidate longevity genes to individual gene marker data from Danish and Italian centenarian studies. First, a short review on genes which have been found relevant to longevity is given, then the gene marker data are analyzed by applying models with and without consideration of individual's unobserved frailty. The results from the two models are compared to see influneces from heterogeneity on the conclusion and on the fitness of the model. A summary of the major findings concludes the thesis.

PART I

Modeling the Genetic Influences on Life Span

Chapter 1

The Binomial Frailty Model

1.1 The binomial frailty model

1.1.1 The proportional hazard assumption

According to a Danish twin study, 50% of human life span variation after age 30 can be ascribed to the influence of survival attributes (persistent characteristics, innate or acquired, that affect survival chances) that are fixed by age 30, and for individuals who can expect to survive to age 90, this percentage increases to 80% (Vaupel et al. 1998; Yashin & Iachine 1995). With these fixed attributes, it is possible to borrow the idea of Cox's proportional hazard model (Cox 1972) in order to measure the influence they have on life span. The Cox model assumes that the hazard of death for an individual with one observed covariate is multiplicatively proportional to the hazard of death for an individual without it (or the baseline hazard) and additively proportional to the baseline hazard on the log scale. Empirical support for this assumption comes from evidence from animal studies (Promislow & Tatar 1998). It will be shown that the proportional hazard assumption is a convenient approach, and we will adhere to it throughout the entire application.

1.1.2 The binomial frailty model

As was mentioned above, longevity is a polygenic phenomenon. For the sake of simplicity, let us suppose there are N loci, each hosting at most 2 longevity alleles. Each of the longevity gene alleles reduces the hazard of death by the factor $(1-r)$, where r is the risk reduction for one single allele, and p is the frequency of the allele. According to the binomial distribution, the probability of observing i such alleles in

one individual is $B(p|i,2N)$, so that $p(i)=\binom{2N}{i}p^i(1-p)^{2N-i}$. The hazard of death at age x for an individual with i alleles can be defined as

$$\mu(x|i)=(1-r)^i\mu_o(x), \qquad 0\le i\le 2N, \qquad (1.1)$$

where $(1-r)^i$ is the total risk of carrying the i longevity alleles and $\mu_0(x)$ is the baseline hazard rate. The baseline hazard $\mu_0(x)$ is the hazard of death for an individual with no longevity allele. For a mixed population, the total population survival is

$$\bar{s}(x)=\sum_{i=0}^{2N}p(i)s(x|i). \qquad (1.2)$$

$s(x|i)$ is the survival function for a homogenous population with i alleles, so that

$$s(x|i)=e^{-\int_0^x\mu(t|i)dt}=e^{-\int_0^x(1-r)^i\mu_o(t)dt}=s_o(x)^{(1-r)^i}. \qquad (1.3)$$

Now, (1.2) can be rewritten as

$$\bar{s}(x)=\sum_{i=0}^{2N}p(i)s_o(x)^{(1-r)^i}. \qquad (1.4)$$

From (1.4), we see that the total population survival is the weighted average over the survivals of subpopulations carrying i =0 to 2 N alleles of the longevity genes.

One must notice that the model is highly simplified by assuming all the genes have equal effects and function multiplicatively as well as independently. The real situation could be much more complicated due to gene-gene and gene-environment interactions. However, as Georg Box (1976) remarked, " all models are wrong but some models are useful". The simple model will turn out to be helpful in making interesting insights and applications in the genetic study of longevity.

1.1.3 Modelling heterogeneity

One can see that (1.4) holds only when the subpopulation with i longevity gene alleles is homogenous, both genetically and environmentally. Following Vaupel et al. (1979), when there is unobserved frailty z, the hazard of death for a heterogeneous subpopulation with i longevity alleles can be derived as

$$\bar{\mu}(x|i)=\int_0^\infty\mu(x|i,z)f_x(z)dz=\int_0^\infty z\mu(x|i)f_x(z)dz=\mu(x|i)\int_0^\infty zf_x(z)dz=\mu(x|i)\bar{z}(x).$$

Note that an individual's unobserved frailty z is a multiplicative proportional risk factor. It is assumed that frailty z captures all the effects that raise or lower the hazard of death, $\mu(x|i)$, for individuals with i longevity gene alleles. When z is gamma-distributed at birth with mean 1 and variance σ^2, we have $\bar{z}(x) = (1+\sigma^2(1-r)^i H_0(x))^{-1}$ where $H_0(x)$ is the cumulative function of the baseline hazard $\mu_0(x)$, $H_0(x) = \int_0^\infty \mu_0(x)dx$. Then we have

$$\bar{\mu}(x|i) = \frac{(1-r)^i \mu_0(x)}{1-\sigma^2(1-r)^i \ln s_0(x)}. \qquad (1.5)$$

The average survival of that subpopulation is

$$\bar{s}(x|i) = (1-(1-r)^i \sigma^2 \ln s_0(x))^{-1/\sigma^2}. \qquad (1.6)$$

PROOF: According to the Cox model, the hazard of death for one individual with frailty z and risk R is

$$\mu(x,z,R) = zR\mu_0(x).$$

The survival function is

$$s(x,z,R) = e^{-H(x,z,R)} = e^{-zRH_0(x)} = s_0(x)^{zR},$$

$$s(x,R) = E_z s(x,z,R) = E s_0(x)^{zR} = E e^{-zRH_0(x)},$$

where $z \sim \Gamma(k,\lambda)$.

Let $w = RH_0(x)$,

$$Ee^{-zw} = \int e^{-zw} f(z)dz = \int e^{-zw}\lambda^k z^{k-1}e^{-\lambda z}/\Gamma(k)dz = \int \lambda^k z^{k-1}e^{-(\lambda+w)z}/\Gamma(k)dz$$
$$= \lambda^k/(\lambda+w)^k \int (\lambda+w)^k z^{k-1}e^{-(\lambda+w)z}/\Gamma(k)dz = \lambda^k/(\lambda+w)^k = (1+w/\lambda)^{-k}.$$

Since $1/\lambda = \sigma^2, k = \lambda = 1/\sigma^2$, and replacing w with $RH_0(x)$, we have

$$s(x,R) = E_z s(x,z,R) = Ee^{-zRH_0(x)} = (1+\sigma^2 RH_0(x))^{-1/\sigma^2} = (1-\sigma^2 R\ln s_0(x))^{-1/\sigma^2}.$$

(1.6) is simply the case when $R = (1-r)^i$. **QED**.

Now (1.4) becomes

$$\bar{s}(x) = \sum_{i=0}^{2N} p(i)\bar{s}(x|i) = \sum_{i=0}^{2N} \binom{2N}{i} p^i (1-p)^{2N-i} (1-(1-r)^i \sigma^2 \ln s_0(x))^{-1/\sigma^2}.$$

(1.7)

The proportion of survivors with i longevity alleles at age x is simply

$$p(x|i) = \frac{p(i)\bar{s}(x|i)}{\bar{s}(x)} \tag{1.8}$$

so that proportion of subpopulation carrying i longevity alleles at age x is a function of the initial frequency of the allele p, risk reduction r, variance of heterogeneity σ^2, the number of alleles i and total population survival $\bar{s}(x)$.

With observed $\bar{s}(x)$ from population data, equations (1.4) and (1.7) can be solved numerically to get $s_0(x)$ when N, σ^2, p and r are known, and then the trajectory of $p(x|i)$ can be calculated from (1.8). In the latter chapters, the non-parametric approach will be further used in estimating the relative risks of genes based on a Two-step MLE in a relative risk model to be discussed later in Chapter 5.

1.2 Heritability of life span

The heritability of life span is by definition the percentage of genetic variation in life span σ_G^2 among total life span variance ${\sigma_T}^2$ of the entire population (Falconer & Mackay 1996)

$$h^2 = \frac{\sigma_G^2}{{\sigma_T}^2} = 1 - \frac{\sigma_W^2}{{\sigma_T}^2}, \tag{1.9}$$

where σ_W^2 is the life span variation due to environmental heterogeneity. The total variance in life span can be calculated as

$${\sigma_T}^2 = \int_0^\infty \sum_{i=0}^{2N} p(i)\mu(x|i)s(x|i)x^2 dx - \left(\int_0^\infty \sum_{i=0}^{2N} p(i)\mu(x|i)s(x|i)x dx \right)^2 \tag{1.10}$$

where i stands for the number of alleles for the subpopulations and the environmental variance is

$$\sigma_W^2 = \sum_{i=0}^{2N} p(i) \left[\int_0^\infty \mu(x|i)s(x|i)x^2 dx - \left(\int_0^\infty \mu(x|i)s(x|i)x dx \right)^2 \right]. \tag{1.11}$$

In practice, the total variance can be calculated directly from the observed individual life spans, and the genetic variance is

$$\sigma_G^2 = E(\bar{x}_i^2) - E(\bar{x}_i)^2 = \sum_{i=0}^{2N} \binom{2N}{i} p^i (1-p)^{2N-i} \left(\int_0^\infty s(x|i)dx \right)^2 - e_0{}^2 . \quad (1.12)$$

Here e_0 is the life expectancy at birth for the total population, which can be calculated from observed data. Now, when N, σ^2, p and r are given, $s_0(x)$ can be calculated from (1.7) and σ_G^2 from (1.12). Heritability h^2 can then be estimated using (1.9). The calculation for heritability is necessary in the following simulation studies in order to maintain the genetic influence at a reasonable level.

One must notice that the estimated heritability depends on the scale chosen for the response based on which the estimation is done. For example, using Danish twin data, McGue et al. (1993) estimated that heritability is 0.251 when age at death (above 15 years) was measured while 0.342 when it was measured in percentile. The estimation by Herskind et al. (1996) was based on a quadratic transformation of life span (again above age 15) and which is close to the estimate by McGue et al. (1993) without transformation. In the following simulations, we apply the approximate estimates from the two studies but base our calculation simply on age at death.

Although highly simplified, the binomial frailty model is nevertheless a useful tool for genetic study of longevity. The model is characterised by its direct incorporation of polygenic influences in modelling both individual and population survival. In the later chapters, the model will be modified and elaborated to investigate the genetic nature of life span correlation, estimate the likely number of genes that contribute to individual survival, measure the relative risks of genes that are relevant to longevity, and detect gene-environment and gene-sex interactions.

Chapter 2

Insights from the Binomial Frailty Model

2.1 Life span as a function of individual genetic make-up

Do long-lived individuals have a better genetic make-up than their short-lived counterparts? In both (1.3) and (1.6), the survival of a genetically identical population is defined as a function of the risk reduction of each gene allele and the number of such alleles. In order to examine how the genetic parameters contribute to an individual's life span, we assume that there are a total of N =40 diallelic loci at which longevity gene alleles are located and that each of the alleles has a risk reduction of r =0.25. The baseline survival function in (1.12) is expected from (1.7) by introducing the population survival of Danish male twins. Using (1.9), gene frequency p is adjusted so that heritability h^2 can be approximately 0.2-0.26 (Herskind et al. 1996). For two different situations with unobserved frailty, we set the variance of heterogeneity σ^2 =0.1 and 0.5. By simulation, the number of longevity alleles for people who died aged 60-65, 80-85, and over 100 are counted and then the differences in their individual genetic compositions compared (Table 2.1). Figure 2.1a presents the situation for a less heterogeneous population with σ^2 =0.1. The gene frequency has been adjusted to 0.1 so as to obtain a reasonable h^2 =0.23. The figure indicates clearly that long-lived individuals have a better genetic constitution than the short-lived. Looking at the distances between the peaks of frequency distributions for the 3 groups, we can see that this difference increases for centenarians.

What happens in a more heterogeneous population? In Figure 2.1b, the variance of heterogeneity is increased to 0.5 and the gene frequency had to be set at 0.15 to maintain a reasonable h^2 =0.224. We observe once again an increased average number of gene alleles with higher ages at death. Although the overall pattern in the two figures is the same, a careful comparison reveals interesting distinctions. First, the 3 peaks are closer to each other in Figure 2.1b than in Figure 2.1a, which reflects a

decrease in the importance of the genetic influence on survival although the total genetic component in life span is almost equal. Second, the bases of the 3 bell-shaped distributions are wider in a situation with higher heterogeneity, which indicates a more divergent genetic heterogeneity within each of the 3 groups.

Due to the high degree of environmental heterogeneity in Figure 2.1b, some individuals with a sound genetic composition failed to survive to old age. Some of the people who died aged 60-65 had as many as 20 longevity alleles. On the other hand, there are some individuals with fewer than 10 longevity alleles who managed to survive to age 100. One can conclude that, while genes play an important role in maintaining individual survival, environmental heterogeneity interferes with the genetic influence. As will be discussed in section 2.3, there is risk compensation that acts on environmental heterogeneity in such a way as to allow individuals with lower

Table 2.1 Medians of longevity gene alleles and of environmental frailty for individuals died at ages 60-65, 80-85 and above 100, N =40, r =0.25

Age at death	Number of individuals	Median of alleles	Median of environmental frailty z_e	95% range of z_e
		p=0.1, σ^2=0.1, h^2=0.23		
60-65	1156	7	1.03	0.58-1.60
80-85	1156	9	0.91	0.52-1.49
100+	1156	16	0.67	0.36-1.15
		p=0.15, σ^2=0.5, h^2=0.224		
60-65	2995	11	1.12	0.34-2.64
80-85	2995	13	0.60	0.15-1.75
100+	2995	19	0.07	0.01-0.37

environmental frailty but an unfavourable genetic composition the chance to survive to old age nonetheless. At the same time, some individuals with favourable genetic make-ups fail to reach old age due either to chance or to a high degree of environmental frailty. This point is supported by results in Table 2.1. When variance of environmental frailty is high, σ^2=0.5, medians of longevity alleles for the 3 groups are all higher than that when σ^2=0.1. Although with better genetic make-ups, the short-lived group has upper bound of 95% range for environmental frailty (z_e) as high as 2.64. While in Figure 2.1b some centenarians have got only as few as 10 alleles, we

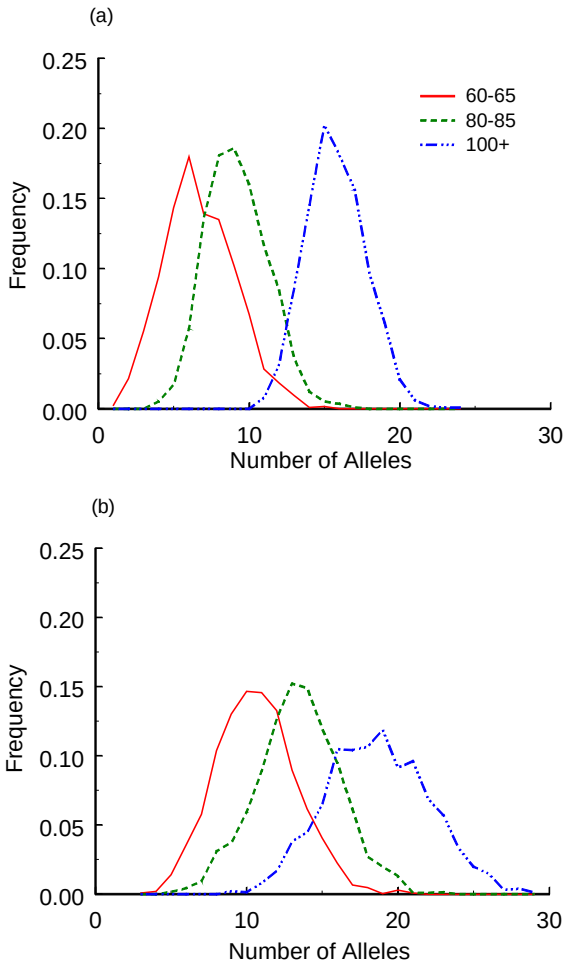

Figure 2.1 Distribution of number of longevity alleles for people died at different ages (a) p=0.1, r=0.25, N=40, σ^2=0.1, h^2=0.23 (b) p=0.15, r=0.25, N=40, σ^2=0.5, h^2=0.224

see that in Table 2.1 the same group has extremely low environmental frailty. We will discuss this compensation effect under the heading of risk compensation. Besides risk compensation, in empirical situations, random death (such as deaths from accidents, violence, etc.) is another important factor that blocks the genetic influence from being realised, especially at old ages.

2.2 Genotype frequency and mortality trajectories by age

When collecting data on individual genotypes, we are certainly interested in finding candidate genes that are of high importance in modulating life span. In order to make inferences about the observed genes, it is important to know how the effect of

one single gene can be perceived in a genetically and environmentally heterogeneous background. Suppose there is one observed gene allele with risk reduction r_1 and frequency p_1, and risk of carrying two such alleles is $(1-r_1)^2$. Like before, we assume there are N additional diallelic loci carrying longevity alleles, each with risk reduction r and frequency p. If environmental heterogeneity follows a gamma distribution with mean 1 and variance σ^2, the total population survival at age x is the weighted sum of survivals for the 3 sub-populations with 0, 1 and 2 alleles of the observed gene.

$$\begin{aligned}\bar{s}(x) &= (1-p_1)^2\sum_{i=0}^{2N} p(i)s(x|i,0) + 2p_1(1-p_1)\sum_{i=0}^{2N} p(i)s(x|i,1) + p_1{}^2\sum_{i=0}^{2N} p(i)s(x|i,2)\\ &= (1-p_1)^2\sum_{i=0}^{2N} p(i)(1+\sigma^2(1-r)^i H_o(x))^{-1/\sigma^2}\\ &+ 2p_1(1-p_1)\sum_{i=0}^{2N} p(i)(1+\sigma^2(1-r)^i(1-r_1)s_o(x))^{-1/\sigma^2}\\ &+ p_1{}^2\sum_{i=0}^{2N} p(i)(1+\sigma^2(1-r)^i(1-r_1)^2 s_o(x))^{-1/\sigma^2}.\end{aligned} \tag{2.1}$$

The genetic variance is now

$$\begin{aligned}\sigma_G^2 &= E(\bar{x}_i^2) - E(\bar{x}_i)^2 = (1-p_1)^2\sum_{i=0}^{2N} p(i)\left(\int_0^\infty s(x|i)dx\right)^2\\ &+ 2p_1(1-p_1)\sum_{i=0}^{2N} p(i)\left(\int_0^\infty s(x|i,r_1)\right)^2 + p_1{}^2\sum_{i=0}^{2N} p(i)\left(\int_0^\infty s(x|i,2r_1)\right)^2 - e_0{}^2.\end{aligned} \tag{2.2}$$

(2.2) can be used with (1.9) and (1.10) to estimate h^2.

2.2.1 The gene frequency trajectory by age

The right-hand side of (2.1) consists of 3 sub-populations carrying 0, 1 and 2 alleles of the gene with risk reduction r_1 and allele frequency p_1. But each of the 3 sub-populations is itself heterogeneous, both genetically and environmentally. This mimics the situation we are facing when looking at one specific gene because it is impossible to have all the biological and non-biological factors measured simultaneously. Given this complicated background, how does the frequency of one observed gene or genotype change with age in a heterogeneous population? From

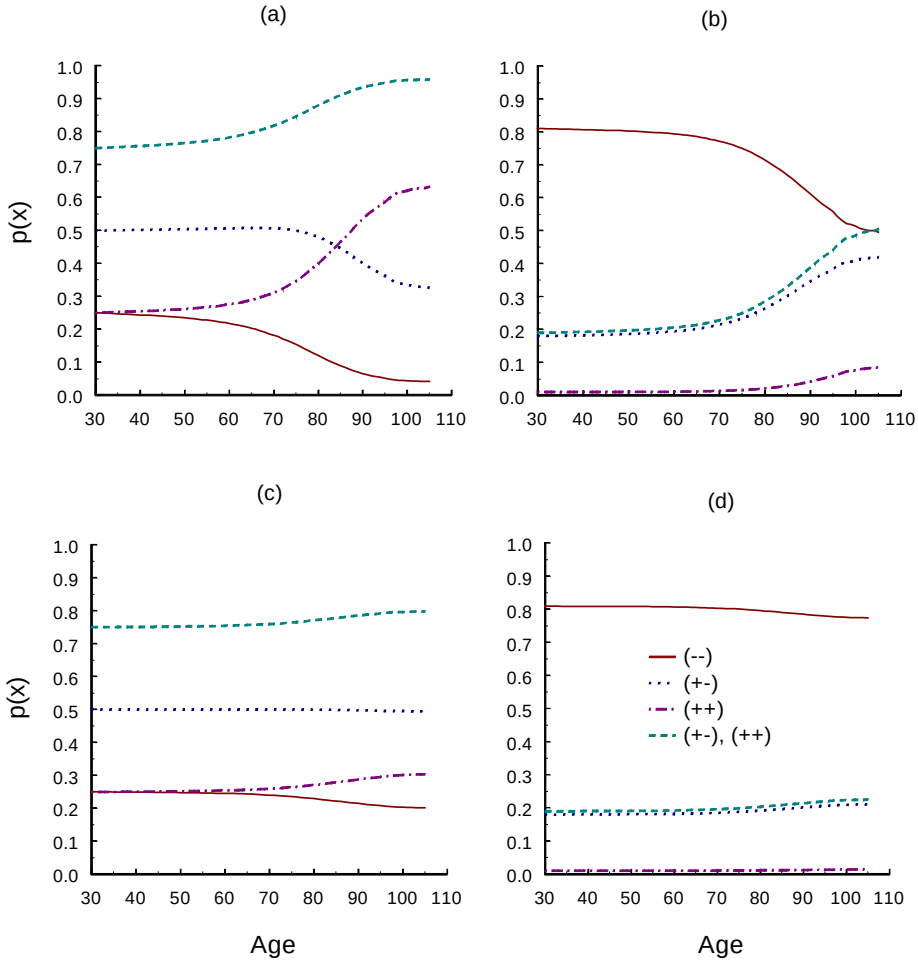

Figure 2.2 Genotype frequency trajectory by ages for one single gene allele (a) r=0.2, r_1=0.5, p_1=0.5; (b) r=0.25, r_1=0.5, p_1=0.1; (c) r=0.25, r_1=0.1, p_1=0.5; (d) r=0.25, r_1=0.1, p_1=0.1.

(2.1), proportions by age for people with 0, 1 and 2 alleles of the observed gene can be calculated as

$$\pi_0(x) = (1-p_1)^2 \sum_{i=0}^{2N} p(i)(1+\sigma^2(1-r)^i H_o(x))^{-1/\sigma^2} / \bar{s}(x),$$

$$\pi_1(x) = 2p_1(1-p_1) \sum_{i=0}^{2N} p(i)(1+\sigma^2(1-r)^i (1-r_1)s_o(x))^{-1/\sigma^2} / \bar{s}(x),$$

$$\pi_2(x) = p_1^2 \sum_{i=0}^{2N} p(i)(1+\sigma^2(1-r)^i (1-r_1)^2 s_o(x))^{-1/\sigma^2} / \bar{s}(x). \qquad (2.3)$$

To examine the frequency trajectories for various situations, the genetic parameters are set as $N=40$, $p=0.1$, $\sigma^2=0.5$, r_1=0.5 or 0.1, p_1=0.5 or 0.1 and $\pi_0(x)$, $\pi_1(x)$ and $\pi_2(x)$ are calculated for their different combinations. The genetic

component in life span has been maintained at around $h^2 \approx 0.2$ by adjusting risk reduction for each of the longevity alleles. In Figure 2.2, the four plots correspond to high effect-high frequency (a), high effect-low frequency (b), low effect-high frequency (c), and low effect-low frequency (d). The patterns in Figure 2.2 indicate that only genes with strong effects manifest changes in their frequencies while the frequencies of genes with small effects remain almost constant with age regardless of their initial frequencies. In Figure 2.2a, the proportion of genotype $(-+/+-)$ decreases because of a rapid increase in frequency of genotype (++) in the population. But the proportion of all carriers of the allele, which is the sum of (++) and $(-+/+-)$, increases constantly (Figure 2.2a, b).

In empirical studies, the genotypic frequency changes can be observed from data containing individual genotype information and age at participation. This frequency change provides a clue from which inferences about frequency and risk of the gene can be made, based on the binomial frailty model.

2.2.2 The genotype specific mortality trajectories

Corresponding to (2.3), the hazards of death for the 3 genotypes are

$$\begin{aligned}
\tilde{\mu}_0(x) &= (1-p_1)^2 \sum_{i=0}^{2N} p_i(x)(1-r)^i \mu_o(x) \\
&\quad / (1+\sigma^2(1-p_1)^2 \sum_{i=0}^{2N} p_i(x)(1-r)^i H_0(x))^{-1/\sigma^2}, \\
\tilde{\mu}_1(x) &= 2p_1(1-p_1) \sum_{i=0}^{2N} p_i(x)(1-r)^i (1-r_1)\mu_o(x) \\
&\quad / (1+\sigma^2 2p_1(1-p_1) \sum_{i=0}^{2N} p_i(x)(1-r)^i (1-r_1) H_o(x))^{-1/\sigma^2}, \\
\tilde{\mu}_2(x) &= p_1^2 \sum_{i=0}^{2N} p(i)(1-r)^i (1-r_1)^2 \mu_o(x) \\
&\quad / (1+\sigma^2(1-p_1)^2 \sum_{i=0}^{2N} p_i(x)(1-r)^i (1-r_1)^2 H_o(x))^{-1/\sigma^2}.
\end{aligned} \tag{2.4}$$

In Figure 2.3, the parameters are set as $N = 40$, $r_1 = 0.5$, p_1=0.1, p=0.1, σ^2=0.5 (a) and σ^2=0 (b) to see how the mortality trajectory for carriers of one strong-effect gene

changes with age and, at the same time, to examine the influence of genetic and environmental heterogeneity. Risk reduction for the $2N$ gene alleles has to be adjusted to 0.21 (a) and 0.2 (b) so as to achieve a reasonable value for h^2. Figure 2.3a

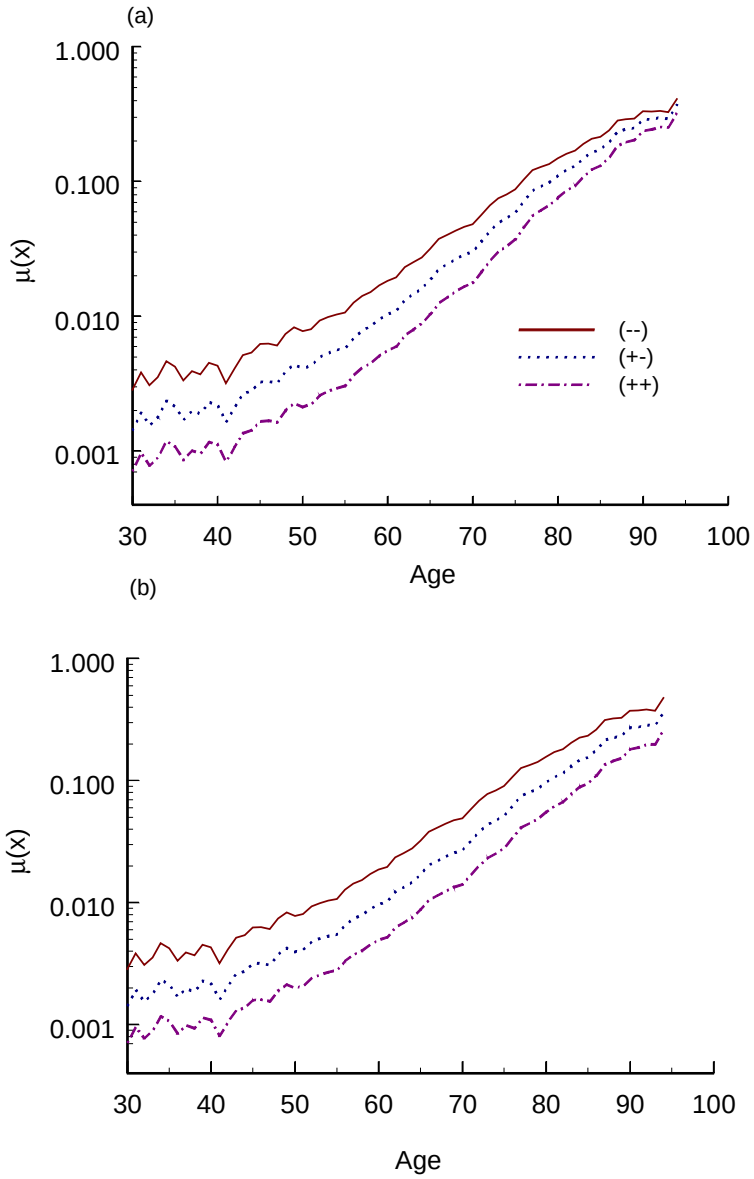

Figure 2.3 Genotype specific mortality trajectory for one gene with r_1=0.5 and p_1=0.1 assuming there are 2N=80 identical gene alleles each with p=0.1 (a) r=0.21, σ^2=0.5; (b) r=0.2, σ^2=0. h^2 is adjusted to about 0.2

presents the situation when individuals carrying each of the 3 genotypes $(++,-+/+-,--)$ are heterogeneous both genetically and environmentally. Although the single gene in question is beneficial, the mortality curves of the 3 groups converge

at very old ages. In Figure 2.3b, σ^2=0, which means that individuals in each of the 3 groups are identical except for their genetic make-ups, i.e., there are no differences due to environmental variation. Without environmental heterogeneity, a strong convergence like in Figure 2.3a is no longer observed, but a slight convergence due to genetic heterogeneity does exist. The relative significance of the phenomenon could be important for two reasons. First, the mortality trajectory of certain genes or genotypes is influenced not only by environmentally induced heterogeneity but also by the existence of differences in the genetic composition of individuals. Second, the influence of one particular gene decreases with advancing age as the mortality curves for populations with and without the gene converge. In general, the two terms, genetic and environmental variations, can not be discriminated with data on independent individuals. The split is extremely model-sensitive as illustrated by Hougaard (1991, 1995). In our simulation approach here, we are assuming that individual's total genetic make-up is known. The gamma distributed environmental frailty is just another layer of frailty variation imposed in order to create an overall picture of heterogeneity. Even though we are not trying to identify the environmental frailty but instead creating it in the simulation, one must keep in mind that the model assumption does matter. We just use the gamma frailty model in the simulation for convenience consideration.

2.3 Risk compensation

One puzzling phenomenon is the fact that there are long-lived individuals who might nevertheless exhibit some harmful attributes such as smoking. Such an incidence raises two questions. First, is this due to a compensation effect? Second, are such individuals genetically different from the others? To answer these questions, we assume that there is one attribute, smoking, that increases the hazard of death by 1.5 and the proportion of people who smoke is 10%. In addition, let us assume that there is one beneficial gene with a risk reduction of 0.3 and allele frequency 0.1. Let us assume further that there are N=40 diallelic loci hosting at most 2N=80 gene alleles, each with risk reduction r=0.15 and frequency p=0.1, for an environmentally heterogeneous population with the variance of unobserved environmental frailty σ^2=0.5 and mean 1, so that the genetic component of life span is controlled at h^2=0.237. A simulation experiment generated 3,780 individuals who survived to

above the age of 100. We grouped people into the categories of smoker and non-smoker, carrier and non-carrier of the strong-effect gene and calculated proportion of carriers of the gene allele among smokers and non-smokers (Table 2.2). The result

Table 2.2 Gene carriers among smokers and non-smokers

Smoking	Longevity gene			
	+	–	Sum	Prop.
+	68	108	176	0.386
–	1143	2461	3604	0.317
Sum	1211	2569	3780	0.320

shows that the proportion of carriers of the beneficial allele among smokers (38.6%) is higher than that among the non-smokers (31.7%), which means that some long-lived smokers are simply lucky people with genetic make-ups better than the long-lived non-smokers. The better genetic composition compensates for the harmful attributes and provides a better opportunity for people to survive to old age even if some of them smoke. Notice that the proportions of the allele carriers among both centenarian smokers and centenarian non-smokers are much higher than the initial frequency (10%). Meanwhile, the proportion of smokers among centenarians has decreased from 10% at the beginning till less than 5% (176 out of 3,780).

Risk compensation is a common phenomenon that exists in any consequence resulted from a "complex" organisation of causal agents. For example, Fox and Collier (1976) reported that survival of the healthier men in the industry is a major factor that contributes to the low mortality rates in the workers in manufacture of polyvinyl chloride in Great Britain, which is just another example of risk compensation. However, the present simulation study points out, for the first time, that such an mechanism is also involved in the context of longevity study and emphases the importance of heterogeneity.

Since the compensation effect is due to the existence of individual differences, one would be highly advised to take heterogeneity into account when attempting to infer the risk of some biological or social attribute that contributes to longevity. In addition, the result also points to the importance of conducting centenarian studies. Since the centenarians constitute a special population representing successful aging, such investigation could help to find out the key attributes that contribute to longevity.

Summary

In Part one, a highly simplified binomial frailty model was developed. The model assumes that there are a number of diallelic longevity loci. Gene alleles, each of which lowers the risk of death at all ages by a factor $(1-r)$, are located at the loci with binomial probability p. If a person has n such alleles, then the risk of death is lowered by the factor $(1-r)^n$ at all ages. Formulas are derived for estimating the baseline age-trajectory of mortality and life span heritability h^2 in Chapter 1. Some interesting insights based on the model are presented in Chapter 2. The key conclusions from Part one are:

(1) The relationship between life span and individual genetic make-up is non-linear and complicated due to the existence of non-genetic heterogeneity that compensates for and interferes with the genetic influence on life span (section 2.1).

(2) The observed gene or genotype frequency trajectory can provide clues for making inference on the genetic influence of corresponding gene or genotype (section 2.2.1).

(3) The mortality trajectory of a certain gene allele or genotype can be influenced by both environmental and genetic heterogeneities. The influence of one particular gene allele or genotype decreases with advancing age as the mortality curves for populations with and without the gene converge (section 2.2.2).

(4) The genetic and the environmental frailties can compensate each other so that long-survivors with bad non-genetic attributes tend to possess better genetic constitutions (section 2.3).

In the subsequent chapters, the model will be modified and elaborated to yield other theoretical insights as well as some interesting findings.

PART II

Models for Life Span Correlation among Related Individuals

Chapter 3

The Inheritance of Life Span

3.1 Introduction

Many genealogical studies have revealed the correlation of life spans between parents and offspring and among siblings as well. The earliest literature can be traced back to 1899, when Beeton and Pearson first reported life span correlation within families. Among the later studies was Bell (1918), who concluded that the influence of the paternal side predominates over that of maternal side. Pearl (1931) detected a weak positive correlation between parents and children with differing coefficients in different families. A Finnish and Swedish genealogical study by Jalavisto (1951) revealed a predominant correlation between the maternal life span and the life spans of children, which is contrary to Bell. A study of the offspring of nonagenarians by Hawkins (1965), who used data collected between 1922 and 1930 by Pearl and Pearl, indicated a weak familial tendency to longevity. Later, also with data collected by Pearl and Pearl, Abbott (1974) detected a weak but positive correlation between parents and offspring, with a closer relationship between mother and children. Based on a follow-up study of 2,370 middle-aged civil servants and their spouses, Vandenbroucke (1984) pointed out that the parental influence on the life span of the progeny is independent of disease and health variants. In addition, an adoption study by Sørensen et al. (1988) reported a strong genetic background in premature death in adults especially deaths due to infectious and vascular causes. Although some of the findings are controversial, data from well-defined populations support the notion that there exist transmittable familial attributes affecting life span, and a certain portion of them could be of a genetic nature. Life span is a continuously distributed quantitative trait, rather than a dichotomous or categorical trait. There are some distinctive features of quantitative trait as regards life span. First, it is a polygenic phenomenon related to

the actions of multiple genes, each with a small effect, rather than to one single major gene with a large effect (Martin 1997; McClearn 1997). The polygenic feature of life span makes it difficult to study the pattern of transmission. This is similar to the case of multifactorial diseases, where the paradigm established for the study of single-gene diseases is no longer considered appropriate (Sing et al. 1996). Second, for a polygenic trait, the role of the environment is explicitly included so that aspects of a shared family environment can contribute. Given this complexity, life span correlation should be studied on a quantitative genetic base. The binomial frailty model introduced in Chapter 1 is a suitable model in this regard. In this chapter, the binomial frailty model will be elaborated to study the transmission of longevity genes as well as of life span. By simulation, life span correlation will be compared with correlation of individual frailty and then the age pattern of the correlation will be examined to see how the genetic component in life span changes over age.

3.2 Inheritance of longevity genes

The binomial frailty model can be extended to study the inheritance of genetic frailty and longevity. In this section, we first model the probability of gene transmission and then study the relationship between frailty correlation and correlation in life span.

If the father has n_1 longevity alleles and the mother has n_0, the conditional probability for the child to have n alleles is

$$P(n|n_0,n_1)=\sum_{i=0}^{n} Hyper(i|n_o,N,2N)\,Hyper(n-i|n_1,N,2N) \quad 0\le n_0,n_1\le 2N \tag{3.1}$$

where $$Hyper(i|n_o,N,2N)=\frac{\binom{n_0}{i}\binom{2N-n_0}{N-i}}{\binom{2N}{N}}$$

is a hypergeometric distribution because, among all the alleles for one individual, half are from the father and half are from the mother without replacement.

According to Bayes' theorem, the conditional probability of the father having n_1 and the mother n_0 alleles, given the child has n is

$$P(n_0,n_1|n)=\frac{P(n|n_0,n_1)P(n_0)P(n_1)}{\sum_{n_0=0}^{2N}\sum_{n_1=0}^{2N}P(n|n_0,n_1)P(n_0)P(n_1)}\ . \qquad (3.2)$$

The denominator in (3.2) is the probability that an individual (here the offspring) carries n alleles, which is by assumption of section 1.1.2 given by the binomial probability $B(n|p,2N)$, i.e. $\sum_{n_0=0}^{2N}\sum_{n_1=0}^{2N}P(n|n_0,n_1)P(n_0)P(n_1)=B(n|p,2N).$

Replacing $p(n|n_0,n_1)$ in the numerator of (3.2) with (3.1), and (3.2) becomes

$$P(n_0,n_1|n)=\frac{\sum_{i=0}^{n}Hyper(i|n_0,N,2N)Hyper(n-i|n_1N,2N)B(n_0|p,2N)B(n_1|p,2N)}{B(n|p,2N)},$$

$$0\le n\le\frac{1}{2}(n_0+n_1)\ . \qquad (3.3)$$

Given the total number of alleles an individual has, the probability the mother has n_0 and father has n_1 alleles can now be easily calculated using (3.3) when p and N are known.

Formula (3.3) is important because it is the probability of the father and mother having a given number of alleles, given n for the offspring. For a fixed n, we can calculate $P(n_0,n_1|n)$ and its cumulated probability so that the parents' number of alleles can be estimated.

When $P(n|n_0,n_1)$ is known from (3.1), the probability density function of observing the life-span pattern for a triple (mother, father and child) is

$$f(x_0,x_1,x)=\sum_{n_0=0}^{2N}\sum_{n_1=0}^{2N}\sum_{n=0}^{(n_0+n_1)/2}f(x_0|n_0)P(n_0)f(x_1|n_1)P(n_1)f(x|n)P(n|n_0,n_1) \qquad (3.4)$$

where x_0, x_1 and x are the ages at death of mother, father, and child. The functions $f(x_0|n_0)$, $f(x_1|n_1)$ and $f(x|n)$ are the probability denstity functions of the conditional survival distributions for the mother, the father and the child carrying n_0, n_1 and n longevity alleles. (3.4) can be applied to empirical observations and used as the likelihood function for parameter estimation.

Specification of a beginning age (after the age of reproduction) is needed when applying (3.4) to real observations since the data is conditional on the parental survival to allow them for reproduction. In addition, (3.4) holds only when mating is random, which is a popular assumption in population genetics.

3.3 Inheritance of life span

As a polygenic phenotype, life span is modulated by both genetic and environmental factors (Schachter et al. 1993; Vaillant 1991; Christensen & Vaupel 1996). What kind of conclusion can be deduced by studying life spans of related individuals using genealogy, twin data, or adoption data (Petersen, PhD thesis) without any knowledge about individual genetic make-up? To answer this question, we use simulation studies based on the binomial frailty model to show the association between correlation of life span and correlation of individual frailty. Using the same parameters as in Figure 2.1b, i.e., $p = 0.15$, $r = 0.25$, $N = 40$, $\sigma^2 = 0.5$, we simulated 150,000 nuclear families, each with four members (parents and two children). The correlation coefficients of life span and individual frailty between siblings and between parents and children are calculated for 150,000 sib-pairs and 300,000 father-child/mother-child pairs (Table 3.1). As we see in Figure 2.1b, the heritability estimate is 0.224, which is about twice the correlation of siblings' life spans since we assume that all the genes act multiplivatively and independently on the baseline hazard without intercations (Falconer & Mackay 1996).

Table 3.1 Life span correlation vs frailty correlation

Correlation	Siblings	Parent-child
Life span	0.124	0.122
Frailty	0.210	0.206
Genetic frailty	0.500	0.500
Environmental frailty	0.002	0.006
Sample size	150000	300000

In the simulation, frailty for individual i is calculated as the product of genetic frailty (z_g) due to his or her genetic make-up and environmental frailty (z_e), i.e., $z_i = z_{g_i} z_{e_i} = (1-r)^{n_i} z_{e_i}$. z_e is gamma-distributed with mean 1 and variance σ^2. It is

interesting to see that the correlation of individual frailty both between the siblings and between the parent-child pairs is higher than the life span correlation (Table 3.1). This phenomenon, which was first described by Vaupel (1988), indicates that it would be more useful to apply a model based on frailty correlation rather than on life span correlation since a trivial correlation of life span does not necessarily mean a low correlation of frailty. Models based on frailty correlation have been developed and applied to studies on related individuals such as twins (Yashin & Iachine 1995, 1997; Yashin et al. 1999) and kinship-pairs (Tan et al. submitted). The results from these empirical studies support the above conclusion. The correlation of genetic frailty is 0.5 for all kinship pairs from the simulation. The correlation of environmental frailty is about zero, since we assume there is no influence due to a shared environment.

One question concerning life span correlation is whether relatives of long-lived individuals tend to live long as well? To find out the answer, we first converted individual life spans into survival probabilities and then selected sib-pairs from the 150,000 families with siblings who were in the top 10% and 1% of survivors in the population. As a result, we got 15,000 sib-pairs, each with at least one member who belonged to the top 10% of survivors and 1,500 sib-pairs, each with at least one member belonging to the top 1% of survivors. The distribution of survival probability for siblings among the top 10% and 1% of longest-lived individuals are plotted in Figure 3.1a. As can be seen, siblings of long-lived individuals have a higher probability of living a long life and a lower probability of dying early than “ordinary” people as a whole, an indication that the genetic influence on life span matters – especially at very old ages. Furthermore, as the percentage level increases from 10% to 1%, we observe a wider outspreading of the frequency (Figure 3.1b). This tells us that the more selected the individuals are, the greater the survival time of their co-sibs will differ from that of the ordinary people. This result is supported by empirical observations of family genealogies of the European nobility (Tan et al. submitted). The result does not mean, however, that the genetic component of life span gains in importance with increasing age. On the contrary, it becomes less significant due to environmental heterogeneity and chance. We shall discuss this problem in the next section.

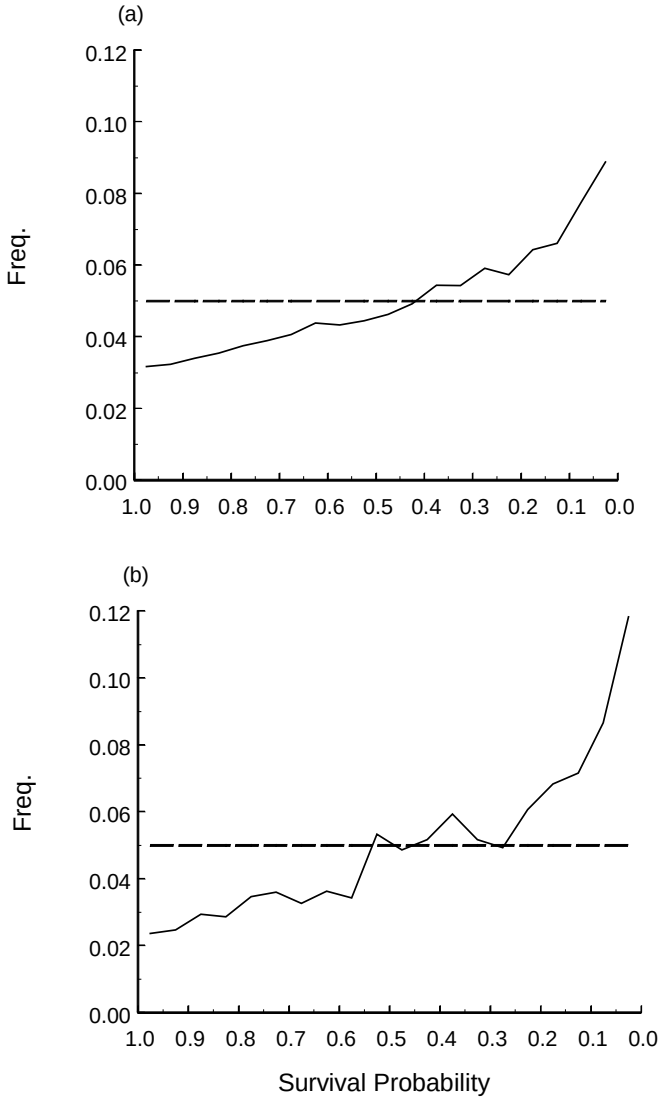

Figure 3.1 Distribution of survival probability for siblings of the top 10% (a) and 1% (b) longest-lived individuals.

3.4 The genetic components in life span as a function of age

Given the existence of family life-span correlation, an interesting question arises. Does the correlation on remaining life span become stronger, weaker, or remain the same with increasing age? The answer to this question is uncertain as yet, although there have been several studies touching on the topic (Becquet-Appel & Jakobi 1990; Gavrilova et al. 1998). Unfortunately these studies are confronted with two

difficulties. The first has simply to do with data resources. One needs large sample sizes to arrive at reliable parameters for the old ages when most individuals have died. The second difficulty can be ascribed to methodology, but it originates from the distribution of life spans. The distribution of life spans is skewed, such that conventional methods produce estimates of parameters that contain a scale effect (Falconer & Mackay 1996), and the significance of parameters cannot be tested statistically since the testing procedures are usually based on the assumption of a normal distribution.

Using simulated data from Table 3.1 and Figure 3.1, pairs in which the two sibs survived to different ages, ranging from survival probability 1 to 0.1, are first selected and then the correlation coefficients of ages at death and of frailty between the selected sib-pairs are estimated. In Figure 3.2, the correlation coefficients are plotted as a function of survival, with median and range calculated from 100 repeats for each point. As can be seen, both the correlation of life span and the correlation of frailty decrease with age. This means that the genetic similarity among siblings becomes less and less important in determining their remaining life spans. Although these results are from a simulation based on the simple proportional hazard and the binomial genetic frailty assumptions, they are supported by an empirical study of the genealogies of European nobility by Tan and Vaupel (unpublished data). The study reported a declining pattern of life span correlation between sib-pairs and between parents and children with increasing age.

Two phenomena became apparent in this study that seem to contradict each other. While Figure 3.1 indicates that a long-lived person is likely to have long-lived siblings, the age pattern of life span or frailty correlation indicates a declining trend with increasing age. In order to clarify the contradiction, we compare the results from Figure 3.1 with those of Figure 3.2a. The intuition from Figure 3.1 is straightforward: siblings of a long-lived individual tend to live long as well. This can be explained by the increased probability that siblings of a long-lived sib inherit more longevity-attributes from their parents than other individuals (Vaupel 1988). But in Figure 3.2a we are testing the importance of biological and social similarities in determining the remaining life spans. The conclusion from Figure 3.2a is conditional on the fact that both members of the sib-pair have survived up to that point, while, at the same time, surviving to that stage is a consequence of, among other things, existing similarities.

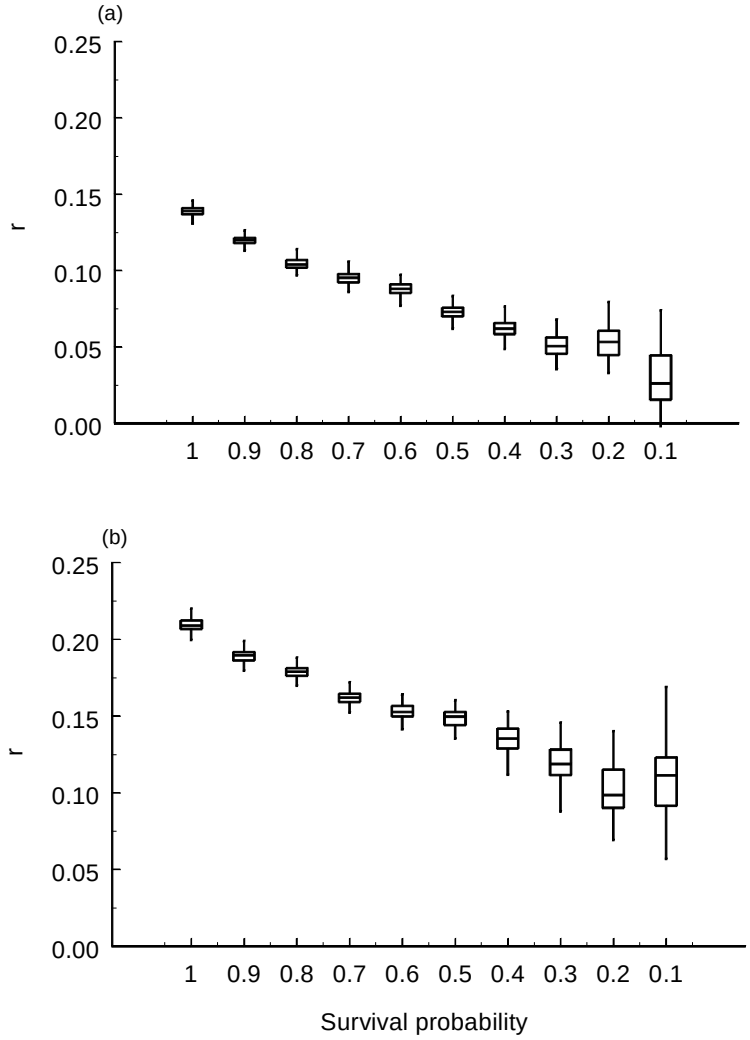

Figure 3.2 The age patterns of life span (a) and frailty (b) correlation among siblings. The figures show boxes contain half the r values obtained from 100 repeats as well as the whole range of r values.

This means what we see, in Figure 3.1, is the accumulated consequence brought about by the similarities. In Figure 3.2a, on the other hand, we see the conditional importance of these similarities in maintaining the rest of their survival. The declining trend we observe does not mean, however, that the importance of the genetic influence decreases, since we assume a constant genetic effect over all ages at the individual level.

Hougaard (1995) pointed out that the pattern revealed by Figure 3.2 is dependent on the assumption of frailty distribution. However, since the pattern gains supports from empirical observations. We suppose that the binomial genetic frailty is a convenient and useful approach in this regard.

3.5 Summary

The simulation studies on family correlation of life span in this chapter show

(1) Behind the modest correlation of life span, there is a strong correlation of individual frailty among first-degree relatives (section 3.3).

(2) Siblings of long-lived individuals have higher survival than the ordinary population (section 3.3).

(3) The observed life span correlation among related individuals decreases with increasing age given the binomial genetic frailty assumption (section 3.4).

These conclusions depend on the strict and simplistic assumptions that underlie the binomial frailty model. Empirical research (Tan et al. submitted; Tan & Vaupel unpublished data; Vaupel 1988; Yashin & Iachine 1995, 1997; Yashin et al.1999), however, has produced some similar findings that are consistent with the above conclusions.

Chapter 4

Estimating the Number of Human Longevity Genes

4.1 Introduction

The genetic contribution to human longevity has been explicitly ascertained by studies of twins (Wyskak 1978; Carmelli 1982; Harris et al. 1992; Hayakawa et al. 1992; McGue 1993; Herskind 1996) and by studies of candidate genes (Bladbjerg et al. 1999; Bathum et al. 1998; De Benedictis et al. 1997, 1998a, 1998b; Ivanova et al. 1998; Yashin et al. 1998, 1999b; Toupance et al. 1998). Like any particular phenotype that is continuously distributed, life span is a complex trait influenced by a combination of both multiple genes and environment with possible interactions (Vaillant 1991; McClearn 1997). The genetic contribution to life span has been estimated as accounting for about 25% of the total variance in life span (McGue 1993; Herskind 1996). Given the polygenic feature of life span and its percentage of genetic contribution to it, an important question remains to be answered either explicitly or implicitly is: How many genes are involved (Wachter 1997)? Assuming that there are 100,000 genes in the human genome, Martin (1987, 1997) estimated, based on phenotype analysis, that there could be about 7,000 genes associated with longevity if such allelic variation or mutation took up to 7% of the genome. He also predicted there could be a smaller sub-set of the genes that may manifest large effects. In this chapter, estimates based on another approach are to be presented using survival data of Danish twins. The estimation will be based on the binomial frailty model with unobserved environmental heterogeneity introduced in the previous chapters. The strategy also relies on the heritability estimates from previous studies on the same data (McGue et al. 1993; Herskind et al. 1996). Two other data sets that provide life span correlation for siblings, the Quebec genealogy and the European noble family genealogy, are also examined in order to compare the differences in the estimates.

4.2 Materials and methods

4.2.1 Data source

The Danish twin data is taken from The Danish Twin Registry established in 1954. The data consists of 5,072 male twins and 5,264 female twins born between 1870 and 1900 who survived above age 30 drawn from a total of 5,465 male twins who died at ages 6-105 and a total of 5,735 female twins who died at ages 6-103. Life expectancy at age 30 is 41.5 for male twins and 43.7 for female twins. Heritability has been estimated as 0.23 from male twins and 0.20 from female twins (Herskind et al. 1996). Using the same data but for twins born between 1870-1880, McGue et al. (1993) estimated a heritability around 0.25 both for males and for females.

The Quebec genealogy contains 13,554 individuals born from 1620 to 1705 who were descendants of French immigrants and who lived in the St. Lawrence Valley of Quebec their whole lives. There are 5,336 males and 6,440 females who survived above age 20. The mean life span is 63 for males and 60.5 for females. A total of 4,443 pairs of full-brothers and 6,269 pairs of full-sisters are found. Correlation coefficient for full-brothers is 0.083 and for full-sisters is 0.075. Heritability can be roughly estimated as 0.15 for both sexes when assuming a predominant additive genetic inheritance.

Computerized at Max-Planck Institute for Demographic Research in Rostock, Germany, the European Noble Family Genealogy Database (EuroGen) is a large data resource for study of family longevity. In this study, only siblings born between 1870-1900 and who survived above age 30 are chosen. This is based on two considerations, (1) to create a time-period consistent with that of the Danish twin data and (2) to minimize any cohort effect introduced by improved survival over time. A total of 2,610 male sib-pairs and 984 female sib-pairs were compiled (more males were recorded). Life span correlation for brothers is 0.11 and for sisters is 0.20. Since the estimate for sisters is less reliable due to small number of observations, only male sib-pairs are used for the calculation.

4.2.2 Estimation strategy

In section 1.2 of Chapter 1, the heritability of life span was derived following the definition from quantitative genetics using the basis of a binomial frailty model

and incorporating a polygenic modulation on survival. Recalling (1.9), (1.11) and (1.12), the heritability of life span can be defined as

$$h^2 = \sigma_G^2 / {\sigma_T}^2 = \frac{E(\bar{x}_i^{\,2}) - E(\bar{x}_i)^2}{\sigma_T^2} = \frac{\sum_{i=0}^{2N} \binom{2N}{i} p^i (1-p)^{2N-i} \left(\int_0^\infty s(x|i)dx \right)^2 - {e_0}^2}{\sigma_T^2}. \tag{4.1}$$

In (4.1), σ_T^2 is the total variance of life span. It can be estimated from individual observations when data on individual life spans are available with e_0 being the life expectancy at birth which again can be estimated from the sample and $s(x|i)$ is the survival function for the subpopulation carrying i longevity alleles. Given the risk reduction r and frequency p for each of the alleles, $s(x|i)$ can be calculated using (1.3) when assuming no environmental heterogeneity, and using (1.6) when heterogeneity is considered. In both cases, the baseline survival function $s_0(x)$ can be obtained using the non-parametric approach discussed in section 1.1.3.

Now in (4.1), h^2 is a function of number of loci N, risk reduction of the allele r and allele frequency p. When h^2 is known from previous studies, the number of loci where longevity alleles can be observed can be estimated given the risk reduction of each allele and its frequency.

4.3 Results

In Figure 4.1, the calculated heritability h^2 is shown as a function of number of loci N for given genetic parameters (risk reduction and frequency) and variances of environmental heterogeneity using Danish female twin survival data. The overall pattern in Figure 4.1 tells us:

(1) When genetic parameters are the same, the existence of environmental heterogeneity reduces heritability (comparing the 3 curves at the bottom).

(2) High allele frequency can result in high h^2 (comparing the bold-solid and the dash-dotted lines in the middle).

(3) The higher the effect of the allele, the higher the heritability (comparing the dash-dotted and the solid lines) of life span.

(4) Heritability increases with number of longevity loci.

To get a reasonable number of loci for a given set of parameters, only these values of N are considered when $h^2 \approx 0.25$. From Figure 4.1, four such values of N corresponding to the four sets of parameters are obtained. Do the same for all the combinations of $r \in \{0.05, 0.1, 0.25, 0.5\}$, $p \in \{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$

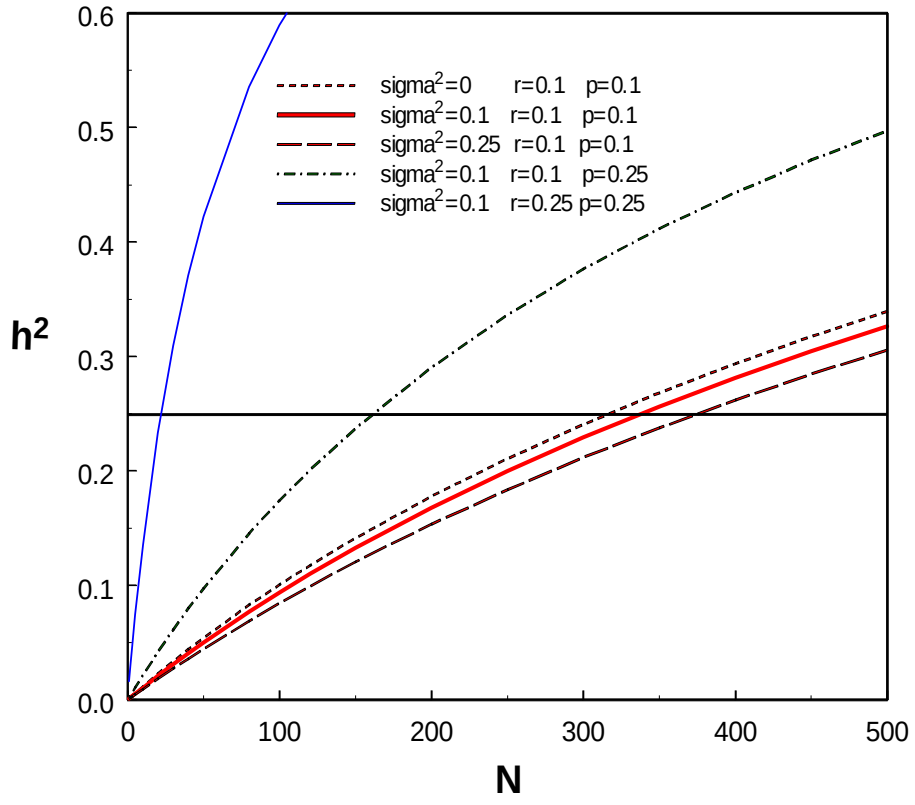

Figure 4.1 Heritability h^2 as a function of number of loci N for given genetic parameters (riskreduction and frequency of allele) and variances of environmental heterogeneity using Danish female twin survival data.

and $\sigma^2 \in \{0, 0.1, 0.5\}$, the likely corresponding numbers of longevity diallelic loci are estimated (Table 4.1). The specification for each parameter in Table 4.1 covers a wide range of variations. For example, risk reduction r begins from 0.05, which is small, to 0.5 when one allele alone can cut the hazard of death by half. As a result, a mere dozen gene alleles with strong influence can have an equivalent influence on maintaining heritability as perhaps several thousands of genes with lesser effect. For the same genetic parameters r and p, the estimated number increases when

heterogeneity is included and it goes up further when σ^2 is increased from 0.1 to 0.5. This conclusion illustrates a well-known result from Vaupel (1979) who noted that the effect of a covariate was reduced when frailty is included in the model. The numbers in each of the three blocks in Table 4.1 are nearly symmetric on the allele frequency as

Table 4.1 Upper bounds, when $h^2 \approx 0.25$, for the number of diallelic loci with allele frequency $\geq p$ and risk reduction $\geq r$

Proportion p	Risk reduction r			
	0.05	0.1	0.25	0.5
σ^2=0:				
0.05	2500	600	82	16
0.1	1300	320	43	8
0.25	635	150	20	4
0.5	472	110	15	3
0.75	634	146	20	4
0.9	1300	307	40	7
0.95	2470	580	75	13
σ^2=0.1:				
0.05	2700	650	90	18
0.1	1420	330	48	9
0.25	700	160	22	4
0.5	500	125	16	3
0.75	670	158	21	4
0.9	1400	330	43	7
0.95	2600	620	81	13
σ^2=0.5:				
0.05	3500	830	115	21
0.1	1820	440	60	11
0.25	880	210	28	5
0.5	650	155	21	4
0.75	866	200	27	5
0.9	1780	424	55	9
0.95	3460	800	105	18

a result of binomial distribution. Since the numbers are numbers of diallelic loci, the number of longevity genes they could carry is doubled, since we assume there can be at most two alleles occupying one locus. The most important information of Table 4.1 is that, if longevity is due to genes with very strong effects, there could be only about a dozen with a prevalence of at least 5%; but if longevity is related to genes with small effects, there might be several thousands. If both strong and weak genes are involved, we wouldn't expect to find many genes with strong influence. In a very extreme case when the allele has very small risk reduction (r=0.05), very low frequency (p=0.05)

and large variance of heterogeneity (σ^2=0.5), we get an estimated number of about 7,000 genes. This number is interesting because it coincides with the number Martin (1997) had predicted.

The same calculations have been done on two other data but only for one set of parameters (r=0.1, σ^2=0.1, p=0.1) in order to check how the estimated numbers of loci vary between different data sources. The curves in Figure 4.2 represent siblings from Quebec genealogy (male: dashed; female: dotted), Danish twins (male: dash-dotted; female: solid) and male sibs from European noble families (light dashed). Symbols on the curves mark the numbers of loci estimated for given heritability from

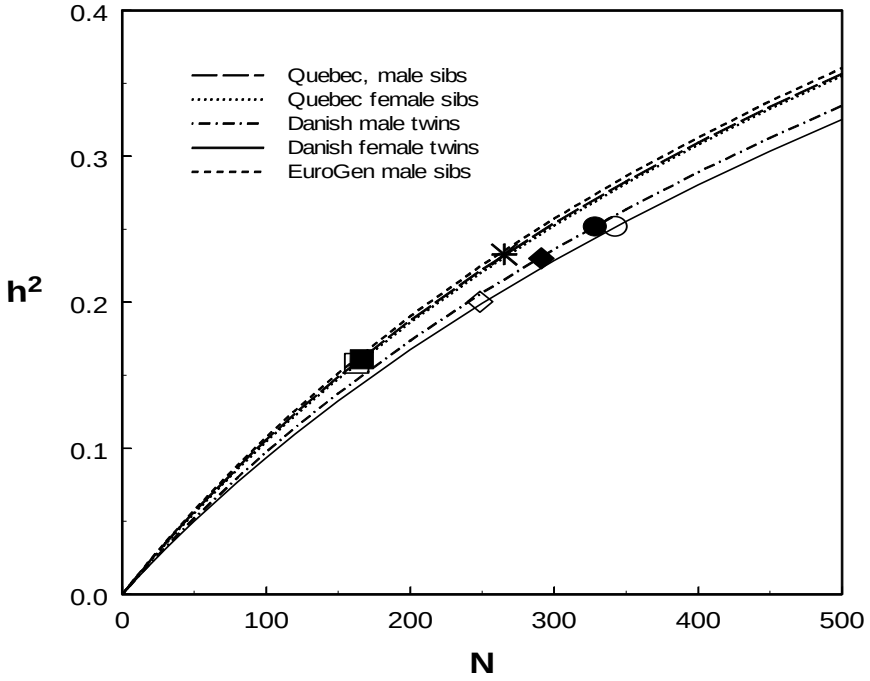

Figure 4.2 The estimated numbers of loci for different data sets for corresponding heritabilities when σ^2=0.1, r=0.1 and p=0.1.

corresponding data. For Quebec, the estimate is about 170 for males (solid rectangle) and 165 for females (empty rectangle). On the Danish curves, we get the estimated number of loci 320 for males (solid circle) and 330 for females (empty circle) when h^2=0.25 for both sexes (McGue et al. 1993). Based on another heritability estimation by Herskind et al. (1996), the numbers are about 300 for males (solid diamond, h^2=0.23) and about 260 for females (empty diamond, h^2=0.2). The star on Figure 4.2

indicates the estimated number (about 250) from EuroGen male sibs. Comparing results from the three different data sets on Figure 4.2, the numbers estimated from Quebec data is lower as a result of low heritability of life span during a harsh period approximately 400 years ago. The encouraging point is that the estimated numbers do not differ drastically for different data sets. Since the heritability estimated from twin data is more reliable than that from the genealogical studies due to difficulty in separating genetic and environmental components in family correlation, we can trust our conclusion based on the results drawn from the Danish female twins data (Table 4.1) although large differences are not expected.

4.4 Conclusions

This study, for the first time, provides a quantitative estimate of the number of genes relevant to human longevity using demographic data as the empirical basis. However, several points should be noted when interpreting the results. 1. The estimated numbers are upper-bound estimates since we are assuming similar effect alleles. 2. The model assumes that all genes function independently and ignores the existence of interactions (gene-gene and gene-environment interactions) so that the numbers could be overestimated. 3. The model is based on a proportional hazard assumption by which allele effects are assumed to be constant over time. This may not be an accurate assumption. However, since we intend only to give a rough answer, such assumptions are helpful in order to make the model applicable.

As we can see from Table 4.1, the estimated upper-bound of 7,000 gene alleles is consistent with Martin's prediction (Martin 1997). The number could be drastically reduced when there are other genes which manifest strong effects. Although a small number of large effect genes could serve to maintain the genetic component at the same level as a couple of thousands small effect genes, the real situation could be a combination of them. Sacher (1975) reckoned that a considerably small number of genes might be able to produce a notable improvement in the general vigor or intelligence of *Homo sapiens* in an evolutionarily natural way. By studying the evolutionary history of mammals, a remarkable increase in maximum life span was observed from our sibling species the chimpanzee (about 50 years) to man (Cutler 1979). It suggests that relatively few genetic alterations in the genome were necerssary

during the recent evolutionary history of man to significantly extend his innate ability to maintain mental and physical health. Given the above consideration, one could expect that there are quite a few large effect genes working together with many other small effect genes in modulating the human life span.

The parameter specifications in Table 4.1 ($r \in \{0.05, 0.1, 0.25, 0.5\}$, $p \in \{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$ and $\sigma^2 \in \{0, 0.1, 0.5\}$) are based on experiences from empirical applications of the binomial frailty model on the Danish and Italian genetic data discussed in Chapter 6. The different parameter combinations cover a wide range of possible situations. Influences on estimation from genetic parameters (risk reduction and frequency) can be seen as straightforward, but there are also non-genetic factors that can interfere. As seen in Table 4.1, larger numbers of alleles are needed for bigger values of the variance in environmental frailty, σ^2, in order to maintain the heritability level at 0.25.

The baseline hazard functions for individuals without any longevity allele can be estimated from (1.4) and (1.7) non-parametrically for different populations when all parameters are given. In Figure 4.3, the baseline hazards for the five populations in log scale corresponding to parameters in Figure 4.2 together with estimated numbers of loci are shown. There are two interesting points on this figure. First, they all increase with age faster than Gompertz (if it was Gompertz, a straight line is expected). This is consistent with Yashin and Iachine (1997) who observed the same trend for baseline hazard functions calculated semi-parametrically from bivariate survival data on Danish twins. This unexpected distribution shows the advantage of using a non-parametric approach in the application. Second, while high hazards of death can be observed for individuals without any longevity alleles, reflecting the importance of genetic contribution in maintaining individual survival, the order of the curves in Figure 4.3 represents the relative importance of genetic influence on survival for the five populations. The order (Danish twin: highest; EuroGen: middle; Quebec: lowest) is in accordance with the heritability estimations (Danish twins 0.25>EuroGen 0.22>Quebec 0.15) and may, as noted, be a result of a secular trend in life span inheritance.

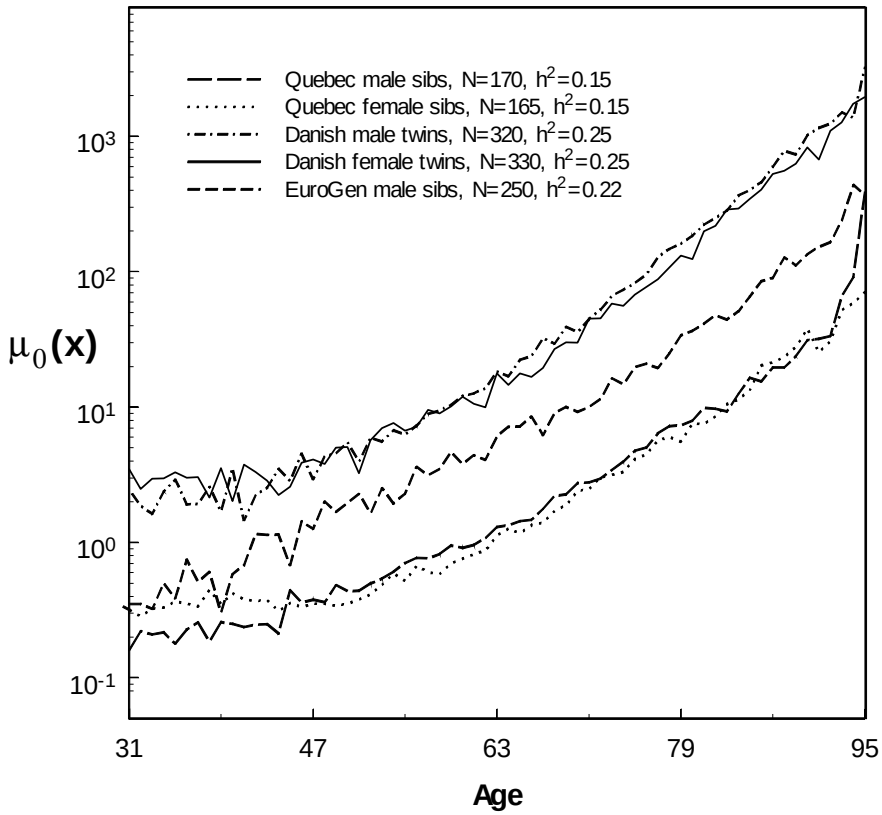

Figure 4.3 The baseline hazards for the five populations in log scale corresponding to parameters σ^2=0.1, r=0.1, p=0.1 and estimated numbers of loci in Figure 4.2

4.5 Summary

Applying the binomial frailty model to estimate the number of longevity genes using vital-statistic data on twins and genealogies indicates

(1) There could be as many as 7,000 genes relevant to longevity with small effects and low frequencies.

(2) Alternatively, there could be several dozens of strong effect genes with high frequencies that are responsible for the genetic variation of life span.

(3) One can also postulate that there could be a couple of thousands of genes with small effects combined with a limited number of high impact genes that together contribute to variation in individual life span.

PART III

Models for Gene Marker Data on Unrelated Individuals

Chapter 5

The Binomial Frailty Model for Gene Marker Data

5.1 Introduction

The tremendous advances in molecular genetics have spurred on the genetic study of longevity. In contrast to the rapid development in biological techniques involved, the statistical methods engaged in data analysis have remained the same, mainly a simple χ^2-test. The logic of deploying a χ^2-test is to compare the frequency change of a certain gene allele or genotype among different age groups or between control (usually young people) and centenarian groups in order to see if any departure exists. For this reason, we call such an approach the gene frequency method. Any effect on the longevity of a certain genotype or gene allele can be confirmed when a significant difference is found. For example, in a Japanese study on association of DNA polymorphism of HLA class II genes and human longevity (Akisaka et al. 1997), significantly higher frequencies for alleles HLA-DRB1-0101, 1201, 1401, HLA-DQA1-0101, 05 and HLA-DQB1-0503 were found among centenarians than among ordinary people and thus are favourable to longevity. Meanwhile HLA-DRB1-0403, 1302, HLA-DQA1-0102, 0103 and HLA-DQB1-0604 show significantly lower frequencies in the centenarian group than in the young group and are thus unfavourable to survival. Similar applications in the studies on genes and longevity or on gene associated health disorders can be found in the literature. Examples include the intensive study on apolipoprotein gene variations and their relationship to cardiovascular diseases (Aburatani et al. 1988; Myant et al. 1989; Paulweber et al. 1990; Sandholzer et al. 1992) and to longevity (Kervinen et al. 1994; Schachter et al. 1994; De Benedictis et al. 1996, 1998; Pepe et al. 1998; Klaver et al. 1998; Zhang et al. 1998). Although popular in use, the gene frequency method has many disadvantages:

1. Life span is a continuous quantitative trait. It is not reasonable to simply group it into young and old survivors as is done for the gene frequency method. In this sense, the gene frequency method is only a rough approach that does not fully make use of the individual survival information available in the analysis and thus has a lower efficiency for making inferences.

2. The gene frequency method has difficulty in controlling for confounding factors. Life span can be affected by factors like the individual's geographical location or sex, because of the existing regional difference in mortality, due to heterogeneity in social and economical environment, and sex differential mortality. In order to make inferences for a certain gene on its influence in survival, it is necessary to consider all the possible confounding factors that affect life span. By stratifying the sample, it is possible to keep control over the confounding factors, but this usually requires large sample sizes because the data have to be divided into smaller subsets and this consequently reduces the efficiency of the study.

3. Similar to the confounding problem, the gene frequency method is also not a good way to deal with interactions. Evidences of gene-environment and gene-sex interactions have been found in previous studies (De Benedictis et al. 1998, 1999; Ivanova et al. 1998; Nuzhdin et al. 1997). Interaction can be detected by making separate conclusions on different sexes or regions when the sample is accordingly grouped. Again inference has to be made on considerably smaller subset of the data.

4. As a continuous trait, life span is affected by factors that both can be biological and environmental. This results in individual differences or heterogeneity in their frailty composition. It will be shown that consideration of such differences is crucial for evaluating the influence of both genetic and environmental attributes in modulating life span. Unfortunately, the gene frequency method is incapable of integrating unobserved heterogeneity.

5. In cross-sectional studies, participants are taken from different birth cohorts. They exhibit heterogeneous patterns of survival due to secular trends in mortality improvement (Vaupel et al. 1998). However, the differences in individual survival probability are completely ignored by the gene frequency method and the conclusions made from such an approach could be biased.

As an extension of the gene frequency method, the logistic regression model can be used to help to account for confounding or interactions. However, such an

extension is still incapable of making full use of the individual information and neither of modelling heterogeneity. The Cox's proportional hazard model is a popular method in doing survival analysis aimed at finding risk factors that contribute to the duration of survival, but it is not applicable in the situation when dealing with data from cross-sectional studies because all participants are completely censored as regards their life spans. The Cox model is a useful tool when dealing with data from longitudinal studies, but one must bear in mind that longitudinal studies on longevity are both expensive and time consuming. In addition, there could be a selection bias because longitudinal study requires long-term cooperation of the participants and some people may decide to drop out, leaving highly selected individuals in the sample. Due to all of these problems and difficulties, the cross-sectional design is an economical and quick way to do longevity study. However the analytical drawback of the design calls for new methods to analyse data collected from increasingly numerous emerging studies. In this chapter, the binomial frailty model will be further employed for this purpose.

5.2 The relative risk model

5.2.1 The model

We consider the simple situation of just one genotype. Suppose the genotype frequency at birth is p, then in a mixed population consisting of both carriers and non-carriers of the genotype, the mean survival at age x is the weighted average of the two survival distributions for carriers and non-carriers with the weights p and $(1-p)$ respectively. If $\mu_0(x)$ is the individual baseline hazard of death of non-carriers and $s_0(x)$ the corresponding survival, then we will have (Vaupel & Yashin 1985)

$$\bar{\mu}(x) = p(x)\mu_1(x) + (1-p(x))\mu_0(x) \qquad (5.1)$$

$$\bar{s}(x) = ps_1(x) + (1-p)s_0(x) \quad . \qquad (5.2)$$

$p(x)$ is the proportion of people with the genotype at age x. Similar to (1.8), $p(x)$ can be calculated as

$$p(x) = \frac{ps_1(x)}{\bar{s}(x)}. \qquad (5.3)$$

In (5 .1) and (5.2), $s_1(x)$ and $\mu_1(x)$ are survival and hazard functions for the group of individuals carrying the genotype. Defining relative risk of the genotype as R and assuming individuals are homogeneous except for the observed genotype, we have according to (1.1), (1.3), $\mu_1(x) = R\mu_0(x)$ and $s_1(x) = s_0(x)^R$. Here relative risk R equals the factor $(1-r)$ in the previous chapters. However, risk reduction r in the new context can be positive or negative so that R can be smaller or bigger than 1 corresponding to robust or frail genotypes. When considering individual heterogeneity and assuming that the unobserved frailty is gamma-distributed with mean 1 and variance σ^2, we have according to (1.5) and (1.6),

$$\mu_1(x) = \frac{R\mu_0(x)}{1-\sigma^2 R \ln s_0(x)}, \qquad (5.4)$$

$$s_1(x) = (1 - R\sigma^2 \ln s_0(x))^{-1/\sigma^2}. \qquad (5.5)$$

When genotypes are observed for survivors at different ages, the likelihood at age x based on the binomial distribution can be constructed as

$$L(x|s_0(x), p, R) \propto p(x)^{n(x)}(1-p(x))^{N(x)-n(x)} \qquad (5.6)$$

where $n(x)$ is the number of carriers at age x and $N(x)$ is the total number of individuals at age x from the sample. Note (5.6) is obtained by omitting the constant term from the binomial distribution, $\binom{N(x)}{n(x)}$, because it is irrelevant to the parameters of interest.

If we are interested in a single allele instead of a genotype, then the frequency p to be estimated is no longer genotype frequency but frequency of the allele. Since there could be 3 genotypes $(--), (-+/+-)$ and $(++)$, according to the binomial distribution of the allele, we can calculate the initial frequencies for the three genotypes as

$p(--) = (1-p)^2$

$p(+-) = p(-+) = p(1-p)$

$p(++) = p^2$

The relative risks of the three genotypes can be defined as

$R(--) = 1$

$R(+-) = R(-+) = R_1$

$R(++) = R_2$

Two different risk parameters are assigned for heterozygous and homozygous genotypes because the effect of having two alleles may differ from that of having one in case of co-dominance or additive effect. Similar to (5.2), we have

$$\bar{s}(x) = (1-p)^2 s_0(x) + 2p(1-p)s_1(x) + p^2 s_2(x) \qquad (5.7)$$

where $s_0(x), s_1(x), s_2(x)$ are survival functions for the 3 groups with zero, one and two alleles. Frequencies by age for the 3 genotypes are

$$p_0(x) = \frac{(1-p)^2 s_0(x)}{\bar{s}(x)} ,$$

$$p_1(x) = \frac{2p(1-p)s_1(x)}{\bar{s}(x)} ,$$

$$p_2(x) = \frac{p^2 s_2(x)}{\bar{s}(x)} . \qquad (5.8)$$

When frequencies of different genotypes by age are observed, the likelihood function for age x based on the multinomial distribution (Hastings & Peacock 1975) can be constructed as

$$L(x|s_0(x), p, R) \propto p_0(x)^{n_0(x)} p_1(x)^{n_1(x)} p_2(x)^{n_2(x)} . \qquad (5.9)$$

Again, like in (5.6), the items that are not dependent on the parameters of interest are omitted in (5.9).

The likelihood of (5.6) or (5.9) is a function of risk R, frequency p and $s_0(x)$. With known parameters, R and p, $s_0(x)$ can be calculated as discussed in section 1.1. In the estimation procedure, a Two-step MLE is introduced. For every new estimates of R and p, a new $s_0(x)$ is calculated and put into the likelihood function for re-estimating the parameters (section 5.2.3).

5.2.2 Combining demographic information in the analysis

Incorporating demographic information into the analysis is possible because the equations of (5.2) and (5.7) involve $\bar{s}(x)$, the population survival function. As discussed above, $\bar{s}(x)$ can be used to calculate the baseline survival function $s_0(x)$. In this way, the survival distribution of the population from which the sample is taken influences the baseline hazard in the likelihood function. When integrating the

population survival into the model, one should take into account the existing sex differential mortality by which female population manifests longer survival than the males (Hazzard 1986; Holden 1987; Keyfitz & Flieger 1990). This can be done by introducing male and female survival function separately so that

$$\bar{s}_m(x) = ps_{1,m}(x) + (1-p)s_{0,m}(x) \text{ and}$$

$$\bar{s}_f(x) = ps_{1,f}(x) + (1-p)s_{0,f}(x). \qquad (5.10)$$

In this case, the likelihood for the sample consisting of both male and female individuals becomes

$$L(x|s_{0,m}(x), s_{0,f}(x), p, R) \propto p_m(x)^{n_m(x)}(1-p_m(x))^{N_m(x)-n_m(x)} \\ p_f(x)^{n_f(x)}(1-p_f(x))^{N_f(x)-n_f(x)} \qquad (5.11)$$

Note that in (5.11), although the baseline survival functions are different for males and for females, the risk and frequency parameters are assumed to be the same for the two sexes. The assumption of similar risk may not hold when there is gene-sex interaction (to be discussed later). It is possible to use population survival from only one sex and specify one parameter for the relative risk of sex in the analysis so that a proportional hazard of sex is assumed. Such an approach may not be appropriate because there could be mortality crossover at old ages indicating non-proportionality of the relative risk of sex. However, this is completely avoided when male and female survival distributions are introduced separately.

5.2.3 The Two-step MLE

Combining demographic information in the analysis is one very important feature of the present approach. In this situation, the parameter estimation depends on both the observed individual genetic data (d) and the observed population survival ($\bar{s}$) as well. That is, we are trying to estimate the parameters by maximising the likelihood $L(\hat{R}, \hat{p}|d, \bar{s})$ given the two kinds of data observed. In order to carry out the estimation, a Two-step MLE is useful. The procedure starts with a guess of the parameters, and then, in step 1, the baseline survival function is calculated from (5.2) or (5.7) by solving the equations using a numerical method. In step 2, the baseline survival so obtained is then introduced into the likelihood function (5.9) or (5.11) for

estimating the parameters by maximisation. The procedure returns the new parameter estimates to the first step in order to get an updated baseline survival function so that at the i th iteration, the baseline survival is a function of the new parameter estimates and the observed population survival, i.e., $s_{0,i} = F(\hat{R}_i, \hat{p}_i, \bar{s})$. A new set of parameter estimates $(\hat{R}_{i+1}, \hat{p}_{i+1})$ is obtained by maximising the likelihood $L'(\hat{R}_{i+1}, \hat{p}_{i+1} | d, s_{0,i})$ in step 2. This iteration continues until the j th iteration when the likelihood converges so that $\hat{R}_j \to \hat{R}, \hat{p}_j \to \hat{p}$, the maximum likelihood estimators from $L(\hat{R}, \hat{p} | d, \bar{s})$, given the two kinds of data (Figure 5.1). Advantages of introducing the Two-step MLE include:

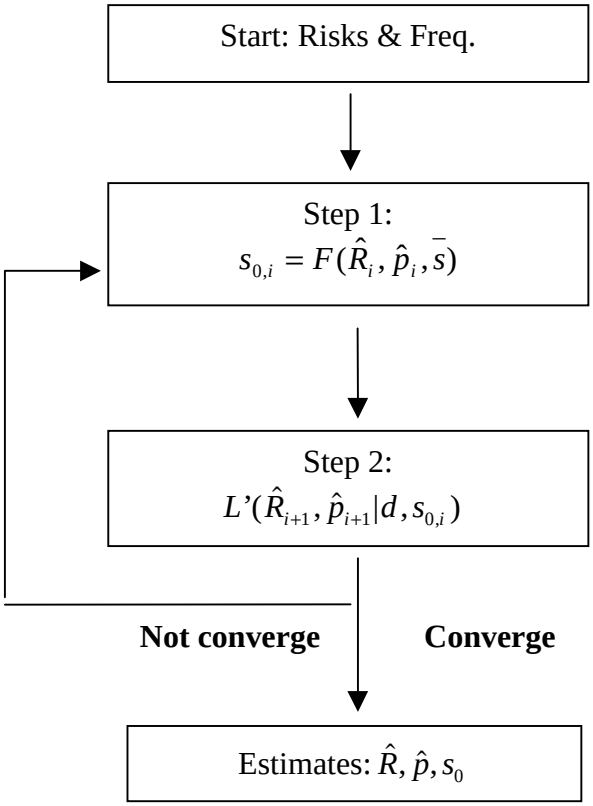

Figure 5.1 The Two-step MLE used for estimating the genetic parameters

(1) It insures the feasibility of combining genetic and demographic data in the analysis. This is important because, first, if the age-specific cohort survival can be compiled, the influence from secular mortality changes is easily avoided; also, as will be discussed later, the period life tables can help in the analysis when age-specific

cohort survival is not available. In both cases, parameter inferences can always be made by survival analysis on genetic data from cross-sectional studies.

(2) The baseline survival functions for the two sexes obtained in this way are non-parametric. This reduces the model assumption and avoids bias in parameter estimates due to model mis-specification.

(3) Introducing extra information from population statistics (the population survival functions) enhances the estimation procedure such that the model can be applicable to data drawn from small-scale investigations. This can be demonstrated using the simulation studies discussed below.

5.3 Simulation studies

In order to study the effectiveness of the model, artificial data are generated with fixed parameters and then the model is applied to the simulated data to recapture the parameters used in the simulation.

5.3.1 Generating the data

Different settings of parameters R and p are specified to generate the data to be used to retrieve the parameters. Figure 5.2 are simulated and theoretical proportion dynamics for 3 genotypes of one allele with R=0.7 and p=0.1. The survival distribution is taken from the period life table calculated from the pooled data for 13 developed countries from ages 80-105 and period 1980-1990 (Thatcher et al. 1998). In this simulation, we have 50 people by each age, so that totally we have 1,300 people in this data set. With this example, initial frequencies for the 3 genotypes, $(--), (-+/+-)$ and $(++)$ are 0.81, 0.18 and 0.01. Their corresponding risks are 1, 0.7 and 0.49 (here we assume $R_2 = R^2$). Genotype $(++)$ is very robust because it reduces the hazard of death by nearly one half. From Figure 5.2, we see that the frequency of genotype $(++)$ goes up from 0.01 at initial age to about 0.3 at age 105, a drastic increase due to favourable selection. At the same time, the genotype $(--)$ declined from 0.81 to about 0.1 because this sub-population receives no benefit from the observed allele.

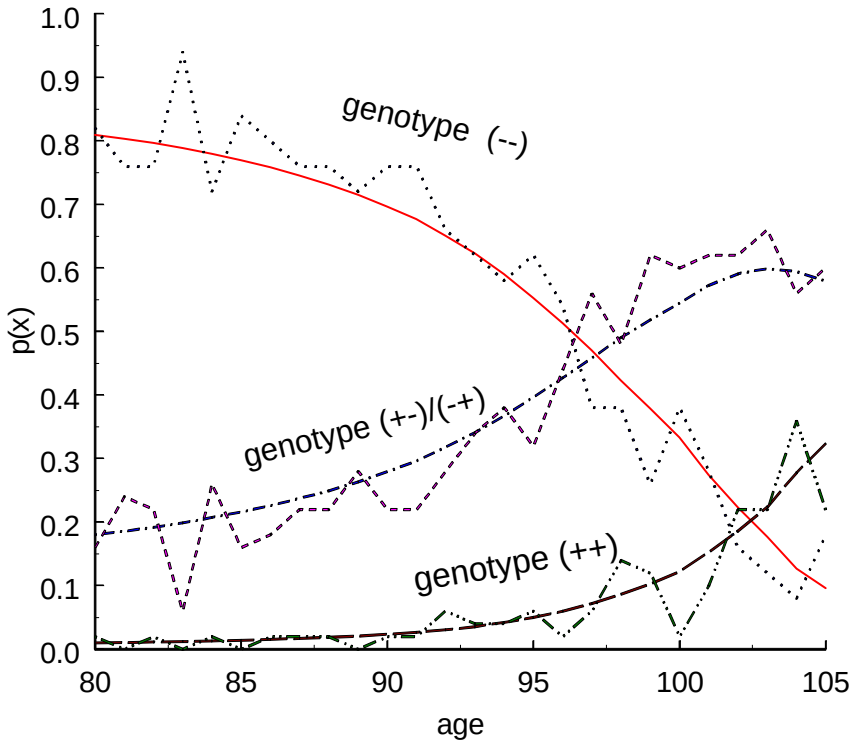

Figure 5.2 Simulated and theoretical proportion dynamics for 3 genotypes of one allele with R=0.7 and p=0.1

5.3.2 Retrieving the parameters

Using the data presented in Figure 5.2, we estimate the parameters by the Two-step MLE. In the estimation process, life table survival function used in the simulation was introduced as the observed population survival function. The estimated initial frequency is $\hat{p}$=0.108 (SE=0.008) and the estimated relative risk is $\hat{R}$=0.712 (SE=0.020). The estimated survival functions for the 3 genotypes are close to their theoretical distributions (Figure 5.3). Both the estimated and theoretical baseline survival functions are close to the population survival since frequency of the observed gene allele is very low (p=0.1).

The standard errors for the parameter estimates are obtained from the Fisher's information matrix. Since the estimation involves the two-step procedure, validation of parameter and standard error estimates should be considered. Such concern is

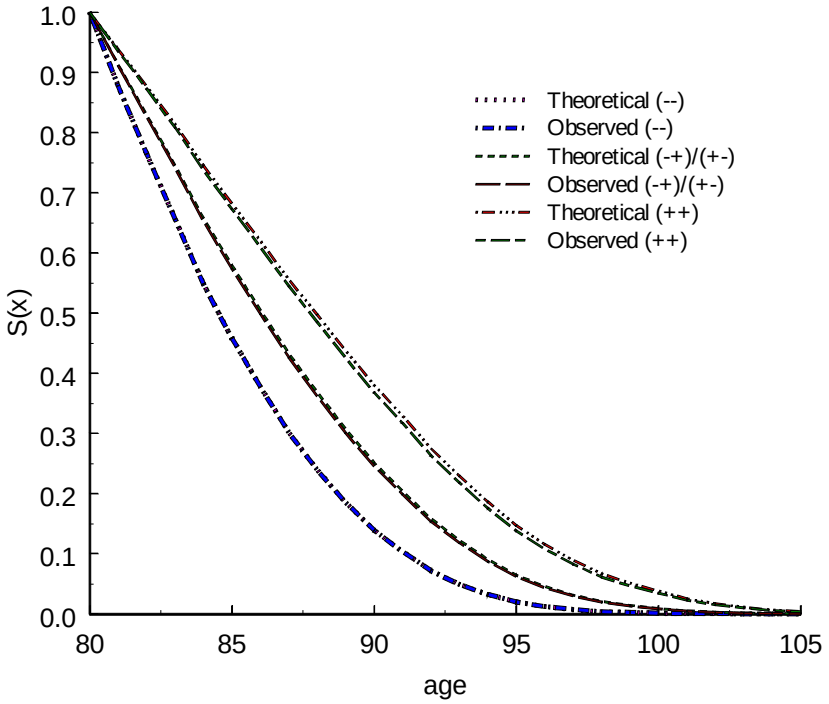

Figure 5.3 The estimated and theoretical survival distributions for the 3 genotypes

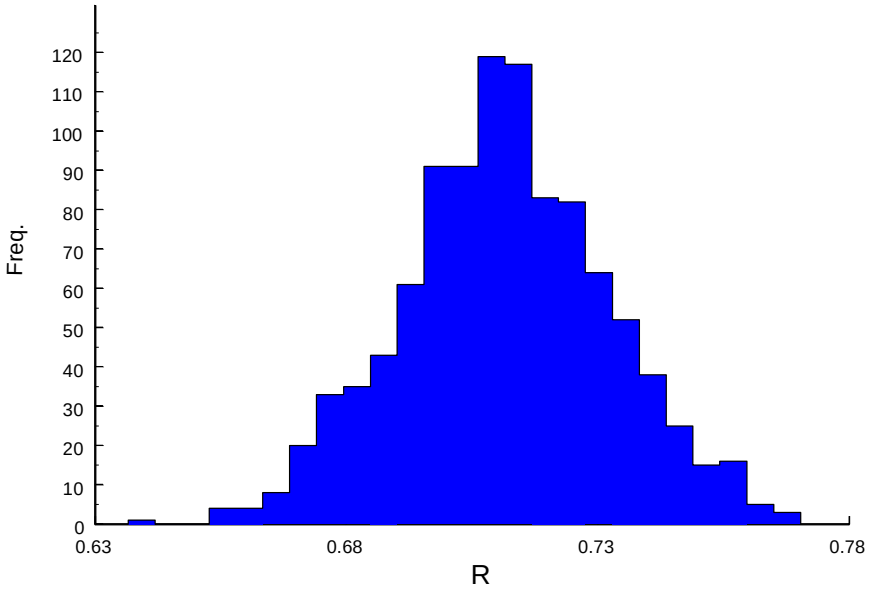

Figure 5.4 Frequency distribution of the estimated R from 1000 simulations

originated from the calculated baseline survival function as described in section 5.2.3. In order to examine if the two-step procedure can produce unbiased parameter

estimates with reliable standard errors, a simulation study was conducted. We simulated 1000 data sets by adopting parameter estimates by the two-step MLE on the data set generated in section 5.3.1 with exactly the same data size and structure. The idea is to check if (1) the parameter estimates from the two-step MLE are unbiased and (2) the corresponding standard errors from the two-step MLE are reliable. Applying the two-step MLE, we got 1000 estimates on the parameters R and p (Figure 5.4). The means and standard deviations for these parameters are then calculated. We got the estimated relative risk R =0.711 with standard deviation 0.021 and frequency p =0.108 with standard deviation 0.010. Take the estimate on relative risk R for example, the 1000 simulations produced a 95% range from 0.672 to 0.752. The 95% CI for R based on variance by applying two-step MLE on the data generated in section 5.3.1 is from 0.673 to 0.751. Both intervals are very close which is a clear indication that the parameters together with their standard errors estimated by the two-step MLE are valid.

5.3.3 Sensitivity study

As a polygenic trait, there could be many genes that are responsible for longevity. While some of the genes have marked influences on survival there could be lots of others with only small effects. In regard to their frequencies, the difference could be very big; some could be common while some others could be very rare. In this section, the model will be studied by simulation to see how it performs under various situations.

a. Sensitivity to magnitude of parameters

Three levels of relative risk of the gene are specified by setting R =1.1, 1.5 and 2. For each level of R, gene allele frequency p varies from 0.05 to 0.95. This situation applies to the simple case of formulae (5.1), (5.2) and (5.3). For each combination of p and R, 100 data sets were simulated by generating 20 individuals for each age from ages 80-99. The means and standard errors of the estimated relative risk R and frequency p are presented in Tables 5.1 and 5.2 where $CV = SE / Est.$ is used to check stability of the estimation. The results from Tables 5.1 and 5.2 show acceptable estimates for the fixed parameters that have been used to generate the data.

The model is sensitive to small risks while large risk factors are estimated with less relative variablity. In term of allele frequency, higher frequency gets more precise estimation (Table 5.2).

Table 5.1 Mean and standard error of estimated risk for different R, p and *data size**

Fixed R for different p	Data Size=400			Data Size=1,600			SE_{400}/SE_{1600}
	R	SE	CV(%)	R	SE	CV(%)	
p=0.05							
1.1	1.066	0.108	10.099	1.109	0.074	6.710	1.446
1.5	1.396	0.173	12.369	1.513	0.107	7.084	1.610
2.0	1.912	0.279	14.579	2.019	0.137	6.774	2.038
p=0.10							
1.1	1.102	0.103	9.375	1.087	0.054	4.980	1.909
1.5	1.479	0.150	10.161	1.502	0.074	4.926	2.023
2	1.988	0.255	12.815	2.000	0.118	5.886	2.164
p=0.25							
1.1	1.105	0.081	7.298	1.100	0.042	3.788	1.936
1.5	1.515	0.121	8.014	1.497	0.059	3.968	2.043
2	2.026	0.176	8.687	2.008	0.088	4.396	1.994
p=0.50							
1.1	1.102	0.078	7.113	1.105	0.040	3.601	1.970
1.5	1.495	0.108	7.230	1.511	0.049	3.252	2.199
2	2.034	0.209	10.267	2.010	0.089	4.416	2.352
p=0.75							
1.1	1.101	0.096	8.729	1.104	0.043	3.932	2.213
1.5	1.518	0.138	9.095	1.509	0.073	4.830	1.895
2	2.074	0.308	14.826	2.002	0.120	5.986	2.566
p=0.90							
1.1	1.117	0.137	12.237	1.117	0.066	5.938	2.062
1.5	1.539	0.242	15.686	1.514	0.119	7.879	2.025
2	2.068	0.507	24.490	2.072	0.194	9.369	2.610
p=0.95							
1.1	1.194	0.226	19.682	1.116	0.090	8.078	2.510
1.5	1.633	0.431	26.365	1.535	0.163	10.648	2.634
2	2.306	0.974	42.250	2.033	0.282	13.890	3.450

**Each case is calculated from 100 data sets*

b. Sensitivity to data size

By increasing the number of observations of each age from 20 to 80, a larger data set consisting of 1,600 people is generated. Calculations have been repeated and

the results are given in Tables 5.1 and 5.2. The larger data sets produce better estimates with smaller errors. The standard errors of the two sample sizes are compared by calculating $SE(400) / SE(1600)$ which, as might be expected, is around 2 for all different settings, both for the estimates of risk R and of frequency p. Though better estimates are obtained for larger data, the model appears to produce reasonable estimates for small samples as well.

Table 5.2 Mean and standard error of estimated frequency for different R, p and *data size**

Fixed p For different R	Data Size=400			Data Size=1,600			SE_{400}/SE_{1600}
	p	SE	CV(%)	p	SE	CV(%)	
R=1.1							
0.05	0.068	0.012	17.725	0.050	0.007	14.510	1.644
0.10	0.105	0.016	15.074	0.102	0.011	10.456	1.479
0.25	0.251	0.034	13.623	0.253	0.018	7.064	1.917
0.50	0.507	0.041	8.090	0.499	0.021	4.237	1.940
0.75	0.750	0.037	4.865	0.747	0.018	2.339	2.087
0.90	0.900	0.025	2.761	0.897	0.014	1.506	1.839
0.95	0.947	0.019	1.971	0.948	0.009	0.944	2.086
R=1.5							
0.05	0.063	0.013	19.847	0.050	0.007	13.319	1.883
0.10	0.107	0.019	17.327	0.100	0.008	8.380	2.202
0.25	0.250	0.033	13.085	0.252	0.015	5.795	2.237
0.50	0.502	0.036	7.128	0.496	0.017	3.375	2.138
0.75	0.751	0.037	4.867	0.748	0.019	2.586	1.891
0.90	0.893	0.027	3.037	0.900	0.014	1.535	1.964
0.95	0.946	0.023	2.420	0.949	0.010	1.050	2.298
R=2							
0.05	0.059	0.014	23.695	0.049	0.006	12.338	2.300
0.10	0.103	0.019	18.069	0.100	0.010	9.894	1.898
0.25	0.248	0.028	11.269	0.251	0.015	5.788	1.929
0.50	0.501	0.044	8.699	0.502	0.022	4.396	1.977
0.75	0.740	0.044	5.915	0.742	0.021	2.868	2.058
0.90	0.899	0.028	3.109	0.893	0.017	1.852	1.688
0.95	0.944	0.024	2.539	0.949	0.012	1.293	1.954

**Each case is calculated from 100 data sets*

c. Sensitivity to data structure

Many problems can arise when collecting data from the elderly due to difficulties like small populations, missing records, a reluctance to response and so on.

In large data sets, it is very difficult to ensure the same number of individuals for each age group. Suppose we have a sample where the number of participants is reduced by 1/5 for each age beginning from age 91. If before 91 we observe 80 people at each age, then at age 91 we have 64 until at the final age 105 we only have 2 survivors. By setting R=0.7 and p=0.1 in the simulation, we get the means of estimated parameters from 100 simulations, R=0.700 and p=0.101 with standard error 0.031 and 0.008

Table 5.3 Confidence interval for estimated parameters

Parameter	Mean	SE	95% Range
R	0.6996	0.0308	0.6392 - 0.7600
p	0.1009	0.0075	0.0862 - 0.1156

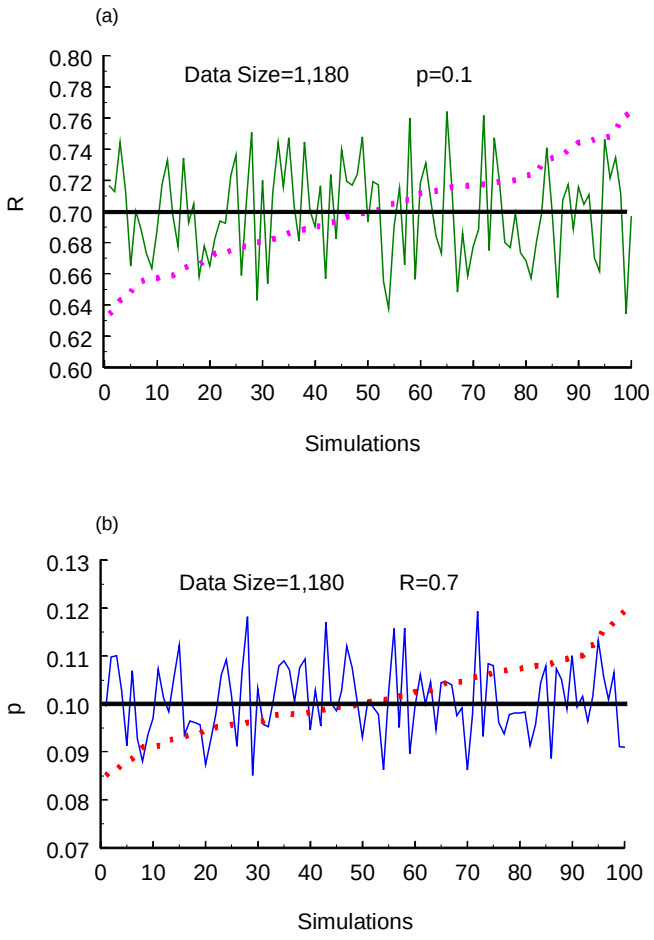

Figure 5.5 Estimated risk (a) and frequency (b) of the allele for uneven sampling from 100 simulations

respectively (Table 5.3). The estimates obtained from uneven sampling appear to be reasonable provided the sample is big enough (Figure 5.5). Comparing Figures 5.5 with 5.4, we see that the precision of the estimates from uneven sampling is decreased since we have a larger fluctuation though the sample sizes are not very different. Different from relative risk R, the estimated frequency does not change much under such a situation. The result tells us that the frequency parameter is more affected by data from the younger ages.

Another alteration in data structure is discrete sampling by age. This may happen when data for some ages are not available. For example, in some age groups, we may not have any observations. To study the effect of such a data structure, the settings in Figure 5.5 will be used but with observations from ages 86-90 eliminated. Such a data structure represents both uneven and discrete sampling (Table 5.4) and is

Table 5.4 Data structure of one simulation

Age	$N(—)$	$N(-+/+-)$	$N(++)$	N
80	61	19	0	80
81	65	14	1	80
82	62	16	2	80
83	61	18	1	80
84	61	19	0	80
85	63	17	0	80
91	43	20	1	64
92	39	11	1	51
93	24	13	3	40
94	20	12	0	32
95	14	11	1	26
96	12	7	1	20
97	8	7	1	16
98	4	7	2	13
99	3	5	2	10
100	2	4	2	8
101	3	3	0	6
102	3	2	0	5
103	0	3	1	4
104	0	1	2	3
105	0	2	0	2
				$\sum N$ =780

Table 5.5 Confidence interval for estimated parameters

Parameter	Mean	SE	95% CI
R	0.6988	0.0301	0.6398 - 0.7578
p	0.0997	0.0082	0.0836 - 0.1158

common in empirical studies. Estimates are shown in Figure 5.6. Although the data size is smaller than before, the estimation is still fairly good. The mean of 100 runs for relative risk R is 0.699 and for frequency p is 0.100 with standard errors 0.030 and 0.008 respectively. Both relative risk R and frequency p are well estimated (Table 5.5).

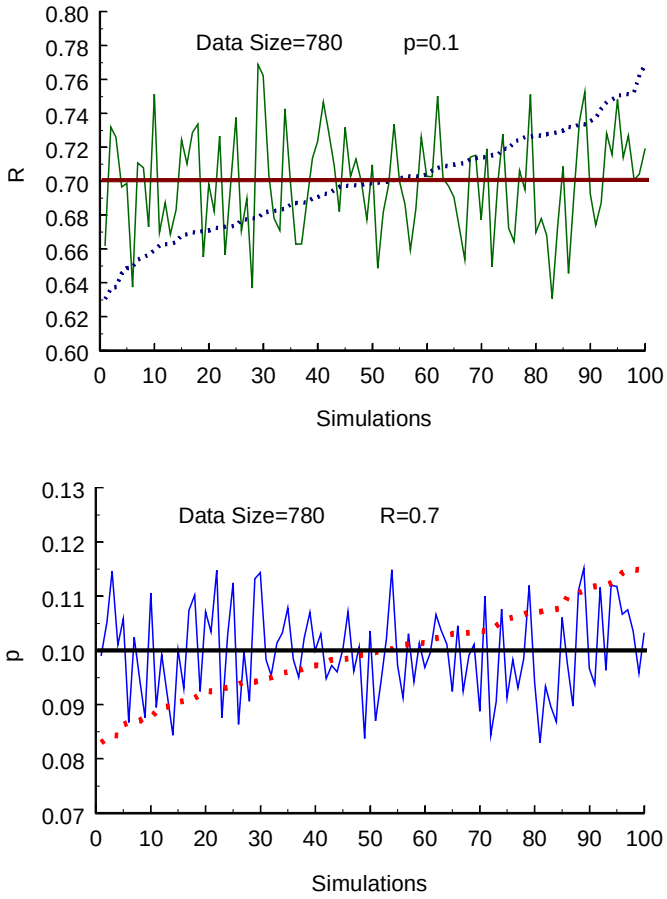

Figure 5.6 Estimated risk (a) and frequency (b) of the allele for uneven and discrete sampling from 100 simulations

5.4 Problems with cross-sectional data

In cross-sectional studies, participants of different ages are taken from different birth cohorts to form a synthetic cohort. They manifest different demographic characteristics since the individuals have experienced quite different social and

environmental changes. At the same time, there could also be changes in the genetic composition due to mixture of people from different ethnic groups. All these changes could potentially influence our evaluation of the effects of genes on survival. For more efficient research design and close-to-real inferences, it is therefore imperative to study how and to what extent these influences affect the results.

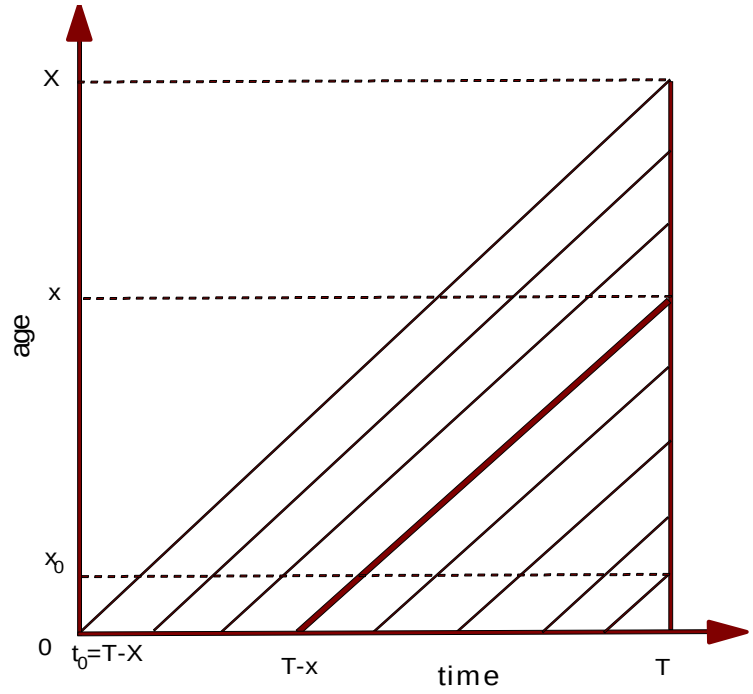

Figure 5.7 Lexis diagram of showing the difference between cohort and cross-sectional data

5.4.1 Secular change in cohort mortality rate

Suppose that at time T we take a sample of individuals aged from x_0 to X. The individuals are born in different cohorts beginning from the first cohort t_0 (Figure 5.7). Then we have $t_0 = T - X$. Assume there is a linear progress in mortality improvement $k(t)$ at time t and we use the first cohort as reference, that is $k(t_0) = k(T - X) = 1$. If for the last cohort $k(T - t_0) = a$, we have

$$k(t) = 1 - \frac{1-a}{T-t_0}(t - t_0) \qquad (5.12)$$

so that hazard and survival functions for cohort t are

$$\mu(x,t) = k(t)\mu(x,t_0) \quad \text{and}$$

$$s(x,t) = s(x,t_0)^{k(t)} \qquad (5.13)$$

Here $\mu(x,t_0)$ and $s(x,t_0)$ are hazard and survival functions for the cohort born at time t_0. Figure 5.8 shows three hazards of death: $\underline{\mu}(x)$ for the first cohort, $\mu^*(x)$ for

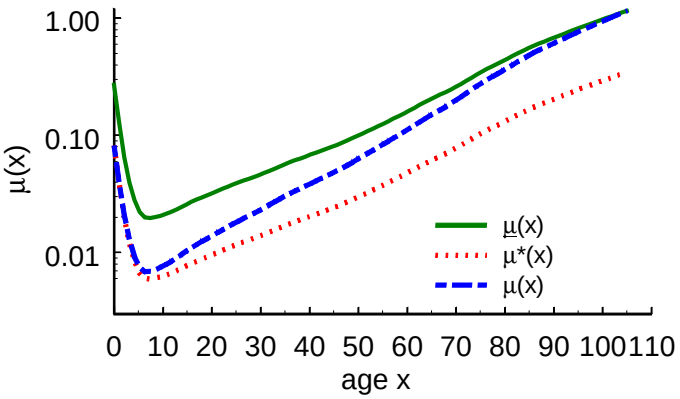

Figure 5.8 Logrithms of mortality rates for the earliest cohort (solid line), latest cohort (dotted line) and the synthetic cohort (dashed line)

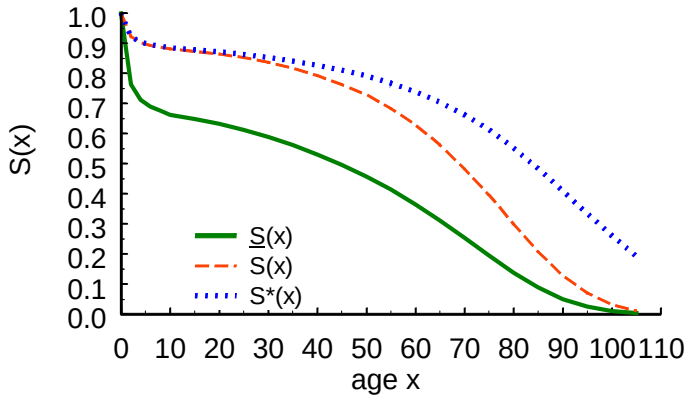

Figure 5.9 Survival functions for the earliest cohort (solid line), latest cohort (dotted line) and the synthetic cohort (dashed line)

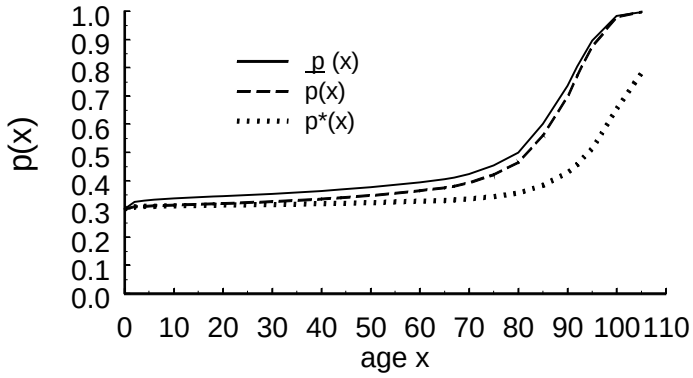

Figure 5.10 Frequencies of carriers of the gene allele for the earliest cohort (solid line), latest cohort (dotted line) and the synthetic cohort (dashed line)

the last cohort, while falling between them, the hazard of the synthetic cohort (a pseudo-cohort that we observe at time T from a cross-sectional study, which consists of subjects born from time t_0 to T) $\mu(x)$ is given. These hazards are estimated by introducing the survival distribution of the 19th century Italian population (Del Panta & Rettaroli 1994) as the first cohort and setting a to 0.3. We can see from Figure 5.8 that the age pattern of mortality trajectory for the synthetic cohort is close to the latest cohort at the beginning ages but moves gradually towards and at late ages approaches the hazard of the first cohort. The same trend applies to the survival of the synthetic cohort $s(x)$ in Figure 5.9 where $\underline{s}(x)$ is survival function for the first cohort and $s^*(x)$ for the last cohort. Note here that $s(x)$ is not a valid survival function since it is not always monotonous. Because for each age x, $s(x)$ is taken from different cohort, there could be cases when $s(x) < s(x+1)$, which can't be true for any form of real survival distribution.

Given the existence of a secular trend in mortality change, what kind of influence does it have on the frequency trajectory for one observed genotype? Assuming there is one genotype with frequency p=0.3 and relative risk R=1.5, then according to (5.2) and (5.3) frequency of carriers at age x is

$$p(x) = \frac{ps_0(x)^R}{\bar{s}(x)} \tag{5.14}$$

and total survival is

$$\bar{s}(x) = ps_0(x)^R + (1-p)s_0(x). \tag{5.15}$$

The survival distributions in Figure 5.9 can be introduced to the right-hand side of (5.15) to get corresponding baseline survival distributions for the given R and p, then the corresponding frequencies can be calculated from (5.14). Figure 5.10 shows the age pattern of the frequency for individuals with the genotype for the earliest cohort, $\underline{p}(x)$, the latest cohort, $p^*(x)$, and the synthetic cohort $p(x)$ which is to be observed from a cross-sectional study. The observed trajectory resembles the latest cohort at the beginning but approaches the case of the earliest cohort as age increases, an indication that the early cohorts strongly influence the observed gene frequency trajectory at later ages.

If survival distributions for all of the cohorts involved in the sample are available, one can easily construct, for the synthetic cohort, an artificial survival from

them. In this case, the secular trend in mortality change is not a problem at all. When age specific cohort survival is not available, one easy choice is to take one survival distribution from a period life table. However, the consequence of using life tables

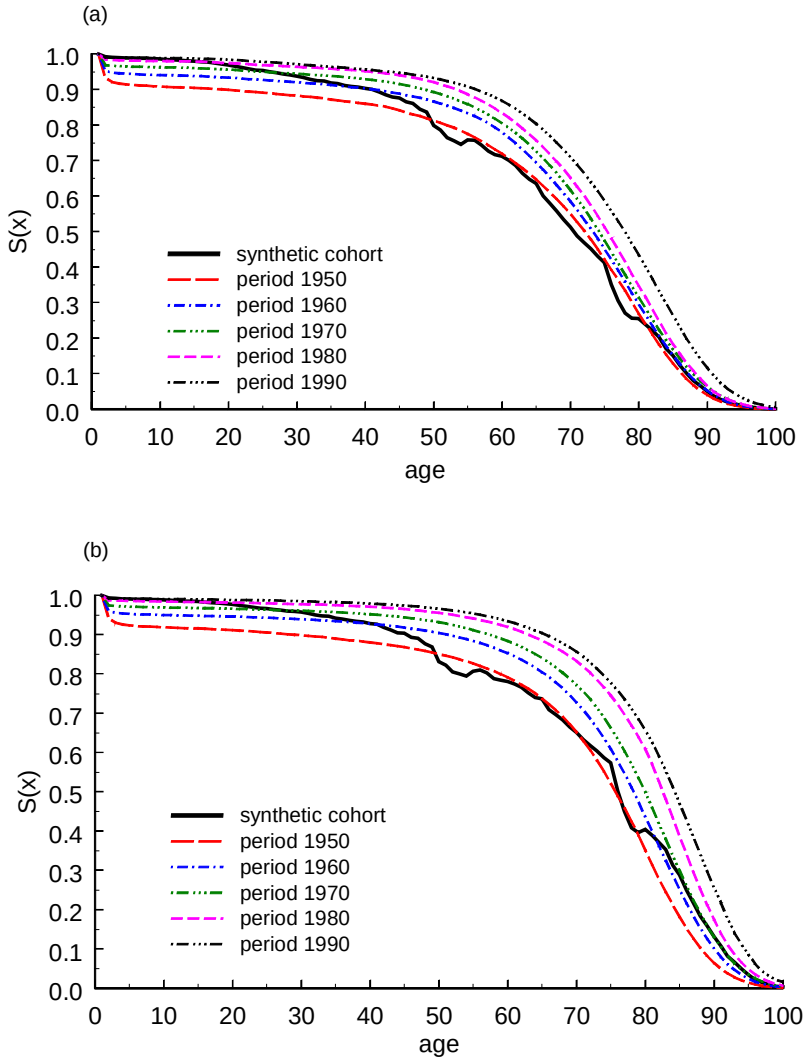

Figure 5.11 Survival distributions for the synthetic cohort and 5 periods, 1950, 1960, 1970, 1980, 1990 for males (a) and females (b)

from different periods for the estimates of genetic parameters (risks and frequencies) have to be studied. Let's assume there is one beneficial genotype with relative risk R=0.5 and frequency p=0.3. A cross-sectional study aimed at this gene conducted in 1995 is simulated with a sample size of 100,000 individuals aged from 1 to 100 (100 individuals at each age). In the simulation, age-specific cohort survival obtained from

the Italian cohort life tables (cohorts 1896 to 1995) was used. A secular reduction on mortality has been reported (Vaupel et al. 1998) for these cohorts. The data are simulated using age-specific cohort survival for males with given genetic parameters. Survival distributions are taken from 5 periods: 1950, 1960, 1970, 1980 and 1990. Figure 5.11a shows the survival functions for the synthetic cohort and for the 5 periods for males. We can see that the curve for the synthetic cohort is neither monotonous nor smooth, with two obvious dips around ages 50 and 80. Since the synthetic cohort is constructed for 1995, the two dips are obviously results from events that happened around 1914 and 1944 which could be the Spanish flu, and the first and second world wars. At early ages of the synthetic cohort, survival is close to the late periods, but at late ages survival becomes close to the early periods. The same pattern can be observed for females as well (Figure 5.11b) but difference at late survival between the artificial cohort and the first period is larger than for males in Figure 5.11a, indicating more achievements in reducing mortality at old ages for females than for males. The different survival functions of Figure 5.11a are used for estimating the parameters leading to the results shown in Table 5.6. The closest

Table 5.6 Parameter estimates using different life tables

Period	R =1.5		p =0.3	
	Est.	S.E.	Est.	S.E.
Synthetic	1.516	0.031	0.297	0.006
1950	1.486	0.030	0.298	0.006
1960	1.524	0.032	0.295	0.005
1970	1.562	0.034	0.293	0.006
1980	1.572	0.035	0.290	0.005
1990	1.723	0.045	0.290	0.005
Period	R =1		p =0.3	
	Est.	S.E.	Est.	S.E.
Synthetic	1.019	0.015	0.312	0.005
1950	1.017	0.013	0.312	0.005
1960	1.019	0.014	0.312	0.005
1970	1.020	0.016	0.312	0.005
1980	1.021	0.016	0.312	0.005
1990	1.025	0.019	0.312	0.005

estimates are from period 1950 when R is estimated as 1.486 and p as 0.298. It is encouraging to see that estimates from the adjacent periods (1960 and 1970) are not significantly different from the true parameters when looking at their standard errors. This means that estimates using life tables from periods 1950-1970 are all acceptable.

When survival distributions from the last two periods (1980 and 1990) are used, relative risk R is overestimated while frequency p underestimated. From these results we conclude that for our purpose it is feasible to use a period life table as a substitute when age-specific cohort survival is not available, but attention should be paid in deciding which period to pick up; obviously we must not just take the period life table for the time point at which the study is performed.

If improper application of a life table could result in biased estimates, could a neutral gene be mis-evaluated as significant in affecting longevity? In another simulation, the relative risk of one genotype is set to $R = 1$ but with the same frequency and data size as before. The estimates for the neutral gene using different period life tables do not show significant deviations from their true values (Table 5.6), and the estimates on frequency are even constant, although both R and p are slightly overe-stimated than their true values. One can conclude from the two examples that while arbitrary application of period life tables could lead to erroneous inference (over or under estimations), misjudgement of neutral attributes is unlikely to happen.

5.4.2 Secular change in gene frequency

Up to now, all calculations have been based on the assumption that the gene frequency is constant in all cohorts wherein individuals participating in the cross-sectional study were born. What are the consequences when this assumption does not hold? Suppose initial genotype frequency for a single gene increases linearly from p_a to p_b within the cohort duration $T - t_0$ (Figure 5.7) covered by the study. Then for the cohort born at time t,

$$\bar{s}(x,t) = p(t)s_0(x,t)^R + (1 - p(t))s_0(x,t) \qquad (5.16)$$

where R is the relative risk of the genotype and $p(t)$ is the initial genotype frequency for the cohort t. According to the linear assumption, we have

$$p(t) = p_a + \frac{p_b - p_a}{T - t_0}(t - t_0). \qquad (5.17)$$

Genotype frequency for this gene at age x for cohort t is

$$p(x,t) = \frac{p(t)s_0(x,t)^R}{\bar{s}(x,t)}. \qquad (5.18)$$

The age-pattern of genotype frequency to be observed from the cross-sectional sample can be obtained from (5.18) by substituting t with $T-x$ for age x. That is

$$p(x,T-x)=\frac{p(T-x)s_0(x,T-x)^R}{\bar{s}(x,T-x)}. \qquad (5.19)$$

In (5.18) and (5.19), the frequency at age x not only depends on relative risk R but

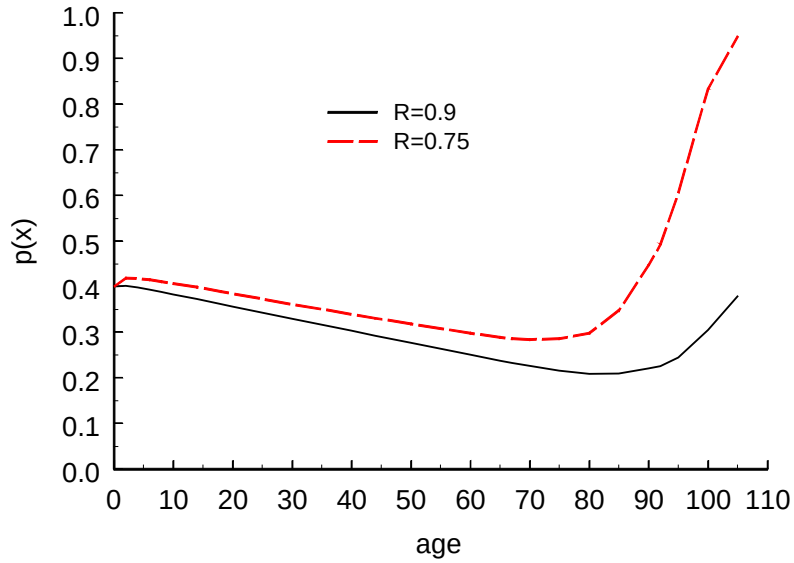

Figure 5.12 Observed frequency trajectories for two genotypes with risks 0.9 and 0.75 when there is a linear increase in genotype frequency

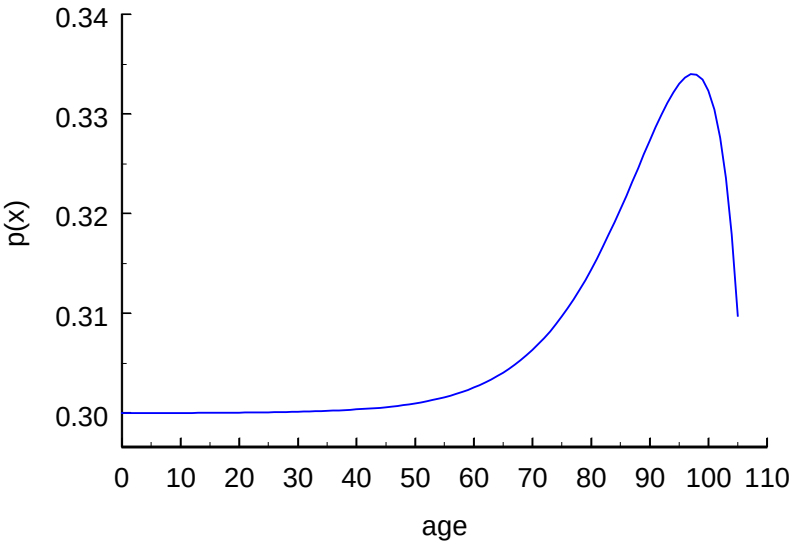

Figure 5.13 Frequency trajectory for one genotype with initial frequency 0.3 when genotypic risk changes linearly from 1 to 0.5

also on the cohort specific initial frequency $p(T-x)$. In this case, inference on the risk R of the genotype based on the age-pattern of observed genotype frequency will definitely be influenced. Figure 5.12 shows the observed frequency trajectories by age for carriers of two genotypes with relative risks R=0.9 and R=0.75 when assuming there is a linear increase in frequency from p_a=0.1 to p_b=0.4 based on the survival distribution from Figure 5.8. Instead of a constant increase as it should be for a beneficial attribute, the observed pattern first unexpectedly decreases until about age 80 and then goes up. Such a pattern can be ascribed to the low genotype frequency in old age cohorts and high frequency in the younger age cohorts. In such a situation, the genotype could be judged as frail if one took a sample only from ages before 80 or even 90. One couldn't expect the initial frequency for a gene to undergo rapid change within the cohort coverage of the sample (something like 100 years) given that the population is in Hardy-Weinberg equilibrium. However, as will be shown in the application, some gene frequencies do differ in different geographical areas due to heterogeneity in ethnic compositions. In this case, it is important to carefully control the sampling process so that ethnic homogeneity over age groups can be assured and possible influences on risk estimation from heterogeneous initial gene frequencies are eliminated.

5.4.3 Secular change in risk of genotype

Another assumption in the model is that risk for the gene of interest does not change with the birth cohort. When this assumption is violated, the situation will be complicated. Suppose that the risk of one genotype changed from R_a to R_b linearly with the birth cohorts covered by the study. Define the genotype risk for cohort born at time t as

$$R(t) = R_a + \frac{R_b - R_a}{T - t_0}(t - t_0). \qquad (5.20)$$

Here, please note $R(t)$ is fixed at birth for the cohort born at time t. It is constant over ages for that cohort. Assuming there isn't secular change in gene frequency and no mortality improvement by cohort, we get genotype frequency at age x for the cohort born at year t as

$$p(x,t) = \frac{p s_0(x,t)^{R(t)}}{\bar{s}(x,t)}. \qquad (5.21)$$

Again, the observed genotype frequency at age x from a cross-sectional study can be calculated by replacing t with $T - x$. The altered pattern of the frequency trajectory is plotted in Figure 5.13 when genotype frequency is p=0.3, R_a=1, R_b=0.5. Here, the risk of the genotype can't be detected when sampling only covers the centenarians and the young group because the frequency peak is missed. Fortunately, like the case for change in gene frequency discussed above, the change in genetic risk is unlikely to happen in a population that is in Hardy-Weinberg equilibrium within a century; but when the sampling process is not well controlled, that is for example, heterogeneous ethnic groups are involved for different ages, one could expect to encounter such a problem.

5.5 Heterogeneity

In term of longevity study, there exist thousands of social and biological attributes that influence survival. It is impossible to have all of them included and measured within a single study. Meanwhile, it is not reasonable to focus on several attributes while disregarding the existence of the others that also contribute. The modelling of heterogeneity with observed covariate has been discussed in section 1.1.3. This part is aimed at revealing how ignorance or misspecification of heterogeneity can alter the parameter estimates.

Given one genotype with relative risk R and frequency p, we simulated data sets by setting the variance of gamma-distributed hidden frailty σ^2 to 0, 0.1, 0.25 and 1. Then the parameters R and p are estimated for the exact as well as for deliberately mis-specified values of σ^2. Tables 5.7, 5.8 and 5.9 are the estimates for data sets simulated with R=2 and p=0.5. In Table 5.7, relative risk R is underestimated when ignoring heterogeneity for heterogeneous populations. Meanwhile, we overestimate the risk of the observed genotype for a homogeneous population by applying a heterogeneous model. On the contrary, frequency of the attribute got overestimated in a heterogeneous population when disregarding σ^2 but underestimated in a homogeneous population by introducing σ^2 (Table 5.8). The highest log likelihood can only be obtained when a proper σ^2 is used in the estimation (Table 5.9), but when

examining their differences using likelihood ratio test, the assumed σ^2 does not cause much difference in the likelihood except when an extremely different σ^2 is introduced. This explains why it is difficult to estimate σ^2 when data size is not big enough. The finding suggests that if we have small data sets, it is better to set σ^2 to different values and estimate the likelihoods and then find the maximum that gives the best estimate. In this situation, we can test the significant level of σ^2 by likelihood

Table 5.7 Average and SE of the estimated risk for different σ^2

Assumed	σ^2=0		σ^2=0.1		σ^2=0.25		σ^2=1	
	R	SE	*R*	SE	*R*	SE	*R*	SE
True								
σ^2=0	**2.0040**	0.1394	2.2623	0.1810	2.7290	0.2634	6.9945	1.4259
σ^2=0.1	1.7863	0.1210	**1.9859**	0.1565	2.3459	0.2244	5.5379	1.0623
σ^2=0.25	1.5954	0.0936	1.7441	0.1198	**2.0134**	0.1711	4.3900	0.8330
σ^2=1	1.2105	0.0745	1.2606	0.0925	1.3509	0.1262	**2.1170**	0.4597

Table 5.8 Average and SE of the estimated frequency for different σ^2

Assumed	σ^2=0		σ^2=0.1		σ^2=0.25		σ^2=1	
	p	SE	*p*	SE	*p*	SE	*p*	SE
True								
σ^2=0	**0.4975**	0.0271	0.4871	0.0276	0.4749	0.0284	0.4531	0.0323
σ^2=0.1	0.5144	0.0249	**0.5043**	0.0251	0.4918	0.0254	0.4656	0.0265
σ^2=0.25	0.5228	0.0252	0.5133	0.0258	**0.5011**	0.0266	0.4702	0.0294
σ^2=1	0.5364	0.0262	0.5313	0.0272	0.5237	0.0288	**0.4953**	0.0345

Table 5.9 Average and SE of the calculated log likelihood for different σ^2

Assumed	σ^2=0		σ^2=0.1		σ^2=0.25		σ^2=1	
	LL	SE	LL	SE	LL	SE	LL	SE
True								
σ^2=0	**-554.993**	12.747	-555.254	12.588	-556.221	12.341	-565.017	11.392
σ^2=0.1	-579.415	13.432	**-579.330**	13.425	-579.670	13.381	-589.309	13.106
σ^2=0.25	-607.937	10.722	-607.610	10.662	**-607.450**	10.572	-610.046	10.294
σ^2=1	-667.586	7.938	-667.343	7.985	-667.028	8.051	**-666.362**	8.235

**Data size=1,000, R=2.0, p=0.5, age from 80-99. Each case is calculated from 100 data sets*

ratio test, i.e. by comparing the likelihood at the peak with the likelihood from the homogeneity model with one degree of freedom. However, one has to notice that such an approach has the potential to underestimate the variances of the other parameters. In the model application, we estimate σ^2 in a setting that it is shared by all the genotypes in order to minimise the weight of it in the overall estimation.

However, since we are dealing with data collected from cross-sectional studies, we are confronted with a new problem when incorporating heterogeneity due to the potential secular changes that might alter the cohort-specific variance of the unobserved frailty. Since individuals participating the study were born into different cohorts, one needs to specify cohort-specific variance of unobserved frailty in order to deal with the situation. This creates a big estimation problem given the large number of variance parameters to be estimated especially when sample size is small. One way to overcome the difficulty is to assume a single variance parameter for all the cohorts as we did in the simulation. This can help to reduce the pain, but on the other hand, there are some side effects that need to be considered. For example, if the true variance is decreasing with time, then as discussed above (Table 5.7), we will overestimate the risk of the gene by imposing a single variance. On the contrary, if the true variance is increasing with time, we could underestimate the risk. However, there is no literature up to now addressing the question concerning the secular trend of heterogeneity. In addition, we can't expect to answer this question in a small-scale study and which primarily focuses on measuring effects of the observed individual genetic variants.

In addition to the difficulty in fitting a frailty model to the empirical data, the estimated results from such a model have to be interpreted with caution. As models with and without heterogeneity can be applied, it has to be pointed out that both approaches are correct, but that they are estimating different parameters. Without taking heterogeneity explicitly into account, *R* describes the change of the marginal survival or hazard function. Taking heterogeneity into account, *R* describes the change of the individual survival or hazard function given the individual frailty. It is important to note that in both situations *R* is a meaningful parameter to describe the effect of a gene.

5.6 Modelling interactions

From (5.10) we see that $\bar{s}(x)$ is the weighted mean survival of the two sub-populations. Formula (5.10) can be naturally extended to include more than two sub-groups on the right-hand side with risk compositions of the observed genetic and non-genetic covariates.

$$\bar{s}(x) = \sum_{i=1}^{k} p_i s(x, R_i) . \qquad (5.22)$$

In (5.22), k is the total number of compositions so that $\sum_{i=1}^{k} p_i = 1$, R_i is the risk for sub-group i. Proportion of sub-group i at age x is

$$p_i(x) = p_i s(x, R_i) / \bar{s}(x) . \qquad (5.23)$$

Based on multinomial distribution (Hastings & Peacock 1975), the likelihood function can be written as

$$L \propto \prod_{x=x_0}^{\infty} \prod_{i=1}^{k} p_i(x)^{n_i(x)} \qquad (5.24)$$

where $n_i(x)$ is number of individuals at age x in sub-group i. $\sum_{i=1}^{k} n_i(x) = N(x)$, is the total number of survivors at age x.

The extension of (5.10) to include multiple groups enables the incorporation of confounding factors as well as interactions in the model. Suppose we observe, additional to the genetic covariates, the individual's sex and region (south or north for example). Combinations of the genetic and non-genetic covariates form subpopulations that are homological to the situation of (5.22). The proportions and total risks for different sub-groups are shown in Table 5.10. In the table, the proportion of individuals from the south is p_s but $1 - p_s$ from the north. Proportion of carriers in the south is p_{gs} but p_{gn} in the north. R_{area} is the risk of a confounding factor area defined as the relative risk of being from the south in reference to being from the north. $R_{g\times a}$ is the relative risk of gene-area interaction; $R_{g\times s}$ is defined as the relative risk for male carriers to female carriers. Given this parameterization, R in Table 5.10 is simply the risk of the gene in females. In order to detect gene-sex interaction, risk compositions are specified for males and for females separately but with shared parameters (Table 5.10). There are two considerations for such an

arrangement. First, male and female survival distributions are different (Hazzard 1986; Holden 1987; Keyfitz & Flieger 1990) with death rate for males usually higher than for females. Secondly, there could be mortality crossover at late ages (Kannisto 1994) indicating the relative risk of sex itself is not proportional. Using both male and female survival functions available from population statistics for the left-hand side of (5.22) instead of introducing only one male or female survival can avoid specifying a proportional risk parameter for sex, and thus improves the estimation while enabling measurement of the risk of gene-sex interaction. In the estimation process, separate likelihood functions are constructed for males and for females, respectively, but with shared parameters. As we can see from Table 5.10, risk of gene-area interaction is defined as the excessive risk of death for carriers from the south to the north. In the same way, risk of gene-sex interaction is the excessive risk of death from male carriers to female carriers.

Table 5.10 Proportions and risks for different sub-groups

Covariate	South		North	
	+	–	+	–
Proportion	$p_s p_{gs}$	$p_s(1-p_{gs})$	$(1-p_s)p_{gn}$	$(1-p_s)(1-p_{gn})$
Risk, males	$RR_{area} R_{g\times a} R_{g\times s}$	R_{area}	$RR_{g\times s}$	1
Risk, females	$RR_{area} R_{g\times a}$	R_{area}	R	1

5.7 Sampling bias, interaction and confounding factor

The non-genetic covariates discussed above, sex and region, are also confounding factors in that they themselves potentially have direct influences on survival because of the existing sex and regional differences in mean life span. A good research design should take into account these important confounding factors and have their interferences on the conclusion controlled or at least minimised. This section investigates the influences on parameter estimates introduced by sampling biases on the confounding factors. This is done by simulating data for given parameters and comparing estimates from biased samples with the true parameters. Different examples are used to illustrate the various situations that might be encountered.

One very common phenomenon in longevity study is that individuals available in the old age groups (for example the centenarian group) are predominantly females, reflecting the strong selection by sex differential mortality. In the sampling process, usually we are taking only a small sample of the total centenarian population. Sometimes the participants in the study differ from those not in the study with respect to the factors that potentially influence survival. What happens to the results if the sample taken does not represent the total population? The following examples are aimed at exploring the answers to this question.

Example 1. Assume there is one neutral gene with relative risk 1 and frequency 0.1. In order to examine the influence of this gene on life span, we take a cross-sectional sample from a population for which initial proportion of males or females is 50% but hazard of death for being a male is 1.5 times higher than being a female. In the sample, we include all individuals aged from 30-80 for both sexes. In the centenarian group, we take all males but only a proportion of females from 95% to 5% (Table 5.11). We can see that the sampling bias results in increased underestimations for the risk of sex as the percentage of females included in the sample decreases. When only 5% of the female centenarians are covered, the risk of sex becomes significantly <1 which is completely contradictory to the true value. Different from the risk of sex, estimates on the genetic parameters are not influenced.

In the above example, although the sampling is biased on sex, the gene function and sex are independent. Because of this independence, estimates for genetic parameters are not affected by the biased sampling; but in practice, we don't know if dependence existed or not, and when there is such dependence, we should know what is going to happen.

Example 2. Similar to example 1, the same sampling bias is encountered but the gene has sex-specific effects that reduce the hazard of death by 0.7 for males. In Table 5.12, results on parameter estimates from two different models are presented. Model I takes into account the existing gene-sex interaction, while model II ignores it. Different from the result in Table 5.11, the estimate of risk of the gene is seriously biased when ignoring gene-sex interaction in model II. Under this model, relative risk of the neutral gene is significantly smaller than 1, meaning it is beneficial to survival. However, one must note that the biased estimation of risk of the gene by model II is mainly due to ignorance of the existing gene-sex interaction, not due to the sampling

bias on female centenarians. Under model I, the full model, true genetic parameters are obtained. As a result of sampling bias on sex, the risk of sex is underestimated

Table 5.11 Parameter estimation for example 1

Percent	Freq. gene		Freq. sex		Risk gene		Risk sex		Data size
	p	SE	p	SE	R	SE	R	SE	
	0.10		0.50		1.00		1.50		
95%	0.10	0.00	0.50	0.00	1.01	0.04	1.59	0.04	13158
70%	0.10	0.00	0.50	0.00	1.02	0.04	1.41	0.04	12456
50%	0.10	0.00	0.49	0.01	1.02	0.05	1.33	0.04	11889
30%	0.10	0.00	0.49	0.01	1.05	0.06	1.21	0.04	11305
5%	0.10	0.00	0.48	0.01	1.03	0.09	0.87	0.05	10623

Table 5.12 Parameter estimation for example 2

Percent	Model	Freq. gene		Freq. Sex		Risk gene		Risk sex		Risk inter.		Data size
		p	SE	p	SE	R	SE	R	SE	R	SE	
		0.10		0.50		1.00		1.50		0.70		
95%	I	0.10	0.00	0.50	0.00	1.00	0.10	1.49	0.05	0.71	0.07	13156
	II	0.10	0.00	0.50	0.00	0.75	0.03	1.58	0.05			13150
70%	I	0.10	0.00	0.50	0.00	1.00	0.09	1.42	0.05	0.71	0.07	12407
	II	0.10	0.00	0.50	0.00	0.75	0.03	1.51	0.05			12459
50%	I	0.10	0.00	0.50	0.01	1.00	0.10	1.35	0.05	0.71	0.08	11868
	II	0.10	0.00	0.50	0.01	0.76	0.03	1.43	0.05			11849
30%	I	0.10	0.00	0.50	0.01	1.00	0.10	1.25	0.05	0.72	0.08	11302
	II	0.10	0.00	0.50	0.01	0.78	0.04	1.31	0.05			11290
5%	I	0.10	0.00	0.49	0.01	1.00	0.14	0.89	0.06	0.75	0.14	10539
	II	0.10	0.00	0.49	0.01	0.83	0.07	0.95	0.06			10555

Table 5.13 Parameter estimation for example 3

Percent	Model	Freq. gene s.		Freq. gene n.		Freq. area		Risk gene		Risk area		Data size
		p	SE	p	SE	p	SE	R	SE	R	SE	
		0.10		0.20		0.50		1.00		1.50		
95%	I	0.10	0.00	0.20	0.00	0.50	0.00	0.99	0.03	1.45	0.04	13147
	II	0.15	0.00			0.50	0.00	0.94	0.03	1.45	0.04	13173
70%	I	0.10	0.00	0.20	0.00	0.50	0.00	1.00	0.03	1.38	0.04	12456
	II	0.15	0.00			0.50	0.00	0.95	0.03	1.38	0.04	12447
50%	I	0.10	0.00	0.20	0.00	0.50	0.01	0.99	0.04	1.31	0.04	11936
	II	0.15	0.00			0.50	0.01	0.97	0.04	1.31	0.04	11891
30%	I	0.10	0.00	0.20	0.00	0.49	0.01	0.99	0.04	1.19	0.04	11338
	II	0.15	0.00			0.49	0.01	0.96	0.04	1.19	0.04	11328
5%	I	0.10	0.00	0.20	0.00	0.49	0.01	1.02	0.08	0.88	0.05	10675
	II	0.15	0.00			0.49	0.01	1.01	0.07	0.86	0.05	10655

both by model I and by model II because of the overrepresentation of males in the centenarian group. The most important conclusion from this example is that, in order to make correct inference on risk of the gene, possible interactions should be taken into account and included in the model.

For some genes, risks and frequencies of their various forms of polymorphism may differ for different geographical regions (to be shown later in the applications). In the following examples, consequences of ignoring such frequency differences and gene-area interaction, together with sampling bias on area will be investigated.

Example 3. Assume the same neutral gene as in the examples above but that frequencies in different regions vary with p=0.1 in the south and p=0.2 in the north. Further, assume that half of the population is from the south and half from the north with the relative risk of being from the south R=1.5 in relation to being from the north. The sample covers all individuals aged from 30-80 and all centenarians from the south but only a proportion of centenarians from the north (Table 5.13). To estimate the parameters, two models are employed. Model I is a full model that assumes different gene frequencies for the two regions but model II ignores it. Results in Table 5.13 show that the risk estimate of the gene from model I is better than that by model II when assuming same frequency for both the south and the north. The risk of being from the region is underestimated by both models I and II due to sampling bias. When only 5% of northerners are in the sample, we even get risk estimation on area smaller than 1, and a frail attribute is erroneously evaluated as robust. The frequency estimate is very stable regardless of the sampling bias, but p is estimated as 0.15, the average of the south (0.1) and the north (0.2). Since in this example, area and gene are independent, the worse estimate of the risk of the gene from model II is solely due to the same gene frequency assumption in model II. The result from this example tells us that when there is a difference in gene frequency by region, ignoring the difference could result in biased estimates of the risk of the gene.

Example 4. In addition to the above example, we now suppose that there is a gene-area interaction that reduces the hazard of death by 0.7 for southerners. Results on parameter estimates from the two models, model I with consideration of gene-area interaction and model II without it, are shown in Table 5.14. In addition to the biased estimate of the risk of area, which is similar to example 3, the estimate of the risk of gene-area interaction is also biased for both models, and both can be ascribed to the biased sampling of area in the centenarian group. The bias of the risk estimate from model II can result from both sampling bias and ignorance of regional frequency differences of the gene. From this example, we conclude that good estimates on

genetic parameters can only be achieved when interaction and regional differences in gene frequency are taken into account.

Table 5.14 Parameter estimation for example 4

Per-cent	Model	Freq. genes.		Freq. Genen.		Freq. area		Risk gene		Risk area		Risk inter.		Data size
		p	SE	p	SE	p	SE	R	SE	R	SE	R	SE	
		0.10		0.20		0.50		1.00		1.50		0.70		
95%	I	0.10	0.00	0.20	0.00	0.50	0.00	1.00	0.04	1.49	0.05	0.69	0.05	13174
	II	0.15	0.00			0.50	0.00	0.93	0.03	1.46	0.05	0.81	0.04	13165
70%	I	0.10	0.00	0.20	0.00	0.50	0.00	1.00	0.04	1.42	0.05	0.68	0.05	12543
	II	0.15	0.00			0.50	0.00	0.92	0.04	1.38	0.05	0.81	0.05	12500
50%	I	0.10	0.00	0.20	0.00	0.50	0.01	1.01	0.05	1.35	0.05	0.66	0.05	11973
	II	0.15	0.00			0.50	0.01	0.94	0.04	1.31	0.04	0.79	0.05	11988
30%	I	0.10	0.00	0.20	0.00	0.50	0.01	1.00	0.05	1.24	0.05	0.66	0.06	11468
	II	0.15	0.00			0.50	0.01	0.92	0.05	1.21	0.04	0.79	0.06	11488
5%	I	0.10	0.00	0.20	0.00	0.49	0.01	1.03	0.10	0.91	0.06	0.58	0.08	10799
	II	0.15	0.00			0.49	0.01	0.94	0.09	0.88	0.05	0.72	0.08	10802

Results from examples 2 and 4 tell us: (1) sampling bias could lead to biased estimates on confounding and interactions and; (2) sampling bias can not prevent us from approaching the true estimate of risk of the gene provided a full model is applied. When sampling is biased on one confounding factor, one would not expect a valid estimate for the risk of that confounding factor, but if our interest is in the genetic parameters, such a bias is usually not problematic.

Example 5. Unlike all of the previous examples, we now have a sample biased by the proportion of individuals from early ages. Given the same parameters as in example 1, we include in our sample only a proportion of females aged from 30-40 from the total population so that the sample is biased for sex for younger ages. Parameter estimates are shown in Table 5.15, from which we observe biased estimates on initial frequency of sex as well as the relative risk of sex; risk and frequency estimates for the gene are not affected. Comparing results from examples 1 and 5, one can see that, while sampling bias on a confounding factor at very old ages mainly influences the risk estimate of the confounding factor, such bias at early ages can alter the estimates of both risk and initial proportion. The result indicates there is a strong influence on the estimation by the initial proportion of observations at early ages.

Table 5.15 Parameter estimation for example 5

Per-cent	Freq. Gene		Freq. Sex		Risk gene		Risk sex		Data size
	p	SE	p	SE	R	SE	R	SE	
	0.10		0.50		1.00		1.50		
95%	0.10	0.00	0.50	0.00	0.98	0.04	1.50	0.04	13199
70%	0.10	0.00	0.53	0.00	0.98	0.04	1.53	0.05	12695
50%	0.10	0.00	0.55	0.00	0.98	0.04	1.57	0.05	12276
30%	0.10	0.00	0.58	0.01	0.98	0.04	1.60	0.05	11830
5%	0.10	0.00	0.62	0.01	0.98	0.04	1.66	0.05	11296

5.8 Summary

This chapter specifies the binomial frailty model in the analysis of observed gene marker data on unrelated individuals. The theoretical approaches in this chapter show

(1) Population survival can be combined with genotype data in determining the influences of observed genes or genotypes on life span (section 5.2).

(2) Both cohort specific survival and carefully chosen period survival can be used in the parameter estimation for data from cross-sectional studies (section 5.4).

(3) Ignorance of individual heterogeneity in unobserved frailty can substantially underestimate the genetic influences on life span (section 5.5).

(4) Interaction and confounding factors should be considered in order to ensure good estimates of genetic parameters and control for sampling bias.

Chapter 6

Findings about Candidate Longevity Genes

6.1 Introduction

Since the 1970s when HLA polymorphism was first studied for a possible association with human longevity (Gerkins et al. 1974; Macurova et al. 1975; Hansen et al. 1977), an interest has developed in the search for candidate genes that contribute to human survival. This interest has intensified in recent years. Although in animal-based research, large effect genes have been identified, for example, the age-1; daf-2; daf-23 genes in the nematode *C. elegans* (Johnson et al. 1999), studies on humans have mainly focused on genetic polymorphic forms that cause diseases or disorders (Table 6.1). This chapter begins with a brief review on the literature of the studies on gene and survival. Then the model developed in Chapter 5 will be applied to data collected from some of the studies reviewed. Data from Danish longevity studies on genetic variations associated with cardiovascular disease (Bladbjerg et al. 1999), polymorphism at ApoB locus, and genotypes of cytochrome P450 enzymes (Bathum et al. 1998) will be employed. The application of the model to the Danish data shows how the model works with empirical data in the search for candidate longevity genes together with gene-sex interaction while considering individual heterogeneity. Several genes that show potential influence on life span and which were not reported as being significant in previous studies have been found using this method. Other data from a multicentral centenarian study in Italy (De Benedictis et al. 1997, 1998, 1999) is used to show how the model can incorporate confounding factors while measuring gene-sex and gene-environment interactions. In collaboration with the authors of the Italian findings, some of the subsets of the Italian data have been used in co-authored publications (Yashin et al. 1998, 1999b, 2000) and one paper has been accepted (Tan et al. 2001).

6.2 A literature review on the genetic study on longevity

6.2.1 The apolipoprotein genes

The apolipoproteins are important components of lipoproteins (Chylomicron, VLDL, IDL, LDL, HDL). Polymorphic variations of apolipoprotein genes have been intensively studied because of their crucial roles in the metabolism of cholesterol. Variations on genes encoding for apolipoprotein E (ε2, ε3, ε4) have been shown to be associated with atherosclerotic cardiovascular disease from Finnish (Kervinen et al. 1994) and Danish (Gerdes et al. 2000) studies. The allele ε4 has been reported to be responsible for higher level of LDL in the serum that increases the risk of ischaemic heart disease (IHD) (Frikke-Schmidt et al. 2000), while the ε2 isoform has been associated with lower level of plasma cholesterol and thus is seen as protective (Helkala et al. 1996). The type 4 allele of ApoE has also been found to be implicated in the incidences of Alzheimer's disease (AD) (Corder et al. 1993). It is reported that carriers of ε4/4 genotype are 9 times more likely to suffer from AD but explanations to causality remain elusive (Marshall 1998). The frequency of ApoE4 among very old people has been found significantly lower than among young people in French (Schachter et al. 1994) and Chinese (Zhang et al. 1998) studies which could be the consequence of the disease association of the ε4 allele.

The Apolipoprotein B is the only protein constituent in low-density lipoprotein or LDL, the major carrier of cholesterol. Early from 1980s, the relationship between ApoB polymorphism and coronary artery disease (CAD) has been investigated by many authors (Hegele et al. 1986; Myant et al. 1989; Paulweber et al. 1989, 1990; Kervinen et al. 1994). De Benedictis (1997, 1998) examined age-related changes in the 3'ApoB-VNTR genotype frequencies with results showing that the frequency of the homozygotes of the 3'ApoB-VNTR alleles with fewer than 35 repeats is increased in the middle ages but then declines to its minimum in centenarians, an indication that the beneficial ApoB alleles at middles ages could convey survival disadvantages at latter ages.

Apolipoproteins A-I and A-II are constituents of high-density lipoprotein (HDL), which functions in returning cholesterol from the artery wall to the liver to be metabolised and secreted. An elevated level of HDL in plasma appears to protect

against cardiovascular disease (CVD). Some studies have reported that over-expression of ApoA-I leads to increased concentration of HDL and appears to confer protection against atherogenesis because of resulted decrease in the plasma cholesterol deposit in arteries (Hoeg 1996; Saha et al. 1995) and which promotes health and enhances longevity (Tybjaerg-Hansen et al. 1993; Luoma 1997).

Another apolipoprotein gene studied for associations with diseases and longevity is the ApoA-IV gene which plays an important role in lipoprotein metabolism, including modulation of triglyceride-rich lipoprotein catabolism, reverse cholesterol transport and cholesteryl ester transfer protein activity. The ApoA-IV codon 360 mutation, ApoA-IV-2 allele, has been found to confer one of the susceptibility markers for Alzheimer's disease (Csaszar et al. 1997) but such association was not detected in a studied Japanese population (Ji et al. 1999). Again, a French study (Merched et al. 1998) failed to link the mutation with AD but instead surprisingly found that the allele could be a marker of aging and longevity with compatible result from an Italian study (Pepe et al. 1998).

6.2.2 Genes coding for angiotensin converting enzyme

The angiotensin converting enzyme (ACE) is a key member of the reninangiotensin system important in the control of blood pressure, salt and water homeostasis, and cell growth. A Japanese study by Nakata et al. (1997) reported that ACE D/D could influence cerebral infarction due to activation of angiotensin II production in cerebral arteries in Japanese population. Recently, the ACE-I allele has been associated with a slightly increased risk of developing late onset Alzheimer's disease (Alvarez et al. 1999). A depletion of ACE I/I genotype in elderly males was found in a British study indicating sex interaction in the expression (Galinsky et al. 1997). Despite of its disease association, an increased frequency of ACE D/D genotype was found among French centenarians compared to younger controls (Schachter et al. 1994). The results of the studies on associations of ApoB and ACE alleles with diseases and longevity support the assumption of antagonistic gene actions in different ages. However, a case-referent and retrospective cohort study based on the Copenhagen City Heart Study reported no frequency change of the ACE D/D genotype as a function of age (Agerholm-Larsen et al. 1997).

6.2.3 Angiotensinogen gene

Angiotensinogen (AGT) gene is associated with the synthesis of different forms of angiotensinogen which are converted into the active form angiotensin II by ACE. The human angiotensinogen gene is subject of many investigations aimed at determining its relationship with essential hypertension (EH) and CVD because plasma AGT levels correlate with raised blood pressure. Polymorphism of AGT gene codon 174, genotype T/T, had been found as a risk factor for CAD and polymorphism of AGT gene codon 235 genotype M/M is negatively associated with CAD in a studied Japanese population (Cong et al. 1998). The T/T genotype of AGT gene condon 235 was also reported as an important risk factor for elevated blood pressure in a Danish study (Sethi et al. 2001) and for coronary heart disease in a Chinese study (Chen et al. 1998). A study in UAE also associated the T235 allele with EH, myocardial infarction (MI) and reduced life span (Frossard et al. 1998). In addition, the genotypes of the angoitensinogen gene were found significantly associated with variation in systolic resting (Hegele et al. 1994) and exercise (Rankinen et al 2000) blood pressure but only in males. The results indicate that the genetic variation at the locus modifies the responsiveness of the blood pressure to endurance training but the functional impact is related to differences in sex (Sethi et al. 2001).

6.2.4 Cytochrome P450 genes

The cytochrome P450 (or CYP450 enzymes) has an important role in metabolism of many drugs and also in the activation and deactivation of carcinogens and other toxic environmental chemicals (Gonzalez 1992). There are 10 families of cytochrome P450 enzymes. Among them, the CYP2D6 and CYP2C19 are most widely studied. Variations of CYP2D6 and CYP2C19 are genetically determined. There are two polymorphic phenotypes, poor metabolizer (PM) and extensive metabolizer (EM) for both the CYP2D6 and CYP2C19 encoding genes. Some studies have associated the PM phenotype of CYP2D6 gene with susceptibility to Parkinson's disease (PD) (Barbeau et al. 1987; Ho et al. 1996). However, no correlation between CYP2D6 PM/EM polymorphism and PD has been found in studied Japanese (Tsuneoka et al. 1998) and Chinese (Chan et al. 1998; Lo et al.1998; Ho et al. 1999) populations where the PM phenotype is extremely rare. In addition, the PM phenotype of CYP2D6 and activity of CYP2C19 are also reported to be protective against lung

cancer for smokers (Benhamou et al. 1996; Stucker et al. 1995; Tsuneoka et al. 1996). A Danish study by Bathum et al. (1998) associated the PM phenotype of CYP2D6 with longevity but such a connection was not found by Yamada et al. (1998) in a Swedish population. Considering the controversial results, the association of CYP2D6 and CYP2C19 polymorphic variations to disease and survival deserves further investigations.

6.2.5 Major Histocompatibility Complex

The Major Histocompatibility Complex (MHC) represents a group of genetically linked loci coding for an extensive variety of surface antigens. These antigens enable the body to recognise between "self" and "non-self" and then to develop and maintain an effective immune response. The genes of the MHC display remarkable polymorphism. The HLA region of MHC has been studied for gene polymorphism and various immune disorders, susceptibility to cancer, and longevity. Among the earliest pioneer studies, Gerkins et al. (1974) reported that there is a likely relationship between heterozygosity of the HLA system, survival to old age, and a decreased susceptibility to cancer, while studies by Macurova et al. (1975), Hansen et al. (1977), Yarnell & Leger (1979) failed to observe such an association. Lack of association between HLA and age in an aging population was also reported by Blackwelder et al. (1982) but with remark that HLA is related to the aging process in a way that is detectable only at very advanced ages. Extensive studies focusing on allelic forms of HLA and human longevity can be found from the late 1980s. For example, association between polymorphism of HLA class II genes and longevity has been confirmed by Takata et al. (1987). An extremely high possibility that HLA-DRw9 is a risk factor and HLA-DR1 is a favourable factor for longevity in a studied Japanese population on the Okinawa island, where proportion of centenarians is more than 3 time higher than that in the whole Japan, was reported. Later another study (Akisaka et al. 1997) on the same population associated polymorphism of the HLA class II genes (HLA-DRB1, DQ) with longevity. Higher frequencies of the HLA-A3(19), B7, Cw7 and DQ1 alleles in the elderly were observed in an Italian study by Ricci et al. (1998). In addition, there are also studies reported that some HLA-DR alleles manifest sex-dependent influence on longevity. Dorak et al. (1994) found a shorter survival associated with the HLA-DR53 haplotype in males but not in females.

Ivanova et al. (1998) reported the correlation between HLA-DR7, DR11, DR13 and longevity with elevated frequency of DR7 in longevous men and increased frequency of DR11 in longevous women. Gene-sex interaction in the expression of HLA haplotypes was also found by Proust et al. (1982). An excess of the HLA-Cw1 antigen was found in the group of elderly females, while an excess of the HLA-Cw7 antigen in the group of elderly males. However, Izaks et al. (1997) reported no association between HLA allelic variation and mortality among those aged 80 and older in a community in the Netherlands.

6.2.6 Poly (ADP-ribose) polymerase

The poly (ADP-ribose) polymerase (PARP) catalyses poly (ADP-ribosyl)ation which is the covalent posttranslational modification of nuclear proteins in response to oxidative and other types of DNA damage. The PARP activity has been found in correlation with species-specific life span (Grube & Burkle 1992; Burkle et al. 1994) and with the longevity of mammalian species (Burkle et al. 1992). Recently, an increased PARP activity was detected in lymphoblastoid cell lines from centenarians (Muiras et al. 1998). As a “guardian of the genome” (Jeggo 1998), PARP forms a critical regulatory component of the cellular response to DNA damage and is speculated to contribute to longevity (Burkle 1998).

6.2.7 Haemostasis genes

The haemostasis genes encode for the coagulation and fibrinolysis proteins, the coagulation factors that form the intrinsic and extrinsic pathways in the process of coagulation. Coagulation factors fibrinogen, factor V, factor VII and fibrinolysis component plasminogen activator inhibitor 1 (PAI-1) are the best established predictors of risk of atherothrombotic disease since common diallelic polymorphic forms of the genes encoding for these proteins influence the plasma levels of these factors. Iacoviello et al. (1998a) found Q/Q and H7/H7 phenotypes of the factor VII RQ353 and intron 7 polymorphic variations are related with decreased risk of myocardial infarction (MI). The gender-related effect in the genetic modulation of coagulation factor VII plasma levels has been studied intensively. The factor VII polymorphism, RQ353-Q allele and 323-p10 allele, were reported to show stronger effect on the reduction of factor VII activity level in male than in female Italians (Di

Castelnuovo et al. 1998). On the other hand, elevated level of factor VII coagulant activity has been observed with RQ353 polymorphism in females (Mennen et al. 1997) and which leads to increased risk of recurrent cardiac events in postinfarction women (Kalaria et al. 2000; Ossei-Gerning et al. 1998). Increased level of factor VII activity as a cardiovascular risk factor was reported by Scarabin et al. (1996) in French and by Ishikawa et al. (1997) in Japanese postmenopausal females. It has been suggested that hormonal changes induced by menopause increase plasma levels of activated factor VII (Scarabin et al. 1990). Given the disease association, no polymorphism contribution to longevity from factor VII has been found in Scottish nonagenarians (Meiklejohn et al. 2000). The 4G/4G homozygous genotype of the plasminogen activator inhibitor 1 gene has been associated with high PAI-1 level (Mannicci et al. 1997) and which is a risk factor for recurrent myocardial infarction in men (Hamsten et al. 1987; Thogersen et al. 1998). It is intriguing that higher plasma level of PAI-1 homozygote of the 4G allele has been found in centenarians than in young individuals (Mannucci et al. 1997). A factor V mutant, Arg506Gln, is a pathogenetic factor of venous thromboebolism (Dahlback 1995) due to its functional resistance to activated protein C (APC). Parallel frequencies for factor V polymorphisms and hypercoagulability have been reported in centenarians (Mari et al. 1995, 1996) which may mean that hypercoagulation could be sometimes profitable at very old ages. However, there are other studies with contradictory results. For example, a Danish study concluded that longevity is independent of the common variations in the genes associated with cardiovascular disease (Bladbjerg et al. 1999).

6.2.8 Genetic variations of tissue plasminogen activator

The tissue plasminogen activator (TPA) is the key enzyme in the initiation of an endogenous fibrinolytic/thrombolytic response for the removal of fibrin from the vascular tree. An impaired blood clotting mechanism characterised by increases in the TPA antigen is an important predictor for increased risk of stroke especially for women (Macko et al. 1999). Increased level of TPA antigen has also been associated with MI (Thogersen et al. 1998). However, no association between TPA insertion /deletion polymorphism and MI has been confirmed from a large population based study by van der Bom et al. (1997). Likewise, no association was found in a Danish study using centenarians and controls (Bladbjerg et al. 1999).

6.2.9 Methylenetetrahydrofolate reductase gene

Methylenetetrahydrofolate reductase (MTHFR) is the key enzyme in the methylation of homocysteine. A mutant allele of the MTHFR gene, Val-allele, is reported to associate with increased mortality among homozygous individuals due to (1) decreased level of MTHFR activity (Frosst et al. 1995) leading to increased plasma homocysteine (Bratstrom et al. 1998) and high risk of cardiovascular disease (Boushey et al. 1995); (2) decreased DNA methylation leading to increased mutation rate (Chen et al. 1998) and cancer (Laird & Jaenisch et al. 1996). Recent studies by Heijmans et al. (1999; 2000) found that homozygosity for the Val-allele is associated with increased mortality in men in the middle and old ages but not in women with consistent results from both cross-sectional and follow-up studies in the Dutch population. Mutation of the MTHFR gene has also been reported as a genetic factor that prevents the attainment of old ages by Matsushita et al. (1997). On the contrary, Brattstrom et al. (1998) found no connection between the mutation and premature death.

6.2.10 Mitochondrial DNA

The mitochondrial DNA (mtDNA) is a circular molecule which contains genes coding for ribosomal RNAs (rRNAs), transfer RNAs (tRNAs) and some mitochondrial proteins. All mitochondrial-encoded proteins have a specific function in oxidative phosphorylation. They are essential components of energy-transducing enzyme complexes of the inner mitochondrial membrane. About 2% of the oxygen reduced by the mitochondrion escapes as superoxide from the electron transport chain in the inner mitochondrial membrane where mtDNA is attached nakedly, making it very vulnerable to oxidative damage. Association of mtDNA damage with ageing and degenerative diseases (CAD, diabetes, PD, AD) has been studied by many authors (Corral-Debrinski et al. 1992; Kadenbach & Muller-Hocker 1990; Wallace 1992; Sont & Vandenbroucke 1993). MtDNA mutation and survival was reported by Linnane et al. (1989) and Ozawa et al. (1995). In recent studies, mtDNA genotype Mt5178C has been associated with susceptibility to adult-onset diseases and genotype Mt5178A with longevity (Tanaka et al. 1998; Gong et al. 1998). MtDNA haplogroup J was

found three times more frequent in centenarian group than in younger individuals in northern Italy (De Benedictis & Franceschi 1998; De Benedictis et al. 1999).

6.2.11 p53 tumor suppressor gene

Since its discovery in 1979, the history of investigations of the p53 tumor suppressor gene has been a paradigm in cancer research (Harris 1996). The p53 protein is involved in several central cellular processes, including gene transcription, DNA repair, cell cycling, genomic stability, chromosomal segregation, senescence, and apoptosis. The clinical research on the association between p53 mutation and cancer has been controversial. Some studies have linked p53 codon 72 polymorphisms with lung cancer (Murata et al. 1998; Wang et al. 1999), hepatocellular carcinoma (Yu et al. 1999), breast cancer (Wang-Gohrke et al. 1998) and colorectal neoplasia (Sjalander et al 1996). However, the p53 codon 72 association with cancer and aging has been questioned by Sun et al. (1999) considering the inconsistent results in the literature. Bonafe et al. (1999a; 1999b) reported that no significant difference in frequencies of the p53 codon 72 polymorphisms (Pro72 and Arg 72 alleles) between control and the centenarians both on continental Italy and in Sardinia. Given the important role of p53 in the cancer development, the lack of association between p53 codon 72 polymorphism and longevity could suggest a complicated scenario in the genetic modulation of life span such as antagonistic pleiotropy. On the other hand, it has been indicated that p53 polymorphisms appear to modulate an individual's risk of cancer only under some peculiar condition (viral infection) in that the proportion of individuals selected out during aging could be small (Bonafe et al. 1999a).

6.2.12 Tyrosine hydroxylase gene

The Tyrosine hydroxylase (THO) gene encodes the rate-limiting enzyme for the synthesis of the catecholamines, the aminoacid-derived molecules that act as both hormones (adrenalin) and neurotransmitters (dopamine and noradrenalin), and is therefore an important stress-responder gene (De Benedictis et al. 2000). De Benedictis et al. (1998) observed a significant decrease in the frequency of the THO large allele group (alleles 9, 10-1, 10) in male Italian centenarians but not in females. The significant influence from THO locus could be relevant to the complex relationship existing between insulin and catecholamins (Natali et al. 1998) in glucose

metabolism, whose regulation in turn affects life span from yeast (Jiang et al. 2000) to humans (Paolisso et al. 1996). Recently, a retrograde response mechanism in human aging is proposed by examining the dependent frequency distributions between the THO genotypes and the mtDNAhapl-U in Italian centenarians (De Benedictis e al. 2000). The mtDNAhapl-U was found more frequent in the THO long alleles homozygotes (LL) than in the non-LL subjects. The overrepresentation of mtDNAhapl-U in the centenarians reflects a compensatory mechanism in protection against the unfavourable THO LL genotype (De Benedictis et al. 2000). However, such a compensation was only observed in females indicating that the two sexes follow different trajectories toward extreme longevity (Franceschi et al. 2000).

6.2.13 Others

In addition to the above studies, there are other studies that reported genetic variations and longevity. Tan et al. (2001) found a significant influence on longevity from superoxide dismutase 2 (SOD2) gene. The frequency of SOD2-T allele carriers among centenarians shows significant increase in the south of Italy but not in north. The result supports the finding that SOD2 polymorphism affects the efficiency of mitochondrial transport (Shimoda-Matsubayashi et al. 1996). In a Japanese study, Shimokata et al. (2000) reported that the frequencies of dihydrolipoamide succinyltransferase (DLST) genotypes A/A and C/C decrease with age. DLST is a key enzyme of the mitochondrial α-ketoglutarate dehydrogenase complex. Its association with survival could be related to the connection with Alzheimer's disease. It is reported that patients with Alzheimer's disease have suppressed activity of the mitochondrial α-ketoglutarate dehydrogenase complex (Nakano et al. 1997).

The survey above is just a brief review of the up-to-date literature relevant to longevity study. It has to be realised that, with the current rate of gene discoveries it is impossible to put together at any time a comprehensive overview of the genes that are associated with individual survival and their possible functions. However, although the literature has been rich, as mentioned in section 5.1, the statistics utilized have generally been simple and conventional. In the following sections, the model derived in Chapter 5 will be applied to gain new insights from some of the data in the above studies.

Table 6.1 A brief review on reported genes relevant to human survival

Gene	Function	Disease or Longevity Associations
ApoE (ε2, ε3,ε4)	Constituents of serum HDL, VLDL. ε2 associated with lower, ε4 with higher levels of plasma cholesterol.	ε4 with CVD, AD. ε2 with longevity
ApoB	Sole apoprotein of LDL.	CAD
ApoA-I, II	Constituent of HDL. Overexpression leads to higher HDL.	Protection against CAD
ApoA-IV	lipoprotein metabolism	AD? ApoA-IV-2 with longevity
ApoC3	Constituent of VLDL, HDL.	Overexpression causes hypertriglyceridaemia
ACE	Key member of the reninangiotensin system controlling blood pressure, salt-water homeostasis, cell growth.	Myocardial and cerebral infarctions, AD and hypertension.
AGT	Angiotensinogen (AGT) gene is associated with synthesis of different forms of angiotensinogen which are converted into the active form angiotensin II by ACE. Regulating blood pressure.	EH, CVD.
CYP2D6, CYP2C19	Activation and deactivation of carcinogens and other toxic environmental chemicals.	PD? PM protective to lung cancer
MHC: HLA class II (DR, DQ), class I (B,C,A) genes	A group of genetically linked loci coding for an extensive variety of surface antigens that enable the body to develop and maintain an effective immune response.	Immune disorder, cancer susceptibility.
P53	Tumor suppressor gene	Cancer?
PARP	Forms a critical regulatory component of the cellular response to DNA damage.	
Haemostasis genes coding for factors V, VII , PAI-1	Some polymorphisms associated with increased plasma level of coagulation and fibrinolysis proteins.	MI, thromboembolia.
TPA	Key enzyme in the initiation of an endogenous fibrinolytic/thrombolytic response for the removal of fibrin from the vascular tree.	MI? Stroke
MTHFR	Methylenetetrahydrofolate reductase (MTHFR) is the key enzyme in the methylation of homocysteine.	CVD, Cancer?
THO	Catecholamine synthesis	Stress response
SOD2	Oxygen free radicals scavenging	Affects efficiency of the mitochondrial transportation
DLST	Key enzyme of the mitochondrial α-ketoglutarate dehydrogenase complex	AD
Mitochondrial DNA	Circular molecule containing genes coding for rRNAs, tRNAs and some mitochondrial proteins. All mitochondrial-encoded proteins have a specific function in oxidative phosphorylation.	CAD, diabetes, PD, AD

6.3 Application to data from Danish studies

6.3.1 DNA polymorphism of selected CVD indicators

Cardiovascular disease accounts for about 50% of all deaths worldwide. In the Scandinavian countries cardiovascular disease (CVD) has become a major cause of death. In order to detect the importance of genetic variations in prevalence of the disease, blood samples were taken both from young and old people in Denmark (Bladbjerg et al. 1999). The group of old people consists of two parts. The first set of the data is from the first centenarian study carried out on Funen island. All living individuals who were born before December 31, 1894 and who were still alive on

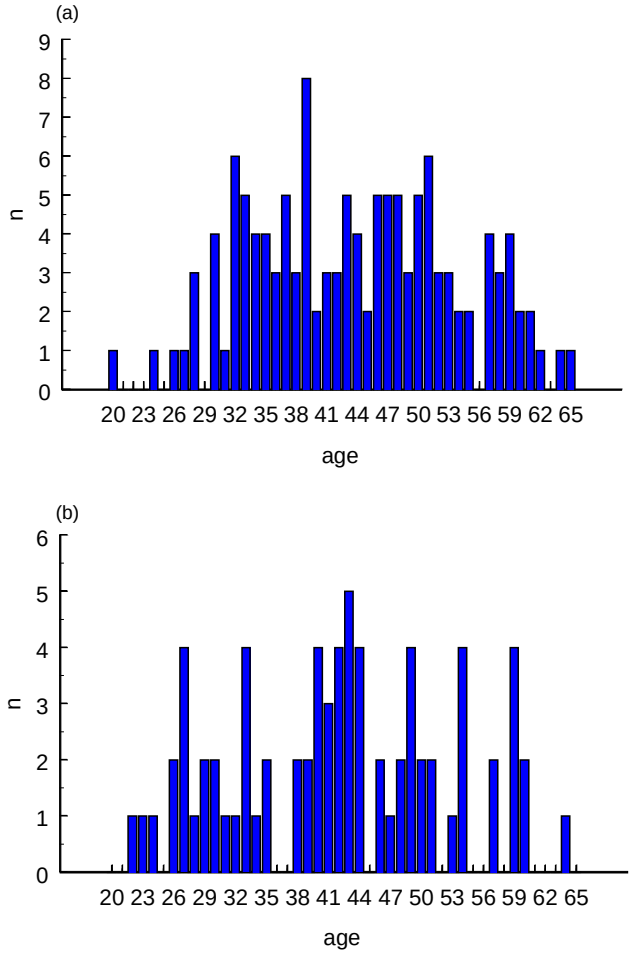

Figure 6.1 Frequency distribution by age for male (a) and female (b) blood donors

May 1, 1994 are included in the study. They were interviewed the same day and 39 out of 51 participants agreed to have blood sample taken. The second set comes from the Danish Longitudinal Centenarian Study, which covered all people who became centenarians during April 1, 1995 and May 31, 1996 throughout Denmark. Blood samples were collected for 148 out of 207 participants. Altogether, blood samples were collected from a total of 187 centenarians, among them 47 males and 140 females. The younger group consists of blood donors at the blood bank of Odense University Hospital. They are healthy people aged from 20-64 years (75 females and 126 males). Figure 6.1 is the frequency distributions by age for male (a) and female (b) blood donors in the younger control group.

Two kinds of candidate genes that influence the developments of CVD are investigated, haemostasis genes and blood pressure regulating genes (Table 6.2). Among the genes tested, FVII forms the extrinsic pathway in the blood coagulation process. According to previous studies, reviewed in section 6.2.1, some of the polymorphic variations (FVII R/Q 353, FVII-323ins10 and FVII intron7(37bp)n) may have higher or lower Factor VII activities. Polymorphic variations of the other genes, PAI-1-675, t-PAintron8ins311, GPIIb/IIIa (glycoprotein), Prothrombin and MTHFR, have been reported relevant to lower or higher CVD risks. Of the two blood pressure regulating genes, angiotensinogen gene is associated with synthesis of different forms of angiotensinogen while ACE (angiotensin converting enzyme) is responsible for the conversion of them into the active form of angiotensin II. Some studies showed that homozygous I/I genotype is associated with higher plasma activities of the enzyme (reviewed in section 6.2.1).

Frequency differences of DNA polymorphism between the young and centenarian groups were first analysed by Else M. Bladbjerg et al. (1999) using conventional statistical methods. As shown in Table 6.2, no differences in frequencies for all the genotypes can be detected between the two groups, which suggests that polymorphic variations of candidate genes of CVD do not contribute to longevity. Since in these data we have for each individual, age at participation and sex, in addition to the genetic information, it is possible to apply the newly developed model and estimate frequency, risk of the genotype and risk of gene-sex interaction for each gene.

Table 6.2 DNA polymorphism of CVD candidate genes

Genotypes	Control group number (frequency)	Centenarian group number (frequency)	χ^2 (p-value)
Haemostasis genes			
FVII R/Q 353			0.82(0.66)
RR353	162(0.81)	157(0.84)	
RQ353	37(0.18)	28(0.15)	
QQ353	2(0.01)	2(0.01)	
FVII-323ins10			1.50(0.48)
P0P0	156(0.78)	153(0.82)	
P0P10	42(0.20)	30(0.16)	
P10P10	3(0.02)	3(0.02)	
FVII intron7 $(37bp)_n$			0.00(0.99)
H5H5	0(0.00)	1(0.005)	
H5H6	10(0.05)	6(0.03)	
H5H7	3(0.02)	6(0.03)	
H6H6	96(0.48)	86(0.47)	
H6H7	75(0.37)	67(0.37)	
H7H7	17(0.09)	15(0.08)	
H7H8	0(0.00)	1(0.005)	
PAI-675(4G/5G)			2.15(0.34)
4G4G	55(0.27)	54(0.29)	
4G5G	106(0.53)	86(0.46)	
5G5G	40(0.20)	47(0.25)	
t-PA intron8ins311			1.85(0.39)
I/I	76(0.38)	71(0.39)	
I/D	94(0.47)	76(0.41)	
D/D	31(0.15)	37(0.20)	
GPIIb/IIIa L/P33			0.21(0.90)
LL	141(0.70)	133(0.71)	
LP	54(0.27)	49(0.26)	
PP	4(0.02)	5(0.03)	
Prothrom 20210G/A			0.84(0.36)
GG	193(0.97)	184(0.98)	
GA	6(0.03)	3(0.02)	
AA	0	0	
MTHFR A/V114			0.60(0.74)
AA	96(0.48)	85(0.46)	
AV	86(0.43)	79(0.43)	
VV	19(0.10)	22(0.12)	
BP regulating genes			
ACE intron16ins287			0.07(0.97)
I/I	46(0.23)	41(0.22)	
I/D	102(0.51)	95(0.51)	
D/D	51(0.26)	49(0.27)	
Angiotens M/T235			0.53(0.77)
MM	81(0.41)	77(0.37)	
MT	98(0.49)	85(0.41)	

Table revised from Bladbjerg et al, Thromb Haemost (1999) 82:1100-5

Before estimation, frequency analysis was done in order to check the frequency distributions of the different gene alleles. Two alleles of FVII intron7(37bp)n, H5 and H8 are not included in the analysis because of their very low frequencies.

For each allele form of the genes, the subjects can be divided into two groups, carriers (with 1 or 2 alleles) and non-carriers. Assume for one gene allele of interest, hazard of death for carriers is R times that of the non-carriers and the initial frequency of carriers in the population is p. In addition, risk of death for male carriers of the allele is $R_{g\times s}$ time that of the female carriers. Note that under such a parameterization, the R is simply the relative risk of the gene in females. According to the relative risk model discussed in section 5.2, the average survival distribution of male and female populations consisting of both carriers and non-carriers are

$$\bar{s}_m(x) = ps_{0,m}(x)^{RR_{g\times s}} + (1-p)s_{0,m}(x) \text{ and}$$
$$\bar{s}_f(x) = ps_{0,f}(x)^{R} + (1-p)s_{0,f}(x). \qquad (6.1)$$

The proportions of male and female carriers at age x can be calculated using (5.3) and (6.1), that is

$$p_m(x) = \frac{ps_{0,m}(x)^{RR_{g\times s}}}{\bar{s}_m(x)} \quad \text{and} \quad p_f(x) = \frac{ps_{0,f}(x)^{R}}{\bar{s}_f(x)}. \qquad (6.2)$$

The likelihood for the sample consisting of both male and female subjects at age x is constructed in accordance with (5.6).

$$L(x|s_{0,m}(x), s_{0,f}(x), p, R, R_{gxs}) \propto p_m(x)^{n_m(x)}(1-p_m(x))^{N_m(x)-n_m(x)} p_f(x)^{n_f(x)}(1-p_f(x))^{N_f(x)-n_f(x)}. \qquad (6.3)$$

Note in (6.3) the frequency is the same for both sexes. The risk of gene-sex interaction is assigned as an extra risk for male carriers in addition to the risk of the allele itself.

The above strategy is primarily aimed at detecting risk of gene-sex interaction and for risks of genes for females. In the application, when no risk of gene-sex interaction is found, we can simply estimate the risk parameters by combining males and females, or in other word, estimate the population risk by deploying (5.10) and (5.11).

In order to carry out the estimation, age specific cohort population survival is calculated from cohort life tables for Danes born between 1894-1974 such that the survival for each age is taken from cohort life table for people who reached that age in 1994. The estimation is done using the Two-step MLE described in section 5.2.3. There are 5 polymorphic forms for FVII intron7(37bp)n, H5, H6, H7, H8, H9, but only estimations on allele H6 and H7 were done. Estimations on the other alleles of this gene were not possible due to the small number of observations.

We first estimted risks and frequency of gene alleles by combining data of the two sexes (assuming no gene-sex interaction) and then we applied the strategy described above to estimate gene-sex interactions. The results on parameter estimates are shown in Table 6.3 for which significant levels for risk parameters (relative risks for the genes and for gene-sex interaction) are determined by testing the statistical differences between the estimated risks and 1 with null hypothesis H_0: R=1. There is no significant risk for each of the single gene allele in Table 6.3, which is consistent with Bladbjerg et al. (1999). However, 3 polymorphic variations, angiotensinogen-M, F7RQ353-R and F7323ins-p10, display significant gene-sex interactions that reduce the hazards of death for male carriers by 0.856, 0.857 and 0.860 when the risk of gene-sex interaction for females is set to 1 as reference. Incorporation of gene-sex interaction by the new model produces results that were missed in the original analysis using only the gene frequency method (Bladbjerg et al. 1999). Ignoring interactions when comparing the gene frequency differences between the control group and the centenarians could be attributed to the lack of efficiency evinced by the gene frequency method when dealing with small samples as discussed in section 5.1. In the new model, the risk of gene-sex interaction is defined as an extra risk for male carriers in comparison to female carriers. Information from both sexes is used in determining the parameter of interaction.

The detected sex-specific influences of the three gene polymorphisms are not totally unexpected as previous studies have found sex-dependent effects of these polymorphisms on other phenotypical traits. For example, the genotypes of the angiotensinogen gene were found to be significantly associated with variation in systolic resting (Hegele et al., 1994) and exercise (Rankinen et al., 2000) blood pressure, but only in males. The results indicate that the genetic variation in the locus modifies the responsiveness of the blood pressure to endurance training, but the

functional impact is related to differences in sex. The gender-related effect in the genetic modulation of FVII plasma levels has been studied extensively. The FVII polymorphisms intron7(37bp)n and –323ins10 were reported to show stronger effect on plasma FVII level in male than in female Italians (Di Castelnuovo et al., 1998).

Table 6.3 Parameter estimates for the CVD associated genes

Gene allele	Freq		Risk			Risk of gene-sex interaction		
	Est.	SE	Est.	SE	p-value	Est.	SE	p-value
MTHFR-A	0.905	0.015	1.062	0.075	0.411	0.930	0.092	0.444
MTHFR-V	0.524	0.026	0.996	0.044	0.930	0.984	0.058	0.778
Angioten-M	0.592	0.025	1.052	0.046	0.260	**0.856**	**0.056**	**0.010**
Angioten-T	0.894	0.016	1.040	0.072	0.578	0.965	0.089	0.691
ACE-I	0.747	0.022	1.020	0.051	0.691	0.968	0.065	0.627
ACE-D	0.768	0.022	0.978	0.053	0.679	1.048	0.071	0.500
F7RQ353R	0.990	0.005	0.939	0.231	0.790	1.219	0.316	0.488
F7RQ353Q	0.195	0.020	1.097	0.063	0.123	**0.857**	**0.063**	**0.023**
t-PA-I	0.845	0.019	1.064	0.062	0.301	0.989	0.073	0.880
t-PA-D	0.625	0.025	1.019	0.046	0.687	0.969	0.059	0.600
PAI-675-4G	0.801	0.020	1.065	0.056	0.240	0.976	0.065	0.710
PAI-675-5G	0.725	0.023	1.008	0.049	0.877	1.015	0.065	0.817
GP2b3a-L	0.980	0.007	1.080	0.155	0.609	0.940	0.180	0.738
GP2b3a-P	0.293	0.023	1.010	0.049	0.840	0.981	0.063	0.767
F7323ins-p0	0.985	0.006	0.994	0.189	0.974	1.072	0.235	0.760
F7323ins-p10	0.220	0.021	1.102	0.060	0.092	**0.860**	**0.061**	**0.021**
F7intron7-H6	0.900	0.015	1.040	0.074	0.586	1.050	0.089	0.576
F7intron7-H7	0.474	0.026	1.000	0.045	0.994	0.941	0.058	0.304
Fabrinogen-G	0.979	0.007	1.015	0.158	0.926	0.563	0.964	0.650
Fabrinogen-A	0.342	0.024	1.067	0.050	0.181	0.918	0.057	0.150

The plasma FVII clotting activity has been suggested as a risk marker of cardiac death (Meade et al., 1993), and also the FVII polymorphisms are predictive of CVD in Italian studies (Iacoviello et al., 1998; Girelli et al., 2000). However, the putative relation between FVII polymorphisms and CVD are turned down in other studies

(Lane et al., 1996; Feng et al., 2000), and it may be speculated that this discrepancy is, at least partly, caused by an effect of gender. In women, FVII levels are increased by menopause (Scarabin et al., 1990, 1996) and hormone replacement therapy (Cushman et al., 1999; Marchien van Baal et al., 2000), and elderly women have higher FVII levels than men of the same age (Scarabin et al., 1996; Ishikawa et al., 1997). Also, women with coronary artery disease have higher FVII levels than men with the disease (Kalaria et al., 2000; Ossei-Gerning et al., 1998). Perhaps the effect of hormones and lifestyle factors drown the effect of genotype on CVD risk in women. Although there have been many studies on disease association, the present study reveals, for the first time, the sex-dependent association of the polymorphisms with longevity.

6.3.2 DNA polymorphism at ApoB locus

The ApoB polymorphic variations have been associated with coronary artery disease (CAD) (reviewed in section 6.2.1). Individuals participating in the CVD study above were also examined on their genetic variations at ApoB locus to see how ApoB allele variations are related to CAD in Danish population. A total of 13 3'ApoB-VNTR alleles have been genotyped with alleles 43, 51, 55 excluded in the analysis because of their very low frequencies (Table 6.4). The same model applied in the analysis of CVD data is introduced. Parameter estimates obtained from separate estimation on risks and frequency for carriers of each allele show no significant influence from any of the 10 alleles on risk of the polymorphism alone or on the risk of gene-sex interaction.

De Benedictis et al. (1998) found age-related changes of the 3'ApoB-VNTR genotype pool on data from Italian centenarian study. In the study, the 3'ApoB-VNTR alleles are pooled into 3 size-classes: small, S, 26-34 repeats; medium, M, 35-39 repeats; large, L, 41-55 repeats. The Italian study reported a significant increase of S/S homozygotes in the observation. Re-examining the results in Table 6.4, we see that the risks on gene-sex interaction for the alleles covered by the L class are all above 1 although not statistically different from 1. One could expect that the results would be improved when alleles are grouped. In accordance to De Benedictis et al. (1998), 6 genotypes, S/S, S/M, S/L, M/M, M/L, L/L, are obtained. Here, our interest is the influence exerted by each of the genotypes resembling the situation described by (5.1), (5.2), (5.3) and (5.6) where p and $p(x)$ are frequencies for each of the 6 genotypes at

birth and at age x. Relative risk R is the risk for carrier of the genotype in respect to non-carriers. Risk of gene-sex interaction $R_{g\times s}$ is the risk for male carriers of the genotype in respect to female carriers as defined in section 5.6. Results on the

Table 6.4 Estimates on allele frequency, risk and gene-sex interaction

ApoB gene allele	Freq.		Risk			Risk of gene-sex interaction		
	Est.	SE	Est.	SE	p-value	Est.	SE	p-value
31	0.144	0.018	1.023	0.065	0.728	0.959	0.080	0.612
33	0.069	0.013	0.918	0.079	0.299	0.997	0.104	0.974
35	0.434	0.025	1.035	0.046	0.441	1.013	0.060	0.834
37	0.652	0.024	1.019	0.047	0.679	1.008	0.061	0.895
39	0.090	0.015	1.046	0.083	0.578	1.001	0.112	0.994
41	0.020	0.007	0.877	0.136	0.367	1.114	0.220	0.604
45	0.015	0.006	1.150	0.233	0.520	1.623	2.430	0.798
47	0.130	0.017	0.942	0.062	0.344	1.089	0.097	0.362
49	0.144	0.018	0.928	0.058	0.216	1.090	0.091	0.327
51	0.026	0.008	1.259	0.223	0.245	1.322	0.950	0.735
53	0.026	0.008	1.034	0.149	0.820	1.703	1.526	0.645

Table 6.5 Estimates on genotype pool frequency, risk and gene-sex interaction

ApoB genotype pool	Freq.		Risk			Risk of gene-sex interaction		
	Est.	SE	Est.	SE	p-value	Est.	SE	p-value
S/S	0.005	0.004	0.700	0.232	0.196	0.818	0.192	0.343
S/M	0.178	0.020	0.987	0.057	0.815	0.945	0.070	0.426
M/M	0.483	0.026	1.084	0.047	0.073	0.940	0.055	0.275
M/L	0.254	0.022	0.925	0.048	0.118	**1.185**	**0.086**	**0.032**
L/L	0.050	0.011	1.113	0.121	0.352	0.849	0.115	0.191

Table 6.6 Estimates on allele class frequency, risk and gene-sex interaction

ApoB allele class	Freq.		Risk			Risk of gene-sex interaction		
	Est.	SE	Est.	SE	p-value	Est.	SE	p-value
S	0.213	0.021	0.955	0.052	0.386	0.958	0.066	0.523
M	0.916	0.014	1.025	0.079	0.756	1.067	0.097	0.490
L	0.334	0.024	0.940	0.045	0.181	**1.158**	**0.076**	**0.038**

estimation are presented in Table 6.5. As in Table 6.4, the risks for the genotypes are not detected as significant but genotype M/L displays a significant sex-dependent influence by which hazards of death for the male carriers are increased by 1.185 (p_{-value} =0.032). Note in Table 6.5, genotype S/L is excluded because of its extreme estimate on gene-sex interaction resulted from missing observation of males in the centenarian groups. The genotype frequency for S/L in the control group is 0.024 for males (3 out of 123) and 0.056 for females (4 out of 71). In the centenarian group, the frequency becomes 0 for males (0 out of 44) and 0.045 for females (6 out of 132). The difference in frequency between centenarians and the control group is larger among males than among females but it is not significant due to the small sample size.

The significant results on genotype-sex interaction for M/L genotype in Table 6.5 could reflect that allele class M or L has a crucial role in determining the influence. In order to test the hypothesis, the model employed above can be elaborated to estimate frequency, relative risk, and risk of gene-sex interaction for carriers of each allele class just as in Table 6.3 and 6.4. But instead of a single allele, we have groups of alleles, S, M and L. Relative risks for allele classes S and M are not significantly different from 1, indicating they are neutral (Table 6.6). Only the risk estimate on class L showed significant sex-dependent influence on survival for males with relative risk 1.158 (p_{-value} =0.038). Comparing the results from Table 6.6 with that from Table 6.5, we see that although the L allele class is important, its effect depends both on sex and on the existence of other allele classes such as class M. This conclusion is also supported by the frequency distribution of genotype S/L discussed above.

6.3.3 Variations of cytochrome P450 genes

The cytochrome P450 (CYP) genes code for cytochrome P450 enzymes which are important in dealing with lipophilic chemicals and in metabolizing therapeutic drugs. Allelic variations of the genes could be responsible for an individual's vulnerability to diseases like cancer that might be associated with environmental chemicals. Aimed at examining the relationship between P450 gene variation and longevity, two genes CYP2D6 and CYP2C19 have been genotyped for their allelic variations in a Danish study by Bathum et al.(1998). The data collected in 1995-1996

consists of a total of 241 individuals aged above 95 (nonagenarians and centenarians). The control group includes 325 healthy Danish volunteers with a median age of 23.7 years (ranging from ages 20-45) for CYP2D6 genotyping and 64 Danish volunteers with median age 27 ranging from 22-52 genotyped for CYP2C19. The centenarians participating the study come from the Danish Longitudinal Centenarian Study (156 blood samples) and from the centenarian study in the county of Funen (31 blood samples). Both studies were mentioned in section 6.3.1. Nonagenarians in the sample (34 blood samples) are part of a feasibility study of a cohort of citizens from Odense in the county of Funen born in 1900 and examined in 1995. The frequencies of different genotypes by age groups are shown in Table 6.7.

Similar to the case of ApoB genotype pool data, estimation can be done on risk and frequency for carriers of each genotype and of each allele of the two genes. However, different from the ApoB data, no gender information on the participants is provided in the data. In this case, the gene-sex interaction can't be measured. This resembles the similar situation described by (5.1), (5.2), (5.3) and (5.6). In the estimation process, age specific cohort survival for females is introduced based on the fact the majority of oldest-old individuals are females. Table 6.8 is the estimated results on risks and frequencies by each genotype. Three genotypes of the CYP2D6 gene, *1/*5, *1/*4, *4/*5, show significant influences on survival, the risks for genotypes *1/*5 (R=0.455, $p_{-value} \approx 0.000$) and *4/*5 (R=0.550, p_{-value}=0.050) are both beneficial which cut individual's hazard of death by about half, but the risk for *1/*4 (R=1.089, p_{-value}=0.023) is deleterious. No genotype of the CYP2C19 gene manifests significant effect on life span.

The 3 significant genotypes in Table 6.8 involves 3 allelic variations of the CYP2D6 gene, allele *1, allele *4 and allele *5. One can suggest that the significant influence could be ascribed to some particular allele or alleles. This uncertainty leads to the consideration of estimating a single allele just as in the case for allele class of ApoB gene in Table 6.6. The estimates on risk and frequency of carriers of each of the 3 alleles are presented in Table 6.9. This time, it is clear that the rare CYP2D6*5 allele, a poor metaboliser, is a very significant beneficial allele that reduces an individual's hazard of death by 0.575 ($p_{-value} \approx 0.000$). It is interesting to see that risk of the allele is very close to the risks of genotypes CYP2D6*1/*5 and CYP2D6*4/*5

(Table 6.8), an indication that CYP2D6*5 allele could be a dominant allele but its effect is dependent on the existence of other alleles. The story for CYP2D6*1/*4 is

Table 6.7 Genotype frequencies for CYP2D6 and CYP2C19

Genotype	Age group					
	95	100	101-104	105+	Total	Control
CYP2D6						
*1/*1	18	93	14	10	135	199
*1/*5	1	12	5	2	20	0
*1/*3	2	6	1	0	9	8
*1/*4	7	34	3	4	48	93
*3/*4	0	3	0	0	3	5
*4/*4	2	9	4	1	16	18
*4/*5	1	2	0	1	4	0
*5/*5	1	0	0	0	1	2
Total	32	159	27	18	236	325
CYP2C19						
*1/*1	23	118	20	10	171	43
*1/*2	10	35	7	7	59	19
*2/*2	1	6	1	1	9	2
Total	34	159	28	18	239	64

Table revised from L. Bathum et al, Eur J Clin Pharmacol (1998) 54:427

Table 6.8 Estimated risks by genotype

Genotype	Frequency	SE	Risk	SE	p-value
CYP2D6					
*1/*1	0.612	0.027	1.033	0.033	0.319
***1/*5**	**0.004**	**0.003**	**0.455**	**0.116**	**0.000**
*1/*3	0.027	0.009	0.951	0.091	0.585
***1/*4**	**0.286**	**0.025**	**1.089**	**0.039**	**0.023**
*3/*4	0.016	0.007	1.042	0.139	0.763
*4/*4	0.054	0.012	0.949	0.065	0.432
***4/*5**	**0.001**	**0.002**	**0.550**	**0.229**	**0.050**
*5/*5	0.007	0.005	1.209	0.286	0.466
CYP2C19					
*1/*1	0.693	0.053	0.984	0.053	0.759
*1/*2	0.279	0.052	1.027	0.056	0.626
*2/*2	0.029	0.019	0.946	0.137	0.690

Table 6.9 Estimated risk for each single allele of CYP2D6

	CYP2D6-1	CYP2D6-4	CYP2D6-5
Freq.	0.923	0.355	0.011
SE	0.014	0.026	0.005
Risk	1.060	1.047	**0.575**
SE	0.058	0.035	**0.086**
p-value	0.303	0.175	**0.000**

different. Although the two alleles are neutral, a combination of these two alleles results in a significantly increased risk of death for carriers possibly due to allele-allele interaction. As reviewed in section 6.2.4, the results are supported by previous studies (Benhamou et al. 1996; Stucker et al. 1995; Tsuneoka et al. 1996), which associated the PM phenotype with low incidence of lung cancer among smokers.

The above applications of the model to empirical data have illustrated how the model can be elaborated to deal with data on the frequency of gene alleles (CVD genes, ApoB genes, P450 genes), genotypes (ApoB genes, P450 genes), genotype pools (ApoB genes) and allele classes (ApoB genes) together with the estimation of gene-sex interactions (CVD genes, ApoB genes). In the next section, the model will be modified to take into account individual heterogeneity in unobserved frailty that influence survival.

6.3.4 Introducing heterogeneity

Estimations for the above data are based on the assumption that individuals carrying each genotype are homogeneous in terms of unobserved hidden frailty. However, as we know, individuals are different with regards to factors that potentially contribute to life span. As already revealed by simulation study in section 5.5, the ignorance of heterogeneity in the estimation may lead to conservative inferences on the risk parameters by which the effects of genes are underestimated. When we assume that unobserved frailty follows gamma distribution, the model discussed in the sections 1.1.3, 5.2.1 and 5.5 can be introduced to deal with the influence of heterogeneity.

In order to get a better assessment of the variance of heterogeneity, σ^2, the three data sets are put together with variance of heterogeneity for all alleles or genotypes assumed to be the same since they are sampled from the same population and the surveys were conducted at almost the same time period (1994-1996). Under this assumption, there is only one σ^2 to be estimated which is a feasible way to carry out the analysis. However, in reality the variance for each genotype may differ. Nevertheless, when effects of the genes in question are not extreme, these variances may not differ from each other dramatically. The single-σ^2 assumption reduces the number of parameters to be estimated and makes calculation easier and more stable

especially when the data size is small as is the case for the Danish data. The log likelihood function for the combined estimation is

$$\begin{aligned}\log L(\sigma^2, R_{CVD}, p_{CVD}, R_{ApoB}, p_{ApoB}, R_{P450}, p_{P450}) &= \log L_1(\sigma^2, R_{CVD}, p_{CVD}) \\ &+ \log L_2(\sigma^2, R_{ApoB}, p_{ApoB}) \\ &+ \log L_3(\sigma^2, R_{P450}, p_{P450}).\end{aligned} \tag{6.4}$$

$R_{CVD}, p_{CVD}, R_{p450}, p_{p450}, R_{ApoB}, p_{ApoB}$ are vectors of risks and frequencies of CVD, ApoB and P450 genes. The genetic parameters (risks and frequencies) are estimated separately while maintaining the same σ^2 in L_1 for CVD data, L_2 for ApoB data and L_3 for P450 data. As shown in section 5.5, estimation for σ^2 is problematic when

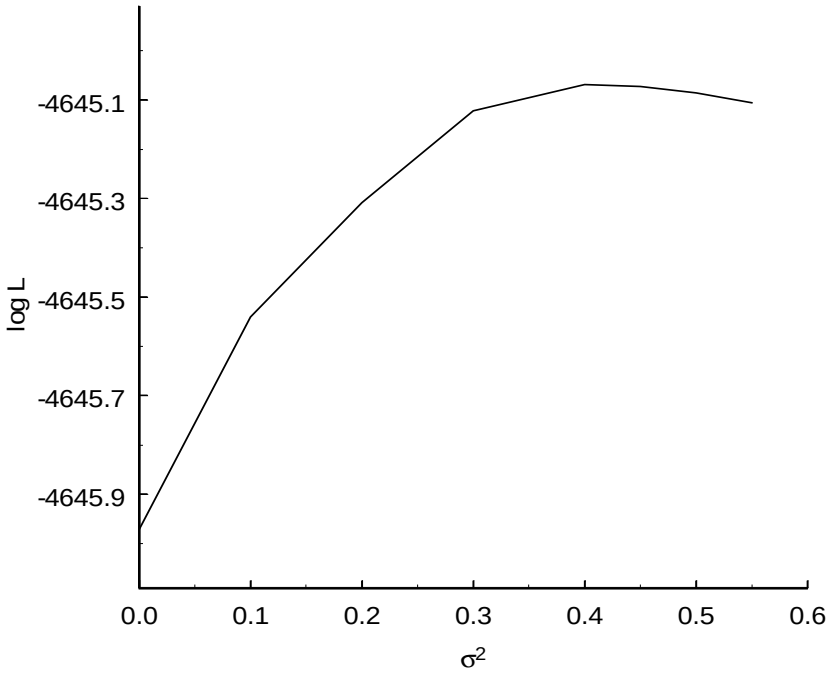

Figure 6.2 The log likelihood plotted against variance of hidden frailty for Danish data

data size is small. A feasible way to get a reasonable estimate for it is to try to fix σ^2 to different values when doing the estimation for genetic parameters and then compare the likelihood values. The point where the highest likelihood is achieved is the proper σ^2. In Figure 6.2, the different values of log likelihood are plotted against

corresponding σ^2. The likelihood reaches its peak (-4645.0689) when σ^2 is about 0.4. Tables 6.10, 6.11 and 6.12 are the new estimates for gene alleles in the CVD data, allele classes in the ApoB data and gene alleles in P450 data. Comparing the new results with estimates in Tables 6.3, 6.6 and 6.9, we see that the estimated effects on significant genes are stronger when taking into account unobserved hidden heterogeneity than these are when heterogeneity is ignored, although there is very little change in the estimated initial frequencies. Risks of gene-sex interaction for angiotensenogen gene allele M, F7RQ353-Q, F7323ins-p10, are 0.67 (p_{-value} =0.003), 0.682 (p_{-value} =0.013), 0.688 (p_{-value} =0.012). Figure 6.3 shows the sex-specific survivals for angiotensenogen M allele carriers. The male carriers have higher survival than the non-carriers. But survivals for female carriers and non-carriers are not different. ApoB allele class L becomes more detrimental with risk of 1.438 (Table 6.11) instead of 1.158 (Table 6.6) but its p-value increased to 0.083. Risk for a carrier of CYP2D6 allele *5 is now 0.233 ($p_{-value} \approx 0.000$) (Table 6.12) instead of 0.575 ($p_{-value} \approx 0.000$) (Table 6.9). The rare allele CYP2D6 *5 can reduce a carrier's hazard of death by about 1/4, a tremendous increase in their survival. Differences in the

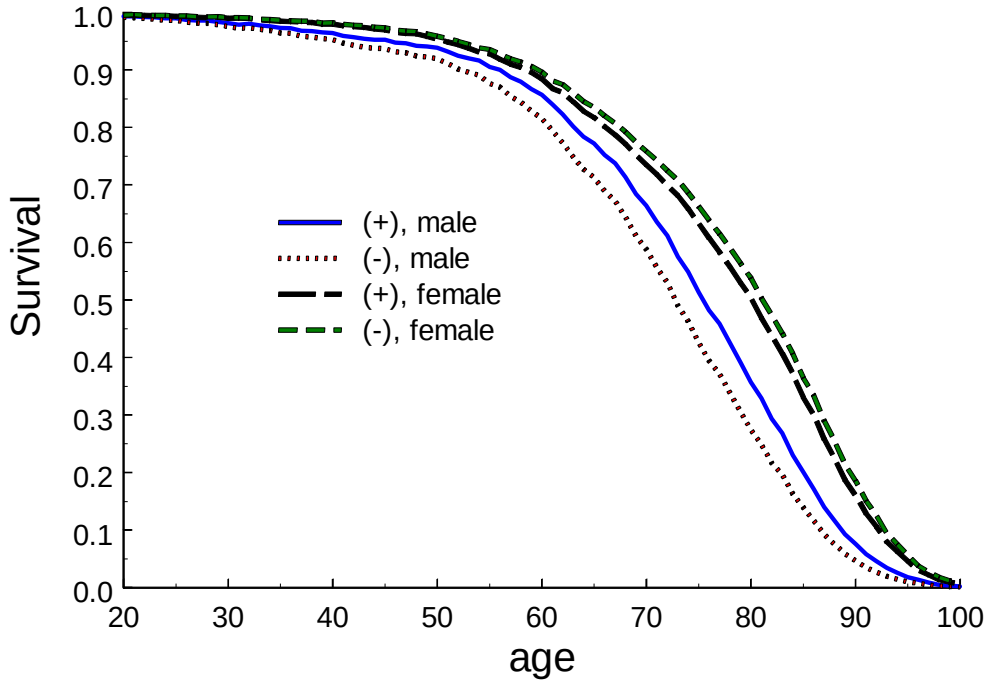

Figure 6.3 The sex-specific survivals for carriers and non-carriers of the angioten-M allele

parameter estimates demonstrate the importance of considering unobserved hidden heterogeneity in individual frailty when making inferences concerning the influences of genetic factors.

Table 6.10 Parameter estimates for the CVD associated genes with heterogeneity

Gene allele	Freq		Risk			Risk of gene-sex interaction		
	Est.	SE	Est.	SE	p-value	Est.	SE	p-value
MTHFR-A	0.905	0.015	1.146	0.190	0.442	0.835	0.216	0.443
MTHFR-V	0.524	0.026	0.989	0.098	0.910	0.953	0.145	0.746
Angioten-M	0.591	0.025	1.121	0.117	0.301	**0.670**	**0.112**	**0.003**
Angioten-T	0.894	0.016	1.096	0.178	0.591	0.912	0.218	0.687
ACE-I	0.747	0.022	1.047	0.123	0.701	0.922	0.161	0.627
ACE-D	0.768	0.022	0.950	0.120	0.673	1.128	0.197	0.517
F7RQ353R	0.990	0.005	0.856	0.494	0.771	1.654	1.062	0.538
F7RQ353Q	0.195	0.020	1.245	0.176	0.164	**0.682**	**0.128**	**0.013**
t-PA-I	0.846	0.018	1.154	0.161	0.339	0.994	0.220	0.977
t-PA-D	0.625	0.025	1.043	0.112	0.699	0.926	0.146	0.615
PAI-675-4G	0.802	0.020	1.159	0.143	0.265	0.963	0.168	0.825
PAI-675-5G	0.725	0.023	1.018	0.119	0.882	1.042	0.172	0.807
GP2b3a-L	0.980	0.007	1.191	0.401	0.633	0.886	0.436	0.794
GP2b3a-P	0.293	0.023	1.022	0.115	0.845	0.955	0.156	0.770
F7323ins-p0	0.985	0.006	0.985	0.393	0.969	1.205	0.655	0.754
F7323ins-p10	0.220	0.021	1.259	0.171	0.129	**0.688**	**0.125**	**0.012**
F7intron7-H6	0.901	0.015	1.097	0.183	0.595	1.150	0.250	0.549
F7intron7-H7	0.473	0.026	0.999	0.106	0.992	0.845	0.131	0.239
Fabrinogen-G	0.978	0.008	0.993	0.372	0.984	0.219	0.806	0.333
Fabrinogen-A	0.341	0.024	1.163	0.131	0.215	0.810	0.129	0.140

However, as discussed in section 5.5, estimating the heterogeneity parameter in the model is problematic when data size is small as is the case for the Danish data. In Figure 6.2, there isn't enough evidence to support adopting a heterogeneity model. Although the likelihood is increased when unobserved heterogeneity is considered, the increase does not reach the significant level as demanded by a likelihood ratio test

with one degree of freedom. We will demonstrate again how consideration of heterogeneity can increase the likelihood with a relatively large sample size in section 6.4.

Table 6.11 Estimates on allele class frequency, risk and gene-sex interaction with heterogeneity

ApoB allele class	Freq.		Risk			Risk of gene-sex interaction		
	Est.	SE	Est.	SE	p-value	Est.	SE	p-value
S	0.208	0.040	0.882	0.110	0.284	0.881	0.151	0.430
M	0.915	0.028	0.988	0.166	0.940	1.280	0.300	0.351
L	0.337	0.047	0.886	0.097	0.241	1.438	0.252	0.083

Table 6.12 Estimated risk for each single allele of CYP2D6 with heterogeneity

	CYP2D6-1	CYP2D6-4	CYP2D6-5
Freq.	0.923	0.357	0.007
SE	0.015	0.027	0.005
Risk	1.151	1.126	**0.233**
SE	0.160	0.095	**0.086**
p-value	0.344	0.184	**0.000**

6.3.5 Conclusions

The applications of the model to the genetic data from Danish studies have illustrated how the model can work under different situations for the estimation of frequency and risk of gene alleles, genotypes and allele classes. In all cases, the gene-sex interaction has been incorporated into the model and thus produced different results for CVD associated genes (angioten-M, F7RQ353-Q, F7323ins-p10) that were not detected in previous studies on the same data by using conventional statistical methods (Bladbjerg et al. 1999). The sex-dependent influence indicates that the effect of a gene on multifactorial trait depends on the physiological background in which the gene is expressed. Therefore, if the age-related physiological scenario changes in

males and females differently, the effect of a certain gene on survival could vary between the sexes. By comparing risk estimates for a single gene allele with that for the corresponding genotypes, it is possible to find gene-gene interaction (M/L genotype for ApoB allele classes and CYP2D6*1/*4) as well as the dominant effect of certain gene allele (CYP2D6*5). In addition to the estimates on risk parameters, the model also provides estimates on gene frequency at initial age. One must notice that the frequency estimates in the above applications are not gene allele frequency but rather the frequency or proportion of carriers of the allele (including both the heterozygous and homozygous genotypes) or genotype. However, corresponding allele frequencies can be calculated since the estimated proportions include individuals carrying one or two copies of the gene allele. Assume p' is the allele frequency. The proportion of carriers of the allele in the population is $p'^2 + 2p'(1-p') = 1-(1-p')^2 = p$ which equals the estimated proportion p in the model. With this relationship, we can calculate allele frequency as $p' = 1-\sqrt{1-p}$. Taking angioten-M allele in Table 6.3 for example, the allele frequency is $p' = 1-\sqrt{1-0.592} = 0.361$. In the same way, allele frequency for angioten-T is calculated as 0.674. Since there are only two polymorphic forms at the locus, frequencies of the two alleles should sum up to about 1 considering the variations in the estimation.

The polygenic feature of life span makes it imperative to take into account the unobserved hidden heterogeneity in individual frailty for two considerations. First there is unobserved heterogeneity in an individual's genetic make-up that accounts for 25% of the variation in life span according to previous studies (McGue et al.1993; Herskind et al. 1996; Yashin & Iachine 1995). Regarding the number of genetic variations that contribute to life span, there could be up to about 7,000 as estimated by Martin (1997). In such a highly polygenic situation, it is recommended that the influences from other unobserved genes be considered when one is making an inference from the observed ones. Second, there is a predominant influence on life span from non-genetic factors that constitute another layer of heterogeneity in individual frailty. As revealed in section 2.2, the existence of genetic and non-genetic heterogeneity in individual frailty affecting life span can influence the inference on an

observed genetic polymorphism. The introduction of a gamma-distributed frailty in the analysis is a convenient and helpful way to deal with the problem.

One concern is that adjustments of significance level for the statistical tests might be needed since we are doing multiple comparisons. However, our interests here are in each gene allele separately with tests on multiple hypotheses instead of doing multiple tests on one single hypothesis when such adjustment is called for (De Benedictis et al. 1999; Weir 1996; Rothman 1990). If one is interested in making an overall conclusion on a single locus with multi-allele, then each test on one allele can be treated as one repeat that contributes to the final result on the hypothesis on the locus. In this case, adjustment is required because existence of any significant allele will result in a positive conclusion. One could argue that the large number of tests conducted could definitely increase the probability that one might come up with a positive p-value just by chance. Take Table 6.3 for example, there are a total of 20 tests as a whole but we only observe 3 alleles with significant gene-sex interactions. With consideration of multiple comparison, we can adjust our significant level to $1-(1-0.05)^{1/20}=0.003$ if the original type I error is set to 0.05. According to this criterion, we could conclude that no significant effects exist in Table 6.3. However, one has to keep in mind that while we are insured to be on the safer side, there is also an error that can occur when an association in the data is not the result of chance. The 3 alleles in Table 6.3 could be examples for this argument since as discussed in section 6.3.1, there had been obvious clinical evidences that indicate the sex-dependent effects of the 3 alleles.

6.4 Application to data from the Italian centenarian study

6.4.1 The Italian centenarian data

In order to examine the association between genetic variation and longevity, a multi-centric longevity study was started in 1995 in Italy. Genetic information was collected from individuals in two groups: centenarians, and a younger group of individuals aged 7 to 84 years. Individuals are recorded by sex and region (southern or northern Italy). The distribution of participants by sex and region is shown in Table 6.13 and distribution of participants in the control group by sex and age is shown in Figure 6.4. The centenarian group consists of people who had reached age 100 or

Table 6.13 Observations by sex and area

Group	Male	Female	Total
Young			
South	311	302	613
North	54	82	136
Total	365	384	749
Centenarian			
South	36	67	103
North	26	83	109
Total	62	150	212

older at the time when blood samples were taken. The oldest in this group was a 109-year-old woman. All participants were clinically healthy people. The number of males and females in the control (i.e. the younger) group are well balanced, but this is not the case for the centenarian group, where there are more than twice as many females as

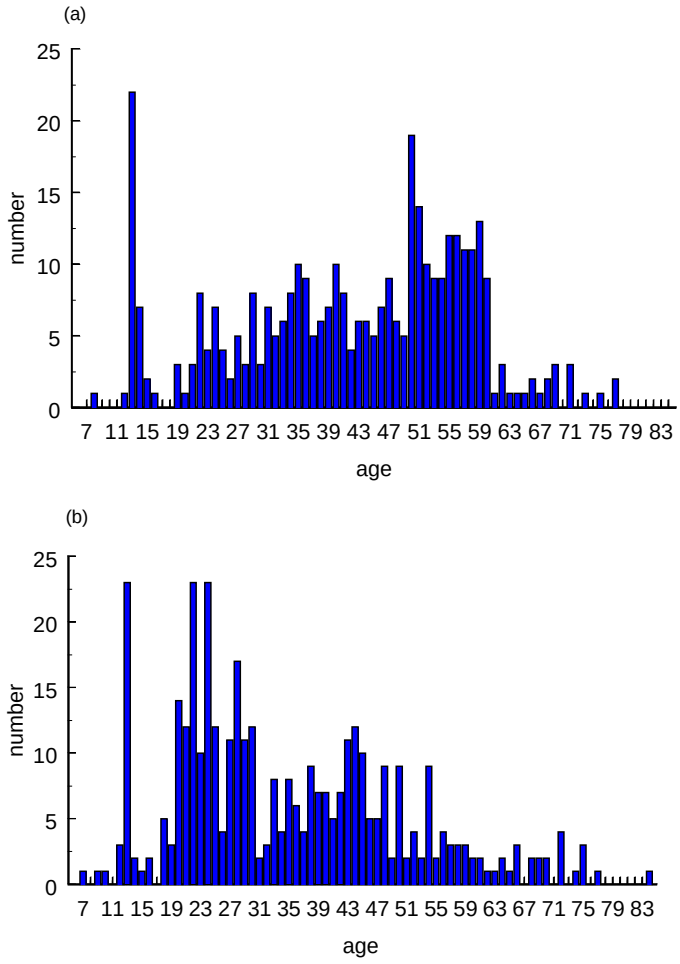

Figure 6.4 Frequency distribution by age for male (a) and female (b) participants in the control group

males. In the younger group, there are more people at younger ages from the south. However, since the new model can control for confounding factors like area and sex, this data structure would not be problematic in the analysis.

Table 6.14 Genes and markers analyzed

LOCUS	Biological role	Chromosome	Marker	Alleles	Number of individuals*
APOB	Major protein of LDL	2p24-23	3'APOB-VNTR	23,26,31,33,34, 35,36,37,39,41, 43,45,47,49,51, 53,55	787
REN	Angiotensin II synthesis	1q32	HUMREN4 (STR)	7,8,10,11,12	375
THO	Catecholamine synthesis	11p.15	HUMTHO1 (STR)	6,7,8,9, 10^{-1} ,10	555
SOD1	Oxygen free radicals scavenging	21q22.1	D21S223 (STR)	1,2,3,4,5,6,7,8, 10	386
SOD2	Oxygen free radicals scavenging	6q25	(T/C)401nt	T,C	354
APOA1	Major protein of HDL. Activator of LCAT	11q13-qter	RFLP-MspI (-78nt)	+,-	328
APOC3	Chylomicrons and VLDL	11q13	RFLP-SstI (3'ter)	+,-	328
APOA4	Newly secreted chylomicrons	11q13	RFLP-HincII (ex3)	+,-	328
INS	Codes insulin	11p.15	RFLP-FokI (1428nt)	+,-	438
PARP	DNA repair	1q41-42	STR (ex1)	83,85,87,89,93, 95,97,99	315
Haplo-group	Oxidative phosphorylation	mtDNA	Associated RFLPs	H,I,J,K,T,U,V, W,X, Others	547
D-Loop	Oxidative phosphorylation	mtDNA	STR	132,134,136, 138,140	393

**for whom information on both gene typization and age at participation was available*

The eleven autosomal genes and the mitochondrial DNA markers shown in Table 6.14 were examined (APOB: De Benedictis et al. 1998a; REN, THO, SOD2, PARP: De Benedictis et al. 1998b; MtDNA haplogroups: De Benedictis et al. 1999;

SOD1, APOA1, APOC3, APOA4, INS, D-Loop MtDNA: unpublished data). Due to the problem of missing values, the valid number of observations for each locus varies (Table 6.14).

6.4.2 Analytic strategy: incorporating interaction and confounding

In the data we observe, except the genetic covariates, an individual's sex and region. This coincides with the situation addressed in section 5.6 and Table 5.10. In the same manner, we can decompose the total population survival into different subgroups in respect to the confounding factor so that

$$\begin{aligned}\bar{s}_m(x) &= p_s p_{gs} s_{0,m}(x)^{RR_{area} R_{g\times a} R_{g\times s}} + p_s(1-p_{gs}) s_{0,m}(x)^{R_{area}} + (1-p_s) p_{gn} s_{0,m}(x)^{RR_{g\times s}} \\ &\quad + (1-p_s)(1-p_{gn}) s_{0,m}(x), \\ \bar{s}_f(x) &= p_s p_{gs} s_{0,f}(x)^{RR_{area} R_{g\times a}} + p_s(1-p_{gs}) s_{0,f}(x)^{R_{area}} + (1-p_s) p_{gn} s_{0,f}(x)^{R} \\ &\quad + (1-p_s)(1-p_{gn}) s_{0,f}(x).\end{aligned} \tag{6.5}$$

Similar to Table 5.10, p_s is the proportion of individuals from the south but $1-p_s$ from the north, p_{gs} is proportion of carriers in the south but p_{gn} in the north. R is the relative risk of having one gene allele or genotype in females. It is defined as the relative risk for carriers with reference to non-carriers. R_{area} is risk of the confounding factor area defined as the relative risk of being from the south versus from the north. $R_{g\times a}$ is the risk of gene-area interaction. $R_{g\times s}$ is the risk of gene-sex interaction. Note that in (6.5) the initial frequencies of carriers are specified for the south and north separately to allow for the distinction of regional differences (section 5.7). Similar to section 6.3, in case of no gene-sex interaction, we report the combined risks of genes for both sexes.

The left-hand side of (6.5) can be obtained from population statistics and applied to the Two-step MLE when one is interested in a non-parametric form for the baseline hazard, as discussed in section 5.2.3. The age-specific cohort survival derived from Italian cohort life tables in section 5.4 fits in the sample of this study since it was carried out around 1995. It has been demonstrated by simulation that such a synthetic survival distribution produces the best estimates for the parameters, although it could be replaced with period life table survival distributions of 1950-1970 in approaching

the true parameters. However, the two dips on both the male and the female curves in Figure 5.11 could under-represent the true survival distributions for the corresponding cohorts considering the reasons for the two events. With this in mind, a more reasonable choice would be the period life tables. According to the simulated results from section 5.4, the period life table of 1960 will be used for the data analysis, though periods 1950 and 1970 are also applicable.

6.4.3 Results

The model was first applied to each single allele at different loci to find candidate alleles that may have potential influence on individual survival. Irrelevant genes were selected out by testing their statistical significance for their relative risks (risk of the gene allele, risks of the gene–area, and gene-sex interactions). Since this is done for each gene allele separately, the estimate of the risk of area is different for different alleles due to missing values. Twelve alleles at 5 loci (APOB, THO, SOD2, INS, mtDNA, both haplogroups and D-loop markers) were selected from the data as showing potential influence on life span. We then put them together into one estimation, with the restriction that they had the same risk of area. Significant levels for risk parameters (relative risks for the genes and for interactions) were determined by testing the statistical differences between the estimated risks and 1, with the null hypothesis H_0: r =1. The probability of a Type I error is $\alpha = 0.05$. The results are shown in Table 6.15. There are 4 gene alleles (APOB39, THO10, mtDNAhapl-J, mtDNAhapl-U) with a potential influence on survival, with risks smaller than one. There are 3 gene alleles that have significant gene–environment interactions (APOB35, APOB39, SOD2-T). For carriers of APOB35 and 39, Southerners have higher risks than Northerners (Table 6.15). But for carriers of SOD2-T, Southerners have lower risk than Northerners. There is no allele with sex-specific influences although the p_{-value} of $r_{g\times s}$ for THO8 allele is 0.059. The overall risk of r_{area} is 1.133 (sd =0.012, $p_{-value} \approx 0.000$), which means that Southerners have a higher risk of death than Northerners. However, none of the 12 gene alleles will show any significant effect when the p-value is adjusted (1-(1-0.05)$^{1/36}$=0.001) according to the number of tests conducted in Table 6.15.

Table 6.15 Parameter estimates without heterogeneity

Genes	Frequency South		Frequency North		Risk of gene			Risk of g×a			Risk of g×s		
	est.	*SE*	*est.*	*SE*	*est.*	*SE*	p_{-value}	*est.*	*SE*	p_{-value}	*est.*	*SE*	p_{-value}
Apob35	0.402	0.017	0.356	0.017	0.918	0.043	0.055	**1.115**	**0.057**	**0.044**	0.997	0.054	0.953
Apob39	0.085	0.010	0.072	0.009	**0.830**	**0.074**	**0.022**	**1.319**	**0.148**	**0.031**	1.147	0.116	0.206
THO7	0.319	0.020	0.361	0.020	1.064	0.048	0.183	0.944	0.051	0.277	0.924	0.048	0.112
THO8	0.239	0.018	0.136	0.015	0.933	0.061	0.268	1.050	0.078	0.525	**0.900**	**0.053**	**0.059**
THO10	0.332	0.020	0.398	0.021	**0.916**	**0.041**	**0.040**	1.082	0.056	0.144	1.093	0.057	0.106
SOD2-T	0.829	0.020	0.801	0.021	0.998	0.051	0.970	**0.915**	**0.038**	**0.023**	1.029	0.074	0.695
INS-	0.985	0.006	0.964	0.009	1.093	0.108	0.392	0.944	0.034	0.094	0.815	0.143	0.195
INS+	0.261	0.021	0.350	0.023	0.953	0.046	0.308	1.049	0.064	0.449	1.106	0.064	0.097
mtDNAhapl-J	0.045	0.009	0.050	0.009	**0.797**	**0.084**	**0.015**	1.182	0.159	0.254	0.989	0.084	0.895
mtDNAhapl-U	0.138	0.015	0.226	0.018	**1.149**	**0.065**	**0.022**	0.873	0.066	0.057	1.041	0.073	0.574
mtDNAstr-136	0.015	0.006	0.060	0.012	0.960	0.098	0.684	0.710	0.148	0.051	1.044	0.115	0.701
mtDNAstr-138	0.034	0.009	0.014	0.006	0.707	0.162	0.070	1.371	0.355	0.296	1.068	0.145	0.639

Table 6.16 Parameter estimates with heterogeneity

Genes	Frequency South		Frequency North		Risk of gene			Risk of g×a			Risk of g×s		
	est.	*SE*	*est.*	*SE*	*est.*	*SE*	p_{-value}	*est.*	*SE*	p_{-value}	*est.*	*SE*	p_{-value}
Apob35	0.405	0.017	0.353	0.017	**0.798**	**0.095**	**0.033**	1.352	0.187	0.060	1.014	0.166	0.932
Apob39	0.087	0.010	0.068	0.009	**0.603**	**0.127**	**0.002**	2.199	0.669	0.073	1.589	0.475	0.215
THO7	0.323	0.020	0.367	0.020	1.202	0.153	0.188	0.866	0.138	0.331	0.838	0.129	0.207
THO8	0.238	0.018	0.141	0.015	0.867	0.148	0.369	1.083	0.219	0.705	**0.719**	**0.118**	**0.018**
THO10	0.327	0.020	0.389	0.021	**0.775**	**0.088**	**0.011**	1.222	0.175	0.204	1.252	0.184	0.171
SOD2-T	0.829	0.020	0.798	0.021	0.977	0.140	0.871	**0.790**	**0.093**	**0.024**	1.096	0.227	0.673
INS-	0.986	0.006	0.967	0.009	1.342	0.376	0.363	0.858	0.086	0.098	0.650	0.317	0.269
INS+	0.259	0.021	0.343	0.023	0.860	0.111	0.204	1.147	0.199	0.459	1.320	0.220	0.146
mtDNAhapl-J	0.044	0.009	0.049	0.009	**0.561**	**0.140**	**0.002**	1.484	0.519	0.351	0.904	0.199	0.628
mtDNAhapl-U	0.135	0.015	0.231	0.018	1.482	0.257	0.061	**0.660**	**0.151**	**0.024**	1.149	0.241	0.537
mtDNAstr-136	0.015	0.006	0.057	0.012	0.863	0.236	0.561	**0.445**	**0.221**	**0.012**	0.989	0.172	0.947
mtDNAstr-138	0.034	0.009	0.012	0.005	**0.395**	**0.212**	**0.004**	2.339	1.512	0.376	1.116	0.373	0.756

In another estimation, we took into account unobserved individual heterogeneity. By introducing different variances of hidden frailty σ^2, we arrived at different values of the likelihood function. The highest likelihood was reached when σ^2 is around 0.42 (Figure 6.5) and when the best fit to the data is obtained (Table 6.16). Among the major changes, the p_{-value} for APOB35 changed from 0.055 to 0.033, the p_{-value} for mtDNAstr-138 from 0.070 to 0.004. On the contrary, the significant effect for mtDNAhapl-U in Table 6.15 disappeared in Table 6.16. The effect of gene– area interactions for APOB35 and 39 become less significant although the risk is higher than in Table 6.15. Meanwhile, the risks of gene-environment interaction for mtDNAhapl-U and mtDNAstr-136 become significant with p_{-value} = 0.024 and $p_{-value} = 0.012$ respectively. In the heterogeneity model, THO8 allele shows a strong sex-dependent influence on survival which reduces the hazard of death for males but not for females ($p_{-value} = 0.018$). The estimates of relative risks in Table 6.16 are all higher when individual heterogeneity is considered. This indicates that if one does not consider heterogeneity, the effect associated with a given gene allele can be systematically underestimated. In Figure 6.6 we present the hazard functions

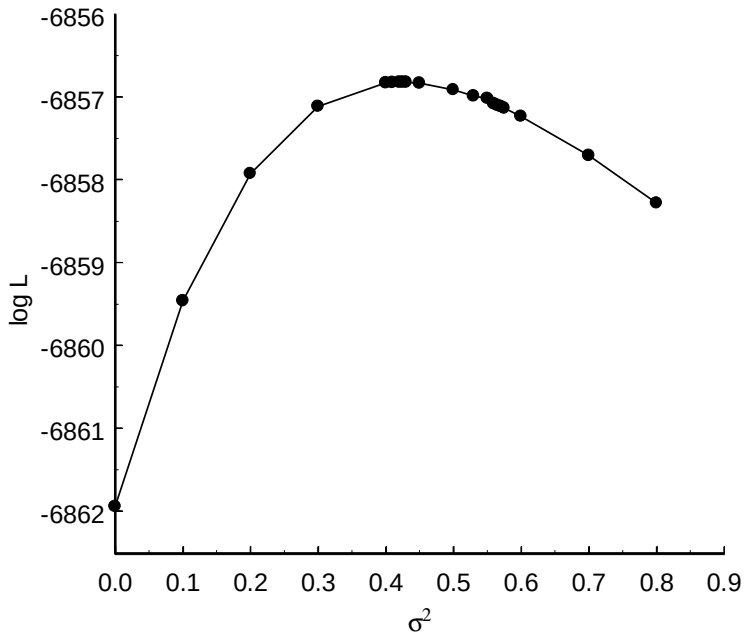

Figure 6.5 The log likelihood plotted against variance of hidden frailty for Italian data

forfemale Northerners with and without the APOB39 allele. The risk of death is substantially reduced when APOB39 is present.

As concerns allele-area interaction, SOD2-T and mtDNAstr-136 have beneficial effects for Southerners although they are neutral genes for Northerners. In Figure 6.7 we plot the mortality curves for mtDNAstr-136 carriers in the South and the North for the two sexes. The risk of death is dramatically lower for Southerners than for Northerners as a result of gene–environment interaction. In the model that considers unobserved heterogeneity, the estimated r_{area} increases from 1.133 in the model without heterogeneity to 1.438 (sd =0.045, $p_{-value} \approx 0.000$).

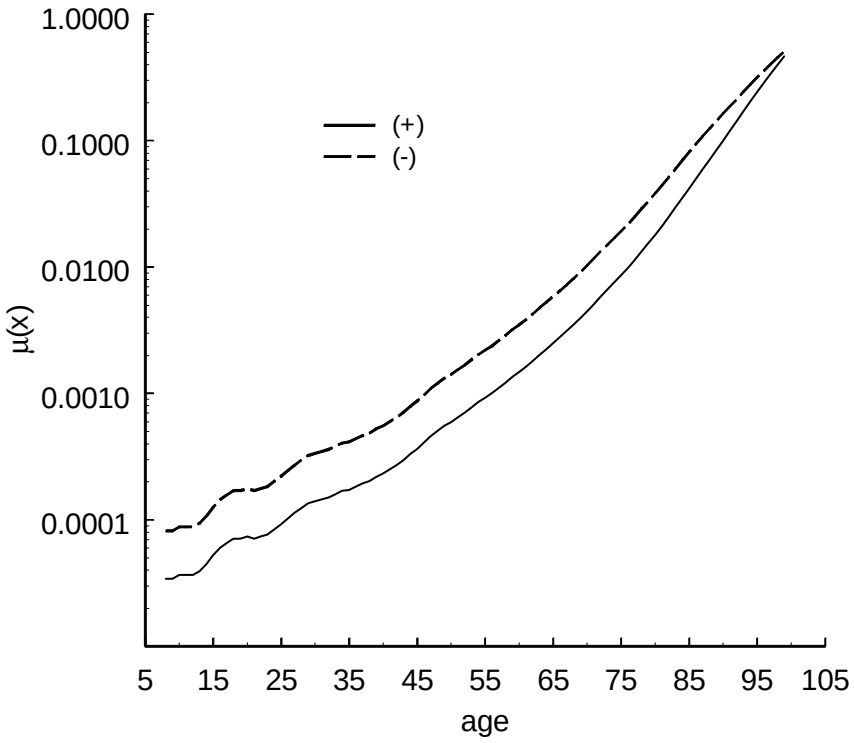

Figure 6.6 Estimated hazard functions for female northerners with (+) and without (-) Apob39 gene in log scale

THO8 is the only gene that shows a sex-specific influence. The gene is neutral in females but it reduces the risk of death ($R_{g\times s} = 0.719$) for males. The mortality curves for Southerners are plotted in Figure 6.8 for both males and females. But only males exhibit a difference between carriers and non-carriers of this allele. In Figure

6.8, the female mortality curves overtake those of males at later ages. The necessity of introducing male and female survival functions in the model is obvious.

While some genes may manifest gene–environment interaction, there are others that exhibit different initial frequencies in different geographic regions. Frequencies for THO8 and mtDNAstr-138 are significantly higher in Southern than in Northern Italy, while the frequencies of INS+, mtDNAhapl-U, mtDNAstr-136 are higher in the North (Tables 6.15, 6.16). Differences in gene frequencies by area are not unexpected, and they may be due to the differing genetic origins of the Southern and Northern Italian populations (Cavalli Sforza et al. 1994). Allowing for different gene frequency by region helps to avoid any possible bias in the estimation of risks, as demonstrated in section 5.7.

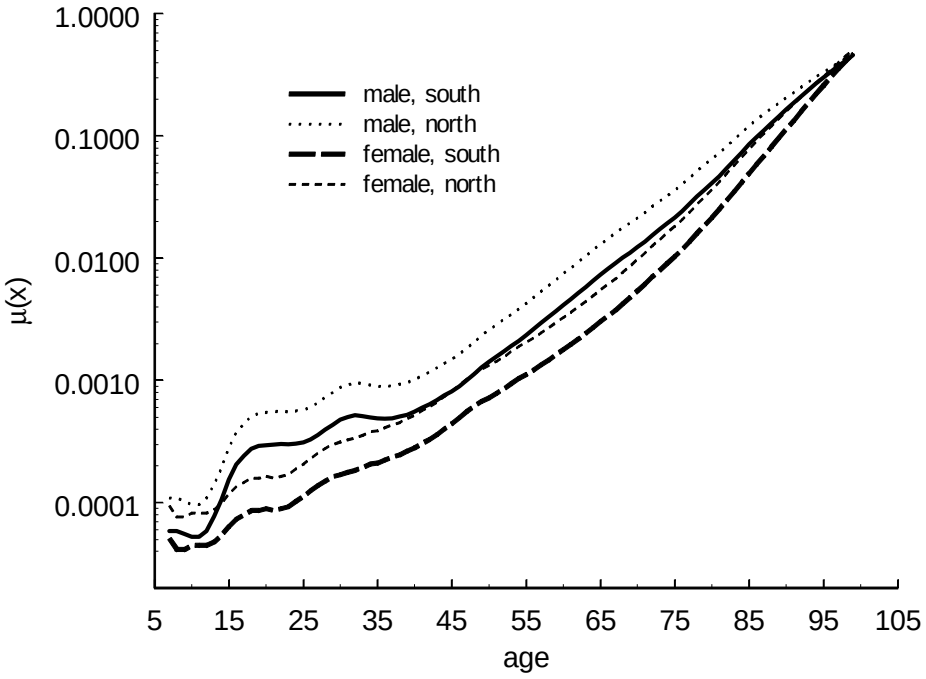

Figure 6.7 Estimated hazard functions for mtDNAstr-136 male and female carriers by area. Southerners have a lower riskof death than northerners for both sexes

6.4.4 Conclusions

This application shows the feasibility of analyzing genotype data in combination with demographic information in order to estimate the relative risk

associated with both a gene itself and a gene–environment interaction, as well as the sex-specific effect on survival. The inclusion of gene–environment interactions is crucial for the following reasons. First, as the results show, gene–environment interactions exist as a common phenomenon in modulating a complex trait such as life span, where environment has an important role to play. Thus the study of gene–environment interaction is an important aspect of genetic research on longevity. Second, ignoring these interactions can result in an incorrect assessment of allele effects. If a gene is beneficial in the south but neutral in the north, for example, it could be assessed as a universally beneficial gene if its interaction with geographic area is ignored (section 5.7).

Besides gene–environment interaction, the environment can itself act as a confounding factor that influences the evaluation of genetic effect (Sellers et al.

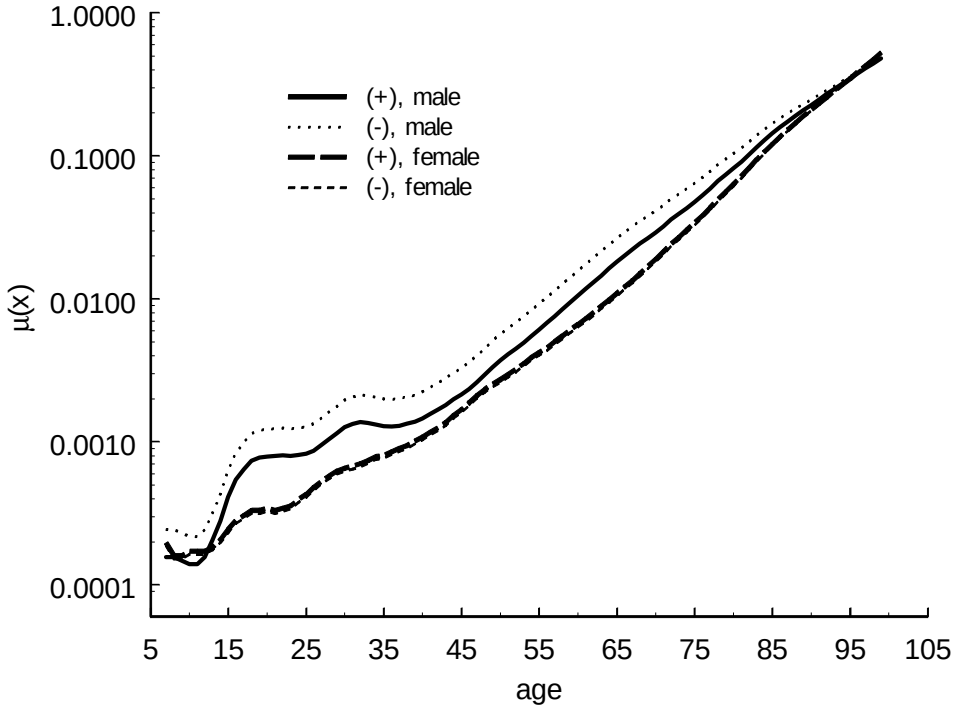

Figure 6.8 Estimated hazards for southerners with (+) and without (-) THO8 for the two sexes. While the death rate does not differ between female carriers and non-carriers, males with the gene have a lower riskof death than those without it

1998). Only when its interference is properly controlled, can the genetic and interactive terms be measured correctly (section 5.7). Sex, as another confounding factor, is successfully avoided when male and female survival functions available from population life table are introduced and the interaction of sex integrated. With this strategy, it is possible to include other confounding factors when necessary, and it is possible to extend this model to explore gene–gene interaction as well. Life span as a complex trait is a polygenic phenotype that involves the co-effect of multiple genes (Vaupel & Tan 1998; Martin 1997). It will be interesting and necessary to discover whether the genes function together, independently or dependently, and if there is any dependency, then the biological significance can be ascertained. Given the fact that there is usually a considerable amount of polymorphism at each locus, a better strategy is to combine the simple gene frequency method with the new approach. In this way, possible interactions can be screened by simply comparing frequencies among different age groups and then examining them carefully and in more detail afterwards by applying new methods.

As a consequence of introducing heterogeneity in the model, hazard functions for different sub-populations merely converge and do not cross – which may not necessarily be the case. On the other hand, the convergence phenomenon (Figures 6.5, 6.6, 6.7) raises an important question regarding the influence of genes on survival at very old ages. It seems that a certain gene becomes unimportant, as the hazards of death for populations with and without it converge. As we know, the risk associated with the gene, as it is assumed, does not change with age at the individual level since we are using a proportional hazard model as described by Cox (1972). The convergence is almost certainly due to unobserved heterogeneity, which compensates for the genetic effect as selection goes on with increasing age in a heterogeneous population of the same genotype (Vaupel & Yashin 1985).

The fact that significant genes were discovered in the present study is not surprising since the candidate genes selected play central roles in crucial metabolic pathways. The APOB gene variations could affect the efficiency in cholesterol metabolism and thus associate with individual's susceptibility to coronary artery disease (Hegele et al. 1986; Myant et al. 1989; Paulweber et al. 1989, 1990; Kervinen et al. 1994) and survival. The significant effects of THO and INS alleles could be relevant to the complex relationship existing between insulin and catecholamins

(Natali et al. 1998) in glucose metabolism, whose regulation in turn affects life span from yeast (Jiang et al. 2000) to humans (Paolisso et al. 1996). The beneficial effect of SOD2-T allele could support the finding that SOD2 polymophisms affect the efficiency of mitochondrial transport (Shimoda-Matsubayashi et al. 1996). Lastly, the biological background for the association between mtDNA variation and longevity is probably relevant to mtDNA haplogroup-specific oxidative phosphorylation efficiency (Ruiz-Pesini et al. 2000).

The application of this model on data collected from genetic studies on aging and longevity should help to detect additional relevant genes that contribute to the process of aging both by prolonging or shortening an individual's life span.

6.5 Summary

This chapter introduces the binomial frailty model to empirical gene marker data from Danish and Italian centenarian studies to estimate relative risks on survival of the observed gene alleles or genotypes. The most important results include

(1) Introducing the binomial frailty model with heterogeneity of unobserved individual frailty can substantially improve the likelihood of the estimation. Ignorance of heterogeneity can lead to systematical underestimation of the parameters and thus result in conservative conclusions.

(2) Three gene alleles from Danish data on CVD associated genes, angiotensinogen gene allele M, F7RQ353-Q, F7323ins-p10, have been found manifesting statistically important sex-dependent influences on survival which were not detected by conventional approach from a previous study.

(3) Allele *5 of the CYP2D6 gene from Danish centenarian study is a very strong beneficial allele that prolongs the carrier's life span.

(4) There are five genes, ApoB35, 39, THO10, mtDNAhapl-J and mtDNAstr138, which were detected as important to longevity in the Italian centenarian study.

(5) There are three genes, SOD2-T, mtDNAhapl-U and mtDNAstr136, in the Italian data exhibit gene-environment interaction but only THO8 gene allele is sex-dependent.

RESUMÉ

A great interest in studying genes and longevity has developed over the past few decades. Two kinds of data are generally being collected: data on related individuals including twin and genealogy data, and data on unrelated individuals but with information on gene markers. As new centenarian studies emerge, more data on individual genotypes will be available. Efficient and powerful statistical models that combine quantitative genetics and survival analysis are needed. This PhD project is aimed at developing new models for data analysis, replacing the conventional statistics used heretofore.

In this presentation, the binomial frailty model derived upon the Cox's proportional hazard assumption and the binomial distribution of gene alleles is introduced. Characterised by its direct incorporation of polygenic influence on individual survival, the model has been applied at the beginning to describe the genetic influence on life span from one observed gene or genotype given the existence of influences from other genetic and environmental heterogeneity. Some interesting insights on topics such as risk compensation are obtained. The model is extended to study family correlation on life span. Through simulation, correlation of life span and correlation of frailty between relatives are compared and the age-patterns of life span correlation and of frailty correlation are explored. The analysis indicates the following: 1. Behind the modest life span correlation there is a considerably strong correlation of frailty; 2. The life span correlation among related individuals decreases as age increases as a result of influences from environmental heterogeneity. The binomial frailty model is further engaged to estimate the number of longevity genes in human beings. The estimation has been done incorporating various assumptions on the genetic and heterogeneity parameters so that the results presented represent a range from normal to extreme situations. A comparison of the results obtained from Danish twin data, Quebec genealogy data and European noble family genealogy data showed relatively stable estimates from the model. The practice of applying the binomial frailty model to related individuals, serves as a bridge linking molecular

genetics with demography in aid of promoting better understanding of the mechanisms of aging and longevity.

A binomial frailty model for gene marker data is specified and this is applied to empirical data in the last part of the thesis. The model combines both the genetic and demographic information together for determining the relative risk of a gene allele or genotype and in estimating the corresponding frequencies. A Two-step MLE has also been introduced, to obtain a non-parametric form of the baseline hazard function. The model has been derived to incorporate gene-environment, gene-sex interactions as well as individual heterogeneity. Detailed studies have been done to examine the sensitivity of the model to the parameter values, data size and data structure. Various aspects of problems arising from sampling bias and confounding as well as problems from introducing period life table survival in the analysis are discussed. The model is then applied to Danish centenarian data to measure the influence on longevity from apolipoprotein genes, genes related to cardiovascular diseases and p450 genes; and to data from Italian centenarian studies used to show the effects of gene-environment and gene-sex interactions. The application of the model to data on cardiovascular disease associated genes and apolipoprotein B genes from the Danish centenarian study reveals genes that manifest significant influences on human life span: the conclusions are supported by previous clinical studies. Inferences on gene-sex and/or gene-environment interactions are supportable for genes examined in both the Danish and the Italian studies. A comparative study has shown remarkable influences from individual heterogeneity in unobserved frailty with risk of genes or genotypes underestimated when ignoring these differences. In addition, the likelihood of the estimation is substantially improved with application of the heterogeneity model.

RESUMÉ

I de seneste årtier er der opstået en tiltagende interesse for at udforske genetiske faktorers indflydelse på lang levetid. To slags data er indsamlet: 1) data over familiemedlemmer inklusiv tvillings- og genealogisk data, 2) data over personer, der ikke er i familie med hinanden, med information om genmarkers. Med ny forskning af hundredårige vil flere og flere individuelle genotype data blive tilgængelige. Effektive og stærke statistiske modeller, der kombinerer kvantitative genetiske analyser med survival analyse, er nødvendige. Nærværende ph.d.-projekt tilstræber at udvikle nye modeller til dataanalyse frem for at benytte konventionelle statistiske metoder.

I det foreliggende arbejde introduceres den binomiale skrøbelighedsmodel afledt af Cox's proportional hazard assumption og den binomiale fordeling af gen alleler. Karakteriseret af dens indirekte inddragelse af polygenetisk indflydelse på individuel overlevelse, anvendes modellen først til at undersøge den genetisk indflydelse af et observeret gen eller en observeret genotype på livslængden under antagelse af indflydelse fra anden genetisk eller miljømæssig heterogenitet. Interessante indblik i bl.a. risiko kompensation er fundet. Herefter er modellen udvidet til at analysere familiekorrelation af livslængde. Via simulation er korrelation af livslængde og korrelation af skrøbelighed blandt familiemedlemmer sammenlignet og aldersmønstret af livslængde-korrelation og skrøbelighedskorrelation undersøgt. Analyserne viser: 1. Ved en lav livslængde-korrelation er der en betragtelig høj korrelation af skrøbelighed; 2. Livslængde-korrelationen blandt familiemedlemmer falder med stigende alder som et resultat af indflydelser fra miljømæssig heterogenitet. Den binomiale skrøbelighedsmodel er yderligere benyttet til at estimere antallet af gener associeret med langt levetid hos mennesker. Estimatet er udført baseret på forskellige antagelser af de genetiske og heterogene parameter, således at resultaterne præsenterer både normale og ekstreme situationer. En sammenligning af resultaterne fra Det danske Tvillingsregister, Quebec genealogy data og European noble familiy genealogy data viser relativt stabile estimater for modellen. Formålet med at anvende den binomiale skrøbelighedsmodel på familiemedlemmer tjener som en måde at

forbinde den molekylære genetik med demografi for at opnå en bedre forståelse af mekanismerne af aldring og et langt liv.

En binomial skrøbelighedsmodel for DNA er specificeret og anvendt på empirisk data i den sidste del af afhandlingen. Modellen kombinerer både genetisk og demografisk information ved at bestemme den relative risiko for et gen allele eller en genotype og ved at estimere tilsvarende frekvens. Tillige bliver en EM algoritme introduceret for at opnå en non-parametisk estimering af baseline hazard funktionen. Modellen er udviklet for at inddrage gen-miljø, gen-køn interaktioner såvel som individuel heterogenitet. Detaljerede studier er udført for at undersøge sensitiviteten af modellen med hensyn til parameterværdierne, datamængde og datastruktur. Forskellige aspekter af problemer i form af sampling bias og confounding såvel som problemer ved introduktion af overlevelsesfunktionen fra en periode-overlevelsestavle i analysen bliver diskuteret. Modellen er derefter anvendt på data fra danske hundredårige for at undersøge indflydelser på lang levetid fra apolipoprotein gener, gener relateret til hjerte-kar-sygdomme og p450 gener samt på data fra italienske hundredårigestudie for at vise effekter af gen-miljø og gen-køn interaktioner. Modellens anvendelse på data over hjerte-kar-sygdomme associerede gener og apolipoprotein B gener fra danske hundredårige viser gener, som manifester en signifikant indflydelse på menneskets livslængde og konklusionen støttes af tidligere kliniske studier. Inferens af gen-køn og/eller gen-miljø interaktioner er baseret på gener undersøgt af både danske og italienske studier. Et sammenlignende studie har desuden vist bemærkelsesværdig indflydelse fra individuel heterogenitet i ikke-observeret skrøbelighed med undervurdering af risko eller indflydelse fra gener eller genotyper når disse forskelle blev ignoreret.

REFERENCES

Abbott MH, Murphy EA, Bolling DR , Abbey H, The familial component in longevity, A study of offspring of nonagenarians, II. Preliminary analysis of the completed study, *Hopkins Med. J.* **1974** 134:1-16.

Aburatani H, Matsumoto A, Itoh H, Yamada N, Murase T, Takaku F, Itakura H, A study of DNA polymorphism in the apolipoprotein B gene in a Japanese population, *Atherosclerosis* **1988** 72(1):71-6.

Agerholm-Larsen B, Nordestgaard BG, Steffensen R, Sorensen TI, Jensen G, Tybjaerg-Hansen A. ACE gene polymorphism: ischemic heart disease and longevity in 10,150 individuals. A case-referent and retrospective cohort study based on the Copenhagen City Heart Study. *Circulation* **1997** 95(10):2358-67.

Akisaka M, Suzuki M, Inoko H, Molecular genetic studies on DNA polymorphism of the HLA class II genes associated with human longevity, *Tissue Antigens* **1997** 50(5):489-93.

Alvarez V, Alvarez R, Lahoz CH, Martinez C, Pena J, Guisasola LM, Salas-Puig J, Moris G, Uria D, Menes BB, Ribacoba R, Vidal JA, Sanchez JM, Coto E, Association between an alpha(2) macroglobulin DNA polymorphism and late-onset Alzheimer's disease, *Biochem Biophys Res Commun* **1999** 264(1):48-50.

Apotech System, Gauss: Mathematical and statistical system. Vol. I: system and graphics mannual. **1996** Apotech system, Maple Valley, WA.

Barbeau A, Roy M, Cloutier T, Plasse L, Paris S, Environmental and genetic factors in the etiology of Parkinson's diseas, *Adv Neurol* **1987** 45:299-306.

Bathum L, Andersen-Ranberg K, Boldsen J, Brosen K, Jeune B, Genotypes for the cytochrome P450 enzymes CYP2D6 and CYP2C19 in human longevitY, Role of CYP2D6 and CYP2C19 in longevity, *Eur J Clin Pharmacol* **1998** 54(5):427-30.

Beeton M, Pearson K, Data for the problem of evolution in man, II, A first study of the inheritance of longevity and the selective death rate in man, *Proceedings of the Royal society of London* **1899** 65:290-305.

Bell AG, The Duration of Life and Conditions Associated with Longevity, A Study of the Hyde Genealog, **1918** Genealogical Records Office, Washington.

Benhamou S, Bouchardy C, Paoletti C, Dayer P, Effects of CYP2D6 activity and tobacco on larynx cancer risk, *Cancer Epidemiol Biomarkers Prev* **1996** 5(9):683-6.

Blackwelder WC, Mittal KK, McNamara PM, Payne FJ, Lack of association between HLA and age in an aging population, *Tissue Antigens* **1982** 20(3):188-92.

Bladbjerg EM, Andersen-Ranberg K, de Maat MP, Kristensen SR, Jeune B, Gram J, Jespersen J, Longevity is independent of common variations in genes associated with cardiovascular risk, *Thromb Haemost* **1999** 82(3):1100-5.

Bocquet-Appel JP, Jakobi L, Familial transmission of longevity, *Ann. Human Biol.* **1990** 17:81-95.

Bonafe M, Olivieri F, Mari D, Baggio G, Mattace R, Sansoni P, De Benedictis G, De Luca M, Bertolini S, Barbi C, Monti D, Franceschi C, p53 variants predisposing to cancer are present in healthy centenarians. *Am J Hum Genet* **1999a** 64(1):292-5.

Bonafe M, Olivieri F, Mari D, Baggio G, Mattace R, Berardelli M, Sansoni P, De Benedictis G, De Luca M, Marchegiani F, Cavallone L, Cardelli M, Giovagnetti S, Ferrucci L, Amadio L, Lisa R, Tucci MG, Troiano L, Pini G, Gueresi P, Morellini M, Sorbi S, Passeri G, Barbi C, Valensin S, et al. P53 codon 72 polymorphism and longevity: additional data on centenarians from continental Italy and Sardinia. *Am J Hum Genet* **1999b** 65(6):1782-5.

Boushey CJ, Beresford SA, Omenn GS, Motulsky AG, A quantitative assessment of plasma homocysteine as a risk factor for vascular disease. Probable benefits of increasing folic acid intakes. *JAMA* **1995** 274(13):1049-57.

Box G, Sciences and statistics, *Journal of the American Statistical Association* **1976** 71:791-802

Brattstrom L, Zhang Y, Hurtig M, Refsum H, Ostensson S, Fransson L, Jones K, Landgren F, Brudin L, Ueland PM, A common methylenetetrahydrofolate reductase gene mutation and longevity, *Atherosclerosis* **1998** 141(2):315-9.

Burkle A, Grube K, Kupper JH , Poly(ADP-ribosyl)ation: its role in inducible DNA amplification, and its correlation with the longevity of mammalian species, *Exp Clin Immunogenet* **1992** 9(4):230-40.

Burkle A, Muller M, Wolf I, Kupper JH , Poly(ADP-ribose) polymerase activity in intact or permeabilized leukocytes from mammalian species of different longevity, *Mol Cell Biochem* **1994** 138(1-2):85-90.

Burkle A, Poly(ADP-ribose) polymerase and aging, *Exp Gerontol* **1998** 33(6):519-23.

Carey JR, What demographers can learn from fruit fly actuarial models and biology, *Demography* **1997** 34(1):17-30.

Carmelli D, Intrapair comparisons of total life span in twins and pairs of sibs, *Hum. Biol.* **1982** 54: 525-537.

Cavalli Sforza LL, Menozzi P, Piazza A, The History and Geography of human genes, **1994** *New Jersey: Princeton University Press*, p 277-280.

Chan DK, Woo J, Ho SC, Pang CP, Law LK, Ng PW, Hung WT, Kwok T, Hui E, Orr K, Leung MF, Kay R , Genetic and environmental risk factors for Parkinson's disease in a Chinese population, *J Neurol Neurosurg Psychiatry* **1998** 65(5):781-4.

Chen D, Zhang M, Fan W, Shi H, Li Y, Chen Q, Zhang J, Gu X, A molecular variant of angiotensinogen gene is associated with myocardial infarction in Chinese, *Chung Hua I Hsueh I Chuan Hsueh Tsa Chih* **1998** 15(3):133-5.

Chen RZ, Pettersson U, Beard C, Jackson-Grusby L, Jaenisch R, DNA hypomethylation leads to elevated mutation rates. *Nature* **1998** 395(6697):89-93.

Christensen K, Vaupel JW, Determinants of longevity: genetic, environmental and medical factors, *Intern Med* **1996** 240:333-341.

Cong ND, Hamaguchi K, Saikawa T, Hara M, Sakata T, A polymorphism of angiotensinogen gene codon 174 and coronary artery disease in Japanese subjects, *Am J Med Sci* **1998** 316(5):339-44.

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance Mav , Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families, *Science* **1993** 261(5123):921-3.

Corral-Debrinski M, Shoffner JM, Lott MT, Wallace DC, Association of mitochondrial DNA damage with aging and coronary atherosclerotic heart disease, *Mutat Res* **1992** 275(3-6):169-80.

Cox DR, Regression models and life-tables, *J R Stat Sco B* **1972** 34:187-220.

Csaszar A, Kalman J, Szalai C, Janka Z, Romics L, Association of the apolipoprotein A-IV codon 360 mutation in patients with Alzheimer's disease. *Neurosci Lett* **1997** 230(3):151-4.

Cushman M, Meilahn EN, Psaty BM, Kuller LH, Dobs AS, Tracy RP. Hormone replacement therapy, inflammation, and hemostasis in elderly women. *Arterioscler Thromb Vasc Biol* **1999** 19(4):893-9.

Cutler RG, Evolution of human longevity: a critical overview, *Mechanisms of Aging and Development* **1979** 9:337-354.

Dahlback B, Inherited resistance to activated protein C, a major basis of venous thrombosis, is caused by deficient anticoagulant cofactor function of factor V. *Haematologica* **1995** 80(2 Suppl):102-13.

De Benedictis G, Falcone E, Rose G, Ruffolo R, Spadafora P, Baggio G, Bertolini S, Mari D, Mattace R, Monti D, Morellini M, Sansoni P, Franceschi C, DNA multiallelic

systems reveal gene/longevity associations not detected by diallelic systems: The APOB locus, *Hum Genet* **1997** 99:312-318.

De Benedictis G, Carotenuto L, Carrieri G, De Luca M, Falcone E, Rose G, Cavalcanti S, Corsonello F, Feraco E, Baggio G, Bertolini S, Mari D, Mattace R, Yashin AI, Bonafe M, Franceschi C, Gene/longevity association studies at four autosomal loci (REN, THO, PARP, SOD2), *Eur J Hum Genet* **1998**a 6:534-541.

De Benedictis G, Carotenuto L, Carrieri G, De Luca M, Falcone E, Rose G, Yashin AI, Bonafe M, Franceschi C, Age-related changes of the 3'APOB-VNTR genotype pool in ageing cohorts, *Ann Hum Genet* **1998**b 62:115-122.

De Benedictis G, C Franceschi, The genetics of successful aging, *Aging Clin. Exp. Res.* **1998** 10(2)147:148.

De Benedictis G, Rose G, Carrieri G, De Luca M, Falcone E, Passarino G, Bonafè M, Monti D, Baggio G, Bertolini S, Mari D, Mattace R, Franceschi C, Mitochondrial DNA inherited variants are associated with successful aging and longevity in humans, *The FASEB Journal* **1999** 13: 1532-1536.

De Benedictis G, Carrieri G, Garasto S, Rose G, Varcasia O, Bonafe M, Franceschi C, Jazwinski SM, Does a retrograde response in human aging and longevity exist? *Exp Gerontol* **2000** 35(6-7):795-801.

Del Panta L, Rettaroli R, Introduzione alla demografia storica. Editori Laterza, **1994** Roma, Bari.

Di Castelnuovo A, D'Orazio A, Amore C, Falanga A, Kluft C, Donati MB, Iacoviello L, Genetic modulation of coagulation factor VII plasma levels: contribution of different polymorphisms and gender-related effects. *Thromb Haemost* **1998** 80(4):592-7.

Dorak MT, Mills KI, Gaffney D, Wilson DW, Galbraith I, Henderson N, Burnett AK, Homozygous MHC genotypes and longevity, *Hum Hered* **1994** 44(5):271-8.

Falconer DS, Mackay, TFC, Introduction to Quantitative Genetics, **1996** London: Longman

Feng D, Tofler GH, Larson MG, O'Donnell CJ, Lipinska I, Schmitz C, Sutherland PA, Johnstone MT, Muller JE, D'Agostino RB, Levy D, Lindpaintner K. Factor VII gene polymorphism, factor VII levels, and prevalent cardiovascular disease: the Framingham Heart Study. *Arterioscler Thromb Vasc Biol* **2000** 20(2):593-600.

Fox AJ, Collier PF, Low mortality rates in industrial cohort studies due to selection for work and survival in the industry. *Br J Prev Soc Med* **1976** 30(4):225-30.

Franceschi C, Motta L, Valensin S, Rapisarda R, Franzone A, Berardelli M, Motta M, Monti D, Bonafe M, Ferrucci L, Deiana L, Pes GM, Carru C, Desole MS, Barbi C, Sartoni G, Gemelli C, Lescai F, Olivieri F, Marchegiani F, Cardelli M, Cavallone L, Gueresi

P, Cossarizza A, Troiano L, Pini G, Sansoni P, Passeri G, Lisa R, Spazzafumo L, Amadio L, Giunta S, Stecconi R, Morresi R, Viticchi C, Mattace R, De Benedictis G, Baggio G, Do men and women follow different trajectories to reach extreme longevity? Italian Multicenter Study on Centenarians. *Aging (Milano)* **2000** 12(2):77-84.

Frikke-Schmidt R, Nordestgaard BG, Agerholm-Larsen B, Schnohr P, Tybjaerg-Hansen A. Context-dependent and invariant associations between lipids, lipoproteins, and apolipoproteins and apolipoprotein E genotype. *J Lipid Res* **2000** 41(11):1812-22.

Frossard PM, Hill SH, Elshahat YI, Obineche EN, Bokhari AM, Lestringant GG, John A, Abdulle AM, Associations of angiotensinogen gene mutations with hypertension and myocardial infarction in a gulf population, *Clin Genet* **1998** 54(4):285-93.

Frosst P, Blom HJ, Milos R, Goyette P, Sheppard CA, Matthews RG, Boers GJ, den Heijer M, Kluijtmans LA, van den Heuvel LP, et al. A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. *Nat Genet* **1995** 10(1):111-3.

Galinsky D, Tysoe C, Brayne CE, Easton DF, Huppert FA, Dening TR, Paykel ES, Rubinsztein DC, Analysis of the apo E/apo C-I, angiotensin converting enzyme and methylenetetrahydrofolate reductase genes as candidates affecting human longevity, *Atherosclerosis* **1997** 129(2):177-83.

Gerdes LU, Gerdes C, Kervinen K, Savolainen M, Klausen IC, Hansen PS, Kesaniemi YA, Faergeman O, The apolipoprotein epsilon4 allele determines prognosis and the effect on prognosis of simvastatin in survivors of myocardial infarction : a substudy of the Scandinavian Simvastatin survival study, *Circulation* **2000** 101(12):1366-71.

Gerkins VR, Ting A, Menck HT, Casagrande JT, Terasaki PI, Pike MC, Henderson BE , HL-A heterozygosity as a genetic marker of long-term survival, *J Natl Cancer Inst* **1974** 52(6):1909-11.

Girelli D, Russo C, Ferraresi P, Olivieri O, Pinotti M, Friso S, Manzato F, Mazzucco A, Bernardi F, Corrocher R. Polymorphisms in the factor VII gene and the risk of myocardial infarction in patients with coronary artery disease. *N Engl J Med* **2000** 343(11):774-80.

Gong J-S, Zhang J, Yoneda M, Sahashi K, Miyajima H, Yamauchi K, Yagi K, Tanaka M, Mitochondrial Genotype Frequent in Centenarians Predisposes Resistance to Adult-Onset Diseases, *J. Clin. Biochm. Nutr.* **1998** 24:105-111.

Gonzalez FJ, Human cytochromes p450: problems and prospects, *Trends Pharmacol Sci* **1992** 13: 346-352.

Grube K, Burkle A, Poly(ADP-ribose) polymerase activity in mononuclear leukocytes of 13 mammalian species correlates with species-specific life span, *Proc Natl Acad Sci U S* A **1992** 89(24):11759-63.

Hamsten A, de Faire U, Walldius G, Dahlen G, Szamosi A, Landou C, Blomback M, Wiman B, Plasminogen activator inhibitor in plasma: risk factor for recurrent myocardial infarction. *Lancet* **1987** 2(8549):3-9.

Hansen HE, Sparck JV, Larsen SO, An examination of HLA frequencies in three age groups, *Tissue Antig*ens **1977** 10(1):49-55.

Harris J, Age differences in genetic and environmental influences for health from the Swedish adoption/twin study of aging, J. Gerontol. *Psychological science* **1992** 47: 213-220.

Harris JR, Lippman ME, Veronesi U, Willett W, Breast cancer (1), *N Engl J Med* **1992** 327:319-328.

Hastings N, Peacock JB, Statistical distributions, **1975** London: Butterworths, p90.

Hawkins MR, Murphy EA, Abbey H, The familial component of longevity. A study of the offspring of nonagenarians, I. Methods and preliminary report. Bull, *Johns Hopkins Hospital* **1965** 117:24-36.

Hayakawa K, Shimizu T, Ohba Y, Tomioka S, Takahasi S, Amano K, Yura A, Yokoyama Y, Hayakata Y, Intrapair differences of phisical aging and longevity in identical twins, *Acta Genet. Med. Gemellol*.**1992** 41: 177-185.

Hayflick L, How and why we age? **1994** New York: Ballantine Books.

Hazzard WR, Biological basis of the sex differential in longevity, *Journal of the American geriatrics society* **1986** 34: 455-471.

Hegele RA, Huang LS, Herbert PN, Blum CB, Buring JE, Hennekens CH, Breslow JL, Apolipoprotein B-gene DNA polymorphism associated with myocardial infarction, *N Engl J Med* **1986** 315(24):1509-15.

Hegele RA, Brunt JH, Connelly PW, A polymorphism of the angiotensinogen gene associated with variation in blood pressure in a genetic isolate. *Circulation* **1994** 90(5):2207-12.

Heijmans BT, Gussekloo J, Kluft C, Droog S, Lagaay AM, Knook DL, Westendorp RG, Slagboom EP, Mortality risk in men is associated with a common mutation in the methylene-tetrahydrofolate reductase gene (MTHFR), *Eur J Hum Genet* **1999** 7(2):197-204.

Heijmans BT, Westendorp RG, Slagboom PE, Common gene variants, mortality and extreme longevity in humans. *Exp Gerontol* **2000** 35(6-7):865-877.

Helkala EL, Koivisto K, Hanninen T, Vanhanen M, Kervinen K, Kuusisto J, Mykkanen L, Kesaniemi YA, Laakso M, Riekkinen P Sr, Memory functions in human subjects with different apolipoprotein E phenotypes during a 3-year population-based follow-up study, *Neurosci Lett* **1996** 204(3):177-80.

Herskind AM, McGue M, Holm NV, Sorensen TI, Harvald B, Vaupel JW, The heritability of human longevity, a population-based study of 2872 Danish twin pairs born 1870-1900, *Hum. Genet.* **1996** 97: 319-23.

Ho SL, Kung MH, Li LS, Lauder IJ, Ramsden DB, Cytochrome P4502D6 (debrisoquine 4-hydroxylase) and Parkinson's disease in Chinese and Caucasians, *Eur J Neurol* **1999** 6(3):323-9.

Ho SL, McCann KP, Bennett P, Kapadi AL, Waring RH, Ramsden DB, Williams AC, The molecular biology of xenobiotic enzymes and the predisposition to idiopathic Parkinson's disease, *Adv Neurol* **1996** 69:53-60.

Hoeg JM, Can genes prevent atherosclerosis? *JAMA* **1996** 276(12):989-92.

Holden C, Why do women live longer than man? *Science* **1987** 238:158-160.

Hougaard P, Modelling multivariate survival, *Scandinavian Journal of Statistics* **1987** 14:291-304.

Hougaard P, Modelling heterogeneity in survival data, *J.Appl. Prob.* **1991** 28:695-701.

Hougaard P, Frailty models for survival data. *Lifetime Data Anal* **1995** 1(3):255-73.

Iacoviello L, Di Castelnuovo A, De Knijff P, D'Orazio A, Amore C, Arboretti R, Kluft C, Donati MB, Polymorpisms in the coagulations factor VII gene and the risk of myocardial infarction, *New England Journal of Medicine* **1998a** 338(2):79-85.

Iacoviello L, Burzotta F, Di Castelnuovo A, Zito F, Marchioli R, Donati MB, The 4G/5G polymorphism of PAI-1 promoter gene and the risk of myocardial infarction: a meta-analysis, *Thromb Haemost* **1998b** 80(6):1029-30.

Ishikawa S, Kario K, Nago N, Kayaba K, Hiraoka J, Matsuo H, Goto T, Miyamoto T, Tsutsumi A, Nakamura Y, Shimada K, Inoue K, Igarashi M, Factor VII and fibrinogen levels examined by age, sex, and other atherosclerotic risk factors in a Japanese population. The Jichi Medical School Cohort Study. *Thromb Haemost* **1997** 77(5):890-3.

Ivanova R, Henon N, Lepage V, Charron D, Vicaut E, Schachter F, HLA-DR alleles display sex-dependent effects on survival and discriminate between individual and familial longevity, *Hum Mol Genet* **1998** 7:187-194.

Izaks GJ, van Houwelingen HC, Schreuder GM, Ligthart GJ, The association between human leucocyte antigens (HLA) and mortality in community residents aged 85 and older, *J Am Geriatr Soc* **1997** 45(1):56-60.

Jalavisto E, Inheritance of longevity according to Finnish and Swedish genealogies, *Ann. Med. Intern. Fenn*. **1951** 40:263-74.

Jeggo PA, DNA repair: PARP - another guardian angel? *Curr Biol* **1998** 8(2):R49-51.

Ji Y, Urakami K, Adachi Y, Nakashima K, No association between apolipoprotein A-IV codon 360 mutation and late-onset Alzheimer's disease in the Japanese population. *Dement Geriatr Cogn Disord* **1999** 10(6):473-5.

Jiang JC, Jaruga E, Repnevskaya MV, Jazwinski SM, An intervention resembling caloric restriction prolongs life span and retards aging in yeast. *FASEB J* **2000** 14: 2135-2137.

Johnson FB, Sinclair DA, Guarente L, Molecular biology of aging, *Cell* **1999** 96: 291-302.

Kadenbach B, Muller-Hocker J, Mutations of mitochondrial DNA and human death, *Naturwissenschaften* **1990** 77(5):221-5.

Kalaria VG, Zareba W, Moss AJ, Pancio G, Marder VJ, Morrissey JH, Weiss HJ, Sparks CE, Greenberg H, Dwyer E, Goldstein R, Watelet LF, Gender-related differences in thrombogenic factors predicting recurrent cardiac events in patients after acute myocardial infarction. The THROMBO Investigators. *Am J Cardiol* **2000** 85(12):1401-8.

Kannisto V, Development of oldest-old mortality, 1950-1990: evidence from 28 developed countries, **1994** Odense: Odense University Press, p59-66.

Kervinen K, Savolainen MJ, Salokannel J, Hynninen A, Heikkinen J, Ehnholm C, Koistinen MJ, Kesaniemi YA, Apolipoprotein E and B polymorphisms--longevity factors assessed in nonagenarians, *Atherosclerosis* **1994** 105(1):89-95.

Keyfitz N, Flieger W, World population growth and aging, **1990** Chicago : University of Chicago Press.

Klaver CC, Kliffen M, van Duijn CM, Hofman A, Cruts M, Grobbee DE, Van Broeckhoven C, de Jong PT, Genetic association of apolipoprotein E with age-related macular degeneration, *Am J Hum Genet* **1998** 63(1):200-6

Laird PW, Jaenisch R, The role of DNA methylation in cancer genetic and epigenetics. *Annu Rev Genet* **1996** 30:441-64.

Lane A, Green F, Scarabin PY, Nicaud V, Bara L, Humphries S, Evans A, Luc G, Cambou JP, Arveiler D, Cambien F. Factor VII Arg/Gln353 polymorphism determines factor VII coagulant activity in patients with myocardial infarction (MI) and control

subjects in Belfast and in France but is not a strong indicator of MI risk in the ECTIM study. *Atherosclerosis* **1996** 119(1):119-27.

Linnane AW, Marzuki S, Ozawa T, Tanaka M, Mitochondrial DNA mutations as an important contributor to ageing and degenerative diseases, *Lancet* **1989** 1(8639):642-5.

Lo H-S, Chen C-H, Hogan EL, Kao K-P, Wang V, Yan S-H, Genetic polymorphism and Parkinson's disease in Taiwan: Study of debrisoquine 4-hydroxylase (CYP2D6), *Journal of the Neurological Science* **1998** 158:38-42.

Macko RF, Kittner SJ, Epstein A, Cox DK, Wozniak MA, Wityk RJ, Stern BJ, Sloan MA, Sherwin R, Price TR, McCarter RJ, Johnson CJ, Earley CJ, Buchholz DW, Stolley PD, Elevated tissue plasminogen activator antigen and stroke risk: The Stroke Prevention In Young Women Study, *Stroke* **1999** 30(1):7-11.

Macurova H, Ivanyi P, Sajdlova H, Trojan J, HL-A antigens in aged persons Tissue Antigens **1975** 6(4):269-271.

Mannucci PM, Mari D, Merati G, Peyvandi F, Tagliabue L, Sacchi E, Taioli E, Sansoni P, Bertolini S, Franceschi C, Gene polymorphisms predicting high plasma levels of coagulation and fibrinolysis proteins, A study in centenarians, *Arterioscler Thromb Vasc Biol* **1997** 17(4):755-9.

Marchien van Baal, Kooistra T, Stehouwer CD. Cardiovascular disease risk and hormone replacement therapy (HRT): a review based on randomised, controlled studies in postmenopausal women. *Curr Med Chem* **2000** 7(5):499-517.

Mari D, Mannucci PM, Coppola R, Bottasso B, Bauer KA, Rosenberg RD, Hypercoagulability in centenarians: the paradox of successful aging, *Blood* **1995** 85(11):3144-9.

Mari D, Mannucci PM, Duca F, Bertolini S, Franceschi C. Mutant factor V (Arg506Gln) in healthy centenarians. *Lancet* **1996** 347(9007):1044.

Marshall E, The Alzheimer's Gene Puzzle, *Science* **1998** 280: 1002.

Martin GM, Genetics and the pathobiology of ageing, *Philos Trans R Soc Lond B Biol Sci* **1997** 352:1773-1780.

Martin GM, Interactions of aging and environmental agents: the gerontological Perstectives, *Prog clin Bio R*es **1987** 228:5-80.

Matsushita S, Muramatsu T, Arai H, Matsui T, Higuchi S, The frequency of the methylenetetrahydrofolate reductase-gene mutation varies with age in the normal population, *Am J Hum Genet* **1997** 61(6):1459-60.

Mayer PJ, Inheritance of longevity evinces no secular trend among members of six New England families born 1650-1874, *Am. J. Hum. Biol.* **1991** 3:49-58.

McClearn GE. Prospects for quantitative trait locus methodology in Gerontology. *Experimental Gerontology* **1997** 32:49-54.

McGue M, Vaupel JW, Holm N, Harvald B, Longevity is moderately heritable in a sample of Danish twins born 1870-1880, *J. Gerontol* **1993** 48, B237-B244.

Meade TW, Ruddock V, Stirling Y, Chakrabarti R, Miller GJ. Fibrinolytic activity, clotting factors, and long-term incidence of ischaemic heart disease in the Northwick Park Heart Study. *Lancet* **1993** 342(8879):1076-9.

Meiklejohn DJ, Riches Z, Youngson N, Vickers MA, The contribution of factor VII gene polymorphisms to longevity in Scottish nonagenarians, Thromb Haemost **2000** 83(3):519.

Mennen LI, de Maat MP, Schouten EG, Kluft C, de Jong PT, Hofman A, Grobbee DE, Coagulation factor VII, serum-triglycerides and the R/Q353 polymorphism: differences between older men and women. *Thromb Haemost* **1997** 78(3):984-6.

Merched A, Xia Y, Papadopoulou A, Siest G, Visvikis SApolipoprotein AIV codon 360 mutation increases with human aging and is not associated with Alzheimer's disease. *Neurosci Lett* **1998** 242(2):117-9.

Mittal KK, Immunobiology of the human major histocompatibility complex: association of HLA antigens with disease, *Acta Anthropogenet* **1984** 8(3-4):245-68.

Muiras ML, Muller M, Schachter F, Burkle A, Increased poly(ADP-ribose) polymerase activity in lymphoblastoid cell lines from centenarian, *J Mol Med* **1998** 76(5):346-54.

Murata M, Tagawa M, Kimura H, Kakisawa K, Shirasawa H, Fujisawa T, Correlation of the mutation of p53 gene and the polymorphism at codon 72 in smoking-related non-small cell lung cancer patients, *Int J Oncol* **1998** 12(3):577-81.

Myant NB, Gallagher J, Barbir M, Thompson GR, Wile D, Humphries SE, Restriction fragment length polymorphisms in the apo B gene in relation to coronary artery disease, *Atherosclerosis* **1989** 77(2-3):193-201.

Nakano K, Ohta S, Nishimaki K, Matsuda S, Alzheimer's disease and DLST genotype, *Lancet* **1997** 350:1367-1368.

Nakata Y, Katsuya T, Rakugi H, Takami S, Sato N, Kamide K, Ohishi M, Miki T, Higaki J, Ogihara T, Polymorphism of angiotensin converting enzyme, angiotensinogen, and apolipoprotein E genes in a Japanese population with cerebrovascular disease, *Am J Hypertens* **1997** 10(12 Pt 1):1391-5.

Natali A, Gastaldelli A, Galvan AQ, Sironi AM, Ciociaro D, Sanna G, Rosenzweig P, Ferrannini E, Effects of acute alpha 2-blockade on insulin action and secretion in humans. *Am J Physiol* **1998** 274: E57-64.

Nuzhdin SV, Pasyukova EG, Dilda CL, Zeng ZB, Mackay TF, Sex-specific quantitative trait loci affecting longevity in Drosophila melanogaster, *Proc Natl Acad Sci U S A* **1997** 94(18):9734-9.

Ossei-Gerning N, Wilson IJ, Grant PJ, Sex differences in coagulation and fibrinolysis in subjects with coronary artery disease. *Thromb Haemost* **1998** 79(4):736-40.

Ozawa T, Katsumata K, Hayakawa M, Yoneda M, Tanaka M, Sugiyama S, Mitochondrial DNA mutations and survival rate, *Lancet* **1995** 345(8943):189.

Paolisso G, Gambardella A, Ammendola S, D'Amore A, Varricchio M, Glucose tolerance and insulin action in healthy centenarians. *Am J Physiol* **1996** 270: E890-896**.**

Paulweber B, Friedl W, Holzl B, Sandhofer F, Genetics of coronary heart disease, *Lancet* **1989** 2(8659):384.

Paulweber B, Friedl W, Krempler F, Humphries SE, Sandhofer F, Association of DNA polymorphism at the apolipoprotein B gene locus with coronary heart disease and serum very low density lipoprotein levels, *Arteriosclerosis* **1990** 10(1):17-24.

Pearl R, Studies on human longevity, IV. The inheritance of longevity. Preliminary report, *Hum. Biol.* **1931** 3:245-69.

Pepe G, Di Perna V, Resta F, Lovecchio M, Chimienti G, Colacicco AM, Capurso A, In search of a biological pattern for human longevity: impact of apo A-IV genetic polymorphisms on lipoproteins and the hyper-Lp(a) in centenarians, *Atherosclerosis* **1998** 137(2):407-17.

Promislow DEL, Tatar M, Mutation and senescence: where genetics and demography meet, *Genetica* **1998** 102/103:299-313.

Proust J, Moulias R, Fumeron F, Bekkhoucha F, Busson M, Schmid M, Hors J, HLA and longevity, *Tissue Antigens* **1982** 19(3): 68-73.

Rankinen T, Gagnon J, Perusse L, Chagnon YC, Rice T, Leon AS, Skinner JS, Wilmore JH, Rao DC, Bouchard C, AGT M235T and ACE ID polymorphisms and exercise blood pressure in the HERITAGE Family Study. *Am J Physiol Heart Circ Physiol* **2000** 279(1):H368-74.

Ricci G, Colombo C, Ghiazza B, Illeni MT, Association between longevity and allelic forms of human leukocyte antigens (HLA): population study of aged Italian human subjects, *Arch Immunol Ther Exp (Warsz)* **1998** 46(1):31-4.

Rothman KJ, No Adjustments are needed for multiple comparisons, *Epidemiology* **1990** 1:43-46.

Ruiz-Pesini E, Lapena AC, Diez-Sanchez C, Perez-Martos A, Montoya J, Alvarez E, Diaz M, Urries A, Montoro L, Lopez-Perez MJ, Enriquez JA, Human mtDNA haplogroups

associated with high or reduced spermatozoa mobility. *Am J Hum Genet* **2000** 67: 682-696.

Sacher GA, Maturation and longevity in relation to cranial capacity in hominid evolution. Tuttle, R. Ed. Antecedents of man and after: Primates; functional morphology and evolution. Mouton: The Hahue; **1975**; p.417-441.

Saha N, Tay JS, Low PS, Basair J, Hong S, Five restriction fragment length polymorphisms of the APOA1-C3 gene and their influence on lipids and apolipoproteins in healthy Chinese, *Hum Hered* **1995** 45(6):303-10.

Sandholzer C, Saha N, Kark JD, Rees A, Jaross W, Dieplinger H, Hoppichler F, Boerwinkle E, Utermann G, Apo(a) isoforms predict risk for coronary heart disease, A study in six populations. *Arterioscler Thromb* **1992** 12(10):1214-26.

Scarabin PY, Aillaud MF, Amouyel P, Evans A, Luc G, Ferrieres J, Arveiler D, Juhan-Vague I, Associations of fibrinogen, factor VII and PAI-1 with baseline findings among 10,500 male participants in a prospective study of myocardial infarction--the PRIME Study, Prospective Epidemiological Study of Myocardial Infarction, *Thromb Haemost* **1998** 80(5):749-56.

Scarabin PY, Vissac AM, Kirzin JM, Bourgeat P, Amiral J, Agher R, Guize L, Population correlates of coagulation factor VII. Importance of age, sex, and menopausal status as determinants of activated factor VII. *Arterioscler Thromb Vasc Biol* **1996** 16(9):1170-6.

Scarabin PY, Bonithon-Kopp C, Bara L, Malmejac A, Guize L, Samama M, Factor VII activation and menopausal status.*Thromb Res* **1990** 57(2):227-34.

Schachter F, Cohen D, Kirkwood T, Prospects for the genetics of human longevity, *Hum Genet* **1993** 91:519-526.

Schachter F, Faure-Delaneff L, Guenot F, Rouger H, Froguel P, Lesueur-Ginot L, Cohen D, Genetic associations with human longevity at the APOE and ACE loci, *Nature Genetics* **1994** 6: 29-32.

Sellers TA, Weaver TW, Phillips B, Altmann M, Rich SS, Environmental factors can confound identification of a major gene effect: results from a segregation analysis of a simulated population of lung cancer families, *Genet Epidemiol* **1998** 15(3):251-62.

Sethi AA, Nordestgaard BG, Agerholm-Larsen B, Frandsen E, Jensen G, Tybjaerg-Hansen A. Angiotensinogen polymorphisms and elevated blood pressure in the general population: the Copenhagen City Heart Study. *Hypertension* **2001** 37(3):875-81.

Sethi AA, Tybjaerg-Hansen A, Gronholdt ML, Steffensen R, Schnohr P, Nordestgaard BG. Angiotensinogen mutations and risk for ischemic heart disease, myocardial infarction, and ischemic cerebrovascular disease. Six case-control studies from the Copenhagen City Heart Study. *Ann Intern Med* **2001** 134(10):941-54.

Shimoda-Matsubayashi S, Matsumine H, Kobayashi T, Nakagawa-Hattori Y, Shimizu Y, Mizumo Y, Structural dimophism in the mitochondrial targeting sequence in the human manganese superoxide dismutase gene, *Biochem Biophys Res Comm* **1996** 226561-565.

Shimokata H, Yamada Y, Nakagawa M, Okubo R, Saido T, Funakoshi A, Miyasaka K, Ohta S, Tsujimoto G, Tanaka M, Ando F, Niino N, Distribution of geriatric disease-related genotypes in the National Institute for Longevity Sciences, Longitudinal Study of Aging (NILS-LSA), *J Epidemiol* **2000** 10:S46-55.

Sing CF, Haviland MB, Reilly SL, Genetic architecture of common multifactorial diseases, *Ciba Found Symp* **1996** 197:211-29.

Sjalander A, Birgander R, Athlin L, Stenling R, Rutegard J, Beckman L, Beckman G P53 germ line haplotypes associated with increased risk for colorectal cancer. *Carcinogenesis* **1995** 16(7):1461-4.

Sont JK, Vandenbroucke JP, Life span expectancy and mitochondrial DNA, Do we inherit longevity from our mother's mitochondria?, *J Clin Epidemiol* **1993** 46(2):199-201.

Sorensen TI, Nielsen GG, Andersen PK, Teasdale TW. Genetic and environmental influences on premature death in adult adoptees. *N Engl J Med* **1988** 24;318(12):727-32.

Stucker I, Cosme J, Laurent P, Cenee S, Beaune P, Bignon J, Depierre A, Milleron B, Hemon D, CYP2D6 genotype and lung cancer risk according to histologic type and tobacco exposure, *Carcinogenesis* **1995** 16(11):2759-64.

Sun Yu, Channa Keshava, Dan S. Sharp, Ainsley Weston, and Erin C. McCanlies DNA Sequence Variants of *p53:* Cancer and Aging, *Am. J. Hum. Genet.* **1999** 65:1779-1782.

Takata H, Suzuki M, Ishii T, Sekiguchi S, Iri H, Influence of major histocompatibility complex region genes on human longevity among Okinawan-Japanese centenarians and nonagenarians, *Lancet* **1987** 2(8563):824-6.

Tan Q, Vaupel JW, Pattern of longevity inheritance in families of the European nobility: frailty models applied to related individuals, submitted.

Tan Q, De Benedictis G, Yashin IA, Bonafe M, DeLuca M, Valensin S, Vaupel JW, Franceschi C, Measuring the genetic influence in modulating human life span: Gene-environment and gene-sex interactions, *Biogerontology*, **2001**, 2(3), in print.

Tanaka M, Gong JS, Zhang J, Yoneda M, Yagi K, Mitochondrial genotype associated with longevity, *Lancet* **1998** 351(9097):185-6.

Thatcher AR, Kannisto V, Vaupel JW, The force of mortality at ages 80 and 120, **1998** Odense: Odense University Press.

Thogersen AM, Jansson JH, Boman K, Nilsson TK, Weinehall L, Huhtasaari F, Hallmans G, High plasminogen activator inhibitor and tissue plasminogen activator levels in plasma precede a first acute myocardial infarction in both men and women: evidence for the fibrinolytic system as an independent primary risk factor, *Circulation* **1998** 98(21):2241-7.

Toupance B, Godelle B, Gouyon PH, Schachter F, A model for antagonistic pleiotropic gene action for mortality and advanced age, *Am J Hum Genet* **1998** 62(6):1525- 1534.

Tsuneoka Y, Matsuo Y, Ichikawa Y, Watanabe Y, Genetic analysis of the CYP2D6 gene in patients with Parkinson's disease, *Metabolism* **1998** 47(1):94-6.

Tsuneoka Y, Fukushima K, Matsuo Y, Ichikawa Y, WatanabeY, Genotype analysis of the CYP2C19 gene in the Japanese population, *Life Sci* **1996** 59(20):1711-5.

Tybjaerg-Hansen A, Nordestgaard BG, Gerdes LU, Faergeman O, Humphries SE. Genetic markers in the apo AI-CIII-AIV gene cluster for combined hyperlipidemia, hypertriglyceridemia, and predisposition to atherosclerosis. *Atherosclerosis* **1993** 100(2):157-69.

Vaillant GE, The association of ancestral longevity with successful aging, *Journal of Gerontology: psychological sciences* **1991** 46(6):292-298.

Van der Bom JG, de Knijff P, Haverkate F, Bots ML, Meijer P, de Jong PT, Hofman A, Kluft C, Grobbee DE, Tissue plasminogen activator and risk of myocardial infarction. The Rotterdam Study, *Circulation* **1997** 95(12):2623-7.

Vandenbroucke JP, Matroos AW, van der Heide-Wessel C, Van der Heide R, Parental survival, an independent predictor of longevity in middle-aged persons, *Am. J. Epidemiol* **1984** 119:742-50.

Vaupel JW, Inherited Frailty and Longevity, *Demography* **1988** 25:277-287.

Vaupel JW, Yashin AI, Deviant dynamics of death in heterogeneous populations, *In N. Tuma, Ed. Sociological methodology* **1985** 179-211, San-Francisco, Jossey-Bass.

Vaupel JW, Kindred Lifetimes, Frailty Models in Population Genetics in Convergent Questions in Genetics and Demography (ed. Adams, J. et al.), **1991a** London: Oxford University Press. pp. 122-131.

Vaupel JW, Relatives' risks, Frailty models of life history data, *Theor. Popul. Biol.* **1991**b 37(1):220-234.

Vaupel JW, Tan Q, How many longevity genes are there? Paper presented at annual meeting of Population Association of America, **1998** Chicago.

Vaupel JW, Manton KG, Stallard E, The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography* **1979** 16(3):439-454.

Vaupel JW, Carey JR, Christensen K, Johnson TE, Yashin AI, Holm NV, Iachine IA, Kannisto V, Khazaeli AA, Liedo P, Longo VD, Zeng Y, Manton KG, Curtsinger JW, Biodemographic trajectories of longevity, *Science* **1998** 280:855-860.

Vaupel JW, Yashin AI, Heterogeneity's ruses: some surprising effects of selection on population dynamics, *Am Stat* **1985** 39:176-185.

Vaupel JW, Wang Z, Andreev KF, and Yashin AI, Population Data at a Glance: Shaded Contour Maps of Demographic Surfaces over Age and Time, **1998** Odense: Odense University Press.

Vaupel JW, Trajectories of mortality at advanced ages, in Wachter KW, Finch CE (eds), Between Zeus and the salmon, **1997** National Academy Press, Washington.

Wachter KW, Between zeus and the salmon: introduction, In: Wachter KW, Finch CE (eds), Between zeus and the salmon, **1997** National Academy Press, Washington, D. C., pp 1-16.

Wallace DC, Mitochondrial genetics: a paradigm for aging and degenerative diseases? *Science* **1992** 256(5057):628-32.

Wang YC, Chen CY, Chen SK, Chang YY, Lin P, p53 codon 72 polymorphism in Taiwanese lung cancer patients: association with lung cancer susceptibility and prognosis *Clin Cancer Res* **1999** 5(1):129-34.

Wang-Gohrke S, Rebbeck TR, Besenfelder W, Kreienberg R, Runnebaum IB, p53 germline polymorphisms are associated with an increased risk for breast cancer in German women. *Anticancer Res* **1998** 18(3B):2095-9.

Weir BS, Genetic data analysis II **1996** Massachusetts: Sinauer Associates, Inc. Publishers, p 133-135.

Wyskak G, Fertility and longevity of twins, sibs, and parents of twins, *Soc. Biol.* **1978** 25: 315-30.

Yamada H, Dahl ML, Lannfelt L, Viitanen M, Winblad B, Sjoqvist F, CYP2D6 and CYP2C19 genotypes in an elderly Swedish population, *Eur J Clin Pharmacol* **1998** 54(6):479-81.

Yarnell JWG, Leger ASST, Balfour C, Russell RB, The distribution, age effects and diseases association of HLA antigens and other blood group markers in a random sample of an elderly population, *J Chron Dis* **1979** 32: 555-561.

Yashin AI, Iachine IA, Genetic analysis of durations, Correlated frailty model applied to survival Danish Twins, *Genetic Epidemiology* **1995** 12: 529-538.

Yashin AI, Iachine IA, How frailty models can be used for evaluating longevity limits, taking advantage of an interdisciplinary approach, *Demography* **1997** 34: 31-48.

Yashin AI, Iachine IA, Harris JR, Half of the variation in susceptibility to mortality is genetic: findings from Swedish twin survival data, *Behav Genet* **1999a** 29:11-19.

Yashin AI, De Benedictis G, Vaupel JW, Tan Q, Andreev KF, Iachine IA, Bonafe M, DeLuca M, Valensin S, Carotenuto L, Franceschi C, Genes, Demography, and Life Span: The Contribution of Demographic Data in Genetic Studies on Aging and Longevity. *Am J Hum Genet* **1999b** 65:1178-1193.

Yashin AI, Vaupel JW, Andreev KF, Tan Q, Iachine IA, Carotenuto L, De Benedictis G, Bonafe M, Valensin S, Franceschi C, Combining genetic and demographic information in population studies of aging and longevity, *Journal of Epidemiology and Biostatistics* **1998** 3: 289-294.

Yashin AI, De Benedictis G, Vaupel JW, Tan Q, Andreev KF, Iachine IA, Bonafe M, DeLuca M, Valensin S, Carotenuto L, Franceschi C, Genes and longevity: Lessons from studies on centenarians. *Journal of Gerontology* **2000** 55a:B1-B10.

Yu MW, Yang SY, Chiu YH, Chiang YC, Liaw YF, Chen CJ, A p53 genetic polymorphism as a modulator of hepatocellular carcinoma risk in relation to chronic liver disease, familial tendency, and cigarette smoking in hepatitis B carriers, *Hepatology* **1999** 29(3):697-702.

Zhang JG, Ma YX, Wang CF, Lu PF, Zhen SB, Gu NF, Feng GY, He L, Apolipoprotein E and longevity among Han Chinese population. *Mechanisms of Aging and Development* **1998** 104:159-167.

Appendix A:

A Gauss program with instructions

Attached here is a computer program written in Gauss for the estimation of frequency and risk for one gene allele. The program is designed to work with ApoB allele data from a Danish study on centenarians (Section 5.2.2). But it can be easily modified to cope with other data or extended to incorporate other genetic and confounding factors.

1. Data manipulation

Statements 4 to 8 are responsible for loading the data and form it into the structure needed for the calculation. The data is an ASCII file named danesi.dat. It has 5 columns containing individual ID (Col.1), sex (Col.2), age (Col.3) and allele information (Col. 4 and 5). Statement 7 assigns 2 for females and 1 for males as indicator for sex. In line 8, the allele 31 is chosen for estimation. Estimation for other alleles can be done by changing the allele number in line 8.

2. Gene counting

Statement 9 activates the gene counting procedure, *nage( )*. This procedure counts for each age, the number of carriers of the gene allele in males and in females. Output of the procedure is a 5 column matrix arranged as age (Col.1), number of non-carriers in males (Col.2), number of carriers in males (Col.3), number of non-carriers in females (Col.4), number of carriers in females (Col.5). In line 9, a vector is defined as age to cover all the ages for individuals in the data.

3. The Two-step MLE

The Two-step MLE works as illustrated in Figure 5.1. The calculation involves two procedures, *SO( )* in line 21 for expecting a non-parametric baseline survival for given parameters, *maxlik( )* in line 25 for estimating the parameters by maximizing the likelihood for given baseline survival function. In the procedure *SO( ),* a numerical dichotomous method is engaged in solving the non-linear equation of (5.10) in Section 5.2.2. The whole calculation is controlled by a loop conditioned by line 27. Statements 10 to 17 load the population survival distributions

and assign them into each age. Line 18 is initial starting point for the calculation. When iteration goes on, the starting point is updated in line 23 with new values from line 26.

4. The maximum likelihood procedure

The maximum likelihood function is put into a procedure named *llik( ).* It calculates and returns the value of the likelihood function to its parent procedure *maxlik( )* in line 25. The *maxlik( )* is a built-in procedure that works on the defined likelihood function given by *llik( ).* One must notice that *llik( )* is self-defined and it should be modified to fit in different data structure.

5. Presenting the results

Results from the estimation are displayed by statements 33 to 35. From the screen, one can read the parameter estimates with standard error and p-value for risk of allele, and risk of gene-sex interaction. The program ends in line 36.

6. The program in Gauss

In order to run the program, we assume the Gauss software has been installed on your computer's hard driver with directory C:\gauss. Installation of the maximum likelihood procedure is required. For more information, please visit http:\\ www.Aptech.com.

```
1 new;
2 library
         c:\gauss\lib\PGRAPH.LCG,
         c:\gauss\lib\GAUSS.LCG,
         c:\gauss\lib\maxlik.LCG;
3 #include c:\gauss\src\maxlik.ext;
4 load a[]=c:\denmark\apob\danesi.dat;
5 a=reshape(a,rows(a)/5,5);
6 a=a[.,3 2 4 5];
7 a[.,2]=2.*(a[.,2].=="F")+(a[.,2].=="M");
8 a=a[.,1 2]~(sumc((a[.,3 4].==31)').>0);
9 gg=nage(a); n=rows(a); age=gg[.,1];

10 load s=c:\denmark\lft;
11 i=1;
12 ssm={};ssf={};
13 do while i<=rows(age);
14 ssm=ssm|selif(s[.,2],s[.,1].==age[i]);
15 ssf=ssf|selif(s[.,3],s[.,1].==age[i]);
```

```
16 i=i+1;
17 endo;
18 p=0.5; rg=1; rgs=1;
19 vv=0;mk=0;
20 hh:
21 sm=SO(ssm,p,rg,rgs); sf=SO(ssf,p,rg,1);
22 maxset;
23 start=ln((p)./(1-(p)))|ln(rg|rgs);
24 var=seqa(1,1,cols(gg));
25 {y,f,g,cov,retcode}=maxlik(gg,var,&llik,start);
26 p=exp(y[1])./(1+exp(y[1])); rg=exp(y[2]); rgs=exp(y[3]);
27 if abs(mk-vv).>=1e-5;vv=mk;goto hh;endif;
28 sp=sqrt(p.*(1-p)./n');
29 srg =sqrt( diag(cov[2,2]).*exp(y[2]).^2 );
30 srgs=sqrt( diag(cov[3,3]).*exp(y[3]).^2 );
31 prg=2.*(1-cdfn(abs((rg-1)./srg)));
32 prgs=2.*(1-cdfn(abs((rgs-1)./srgs)));
    "Frequency     Est.          SE";
33 p~sp;
    "Risk              Est.          SE           p-value";
34 rg~srg~prg;
    "Risk of GxS   Est.          SE           p-value";
35 rgs~srgs~prgs;
36 End;


/**************/
proc(1)=nage(d);
/**************/
local i,ggm,ggf,age,td,g0m,g0f,g1;
i=minc(a[.,1]);ggm={};ggf={};age={};
do while i<=maxc(a[.,1]);
td=selif(d,d[.,1].==i);
if sumc(d[.,1].==i).==0;
g0m=zeros(1,2);g0f=zeros(1,2);goto kk;endif;
g1=(td[.,2].==1  .and  td[.,3].==0)~(td[.,2].==1  .and  td[.,3].==1);
g0m=sumc(g1);
g1=(td[.,2].==2  .and  td[.,3].==0)~(td[.,2].==2  .and  td[.,3].==1);
g0f=sumc(g1);
kk:
ggm=ggm|g0m; ggf=ggf|g0f; age=age|i;
i=i+1;
endo;
retp(age~ggm~ggf);
```

```
endp;

/*********************/
proc(1)=SO(ss,p,rg,rgs);
/*********************/
local v,s,i1,i2,inc,x,rr,r1,pp,i,dd,po,sp,sb,f,f1,f2;
v=1;s={};
do while v<=rows(ss);
i1=-1e+30;i2=0;inc=1;
do while inc>1e-13;
 x=(i1+i2)/2;
rr=1~rg*rgs; pp=(1-p)~p;
sp=pp.*exp(i1)^(rr);
sb=sumc(sp');
f1=sb-ss[v];
sp=pp.*exp(x)^(rr);
sb=sumc(sp');
f2=sb-ss[v];
f=f1*f2;
if f<0;i2=x;else;i1=x;
endif;
inc=i2-i1;
endo;s=s|x;
v=v+1;
endo;
retp(s);
endp;

/*****************/
proc(1)=llik(q,dat);
/*****************/
local p,rg,rgs,rrm,rrf,pp,spm,spf,sb,ml;
p=exp(q[1)./(1+exp(q[1]));
rg=exp(q[2);
rgs=exp(q[3]);
rrm=1~rg*rgs; rrf=1~rg; pp=(1-p)~p;
spm=pp.*exp(sm)^(rrm); sb=sumc(spm'); spm=spm./sb+1e-300;
spf=pp.*exp(sf)^(rrf); sb=sumc(spf'); spf=spf./sb+1e-300;
ml=gg[.,2].*ln(1-spm[.,1])+gg[.,3].*ln(spm[.,2])+
gg[.,4].*ln(1-spf[.,1])+gg[.,5].*ln(spf[.,2]);
retp(ml);
endp;
```

Appendix B:

Biological glossary

allele one of the several alternative forms of a particular gene.

allele frequency (gene frequency) for any given gene, the relative proportion of each allele of that gene found in a population.

chromosome a linear strand composed of DNA and protein, found in the nucleus of a cell, that contains the genes.

codominant two alleles whose phenotypic effects are both expressed in the heterozygote.

DNA (deoxyribonucleic acid) a polymer composed of deoxyribonucleotides linked together by phosphodiester bonds. The material of which most genes are made.

dominant an allele or trait that expresses its phenotype when heterozygous with a recessive allele.

gene the basic unit of heredity encoding the information needed to specify the amino acid sequence of proteins and hence particular traits. A gene is a segment of DNA located at a particular place on a chromosome.

gene-environment interaction the effect on a phenotype value that results from a specific genotype and a specific environment and that is not predictable from either separately.

genetic marker a mutant gene or other peculiarity in a genome that can be used to "mark" a spot in a genome for mapping purposes.

genetic variance the variation of a phenotype in a population that results from genetic causes.

genome All the genes contained in a single set of chromosome. It is one complete set of genetic information from a genetic system.

genotype the allelic constitution of a given individual.

haplotype a set of linked genes or DNA sequences that tend to be inherited together.

Hardy-Weinberg equilibrium the balance in the realtive number of alleles that is maintained within a large population over a period of time assuming: (1) mating is

random; (2) there is no natural selection; (3) there is no migration; (4) there is no mutation.

heritability the proportion of phenotypic variance that is the additive genetic component. Symbolized as h^2. Realized heritability is an estimate of h^2 from a selected experiment.

heterozygous carrying two different alleles of a given gene.

homozygous carrying two copies of the same allele of a given gene.

locus the physical location of a gene on a chromosome.

mitochondrial DNA (mtDNA) a circular ring of DNA found in mitochondria, the structures within cytoplasm that carry out aerobic respiration.

phenotype the physical characteristics of an organism.

polygenic trait a phenotypic trait determined by a number of genes.

polymorphism the existence of two or more genetically determined forms (alleles) in a population in substantial frequency.

quantitative trait a trait that generally has a continuous distribution in a population and is usually affected by many genes and many environmental factors.

recessive an allele or trait that does not express its phenotype when heterozygous with a dominant allele.

RNA (ribonucleic acid) a polymer composed of ribonucleotides linked together by phosphodiester bonds.