Longevity and Genetic Data in Twin and Family Studies

Inaugural - Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Alexander Begun

aus Moskau

Düsseldorf, Juli 2014

Aus dem Institut für Biometrie und Epidemiologie des
Deutschen Diabetes-Zentrums, Leibniz-Institut an
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. G. Giani

Korreferent: Prof. Dr. A. Janssen

Tag der mündlichen Prüfung: 11. Dezember 2014

Contents

Abstract

Aging and survival are caused by a complex and non-observable interaction between genetic and environmental factors. To reveal regularities of this interaction the traditional methods of survival analysis combined with ones of quantitative genetics data are needed. The data used by statistical analysis of longevity usually have a number of peculiarities and drawbacks such as selective sampling and incompleteness caused by censoring and truncation. Genetic data can include information on genes with a known location (genetic markers) for related individuals (e.g. twins, sibs or members of a family).

Finding genes that are differentially expressed under two or more conditions is a main object in experiments with microarrays. Searching for such genes is usually based on statistical methods involving t-statistics and multiple testing and uses datasets with information about thousands of genes, but a relatively small number of individuals. Correlations between individuals are usually not taken into account in these studies.

Phenotypic traits such as length of life and gene expressions can correlate for related individuals because such individuals share genetic and environmental factors. If we do not take into account these correlations the estimates obtained in the studies can be biased and conclusions are wrong. In this work we develop statistical models that combine the strength of the methods of the bi- and multivariate (survival) analysis with methods of genetic analysis and analysis of gene expression data.

In the analysis of survival data we use the concept of frailty assuming that non-observable susceptibility to death can contain both genetic and environmental components. Additional randomness in death process is caused by underlying hazard. Observed covariates in the form of a Cox-like regression are also included in the survival models. We discuss the methods and the problem of identifiability of such models. We show how genetic markers data can be used to locate the position of longevity or frailty genes.

We also discuss how the mixed model method for detecting genes with differential gene expression can be adapted for twin data. All models are illustrated with examples based on analysis of real or simulated data.

Zusammenfassung

Altern und Überleben werden durch komplexe, nicht beobachtbare Wechselwirkungen zwischen genetischen und Umweltfaktoren verursacht. Zur Untersuchung dieser Interaktionen werden traditionelle Methoden der Überlebensanalyse in Kombination mit solchen der quantitativen Genetik benötigt. Die Überlebenszeitdaten, die bei der statistischen Analyse der Langlebigkeit verwendet werden, haben in der Regel eine Reihe von Besonderheiten und Nachteilen. Sie entstammen nicht selten einer selektiven Stichprobe, sind unvollständig und weisen Zensierung und Trunkierung auf. Die eingehenden genetischen Daten beinhalten Informationen über Gene mit bekannter Lokalisation (genetische Marker) für verwandte Personen.

Das Auffinden von Genen, die differentiell unter zwei oder mehr Bedingungen exprimiert sind, ist ein Hauptliegen von Microarray-Experimenten. Die Suche nach solchen Genen basiert in der Regel auf den statistischen Verfahren, die t-Statistiken und multiples Testen verwenden und nutzt Datensätze mit Informationen über Tausende von Genen, die jedoch an einer nur kleinen Anzahl von Individuen erhoben werden. Korrelationen zwischen Individuen werden in diesen Studien in der Regel nicht berücksichtigt.

Phänotypische Merkmale wie Lebensdauer und Genexpressionen können für verwandte Individuen wegen gemeinsamer genetischer Faktoren und/oder Umweltfaktoren korrelieren. Werden diese Korrelationen nicht adäquat berücksichtigt, können die Schätzungen verzerrt und die Schlussfolgerungen falsch werden. In dieser Arbeit werden innovative statistische Modelle entwickelt, die die Stärke moderner Methoden der bi- und multivariaten (Überlebenszeit-)Analyse mit Methoden der genetischen Analyse und der Analyse von Genexpressionsdaten kombinieren.

Zur Analyse von Überlebenszeiten wird in dieser Arbeit das Frailty-Konzept verwendet mit der Annahme, dass nicht beobachtete "susceptibility to death" sowohl genetische als auch umweltbedingte Komponenten enthalten kann. Zusätzliche zufällige Einflüsse auf den Todesprozess werden durch die "baseline hazard" abgebildet. Beobachtete Kovariable lassen sich in Form einer Cox-Regression auch in Überlebenszeitmodellen einbringen. Diskutiert werden in dieser Arbeit die Methoden und das Problem der Identifizierbarkeit solcher Modelle. Gezeigt wird auch, wie genetische Marker zur Lokalisierung der Langlebigkeits- und Frailtygene genutzt werden können.

Ferner wird diskutiert, wie gemischte Effektmodelle zur Detektion von Genen mit differentieller Genexpression an Zwillingsdaten angepasst werden können. Alle Modelle werden an realen oder simulierten Daten illustriert.

Acknowledgements

First, I want to deeply thank my supervisor Prof. Dr. Guido Giani, for support and valuable advice.

My warm thanks go to Prof. Dr. Anatoli Yashin (Duke University, USA) for useful discussions and introducing me to the area of the multivariate survival models with frailty components. I would like to express my gratitude to Prof. Dr. Michael Krawczak (Kiel University, Germany) for discussions and support in research.

I also wish to thank all co-workers in the Max-Planck Institute for Demographic Research (Rostock, Germany) including Prof. Dr. Ivan Iachine, and Prof. Dr. Andreas Wienke for their collaboration.

My special thanks go to Prof. Dr. Arnold Janssen for his great interest in this thesis and for agreeing to act as co-examiner.

Finally, I would like to thank my family for their constant support and love.

Chapter 1

1 Introduction

1.1 Outline

In many epidemiological, biological, demographic and medical studies researchers are interested in the relationship between time-to-death or time-to-onset and factors influencing these characteristics. Usually a lot of such factors - genetic and environmental contribute to mortality and to disease risks and it is very difficult to measure all them directly. Until today not all genes and environmental factors involved in death and disease processes are known.

The concept of frailty as non-observed susceptibility to death has been firstly introduced by Vaupel et al. (Vaupel et al., 1979) to extend classic univariate survival models by taking into account hidden non-observable heterogeneity of mortality in population. Usually frailty is defined as a random variable with known form of the distribution function (e.g. gamma distribution) but maybe with unknown parameters of this distribution. Some authors (Gjessing et al., 2003; Gorfine and Hsu, 2011) define the frailty as a random process. The notion of hazard as an instantaneous rate of death plays a key role in ordinary survival analysis. Usually it is assumed that the frailty component has a multiplicative effect on the baseline hazard function. Observed covariates can be included in the model in the form of a Cox-like regression with unknown regression parameters.

Assuming independence of conditional survival functions for genetically related individuals (e.g. twins or sibs) given their frailties we can consider possible genetic and environmental similarities of these individuals as the components of frailty. This impor-

tant property allows us to use the traditional methods of quantitative genetics and to assess the role of genes in longevity (Iachine, 2002). Using the correlated frailty model applied to Danish twins survival data Yashin and Iachine (1995b) showed that about 50% of the variance of frailty is due to genetic factors. Since the life span is defined by both frailty and hazard function, the heritability of longevity span is much lower.

Another way to take into account genetic factors is to consider a discrete distribution of frailty assuming that the frailty is a function of genotype. A major gene model allows us to estimate the frequency and the relative risk of longevity or frailty genotypes. This model is based on a two-level segregation analysis combining a model of the transmission of genes from parents to offsprings with penetrance functions describing conditional probabilities of phenotypes given the genotype. Studies based on a major gene model show that such models can produce the bivariate survival density function similar to those produced by a correlated gamma-frailty model (Begun et al., 2000a)[**B1**].

It is difficult to adapt the correlated frailty model with continuous distribution of frailty to the data with more than two related individuales. Fortunately, the bivariate major gene model for twins can be easily extended to use genealogical data. This approach has been used to estimate the effects of familial environment (parental reproductive age and parental longevity) on the enfant/child mortality using genealogical data from Canada (Begun et al., 2000b)[**B2**].

We cannot locate the position of longevity gene if additional genetic information such as genetic markers data are not available. If longevity gene is in linkage disequilibrium with markers, the ordinary regression for univariate survival data with unknown regression coefficients for markers as covariates can be used to find significant regression coefficients and to approximately locate the position of longevity gene near respective markers. This method can be considered in the context of association analysis when a genetic variant associated with longevity is either in close proximity to longevity locus or longevity allele itself (Begun, 2007; Begun, 2009)[**B3, B4**]. If longevity gene is in linkage equilibrium with markers, then a combination of the segregation analysis based on the major gene model with linkage analysis using a hidden Markov chain algorithm can be implemented to locate the position of the longevity gene (Begun and Yashin, 2004; Begun, 2013)[**B5, B6**].

Nowadays, the experiments with microarrays already having begun in 1990s are widely used for determining differentially expressed genes under two or more conditions. In me-

thods used in these researches it was mostly assumed that individuals are chosen independently. However, it was not analyzed what will be if we take related individuals for these studies. The mixture model method (Pan, 2003) can be extended to treat the data on twins too (Begun, 2006)[**B7**]. The comparison of the uni- and bivariate model for gene expression data shows that the power and the number of false positives can depend on the concordance-discordance status of twin pairs and also on the correlation between gene expression levels for twins (Begun, 2006; Begun 2008)[**B7, B8**].

1.2 Background

The present thesis resumes my work in the areas of bi- and multivariate genetic analysis of longevity and of bivariate gene expression analysis during last decade. These studies aimed in developing new statistical models for genetic analysis of longevity and in searching for genes differentially expressed under two conditions using twin data. Near the correlated frailty models with continuously distributed frailty the models with discretely distributed frailty (models based on the major gene approach) have also been studied (Begun et al., 2000a; Begun et al., 2000b) [**B1, B2**]. Some properties of the data including left truncation, right censoring, and presence of observed covariates have been taken into account.

Second, a relative risk model with flexible parameterization of the cumulative hazard function has been investigated (Begun, 2008; Begun, 2009)[**B3, B4**]. This model allows us for including the antagonistic pleiotropic effect and for taking into account both continuous cohort trends and sudden changes in longevity allele frequency by searching the gene-longevity association.

Third, the methods aiming in locating a position of longevity or frailty genes using bivariate survival data and information about genetic markers have been tested (Begun and Yashin, 2004; Begun, 2013)[**B5, B6**].

Fourth, an extension of the mixed model method for detecting differential gene expression using bivariate data has been proposed and its properties have been studied (Begun, 2006; 2008)[**B7, B8**].

The thesis consists of nine chapters. In each chapter the brief overview of the most important methods relating to theme of study is given and some aspects of my study in this area are discussed. Some details of the methods are given in Appendices. For more

details I refer the readers to the original papers and to the reference list.

Chapter 2

2 Data Sources

2.1 Human mortality Database

The Human Mortality Database began in the year 2000 as a collaborative project involving research teams in the Department of Demography at the University of California, Berkeley (USA), and at the Max Planck Institute for Demographic Research (MPIDR) in Rostock (Germany). Its aim was to provide comprehensive mortality and population data to researchers. It contains uniform death rates and life tables (e.g., life expectancy) for various populations and include data for more than 30 countries. In papers listed in this thesis we have used mortality data for males and females for one-year age groups in France, Japan, Sweden, and the United States. The data can be obtained from the web of MPIDR.

2.2 Twin survival data from Denmark

The Danish Twin Registry (DTR) was created in the 1950s and is one of the oldest population-based registries in the world. It contains information about twins born in Denmark since 1870 and survived to age 6. Multiple births were manually ascertained in birth registers from all 2,200 parishes in Denmark. As soon as twin was traced, a questionnaire was mailed to the twin, to her/his partner or to their closest relatives (if neither of the twin partners were alive). Zygosity was assessed on the basis of the questions about phenotypic similarities. The reliability of the zygosity diagnosis was validated by comparison with laboratory methods based on the blood, serum, and enzyme groupe determination. It was found that missclassification rates were less than 5%. Other information includes the data on sex, birth, causes of death, health, and life style. An important feature of

the Danish twin survival data is their right-censoring and left truncation. In our study we used the data on same-sex twins with known zygosity born between 1870 and 1900. These data includes 470 male monozygotic (MZ) twin pairs, 475 female MZ twin pairs, 780 male dizygotic (DZ) twin pairs and 835 male DZ twin pairs. More details about the Danish Twin Registry can be found in Hauge (1981).

2.3 Family survival data from Canada

Immigration from France to Canada began in the 17th century. The peopling started from the Quebec city area and expanded rapidly upstream and downstream the St.Lawrence river. Almost whole population (85%) in Quebec derives from about 8,000 original French founders and is genetically homogeneous. The Population Register in Quebec is supported now by inter-university research center and contains around 700,000 vital records of virtually every individual who ever lived in Quebec. The main file is related with various topics or sub-populations (e.g. patients suffering from genetic disorders). A computer linkage system can process married couples and links the family records together. More information about the Population Register in Quebec can be found in Bouchard (1989). From 13544 records, relating to French-Canadian children born in Quebec between 1623 and 1705, 2066 children (1016 boys and 1050 girls) were chosen with valid birth/death dates, who survived until 30 and overlived their parents. The number of children in families fluctuates between 1 and 10 with a mean 2.6.

2.4 Simulated data

If real data was not available in the studies we simulated datasets for given vector parameters. Details of these simulations can be found in the text and in the original papers.

Chapter 3

3 The univariate survival models with proportional hazard

3.1 Definition of survival and hazard functions

Define a nonnegative random variable T representing the lifetime (the time to death) of an individual in a homogeneous population (Kalbfleish and Prentice, 1980; Cox and Oakes, 1984). Assuming that transition to death is a continuous-time Markov process we define an instantaneous rate (hazard) of death as

$$\lambda(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

(if this limit exists). Conditional survival function $S(t|t_0)$ is the probability that the individual does not die in the interval $[t_0, t)$ given that it was alive at age t_0. It is easy to check that $S(t|t_0)$ is a non-increasing left-hand continuous function with $S(t_0|t_0) = 1$ and

$$\lambda(t) = -d \ln S(t|t_0)/dt = -S(t|t_0)^{-1} dS(t|t_0)/dt.$$

It means that $S(t|t_0)$ has a right-hand derivative and is also a right-continuous function. From this and the initial condition $S(t_0|t_0) = 1$ it follows that

$$S(t|t_0) = \exp(-H(t_0, t)),$$

where

$$H(t_0, t) = \int_{t_0}^{t} \lambda(\tau) d\tau$$

is the cumulative hazard. For the absolutely continuous conditional random time $T|t_0$ to death given that the individual was alive at age t_0-, the probability density function

corresponding to $S(t|t_0)$ is

$$f(t|t_0) = -dS(t|t_0)/dt = \lambda(t)S(t|t_0).$$

Taking into account random censoring time C, a vector of time independently observed covariates u, and a non-observed random variable Z (frailty), we define observed time to death as $\min(C, T)$ and the hazard function for the latent time to death as

$$\lambda(t|u, Z) = \lim_{\Delta t \downarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t, u, Z)}{\Delta t}.$$

We will assume that the hazard functions of the latent failure times follow the mixed proportional hazard specification. That is, they depend multiplicatevely on the cause-specific baseline hazard functions, observed regressors, and frailty

$$\lambda(t|u, Z) = \lambda_0(t) \exp(\beta^* u) Z,$$

where symbol "$*$" denotes the matrix transpose and β is the Cox regression parameter vector. To avoid non-identifiability it is usually assumed that $\mathrm{E}Z = 1$.

Since Gompertz in 1825 has proposed the exponential increasing of mortality rates for human population over the segment 30-84 years, there were many attempts to explain this form of dependency and to approximate the force-of-mortality function for other age groups. Strehler and Mildvan (Strehler and Mildvan, 1960) related the exponential increase of human mortality with the linear decline of a vitality index. A lower rate of mortality in the 85-and-older age group of humans than the Gompertzian model predicts (the levelling-off effect) can be explained, for example, by the presence of inhomogeneity in population (Manton, 1982) or in terms of the mutation accumulation theory (Łaszkiewicz et al., 2003).

A multiclonal model (Abernethy, 1998) explains the exponential increase of mortality rates in multicellular organisms. This model is based on the assumption that the cells of such an organism are subject to cellular aging phenomena of limited replicability (Hayflick limit) and mitotic deceleration (exponential increase in time of the mitotic event waiting-time).

The concept of the space of life-cycle-states underlies the approach proposed in **[B9]** (Begun, 2006). The ontogenetic evolution of a multicellular organism is associated with a material point moving in this space under the action of a force field. Such an association seems quite reasonable. In theoretical mechanics the evolution of the material point

follows "the principle of least action". The age evolution of the living organism is also in accordance with some principle of least action allowing biological interpretation. This principle states that the own biological time on the trajectory of the age evolution of the living system is minimal. The own biological time of the multicellular organism (e.g. human being) elapsing between two moments of its calendar age is interpreted as a certain number of generations of the living system at the next lower level of hierarchy (cells). Continuous deceleration with age of the rate of evolution (associated with decreasing of the "kinetic energy") in the space of life-cycle-states leads to an increase of the mitotic event waiting-time and finally to death. The model makes it possible to estimate the dimensionality of the space of life-cycle-states, describes the phylogenetic evolution of this dimensionality, and explains the form of the force-of-mortality function over the different segments of the lifespan of a multicellular organism.

Over the age interval 35-85 years we describe the force-of-mortality function in terms of the Gompertz model by

$$\lambda_0(t) = ae^{bt}.$$

Strehler and Mildvan (Strehler and Mildvan, 1960) theoretically predicted that the parameters a and b are negatively correlated and the changes in the human mortality rate resulting from economical, medical, and other improvements must follow certain regalarities. Namely,

$$\ln a = \ln K - B^{-1}b$$

for some positive constants B and K. The Strehler-Mildvan negative correlation was then confirmed in a number of empirical studies based on the period data for different human populations and often manifested as universal demographic law regulating changes in the age mortality rates. However, later studies (Yashin et al., 2001; Yasin et al., 2002) showed that the Strehler-Mildvan correlation pattern is relatively stable only in certain periods of survival history. That is, the changes in period and cohort hazard functions can follow different patterns with different sets of the coefficients B and K.

3.2 Baseline hazard functions

Now we will discuss briefly different types of the baseline hazard functions used usually to describe times to failure.

Constant hazard. This is the simplest form of the hazard function with probability of

dying within a time interval depending only on length of this interval. Survival function in this case is exponential

$$S_0(t) = \exp(-\lambda t)$$

with $\lambda > 0$, hazard function $\lambda_0(t) = \lambda$, and cumulative hazard $H_0(t) = \lambda t$.

Weibull function. It is a generalization of the exponential distribution and is often used in mortality studies. Two-parameter $\lambda, \gamma > 0$ survival function is

$$S_0(t) = \exp(-\lambda t^\gamma)$$

with hazard function

$$\lambda_0(t) = \lambda\gamma t^{\gamma-1}$$

and cumulative hazard function

$$H_0(t) = \lambda t^\gamma.$$

Gompertz function. This form of hazard is closely related with the Weibull function and is very popular in survival analysis. For exponential hazard function $\lambda_0(t) = a\exp(bt)$ with $a, b > 0$ we have a double exponential survival function

$$S_0(t) = \exp(-(a/b)(e^{bt} - 1))$$

and cumulative hazard function

$$H_0(t) = (a/b)(e^{bt} - 1).$$

The hazard function for the Gompertz-Makeham mortality law includes additionally an age-independent positive component c

$$\lambda_0(t) = c + a\exp(bt).$$

The Gamma-Gompertz-Makeham model with survival function

$$S_0(t) = \left(1 + s^2\left(ct + (a/b)(e^{bt} - 1)\right)\right)^{-1/s^2}$$

is a further generalization of the Gompertz model.

Log-logistic function. This hazard function is more flexible than two parameter hazard functions mentioned above and allows for describing monotone trends as well as bell-shaped ones. The two-parameter survival function is

$$S_0(t) = (1 + at^b)^{-1}$$

with $a, b > 0$, hazard function

$$\lambda_0(t) = \frac{abt^{b-1}}{1 + at^b},$$

and cumulative hazard function

$$H_0(t) = \ln(1 + at^b).$$

Log-normal function. In this model the logarithm of the random time-to-failure T is normally distributed, i.e. $\ln T \sim N(\mu, \sigma^2)$. The probability density function is

$$f_0(t) = \frac{1}{\sqrt{2\pi}\sigma t} \exp\left(-\frac{(\ln t - \mu)^2}{2\sigma^2}\right).$$

More information about hazard functions can be found, for example, in (Wienke, 2011).

3.3 Frailty component

The term frailty has first been suggested in demography to characterize unobserved individual susceptibility to death and to take into account hidden heterogeneity of mortality risk in population and the deviation of mortality rates at advanced ages (Vaupel et al., 1979; Vaupel and Yashin, 1985; Andersen et al., 1993). Frailty models play an important role in analysis of data on related individuals (e.g. twins, sibs) by allowing the assessment of such genetic characteristics as genetic variation and heritability (Yashin and Iachine, 1995a; Yashin and Iachine, 1997).

Different types of frailty distributions are used in survival analysis, including discrete, gamma, log-normal, positive-stable, inverse-gaussian, and power variance function distribution (Hougaard, 1986; Wienke, 2011). Since the univariate survival function is the Laplace transform with respect of frailty distribution, from computational point of view the most attractive distributions are those with simple form of its Laplace transform. Among them are the gamma (most common distribution), inverse-gaussian, compound Poisson, positive-stable, and power variance function distributions.

Discrete distributions. In the simplest situation the population under study consists of two non-observed groups with two different risks of mortality. If the proportion of

individuals in the first group is equal to p, then a randomly selected person belongs to the first group with probability p or to the second group with probability $1-p$. The risk of mortality r is a binary random variable which has the value r_1 with probability p and r_2 with probability $1-p$. This two-point frailty distribution can be easily extended to the k-point discrete one. The discrete frailty can be related with unknown genotypes affecting longevity.

Power variance function distribution $P(\alpha, \delta, \theta)$. This family is exponential for fixed α, δ with natural observation t and parameter θ (Hougaard, 1986). The parameter space is $0 < \alpha \leq 1$, $\delta > 0$, $\theta \geq 0$. The probability density function is

$$f(t; \alpha, \delta, \theta) = -\exp(-\theta t + \delta\theta^{\alpha}/\alpha)(\pi t)^{-1} \sum_{k=1}^{\infty} \frac{\Gamma(k\alpha+1)}{k!}(-\delta t^{-\alpha}/\alpha)^k \sin(\alpha k\pi).$$

The Laplace transform is

$$L(s) = \exp\left[-\frac{\delta}{\alpha}\{(\theta+s)^{\alpha} - \theta^{\alpha}\}\right].$$

Compound Poisson distribution. This distribution can be constructed as the sum of a Poisson distributed number of independent and identically gamma distributed random variables. This is also a three parameter distribution with parameters $\alpha < 0$, $\delta > 0$, $\theta > 0$, with probability density function similar to the density function of the power variance function distribution.

Gamma distribution. The density of the Gamma distribution with mean δ/θ and the variance δ/θ^2 is

$$f(t) = \frac{\theta^{\delta} t^{\delta-1} e^{-\theta t}}{\Gamma(\delta)}, \quad \theta > 0,\ \delta > 0.$$

Its Laplace transform is

$$L(s) = (\theta/(\theta+s))^{\delta}.$$

It is a special case of the power variance function distribution for $\theta > 0$ as $\alpha \downarrow 0$.

Degenerate distribution. The distribution degenerates at zero and corresponds to $\delta = 0$.

Positive-stable distribution. It is a special case of the power variance function distribution given by $P(\alpha, \alpha, 0)$ for $\alpha \in (0, 1]$. Its Laplace transform is

$$L(s) = \exp(-s^{\alpha}).$$

Inverse-gaussian distribution. It is also a special case of the power variance function distribution. For $\alpha = 1/2$ we have the inverse-gaussian distribution with density function

$$f(t) = \delta\pi^{-1/2} \exp(2\delta\theta^{1/2}) t^{-3/2} \exp(-\theta t - \delta^2/t).$$

Log-normal distribution. This distribution defined for positive t does not have the Laplace transform in simple closed form. Its probability density function is

$$f(t; \mu, \sigma) = \frac{1}{\sqrt{2\pi} t\sigma} e^{-\frac{-(\ln t - \mu)^2}{2\sigma^2}}.$$

with mean $\exp(\mu + \sigma^2/2)$ and the variance $\exp(\sigma^2 - 1)\exp(2\mu + \sigma^2)$.

3.4 Identifiability and validation

In parametric approach the parametric form of frailty distribution and of the baseline hazard functions are known. To calculate the Maximum Likelihood (ML) estimates of unknown parameters, we maximize the likelihood function directly with respect to Cox-regression parameters β, the parameters of frailty distribution, and the parameters involved with baseline hazard functions. Alternatively, the Expectation-Maximization (EM)-algorithm can be used for computing iteratively the unknown parameters.

In some cases the distribution assumptions regarding the baseline hazard functions can be incorrect. Therefore, the estimates based on purely parametric approach are inconsistent. In non-parametric approach we do not specify the baseline hazard functions which are regarded as infinite dimensional parameters. The estimates of baseline hazard functions obtained using the non-parametric approach can suffer from some efficiency loss in comparison to the parametric approach, as long as the baseline hazards are correctly specified (Gorfine, Hsu, 2011).

The nonparametric technique of Kaplan and Meier can be used to estimate the hazard $\lambda(t)$ for the failure time data without covariates by

$$\lambda(t_i) = \frac{d_i}{n_i}.$$

Here failure (death) occurs with multiplicity d_i at time t_i, $t_1 < t_2 < ... < t_k$, and n_i is the number subjects at risk just before the moment t_i. The value of $\lambda(t)$ is equal to zero elsewhere. For more details see, for example, (Kalbfleish and Prentice, 1980; Cox and Oakes, 1984).

Similarly, for a known vector u of covariates, the estimates of the background hazard function can be obtained to be

$$\lambda_0(t_i) = \frac{d_i}{\sum_{l \in R(t_i)} \exp(\hat{\beta}^* u_l)},$$

where $R(t_i)$ is the set of all subjects chosen prior moment t_i and the maximum likelihood estimates $\hat{\beta}$ of regression coefficients are obtained from the partial likelihood

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp[\beta^* u_i]}{\sum_{l \in R(t_i)} \exp[\beta^* u_l]}.$$

As above we put $\lambda(t) = 0$ elsewhere.

The identifiability of this univariate model with unspecified functional form of frailty distribution and baseline hazard has been studied elsewhere (Elbers and Ridder, 1982). This model is identifiable given information on T for finite $\mathrm{E}Z$ and is not identifiable when frailty has an infinite mean.

Chapter 4

4 The bi- and multivariate survival models with proportional hazard

4.1 Model description

In Section 3.3 we discussed briefly the problem of heterogeneity in populations due to unobservable factors. Now we shall focus on another aspect of modelling - statistical dependencies in data. We are concerned with it, for example, in analysis of survival data for related individuals (e.g. twins, family data). We shall show how the concept of frailty can be combined with a multivariate approach. The correlated bivariate frailty model has been used in (Begun et al., 2000a; Begun, 2009; Begun and Yashin, 2004; Begun, 2013)[**B1, B4, B5, B6**].

In the bivariate shared frailty model it is assumed that two time-to-death random variables T_1 and T_2 are conditionally independent given shared frailty Z, covariates u_1, u_2, and that the bivariate survival function is defined by

$$S(t_1, t_2|Z, u_1, u_2) = \exp(-Z\exp(\beta^* u_1)H_0(t_1))\exp(-Z\exp(\beta^* u_2)H_0(t_2))$$

for the cumulative baseline hazard function

$$H_0(t) = \int_0^t \lambda_0(\tau)d\tau.$$

The correlated bivariate frailty model is a generalization of the shared frailty model and includes two dependent frailties Z_1 and Z_2. For covariates u_1 and u_2 measured for the two subjects and for two latent random failure times T_1 and T_2 the conditional bivariate

survival function has a form

$$S(t_1, t_2|Z_1, Z_2, u_1, u_2) = \exp(-Z_1 \exp(\beta^* u_1) H_0(t_1)) \exp(-Z_2 \exp(\beta^* u_2) H_0(t_2)).$$

Here we have assumed that T_1 and T_2 are conditionally independent of the vectors of covariates u_1, u_2 and the frailties Z_1, Z_2. In other words,

$$P(T_1 \geq t_1, T_2 \geq t_2|Z_1, Z_2, u_1, u_2) = S(t_1, t_2|Z_1, Z_2, u_1, u_2) = S_1(t_1|Z_1, u_1) S_1(t_2|Z_2, u_2),$$

where $S_1(t_1|Z_1, u_1)$ and $S_1(t_2|Z_2, u_2)$ are the univariate conditional survival functions. Unconditional bivariate survival function is defined by

$$S(t_1, t_2|u) = \mathrm{E} S(t_1, t_2|u_1, u_2, Z_1, Z_2).$$

Although this approach is quite general, for simplicity we will specify the form of hazard functions and frailty distribution. Let $Y_0 \sim \Gamma(k_0, \mu)$, $Y_1 \sim \Gamma(k_1, \mu)$, and $Y_2 \sim \Gamma(k_1, \mu)$ be independent gamma distributed random variables with probability density given by

$$g(y) = \frac{\mu^k y^{k-1} e^{-\mu y}}{\Gamma(k)}$$

for some real positive k_0, k_1, μ. We define

$$Z_1 = Y_0 + Y_1 \sim \Gamma(k_0 + k_1, \mu)$$

$$Z_2 = Y_0 + Y_2 \sim \Gamma(k_0 + k_1, \mu)$$

with $\mathrm{E} Z_1 = \mathrm{E} Z_2 = 1$, $\mathrm{Var}(Z_1) = 1/\mu := \sigma^2$, $\mathrm{Var}(Z_2) = 1/\mu := \sigma^2$, $\mathrm{Corr}(Z_1, Z_2) := \rho$. It is easy to check that

$$k_0 = \rho/\sigma^2, k_1 = \frac{1-\rho}{\sigma^2}.$$

After a number of transformations we get (Wienke, 2011)

$$S(t_1, t_2|u_1, u_2) = \frac{S_1(t_1|u_1)^{1-\rho} S_1(t_2|u_2)^{1-\rho}}{(S_1(t_1|u_1)^{-\sigma^2} + S_1(t_2|u_2)^{-\sigma^2} - 1)^{\frac{\rho}{\sigma^2}}}$$

with

$$S_1(t|u) = (1 + \sigma^2 e^{\beta^* u} H_0(t))^{-1/\sigma^2}.$$

To make the model more flexible we will use a Gamma-Gompertz-Makeham representation of the univariate survival function according to

$$S_1(t|u = 0) = (1 + s^2(ct + (a/b)(e^{bt} - 1)))^{-1/s^2} = (1 + \sigma^2 H_0(t))^{-1/\sigma^2}$$

for some real positive a, b, s. From here we find that

$$H_0(t) = ((1 + s^2(ct + (a/b)(e^{bt} - 1)))^{\sigma^2/s^2} - 1)/\sigma^2.$$

Combining last equalities we get that

$$f(t|u) = \partial S_1(t|u)/\partial t = -S_1(t|u)(1 + \sigma^2 e^{\beta^* u} H_0(t))^{-1} e^{\beta^* u} \lambda_0(t),$$

where

$$\lambda_0(t) = (c + ae^{bt})(1 + s^2(ct + (a/b)(e^{b_t} - 1)))^{\sigma^2/s^2 - 1}.$$

The log-likelihood for the frailty-based bivariate survival model is given by

$$\begin{aligned} \ln L(Data; \theta) = \quad & (1 - \delta_{i1})(1 - \delta_{i2}) \textstyle\sum_i \ln S(t_{i1}, t_{i2}|u_1, u_2) \\ & -\delta_{i1}(1 - \delta_{i2}) \textstyle\sum_i \ln \left(\partial S(t_{i1}, t_{i2}|u_1, u_2)/\partial t_{i1}\right) \\ & -(1 - \delta_{i1})\delta_{i2} \textstyle\sum_i \ln \left(\partial S(t_{i1}, t_{i2}|u_1, u_2)/\partial t_{i2}\right) \\ & +\delta_{i1}\delta_{i2} \textstyle\sum_i \ln \left(\partial^2 S(t_{i1}, t_{i2}|u_1, u_2)/\partial t_{i1}\partial t_{i2}\right). \end{aligned}$$

Here $\theta = (a, b, s^2, \sigma^2, \rho, \beta^*)^*$ is the vector of unknown parameters
and

- $\delta_{i1} = \delta_{i2} = 0$ if both survival times t_{i1} and t_{i2} are censored,
- $\delta_{i1} = 1, \delta_{i2} = 0$ if survival time t_{i1} is uncensored and survival time t_{i2} is censored,
- $\delta_{i1} = 0, \delta_{i2} = 1$ if survival time t_{i1} is censored and survival time t_{i2} is uncensored,
- $\delta_{i1} = \delta_{i2} = 1$ if both survival times t_{i1} and t_{i2} are uncensored.

The expressions for $\partial S(t_{i1}, t_{i2}|u_1, u_2)/\partial t_{i1}$, $\partial S(t_{i1}, t_{i2}|u_1, u_2)/\partial t_{i2}$, and $S(t_{i1}, t_{i2}|u_1, u_2)/\partial t_{i1}\partial t_{i2}$ are given in Appendix A.

The formulas for the bivariate survival function and its first and second partial derivatives in the case when frailties Z_1 and Z_2 have different variances σ_1^2 and σ_2^2 are given in Wienke (2011). If the data is left truncated and both twins enter the population at risk at age $t_i^0 < t_i$ we modify the formulas dividing $S(t_{i1}, t_{i2}|u_1, u_2)$, $\partial S(t_{i1}, t_{i2}|u_1, u_2)/\partial t_{i1}$, $\partial S(t_{i1}, t_{i2}|u_1, u_2)\partial t_{i2}$, and $\partial^2 S(t_{i1}, t_{i2}|u_1, u_2)\partial t_{i1}\partial t_{i2}$ by $S(t_i^0, t_i^0|u_1, u_2)$.

4.2 Identifiability and estimation methods

Identifiability of the shared and the correlated frailty models using data on T_1 and T_2 was proved by Honoré (1993) under assumption of finite means of Z_1 and Z_2. Yashin and

Iachine (1999) proved the identifiability of the correlated frailty model without observed covariates under assumption that Z_1 and Z_2 are gamma-distributed.

In parametrical approach we get the estimates of unknown parameters by maximizing the log-likelihood function directly with respect to θ. An extended version of the EM algorithm was suggested for the analysis of bivariate survival data using the correlated frailty model (Iachine, 1995). This approach allows for estimating the frailty distribution parameters σ^2 and ρ, Cox regression coefficients β, and the hazard profile H.

Chapter 5

5 Genetic analysis of longevity

5.1 Quantitative genetic analysis

Heritability. The main goal of genetic analysis is to give an answer to the question: to what extent genetic variation may account for the variation of some trait, for example, longevity or susceptibility to death. To solve this problem we need to combine the measurement of the phenotype with genetic information. This information is usually available in the form of pedigrees if we study related individuals (sibs, families, twins). In genetic studies of twins the differences in life spans in MZ pair are caused by environmental factors, whereas DZ twins can demostrate less similarity in life spans due to effects associated with genetic differences.

Instead of studying life spans we can analyze frailties within the framework of correlated frailty model. This approach was proposed by Yashin and Iachinne (1995). We decompose the frailty into a sum of of independent genetic and environmental additive components

$$Z = A + D + I + C + E.$$

Here A, D and I are the additive, dominance, and epistasis genetic effects, and C, E stand for shared and non-shared environmental effects. Assuming independence of these components define the associated variance proportions

$$a^2 = \frac{Var(A)}{Var(Z)}, \quad d^2 = \frac{Var(D)}{Var(Z)}, \quad i^2 = \frac{Var(I)}{Var(Z)}, \quad c^2 = \frac{Var(C)}{Var(Z)}, \quad e^2 = \frac{Var(E)}{Var(Z)},$$

where

$$a^2 + d^2 + i^2 + c^2 + e^2 = 1.$$

The correlation coefficient between co-twin's frailties is

$$\rho = \rho_a a^2 + \rho_d d^2 + \rho_i i^2 + \rho_c c^2 + \rho_e e^2,$$

where $\rho_{(.)}$ are the correlations between respective frailty components (A, D, I, C, E) within a twin pair. Under common assumptions of quantitative genetics we get that

$$\rho_a = \rho_d = \rho_i = \rho_c = 1, \quad \rho_e = 0$$

for monozygotic twins, and

$$\rho_a = 0.5, \quad \rho_d = 0.25, \quad \rho_i = k, \quad \rho_c = 1, \quad \rho_e = 0$$

for dizygotic twins with unknown magnitude of epistasis effects k, $0 \leq k < 0.25$ (Neale and Cardon, 1992). All parameters a^2, d^2, i^2, c^2 cannot be found from this decomposition. It is only possible to conclude that the broad sense heritability H^2, characterizing the relative importance of genetic effects and defined as the percentage of variation of the trait (e.g. frailty) explained by the variation of genetic factors, is contained in an interval

$$\rho_{MZ} - \rho_{DZ} \leq H^2 \leq \min\{\rho_{MZ}, 2(\rho_{MZ} - \rho_{DZ})\},$$

where in the absence of epistasis this interval becomes narrower (Iachine, 2002). To make the model identifiable we can reduce it to an ACE model including only additive genetic, common environmental, and uncommon environmental components

$$\begin{aligned} 1 &= a^2 + c^2 + e^2 \\ \rho_{MZ} &= a^2 + c^2 \\ \rho_{DZ} &= 0.5a^2 + c^2 \end{aligned}$$

(Wienke, 2011).

Linkage. If loci are located at the same chromosome there is a chance that they are not transmitted independently between generations. The probability of a single chromosomal crossover between two genes during meiosis (recombination frequency) is a measure of genetic linkage parameter θ. This parameter lies in the interval $[0, 1/2]$ and varies from 0 (completely linked loci, no recombination) to 1/2 (unlinked loci, free recombination). Linkage allows for specifying relative position of genes on the chromosome. Different genetic map distances have been proposed to describe the distance between two genes

(Weir, 1996). A unit of map distance is a centimorgan that describes a recombination frequency of 0.01. The linkage procedure deals with longevity data and information on genetic markers and generally involves calculation of the conditional distribution of life span, the distribution of inheritance vector, and averaging of likelihood with respect to this distribution (Kruglyak and Lander, 1995). Morton suggested linkage analysis using logarithm of the odds (LOD) score (Ott, 1991).

5.2 A major gene model in twin studies

Frailty can include a large number of the risk factors - genetic and environmental. The influence of genes with highest impact on mortality or longevity can be modelled using the major gene approach. Assume that the frailty $Z(g)$ is a function of the genotype in a longevity locus with two alleles a and A and that an individual's instantaneous risk of death $\lambda(t)$ is proportional to the baseline hazard $\lambda_0(t)$, frailty $Z(g)$, and the term $\exp(\beta^* u)$ characterizing the influence of observed covariates u,

$$\lambda(t) = Z(g)\exp(\beta^* u)\lambda_0(t).$$

Specifying the dependency of frailty on the genotype let us assume, for example, that allele a is a beneficial one with multiplicative action $r < 1$, the frequency p and lies in autosomal longevity locus. In Hardy-Weinberg equilibrium the probability $P(g)$ of genotypes aa, $aA + Aa$, and AA is equal to p^2, $2p(1-p)$, and $(1-p)^2$, respectively. Moreover, for these genotypes it holds that $Z(aa) = r^2$, $Z(aA + Aa) = r$, and $Z(AA) = 1$. The monozygotic twins have the similar genotypes and, therefore, the same frailties. The frailties (and genotypes) of the dizygotic twins have less similarity but correlate, since both twins inherit their genes from the same parents. We assume that the genotypes of DZ twins are inherited independently from parents. The conditional probabilities of twins' genotypes can be computed using simple transmission model.

The bivariate survival function for a MZ (DZ) twin pair with survival times t_1 and t_2 for independently chosen parents can be calculated as the sum

$$\begin{aligned}
S_d^{MZ}(t_1,t_2|u_1,u_2) = &\ \textstyle\sum_{g_m,g_f} P(g_m)P(g_f)\sum_g P(g|g_m,g_f) \\
&\times \exp(-Z(g)(\exp(\beta^* u_1)H_0(t_1) + \exp(\beta^* u_2)H_0(t_2))), \\
S_d^{DZ}(t_1,t_2|u_1,u_2) = &\ \textstyle\sum_{g_m,g_f} P(g_m)P(g_f)\sum_g P(g|g_m,g_f)\exp(-Z(g)\exp(\beta^* u_1)H_0(t_1)) \\
&\times \textstyle\sum_g P(g|g_m,g_f)\exp(-Z(g)\exp(\beta^* u_2)H_0(t_2))
\end{aligned}$$

with cumulative hazard $H_0(t) = \int_0^t \lambda_0(\tau)d\tau$. The expressions for conditional genotype probabilities $P(g|g_m, g_f)$ given parental genotypes g_m (mother's genotype) and g_f (father's genotype) can be found in Appendix B. The conditional univariate probability density function

$$-dS(t|Z(g), u)/dt = Z(g)\exp(\beta^* u)\lambda_0(t)\exp(-Z(g)\exp(\beta^* u)H_0(t))$$

in the case of non-censored data and the conditional univariate survival function $S(t|Z(g), u)$ in the case of censored data can be viewed as the so-called penetrance functions in segregation analysis.

5.3 Mixed frailty model

In the mixed discrete-continuous frailty model we assume that the full frailty is a sum of two independent frailty components - discretely distributed frailty (for example, as in the major gene model) and continuously distributed frailty (for example, spread influence of a large number of genes)

$$Z = Z_d + Z_c.$$

Assuming the gamma distribution for continuous part of frailty and the major gene model for the discrete part of frailty we get the bivariate survival in the form

$$S(t_1, t_2|u_1, u_2) = S_d(t_1, t_2|u_1, u_2)S_c(t_1, t_2|u_1, u_2),$$

where

$$S_c(t_1, t_2|u_1, u_2) = S_1(t_1|u_1)^{1-\rho}S_1(t_2|u_2)^{1-\rho}(S_1(t_1|u_1)^{-\sigma^2} + S_1(t_2|u_2)^{-\sigma^2} - 1)^{-\rho/\sigma^2},$$

$\sigma^2 = \mathrm{Var}Z_c$, $\rho = \mathrm{Corr}(Z_{1,c}, Z_{2,c})$, $S_1(t|u)$ was defined in Section 4.1, and $S_d(t_1, t_2|u_1, u_2)$ was defined in previous section.

To compare three approaches - the major gene model, the gamma frailty model, and the mixed model we applied these models to the non-censored Danish twin data with no observed covariates (Begun et al., 2000a)[**B1**]. Marginal univariate survival functions were approximated using the Gamma-Gompertz-Makeham parameterization

$$\begin{aligned} S(t|u=0) = \quad & (1 + s^2(ct + (a/b)(e^{bt} - 1)))^{-1/s^2} = (1 + \sigma^2 H_0(t))^{-1/\sigma^2} \\ & \times \textstyle\sum_{g_m, g_f} P(g_m)P(g_f) \sum_g P(g|g_m, g_f)\exp(-Z(g)H_0(t)). \end{aligned}$$

For the data from the Danish Twin Registry analysis revealed a surprising degree of similarity between models with discrete and those with gamma-distributed frailties. The

similarity was expressed in the likeness of probability density functions, fits of marginal hazards, and maximum likelihood values for all populations we considered (MZ males, MZ females, DZ males, DZ females). The essential difference between the two models involves the behaviour of underlying hazards and the asymptotic behaviour or life-span correlations. But these differences are based on the nature of the frailty distributions. Genetic factors explain about 50% of the frailty variance of the continuous component of frailty. The beneficial allele is spread with a probability of approximately 0.5 and decreases mortality risk by a factor of about 3 for both sexes. Both the model with continuously distributed frailty and the one with discretely distributed frailty are nested in a mixed frailty model. But in accordance with the likelihood ratio test we cannot reject models with purely discrete and purely continuous frailties. This is probably due to the insufficient size of sample - in reality we must take into account both the dominant influence of a major gene as well as spread influence of a large number of genes.

5.4 A major gene model in family studies

In previous section we have used the approach based on the major gene model to analyze the data on twins. This approach can be easily adapted to the family data containing information about more than one generation and more than two members of the family. Assume that we have n sibs in a family. If in the sibship the monozygotic sibs are not contained we can write the n-variate survival function in a form

$$S_d(t_1,...,t_n|u_1,...,u_n) = \sum_{g_m,g_f} P(g_m)P(g_f)\prod_{i=1}^{n}\sum_{g_i} P(g_i|g_m,g_f)\exp(-Z(g_i)\exp(\beta^* u_i)H_0(t_i)).$$

Here t_i is a survival for i^{th} sib.

For the data on French-Canadians we built seven covariates; u_1=(year of birth - 1650), u_2=the age of a child at father's death, u_3=the age of a child at mother's death, u_4 the reproductive age of a father, u_5 the reproductive age of a mother, u_6 the life-span of a father, u_7 the life-span of a mother. The age of child at father's/mother's death were categorized as a follows: 0 if $u_{2,3} \leq 5$, 1 if $5 < u_{2,3} \leq 10$, 2 if $10 < u_{2,3} \leq 15$, 3 - otherwise. The values of the fourth and fifth covariates were put to 0 if $u_{4,5} \leq 35$ and 1 - otherwise. We assumed that the sixth and the seventh covariates were 0 if $u_{6,7} \leq 75$ and 1 otherwise (Begun et al., 2000b)[**B2**].

All estimates of unknown parameters were obtained through the maximization of the likelihood function. It was found that we can use the model with one beneficial allele in Hardy-Weinberg equilibrium with allele frequency 0.406 and multiplicative action 0.485. No cohort effect and no effect of age of a child at parental death have been found. Only two coefficients of Cox's regression were significant: $\beta_4 = \beta_5 = 0.188$ and $\beta_6 = \beta_7 = -0.451$. That is, we can find the beneficial allele in about 41% of cases and the presence of each beneficial allele in the genotype decreases the mortality by about 2.1 times. The greater a parent's life is, the less a child's mortality risk will be. On the contrary, the higher a reproduction age of a parent is, the greater the mortality risk will be. Summarizing these finding we conclude that familial environment may have profound effects not only on infant/childhood mortality, but also on adult mortality. The most important factors of this environment are the parental longevity and the parental reproduction age. The genetic material, which a parent transmits to its offspring might be essentially damaged in the reproductive age after 35, which leads to a shorter child's longevity. But the effect of the parental longevity is stronger. However, it does not mean that only genetic factors play the crucial role in child's longevity. Familial habits and the life-style can affect the life-span to some degree as well.

Chapter 6

6 Searching for genes contributing to longevity using association analysis

6.1 Gene frequency method

In previous section we studied genetic effects on longevity in the case when the genetic data were not available. Correlated frailty model allows us to calculate heritability as a total measure of genetic influence on the life span given longevity data for related individuals (twins, sibs, families). Non-zero heritability can indicate the presence of the non-zero genetic component in frailty. Alternatively we can use the segregation analysis based on the major gene model. Using different assumptions relating to the nature of genetic influence on longevity we can try to detect the existence of a gene with beneficial genotypes decreasing the risk of mortality. Since we assume that the genetic data are not available we average conditional survival function with respect to distribution of unobserved risk of mortality (frailty).

Another approach called the "gene frequency method" has been developed to study the gene-longevity associasions if information on genotype frequencies is available (De Benedictis et al., 1998; Tan et al., 2003; Garastro et al., 2003). In this approach the evaluation is based on the idea that a significant difference in the gene and allele frequencies in distinct age groups can indicate the presence of a genetic influence on life span. However, this method do not answer questions about the estimates of mortality trajectories and survival functions for populations of individuals carrying different genotypes. Most of extensions of the "gene frequency method" suggested later can be classified into 4 groups

- the "parametric method", the "semi-parametric method", the "nonparametric method", and the "relative risk method" (Yashin et al., 1999). In the "relative risk method" survival functions for different genotypes are related. Substantial improvements in the accuracy of statistical estimates based on the "relative risk method" can be achieved by using additional non-genetic data (Yashin et al., 2007). More flexible parameterization applied in a modification of the "relative risk method" relates cumulative hazards parameterically rather than survival functions for different genotypes (Begun, 2007; Begun, 2009)[**B3, B4**]. This method allows us to observe the so-called antagonistic pleiotropic effect when some genotype has an advantage only up to some age.

6.2 A modification of the relative risk model

Let denote a and A the longevity and non-longevity alleles with frequencies p_a and $p_A = 1-p_a$, respectively. In Hardy-Weinberg equilibrium the genotype frequencies are $P_{aa} = p_a^2$, $P_{aA+Aa} = 2(1-p_a)p_a$, and $P_{AA} = (1-p_a)^2$. Assume additionally that for any individual with genotype g ($g \in \{aa, aA + Aa, AA\}$) the risk of mortality at age t is defined by the model with proportional hazard

$$\lambda(t|Z_g) = Z_g \lambda_{0,g}(t),$$

where the gamma distributed random variable Z_g is the individual frailty with mean 1 and variance σ_g^2. The survival function $S_g(t)$ and the cumulative hazard function $H_g(t)$ for individuals with genotype g in this model are related via the equation

$$S_g(t) = (1 + \sigma_g^2 H_g(t))^{-1/\sigma_g^2}.$$

Assume additionally that the cumulative hazards for different genotypes can be parameterized by

$$H_g(t) = c_g t + a_g H_0(t)^{b_g}$$

for some non-negative a_g, b_g, and c_g, and unknown non-decreasing with age baseline cumulative hazard function $H_0(t)$. Without loss of generality we can put $a_{AA} = b_{AA} = 1$. The information about the birth cohort of individuals born in year x can be included in the model by logistic parameterization

$$p_a = 1 - 1/(1 + \exp(\gamma + \delta t + R\phi(t, t_0)))$$

for $t = T - x$. Here T is the year of data collection and the parameters R, γ, and δ are unknown. The value of $R\phi(t, t_0)$ characterizes the sudden change in the allele frequency in the birth cohort $T - t_0$ and the step function $\phi(t, t_0)$ is defined by the interval equations $\phi(t, t_0) = 1$ for $0 \leq t \leq t_0$ and $\phi(t, t_0) = 0$ for $t > t_0$. The value of $\gamma + \delta t$ stands here for the slow cohort effect of the allele frequency.

The genotype frequencies $\pi_g(t)$ and the allele frequency $\pi_a(t)$ for the whole population can be calculated from the formulas

$$\begin{aligned} \pi_g(t) = &\quad P_g S_g(t)/S(t), \\ \pi_a(t) = &\quad [\pi_{aa}(t) + 0.5\pi_{Aa+aA}(t)]/[\pi_{aa}(t) + \pi_{Aa+aA}(t) + \pi_{AA}(t)], \end{aligned}$$

where it holds for the survival function for whole population that

$$S(t) = \sum_g P_g S_g(t).$$

It is assumed that the survival function $S(t)$ is known. The vector of unknown parameters $\theta = (R, \delta, \gamma, a_{aa}, a_{aA+Aa}, b_{aa}, b_{aA+Aa}, c_{aa}, c_{aA+Aa}, c_{AA}, \sigma^2_{aa}, \sigma^2_{aA+Aa}, \sigma^2_{AA})^*$ can be estimated by maximizing the likelihood function

$$Lik = \prod_t \pi_{aa}(t, \theta)^{N_{aa}(t)} \pi_{aA+Aa}(t, \theta)^{N_{aA+Aa}(t)} \pi_{AA}(t, \theta)^{N_{AA}(t)},$$

where $N_g(t)$ is the observed number of individuals with genotype g at age t.

If the data on $S(t)$ in some age intervals are either unreliable or unavailable we can approximate the survival function using the Gamma-Gompertz-Makeham function

$$\tilde{S}(t) = (1 + s^2 \tilde{H}(t))^{-1/s^2})$$

with

$$\tilde{H}(t) = \tilde{c}t + \tilde{a}(\exp(\tilde{b}t) - 1)/\tilde{b}.$$

Unknown non-negative parameters $\tilde{a}$, $\tilde{b}$, $\tilde{c}$, and s^2 can be estimated over the age interval, where $S(t)$ is known. To choose the optimal model we can use the Likelihood Ratio Test for nested models, and either the Akaike Information Criterion or the Bayesian Information Criterion for non-nested models.

Example based on the simulated data showed that, given time of sudden change in allele frequency, survival in whole population, and numbers N_g, all unknown parameters can be identified. Moreover, the antagonistic pleiotropic effect can also be modelled using this approach. For example, the frequency of the beneficial allele increases up to some age

when the survival functions for individuals with beneficial allele and without beneficial allele become equal. Then this frequency falls drastically because after this age the individuals with beneficial allele do not have more advantage in survival (Begun, 2007)[**B3**].

6.3 A bivariate relative risk model

In previous section we discussed the "relative risk method" for independent individuals. This approach can be also adapted for detecting longevity genes for the data set consisting of twin pairs (Begun, 2009)[**B4**]. Using the same assumptions about univariate survivals and parametrical form for cumulative hazard functions $H_g(t)$ and for the frequency p_a of longevity allele a, assume additionally that the information on longevities for twins and their genotypes is also available. Suppose that the life spans of twins T_1 and T_2 are conditionally independent given frailties Z_1 and Z_2 and genotypes g_1 and g_2. If $\mathrm{Corr}(Z_1, Z_2) = \rho$, $\mathrm{Var}Z_1 = \mathrm{Var}Z_2 = \sigma^2$, and $\mathrm{E}Z_1 = \mathrm{E}Z_2 = 1$, then, analogous to the survival function corresponding to the continuous part of frailty in the mixed frailty model of Section 5.3, the bivariate survival function is

$$S(t_1, t_2|g_1, g_2) = P(T_1 \geq t_1, T_2 \geq t_2) = S_{g_1}^{1-\rho} S_{g_2}^{1-\rho} (S_{g_1}^{-\sigma^2}(t_1) + S_{g_2}^{-\sigma^2}(t_1) - 1)^{-\rho/\sigma^2}.$$

For univariate and bivariate survival function in the whole population it holds that

$$\begin{aligned} S(t) = &\quad \textstyle\sum_g P_g S_g(t) \\ S^{MZ}(t_1, t_2) = &\quad \textstyle\sum_g P_g S_{g,g}^{MZ}(t_1, t_2) \\ S^{DZ}(t_1, t_2) = &\quad \textstyle\sum_{g_1,g_2} P_{g_1,g_2}^{DZ} S_{g_1,g_2}^{DZ}(t_1, t_2). \end{aligned}$$

The formulas for P_{g_1,g_2}^{DZ} under assumption of independent transmission of the maternal and paternal alleles to offspring are given in Appendix C. The frequencies $\pi_g^{MZ}(t)$ and $\pi_{g_1,g_2}^{DZ}(t)$ of the genotype g and (g_1, g_2) at age t can be calculated from the formulas

$$\begin{aligned} \pi_g^{MZ}(t) &= P_g S_{g,g}^{MZ}(t,t)/S^{MZ}(t,t) \\ \pi_{g_1,g_2}^{DZ}(t) &= P_{g_1,g_2}^{DZ} S_{g_1,g_2}^{DZ}(t,t)/S^{DZ}(t,t). \end{aligned}$$

If the univariate survival in the whole population is known the vector of unknown parameters $\theta = (R, \delta, \gamma, a_{aa}, a_{aA+Aa}, b_{aa}, b_{aA+Aa}, c_{aa}, c_{aA+Aa}, c_{AA}, \sigma^2, \rho)^*$ can be estimated through maximizing the likelihood function

$$Lik = \prod_{i=1}^{N_g^{MZ}} \pi_{g_i}^{MZ}(t_i, \theta) \prod_{i=1}^{N_g^{DZ}} \pi_{g_{i_1},g_{i_2}}^{DZ}(t_i, \theta).$$

Here t_i is the age of twin pair i at the moment of data collection, N_g^{MZ} and N_g^{DZ} are the observed numbers of MZ and DZ twin pairs in the genetic data set, respectively. As in the case of independent individuals we choose the optimal model using the Likelihood Ratio Test for nested models, and either the Akaike Information Criterion or the Bayesian Information Criterion for non-nested models. Notice that under null hypothesis (no differences between cumulative hazard functions for different genotypes) we assume that $a_{aa} = a_{Aa} = b_{aa} = b_{Aa} = 1$ and $c_{aa} = c_{Aa} = c_{AA} = 0$. Significant deviation from this hypothesis can indicate the gene-longevity association.

To improve the accuracy of statistical estimates and to increase the power we can additionaly use the data on longevity of twins in population. This information can decrease the length of the confidence intervals for σ^2 and ρ. There is a win in statistical power when using the more robust univariate model compared to the bivariate model. Information on longevity included additionally in the dataset can also substantially increase the power (Begun, 2009)[**B4**].

Chapter 7

7 Searching for genes contributing to longevity using linkage analysis

7.1 Introduction

The frequency and the relative risk of mortality of a longevity allele a can be estimated using the major gene model and the data on related individuals such as twins and siblings. To locate the longevity allele in genome we need genetic markers data. The simplest way to find a location of the longevity genes is the standard technique involving a Cox-type proportional hazards univariate model, where markers are considered as observed covariates. If some coefficients of regression are significantly different from zero and all loci are in linkage disequilibrium (the alleles from loci do not occur independently in haplotypes), then respective genetic markers are involved in longevity determination. It could mean that longevity or frailty gene is located in the neighbourhood of respective genetic markers on the chromosome. But if loci are in linkage equilibrium this method will give us an information about the location of longevity or frailty gene only in the case when some genetic marker coincides with this gene. The advantage of bivariate and multivariate survival analysis with genetic markers is that they allow detection not only of the presence of longevity or frailty genes but also determine location of these genes on the chromosome, even if observed genetic markers are in linkage equilibrium (the alleles from loci occur independently in haplotypes).

The methods of linkage analysis usually use approaches based on the maximization of likelihood. Since the construction of the likelihood function involves elements related

to longevity (hazard function, parameters of frailty distribution) and to linkage (recombination distances between longevity locus and genetic markers), observed markers do not influence longevity directly, and position of the longevity gene does not influence the life span, a two-step procedure is useful. In the first step (segregation analysis) we estimate parameters characterizing the bivariate survival function without genetic markers. In the second step (linkage analysis) we determine the location of the longevity or frailty gene on chromosome. The first step involves the usual technique of segregation analysis for the bivariate survival data described above. The linkage analysis involves calculation of the distribution of inheritance vector data, then calculation of the conditional distribution of life span, and finally averaging of likelihood with respect to first distribution.

7.2 Construction of the bivariate survival function

For constructing the bivariate survival function we will use a major gene model with multiplicative action r, $0 \leq r < 1$, of longevity allele and frequency p. If longevity locus is autosomal and in Hardy-Weinberg equilibrium, and parents are independent individuals, the bivariate survival function for dizygotic twins can be calculated as follows (Begun et al., 2000b)[**B2**]

$$
\begin{aligned}
&S_{DZ}(t_1,t_2) = p^4 e^{-r^2 H_0(t_1) - r^2 H_0(t_2)} \\
&+p^3(1-p)(e^{-r^2 H_0(t_1)} + e^{-rH_0(t_1))}(e^{-r^2 H_0(t_2)} + e^{-rH_0(t_2)}) \\
&+p^2(1-p)^2(0.5e^{-r^2 H_0(t_1)} + e^{-rH_0(t_1)} + 0.5e^{-H_0(t_1)})(0.5e^{-r^2 H_0(t_2)} + e^{-rH_0(t_2)} + 0.5e^{-H_0(t_2)}) \\
&+p(1-p)^3(e^{-rH_0(t_1)} + e^{-H_0(t_1)})(e^{-rH_0(t_2)} + e^{-H_0(t_2)}) \\
&+2p^2(1-p)^2 e^{-rH_0(t_1) - rH_0(t_2)} + (1-p)^4 e^{-H_0(t_1) - H_0(t_2)}.
\end{aligned}
$$

To calculate the baseline cumulative hazard function $H_0(t)$ we use the Gamma-Gompertz-Makeham parameterization

$$S(t) = (1 + s^2 \tilde{H}(t))^{-1/s^2} = p^2 e^{-r^2 H_0(t)} + 2p(1-p)e^{-rH_0(t)} + (1-p)^2 e^{-H_0(t)},$$

where $\tilde{\mu}_0(t) = d\tilde{H}(t)/dt = c + a\exp(bt)$, non-negative parameters a, b, c, and s^2 are unknown. This form of parameterization approximates the univariate survival in ages after 30 very good. Notice that unknown cumulative hazard function $H_0(t)$ is the cumulative hazard for frailty equal to 1 and can be calculated using simple bisection procedure. For

construction of the likelihood function we can take into account possible censoring. Maximizing this likelihood function we find the vector of unknown parameters $(p, r, a, b, c, s^2)^*$.

7.3 Construction of the likelihood function for calculating the LOD score profile

In linkage analysis the notion of inheritance vector plays an important role. Let $\gamma_j = (\gamma_{j,1}, \gamma_{j,2}, \gamma_{j,3}, \gamma_{j,4})^*$ be the inheritance vector for the marker number j, $j = 1, ..., l$, with components zero or one. The first and third ones denote the alleles inherited from the mother for the first and for the second twin, respectively (0 if from the grandmother and 1 if from the grandfather). Analogously, the second and the fourth components stand for alleles inherited from the father. Altogether we have $2^4 = 16$ possible inheritance vectors for each locus.

Additionally we assume that

- parental genotypes are independent at each locus with known probabilities in markers' loci and unknown probabilities in longevity locus,
- the probability for each possible inheritance vector at the first step is equal to 1/16,
- probabilities of recombination between marker loci are known,
- observed markers are in linkage equilibrium,
- observed markers do not influence longevity directly,
- only the unobserved major gene may possess this property.

To calculate the probability $P_{g_1,g_2}(M_j)$ that a twin pair has joint genotype (g_1, g_2) in marker M_j we can use the hidden Markov chain algorithm (Lander and Green, 1987) based on the Markov property of a pair (γ_j, M_j). Marker M_j contains information about alleles inherited from the mother and the father by the first twin (the first and the third components) and by the second twin (the second and the fourth component). Because of Markov property of a pair (γ_j, M_j) the stepwise probability of a twin pair having the extended genotype (markers and longevity locus situated on the chromosome on unknown

distance between the observed markers) can be calculated. Then we multiply these probabilities by survival functions (or their derivatives for noncensored data) for observed life spans. The final likelihood is obtained by averaging the result with respect to all possible parental genotypes and twins' genotypes in the longevity locus. Detailed description of the likelihood construction can be found in (Begun and Yashin, 2004)[**B5**].

7.4 Location of the major gene

Although all markers are in linkage equilibrium and do not influence longevity directly the location of longevity gene may influence the joint distribution of the extended markers vector and the life span values for genetically related individuals. We can construct the LOD score profile for the longevity gene calculating the value of $\log_{10}(Likelihood) - \log_{10}(Likelihood_0)$, where $Likelihood_0$ is the value of the likelihood function when longevity gene is situated out of the chromosome (i.e. recombination probabilities of the longevity gene is equal to 0.5), for different positions of the longevity locus on the chromosome. The accuracy of the estimating unknown recombination probability of the longevity gene can be assessed using a specific support interval (Ott, 1991). This support interval must contain all the points where LOD score is higher than or equal to 3. For this interval the linkage is significant. On the contrary, all points where LOD score is less than or equal to -2 must be excluded.

The empirical survival and genetic markers data sets for 10 markers and 1000 DZ twin pairs have been simulated to study the influence of the risk of mortality on the LOD score profile (Begun and Yashin, 2004)[**B5**]. The smaller relative risk r, the greater the correlation coefficients between life spans of siblings. The clear peak of the LOD score profile has been observed only for small values of r. This peak was situated near the real position of the longevity gene. The possibility of the localization of the longevity gene in high degree depends on the action r of the longevity allele.

There are no any principle difficulties to apply this method to two different longevity genes situated on the same or on different chromosomes.

7.5 Joint analysis of bivariate competing risks survival times and genetic markers data

Assume now that there are several types of failure. Define the cause-specific hazard function for subject k from cluster i (twin pair, family or sibship) by

$$\lambda_{ikj}(t|u_{ik}, Z_{ikj}) =$$
$$= \lim_{\Delta t \to 0+} P(t \le T \le t + \Delta t, l_{ik} = j | T \ge t, u_{ik}, Z_{ikj})/\Delta t = \lambda_{0j}(t) Z_{ikj} \exp(\beta_j^* u_{ik}),$$

where l_{ik} is a type of failure, $j = 1, ..., L$, L is a number of the types of failure, Z_{ikj} is an individual frailty for the failure type j, $\lambda_{0j}(.)$ is the underlying cause-specific hazard function, β_j are cause-specific regression coefficients' vectors, and u_{ik} are the vectors of time-independent covariates. If only one of the failure types can occur, the full hazard function for a subject is defined by

$$\lambda_{ik}(t|u_{ik}, Z_{ik1}, ..., Z_{ikL}) = \sum_{j=1}^{L} \lambda_{0j}(t) Z_{ikj} \exp(\beta_j^* u_{ik}).$$

The frailties Z_{ikj} can correlate for subjects from the same cluster and for different types of failure. Dependency between subjects can be caused by the identical alleles transmitted from parents. The polygene inheritance of a failure can lead to the correlations between causes of death or onset.

Consider an example with twin pairs and two competing risks of death. Let two longevity alleles a and b with dominant action and frequencies p_a and p_b, respectivelly, are located in different loci on the same or different chromosomes. Neutral alleles A and B have frequencies $1-p_a$ and $1-p_b$, respectively, and correspond to frailties $Z_{ik1} = Z_{ik2} = 1$. Suppose that the presence of at least one longevity allele a in genotype decreases the risk of the type 1 failure by factor $r_1 < 1$ and the risk of the type 2 failure by factor $q_2 < 1$. Similarly, the presence of at least one longevity allele b in genotype decreases the risk of the type 2 failure by factor $r_2 < 1$ and the risk of the type 1 failure by factor $q_1 < 1$. If both longevity genes are in Hardy-Weinberg and linkage equilibrium and the action r_1, r_2, q_1, q_2 of the longevity genes does not depend on the location on the chromosome, the possible longevity genotypes have the frequencies and frailties given in Table 1 (Appendix D).

Denote the observed data in cluster (twin pair) i, $i = 1, ..., n$, by $(X_i, U_i, L_i, \delta_i)$. Here X_i stands for the time of failure (or censoring), U_i are the vectors of observed covariates, L_i are the types of failure ($L_i \in \{1, 2\}$), and δ_i is the censoring vector. Similarly to

the case with only one cause of death we estimate in the first step the parameters β (Cox regression coefficients), vector-parameter ω characterizing the underlying hazard functions, and parameters $\zeta = (p_a, p_b, r_1, q_1, r_2, q_2)$ characterizing the frailties by maximizing the likelihood function

$$L(X, L, |U, \delta, \beta, \omega, \zeta) =$$
$$= \prod_{i=1}^{n} \mathrm{E}_Z \left(\prod_{k=1}^{2} \exp\left(- \sum_{j=1}^{2} Z_{ikj} e^{\beta_j^* u_{ik}} H_{0j}(X_{ik}|\omega) \right) \left(\lambda_{0L_{ik}}(X_{ik}|\omega) Z_{ikL_{ik}} e^{\beta_{L_{ik}^*} u_{ik}} \right)^{\delta_{ik}} \right).$$

Here $H_{0j}(.)$ is the cause-specific cumulutive hazard and E_Z denotes expectation with respect to distribution of frailty Z.

In the second step we calculate the one- or two-dimensional LOD-score profile for two longevity genes located on the same or different chromosomes (Begun, 2013)[**B6**].

Chapter 8

8 Experiments with microarrays in twin studies

8.1 Main concepts in microarray analysis

The aim of the gene expression profiling is to measure the activity of thousands of genes (the gene expression) at once for creating a picture of cellular function. These profiles can help to compare the gene expression in cells or tissues from two or more experimental conditions (e.g. affected and non-affected patients). Genes contain the information for producing the messenger RNA (mRNA), but not all genes at any moment make mRNA. Whether a gene is "on" (active) or "of" (non-active) depends on many factors such as disease status, sex, local environment, and so on. Usually only a small fraction of genes are differently expressed under changed experimental conditions. A DNA microarray itself is a collection of microscopic spots on a solid surface. There are different technologies for producing the microarray chips. The first experiments with modern form of DNA chips were reported in 1990s.

Analysis of the microarray data consists of a number of steps. The poor-quality and low-intensity features are removed by image analysis. The noise (background) is substracted, then the intensities ratios in the log-scale are calculated and normalized by the data processing step. Then, the differentially expressed genes are detected using statistical tests. Finally, the gene sets (functionally related groups of genes) analysis estimates significance of gene sets.

8.2 Detecting differential gene expression

The classic task in microarray analysis is to detect differential gene expression between two conditions. This may provide useful information on important biological processes or functions and involves not only determining whether there are any differentially expressed genes but also finding those genes with differential expression. The methods based on a modification of the t statistic are mostly used to find genes that are differentially expressed under two conditions. All these methods involve testing a null hypothesis H_0 that there is no differential expression and calculating a test statistic Z for each gene. To estimate the null distribution of the statistic Z, it is auxiliary to construct a suitable null statistic Z^0 whose distribution is the same as the null distribution of Z. This distribution of Z^0 can then be used to detertmine cut-off points. Genes whose Z exceed these cut-off points are claimed significant.

Three non-parametric methods (methods independent on strong parametric assumptions) are often used for detecting differential gene expression. The procedures based on the permutations of arrays and computation order statistics are implemented in the significant analysis of microarray (SAM) (Tusher et al., 2001). The implementation of a t statistic with a Bayesian adjusted denominator is used in the empirical Bayes (EB) method (Efron et al., 2000,2001; Smyth, 2004). The mixture model method (MMM) uses specific permutation of arrays followed by obtaining the sample of Z^0-distributed random variables and fitting a finite normal mixture model to this data (Pan et., al., 2002; Pan, 2003).

As for real data we usually have genes with differential expression, the null distribution Z^0 obtained using the permutation procedure suffers under the overdispersion of the null distribution. Another problem arises if we use data from related individuals such as twins. Twin data are very interesting in genetic studies because of the minimal influence of environment confounders and the absence of the ascertment bias. Correlations between traits of twins can bring distortions into statistical tests. This can lead to wrong inferences, to falsely estimated number of differentially expressed genes, and to uncontrolled changes of the power. It is possible to apply the EB method to gene expression data with correlated replications (the within-array correlation method). A robust method for

treating gene expression data for twins has also been proposed (Begun, 2006)[**B7**]. This method allows avoiding the possible errors in finding differentially expressed genes for twin data and involves usage of a special version of the t statistic, taking into account possible correlations between twins. Null distribution is estimated in this method similarly to the mixture model method.

8.3 Model description for non-related individuals

The univariate MMM aproach is based upon the following assumptions. Let Y_{jk} be the expression level of gene j in array k, $j = 1, ..., G$, $k = 1, ..., K_1, K_1 + 1, ..., K_1 + K_2$. Here, the first K_1 and the last K_2 arrays are obtained under the two conditions, respectively. Assume that

$$Y_{jk} = a_j + b_j x_k + \sigma_j \varepsilon_{jk},$$

where ε_{jk} are independent, identically and symmetrically distributed random errors with mean 0 and the variance 1, $x_k = 1$ for $1 \leq k \leq K_1$ and $x_k = 0$ for $K_1 + 1 \leq k \leq K_1 + K_2$. Detemining whether a gene has differential expression is equivalent to testing for the null hypothesis $H_0 : b_j = 0$ against $H_1 : b_j \neq 0$. To detect differential gene expression for non-related individuals the four-sample equal-variance t-statistic Z_{1j} can be used. This statistic is defined by

$$Z_{1j} = \frac{\hat{Y}_j^{11} + \hat{Y}_j^{12} - \hat{Y}_j^{21} - \hat{Y}_j^{22}}{\sqrt{(1/K_{11} + 1/K_{12} + 1/K_{21} + 1/K_{22})S_{1j}^2}}$$

with $\hat{Y}_j^{11}$, $\hat{Y}_j^{12}$, $\hat{Y}_j^{21}$, $\hat{Y}_j^{22}$, and the pooled sample variance $\hat{S}_{1j}^2$ given in Appendix E.

The null statistic is

$$Z_{1j}^0 = \frac{\hat{Y}_j^{11} - \hat{Y}_j^{12} - \left(\hat{Y}_j^{21} - \hat{Y}_j^{22}\right)}{\sqrt{(1/K_{11} + 1/K_{12} + 1/K_{21} + 1/K_{22})S_{1j}^2}}.$$

Since ε_{jk} are symmetrically distributed, Z_{1j}^0 has the same distribution as Z_{1j} under H_0.

In MMM the probability density of null distribution is approximated by the normal mixture model with g_{01} components

$$f_1(z; \Theta_{g_{01}}) = \sum_{i=1}^{g_{01}} \pi_{1i} \phi(z; \mu_{1i}, V_{1i}),$$

where π_{1i} are the mixing proportions and $\phi(z;\mu_{1i},V_{1i})$ denotes the normal density function with mean μ_{1i} and variance V_{1i}, and $\Theta_{g_{01}}$ is the vector of parameters (μ_{1i},V_{1i}), $i=1,...,g_{01}$. The approximate function $f_1(z;\Theta_{g_{01}})$ allow us to determine the cut-off points for a given Type I error. The unknown parameters of the mixture model can be estimated using the EM algorithm (McLachlan and Basford, 1988). This algorithm is realized in the Fortran program EMMIX, which is freely available on the web http://maths.uk.edu.au/ gjm/emmix//emmix.html.

8.4 Model description for twins

Let $Y_{jk,1}$ and $Y_{jk,2}$ be the expression levels of gene j in array k for the first and the second twin, respectively, with $j=1,...,G$ and $k=1,...,n$. Assume that

$$\begin{aligned} Y_{jk,1} &= a_j + b_j x_k + \sigma_{j,0}\varepsilon_{jk,0} + \sigma_{j,1}\varepsilon_{jk,1}, \\ Y_{jk,2} &= a_j + b_j x_k + \sigma_{j,0}\varepsilon_{jk,0} + \sigma_{j,1}\varepsilon_{jk,2}, \end{aligned}$$

where $\varepsilon_{jk,0}$, $\varepsilon_{jk,1}$, $\varepsilon_{jk,2}$, are independent, identically and symmetrically distributed random errors with mean 0 and the variance 1, $x_k=1$ under condition 1 and $x_k=0$ under condition 2. It is easy to show that the correlation between $Y_{jk,1}$ and $Y_{jk,2}$ for gene j is defined by the formula

$$\rho_j = \sigma^2_{j,0}/(\sigma^2_{j,0}+\sigma^2_{j,1}).$$

As above we test the null hypothesis $H_0: b_j=0$ against $H_1: b_j\neq 0$ to determine whether the gene j has differential expression.

We divide all the twin pairs into three groups with respect to their concordance-discordance status. Similarly to the case of non-related individuals we define statistic Z_{2j} and the null statistic Z^0_{2j} for detecting differential gene expression (Begun, 2006)**[B7]**

$$\begin{aligned} Z_{2j} &= \frac{\bar{Y}^{11}_j+\bar{Y}^{12}_j-\bar{Y}^{21}_j-\bar{Y}^{22}_j}{S^0_{2j}} \\ Z^0_{2j} &= \frac{\bar{Y}^{11}_j-\bar{Y}^{12}_j-(\bar{Y}^{21}_j-\bar{Y}^{22}_j)}{S^0_{2j}} \end{aligned}$$

(for notations in these formulas see Appendix F).

Theorem 1 *If $\varepsilon_{jk,0}$, $\varepsilon_{jk,1}$, $\varepsilon_{jk,2}$, are independent, identically and symmetrically distributed random variables the null statistic Z^0_{2j} has the same distribution as Z_{2j} under H_0 for all $j=1,...,G$.*

The proof of this Theorem is given in (Begun, 2006)[**B7**].

As in the case of non-related individuals we approximate the probability density of null distribution by the normal mixture model with g_{02} components

$$f_2(z;\Theta_{g_{02}}) = \sum_{i=1}^{g_{02}} \pi_{2i}\phi(z;\mu_{2i},V_{2i}),$$

where π_{2i} are the mixing proportions and $\phi(z;\mu_{2i},V_{2i})$ denotes the normal density function with mean μ_{2i} and variance V_{2i}, and $\Theta_{g_{02}}$ is the vector of parameters (μ_{2i},V_{2i}), $i = 1,...,g_{02}$.

8.5 Comparison of the uni- and bivariate approaches

The approximate function $f_i(z;\Theta_{g_{0i}})$, $i = 1,2$, allows us to determine the cut-off points c_{down} and c_{upper} such that the type I error rate α is

$$\alpha = 1 - \int_{c_{down}}^{c_{upper}} f_i(z;\Theta_{g_{0i}})dz \approx p/G,$$

where p is the genome-wide significance level. As the functions $f_i(z)$ should be symmetric about zero, and we are usually interested in both up- and down-regulated genes, we can put $|c_{down}| = c_{upper} = C_\alpha$. Given α and f_i, the value of C_α can be estimated using a simple bisection procedure.

Let ρ_e and ρ_n are coefficients of correlation for differentially and non-differentially expressed genes. Simulation studies for 1000 twin pairs shows that the method for non-related individuals gives approximately the same results as for twins if $\rho_e = \rho_n$. This method finds more (less) false positives for concordant (discordant) twins if $\rho_e > \rho_n$. On the contrary, the method for non-related individuals finds less (more) false positives for concordant (discordant) twins if $\rho_e < \rho_n$. The greater the difference between ρ_e and ρ_n, the greater this effect. In general, results depend on the difference between ρ_e and ρ_n and the concordance-discordance status of twins.

8.6 Power estimation

In (Begun, 2008)[**B8**] it was shown that the standard deviation Ω_j of $\bar{Y}_j^{11}+\bar{Y}_j^{12}-\bar{Y}_j^{21}-\bar{Y}_j^{22}$ can be written in a form

$$\Omega_j(\sigma_j,\rho_j)=\sigma_j\sqrt{\sum_{p,q=1}^{2}\frac{1}{n_{pq}^2}\left(n_{pq}+2K_{pq}^c\rho_j\right)-\sum_{q=1}^{2}\frac{K_q^d\rho_j}{n_{1q}n_{2q}}}$$

with $\sigma_j^2=\sigma_{j,0}^2+\sigma_{j,1}^2$ (for notations in this formula see Appendix F). Note that statistic Z_{2j} differs from statistic Z_{2j}^0 by the value $d_j=2b_j/S_{2j}^0$. Replacing S_{2j}^0 with $\Omega_j(\sigma_j,\rho_j)$ in formula for d_j and taking into account that Z_{2j} under H_0 has the same distribution as Z_{2j}^0, we can estimate the power function $\beta(d_j,\alpha)$ using approximate formula

$$\hat{\beta}(\tilde{d}_j,\alpha)=\int\limits_{-\infty}^{\tilde{d}_j-C_\alpha}f_2(z;\Theta_{g02})dz+\int\limits_{\tilde{d}_j+C_\alpha}^{\infty}f_2(z;\Theta_{g02})dz$$

with $\tilde{d}_j=2b_j/\Omega_j(\sigma_j,\rho_j)$. Simulation studies based on the sample with 1000 twin pairs shows that there is a clear difference in the behaviour of the power function for concordant and discordant twin pairs. The power increases with b_j/σ_j and decreases with ρ_j for concordant twin pairs. For discordant twins, the power increases with both b_j/σ_j and ρ_j. It was not surprising that the power increases with sample size. Discordant twins are more informative than concordant ones and give more power for the same values of b_j/σ_j and ρ_j.

Chapter 9

9 Conclusions

Including frailty in traditional survival analysis extends our possibilities to analyze the longevity data and to interpret results of the studies. The notion of frailty as a random non-observed risk of mortality complements the notion of the hazard function and describes additional source of randomness and variability in mortality process. The third source of variability in mortality are observed covariates. An assumption about conditional independence of mortality risks for related individuals given individual frailties makes it possible to consider genetic factors as the components of frailty and to study the genetic influence on mortality using the traditional methods of quantitative genetics.

The frailty models can be useful in locating of genes contributing to longevity. This is possible if additional genetic information such as genetic markers data are also available. Thereby, searching for parameters relating to frailty and hazard and for parameters defining the location of longevity genes can be carried out separately because longevity does not depend on the position of genes at the chromosome.

Apart from clear advantages of frailty modelling there are some problems and limitations in using the frailty component in survival analysis. It is not easy to adapt the approach based on the correlated frailty model with continuous frailty to groups with more than two related individuals such as family (in the models with discretely distributed frailty this problem can be easily overcome). Secondly, the form of the frailty distribution used in survival analysis may vary and interpretation of results can depend on this form.

Other limitations of frailty modelling relate to the proportional structure of the hazards, assumptions about independence of observed covariates, and about fixed (not chang-

ing with age) frailty. Approaches free of these limitations need to be developed.

Special methods for detecting differentially expressed genes under two or more conditions having been intensively developed in last years. A number approaches and software tools were successfully used by analysis of the data in experiments with microarrays. Such methods include preliminary treatment of the microarray data followed by statistical tests locating differentially expressed genes, and the gene sets analysis. Correlations between individuals may markedly influence the results of search of such genes and the power. To avoid the bias in studies and incorrectness in conclusions we must take into account possiible genetic and environmental similarities between individuals.

Appendix A

$$
\begin{aligned}
S(t_1,t_2|u_1,u_2) &= S_1(t_1|u_1)^{1-\rho}S_1(t_2|u_2)^{1-\rho}(S_1(t_1|u_1)^{-\sigma^2}+S_1(t_2|u_2)^{-\sigma^2}-1)^{-\rho/\sigma^2}
\end{aligned}
$$

$$
\begin{aligned}
\partial S(t_1,t_2|u_1,u_2)/\partial t_1 &= f_1(t_1|u_1)S_1(t_1|u_1)^{-\rho}S_2(t_2|u_2)^{1-\rho} \\
&\times\left[S_1(t_1|u_1)^{-\sigma^2}+S_1(t_2|u_2)^{-\sigma^2}-1\right]^{-\rho/\sigma^2-1} \\
&\times\left[-S_1(t_1|u_1)^{-\sigma^2}-(1-\rho)S_2(t_2|u_2)^{-\sigma^2}+1-\rho\right]
\end{aligned}
$$

$$
\begin{aligned}
\partial S(t_1,t_2|u_1,u_2)/\partial t_2 &= f_2(t_2|u_2)S_1(t_1|u_1)^{1-\rho}S_2(t_2|u_2)^{-\rho} \\
&\times\left[S_1(t_1|u_1)^{-\sigma^2}+S_1(t_2|u_2)^{-\sigma^2}-1\right]^{-\rho/\sigma^2-1} \\
&\times\left[-S_2(t_2|u_2)^{-\sigma^2}-(1-\rho)S_1(t_1|u_1)^{-\sigma^2}+1-\rho\right]
\end{aligned}
$$

$$
\begin{aligned}
\partial^2 S(t_1,t_2|u_1,u_2)/\partial t_1\partial t_2 &= f_1(t_1|u_1)f_2(t_2|u_2)S_1(t_1|u_1)^{-\rho}S_2(t_2|u_2)^{-\rho} \\
&\times\left[S_1(t_1|u_1)^{-\sigma^2}+S_1(t_2|u_2)^{-\sigma^2}-1\right]^{-\rho/\sigma^2-2} \\
&\times((1-\rho)^2\left[S_1(t_1|u_1)^{-\sigma^2}+S_2(t_2|u_2)^{-\sigma^2}-1\right]^2 \\
&+\rho(1-\rho)S_2(t_2|u_2)^{-\sigma^2}\left[S_1(t_1|u_1)^{-\sigma^2}+S_2(t_2|u_2)^{-\sigma^2}-1\right] \\
&+\rho(1-\rho)S_1(t_1|u_1)^{-\sigma^2}\left[S_1(t_1|u_1)^{-\sigma^2}+S_2(t_2|u_2)^{-\sigma^2}-1\right] \\
&+\rho(\rho+\sigma^2)S_1(t_1|u_1)^{-\sigma^2}S_2(t_2|u_2)^{-\sigma^2})
\end{aligned}
$$

Here $f_i(t_i|u_i)=\partial S_i(t_i|u_i)/\partial t_i$, $i=1,2$.

Appendix B

In the major gene model with one beneficial allele a in autosomal locus assume that the first allele of the offspring is inherited from the mother and the second one from the father. Altogether we have 4 genotypes: aa, aA, Aa, and AA. The conditional genotype frequencies of an offspring given parental genotypes can be calculated from formulas

$$
\begin{aligned}
P(g = aa | g_m = aa, g_f = aa) &= 1 \\
P(g = aa | g_m = aa, g_f = aA + Aa) &= 0.5 \\
P(g = aA | g_m = aa, g_f = aA + Aa) &= 0.5 \\
P(g = aA | g_m = aa, g_f = AA) &= 1 \\
P(g = aa | g_m = aA + Aa, g_f = aa) &= 0.5 \\
P(g = Aa | g_m = aA + Aa, g_f = aa) &= 0.5 \\
P(g = aa | g_m = aA + Aa, g_f = aA + Aa) &= 0.25 \\
P(g = AA | g_m = aA + Aa, g_f = aA + Aa) &= 0.25 \\
P(g = aA + Aa | g_m = aA + Aa, g_f = aA + Aa) &= 0.5 \\
P(g = Aa | g_m = AA, g_f = aa) &= 1 \\
P(g = Aa | g_m = AA, g_f = aA + Aa) &= 0.5 \\
P(g = AA | g_m = AA, g_f = aA + Aa) &= 0.5 \\
P(g = AA | g_m = AA, g_f = AA) &= 1
\end{aligned}
$$

Appendix C

$$
\begin{aligned}
P^{DZ}_{aa,aa} &= P^2_{aa} + (1/2)P_{aa}P_{aA+Aa} + (1/16)P^2_{aA+Aa} \\
P^{DZ}_{aa,aA+Aa} &= (1/2)P_{aa}P_{aA+Aa} + (1/8)P^2_{aA+Aa} \\
P^{DZ}_{aA+Aa,aa} &= (1/2)P_{aa}P_{aA+Aa} + (1/8)P^2_{aA+Aa} \\
P^{DZ}_{aa,AA} &= (1/16)P^2_{aA+Aa} \\
P^{DZ}_{AA,aa} &= (1/16)P^2_{aA+Aa} \\
P^{DZ}_{aA+Aa,AA} &= (1/2)P_{AA}P_{aA+Aa} + (1/8)P^2_{aA+Aa} \\
P^{DZ}_{AA,aA+Aa} &= (1/2)P_{AA}P_{aA+Aa} + (1/8)P^2_{aA+Aa} \\
P^{DZ}_{aA+Aa,aA+Aa} &= (1/2)P_{aa}P_{aA+Aa} + 2P_{aa}P_{AA} + (1/2)P_{AA}P_{aA+Aa} + (1/4)P^2_{aA+Aa} \\
P^{DZ}_{AA,AA} &= P^2_{AA} + (1/2)P_{AA}P_{aA+Aa} + (1/16)P^2_{aA+Aa},
\end{aligned}
$$

where $P_{aa} = p_a^2$, $P_{aA+Aa} = 2p_a(1-p_a)$, and $P_{AA} = (1-p_a)^2$.

Appendix D

Table 1: Frailties and genotype frequencies for two dominant longevity genes.

Z_{ik1}	Z_{ik2}	Genotype	Frequency
r_1q_1	r_2q_2	$(aa + Aa + aA) \times (bb + bB + Bb)$	$(1 - (1 - p_a)^2)(1 - (1 - p_b)^2)$
r_1	q_2	$(aa + Aa + aA) \times BB$	$(1 - (1 - p_a)^2)(1 - p_b)^2$
q_1	r_2	$AA \times (bb + bB + Bb)$	$(1 - p_a)^2(1 - (1 - p_b)^2)$
1	1	$AA \times BB$	$(1 - p_a)^2(1 - p_b)^2$

Appendix E

$$Z_{1j} = \frac{\hat{Y}_j^{11} + \hat{Y}_j^{12} - \hat{Y}_j^{21} - \hat{Y}_j^{22}}{\sqrt{(1/K_{11} + 1/K_{12} + 1/K_{21} + 1/K_{22})S_{1j}^2}}$$

where

$$\hat{Y}_j^{11} = \sum_{k=1}^{K_{11}} \frac{Y_{jk}}{K_{11}}; \qquad \hat{Y}_j^{12} = \sum_{k=K_{11}+1}^{K_1} \frac{Y_{jk}}{K_{12}};$$

$$\hat{Y}_j^{21} = \sum_{k=K_1+1}^{K_1+K_{21}} \frac{Y_{jk}}{K_{21}}; \quad \hat{Y}_j^{22} = \sum_{k=K_1+K_{21}+1}^{K_1+K_2} \frac{Y_{jk}}{K_{22}}$$

$$S_{1j}^2 = \frac{\left[\sum_{k=1}^{K_{11}} (Y_{jk} - \hat{Y}_j^{11})^2 + \sum_{k=K_{11}+1}^{K_1} (Y_{jk} - \hat{Y}_j^{12})^2 + \sum_{k=K_1+1}^{K_1+K_{21}} (Y_{jk} - \hat{Y}_j^{21})^2 + \sum_{k=K_1+K_{21}+1}^{K_1+K_2} (Y_{jk} - \hat{Y}_j^{22})^2\right]}{K_1 + K_2 - 4}$$

$$K_{11} = [K_1/2]; \quad K_{12} = K_1 - [K_1/2];$$
$$K_{21} = [K_2/2]; \quad K_{22} = K_2 - [K_2/2];$$

and $[x]$ is the integer part of x.

Appendix F

Denote the groups of twin pairs according their concordance-discordance status by M_1^c, M_2^c, and M^d with the number of pairs equal to l_1, l_2, and l_{12}, respectively. The first group for concordant twins under the first condition, the second group for concordant twins under the second condition, and the third group for discordance twins with the first twin under the first condition and the second twin under the second condition. The full number of twin pairs is equal to $n = l_1 + l_2 + l_{12}$. Denote

$$\begin{aligned} K_{11}^c &= [l_1/2]\,; & K_{12}^c &= l_1 - [l_1/2]\,; \\ K_{21}^c &= [l_2/2]\,; & K_{22}^c &= l_2 - [l_2/2]\,; \\ K_1^d &= l_{12} - [l_{12}/2]\,; & K_2^d &= [l_2/2]\,; \end{aligned}$$

We divide arbitrary groups M_1^c, M_2^c into four groups - M_{11}^c, M_{12}^c, M_{21}^c, and M_{22}^c with number of pairs equal to K_{11}^c, K_{12}^c, K_{21}^c, and K_{22}^c, respectively, and the group M^d into two groups M_1^d and M_2^d with the number of pairs equal to K_1^d, and K_2^d, respectively. Then we form new groups M_{11}, M_{12}, M_{21}, and M_{22} as follows. The group M_{11} includes all twins from M_{11}^c and all first twins from the group M_1^d. The group M_{12} includes all twins from M_{12}^c and all first twins from the group M_2^d. The group M_{21} includes all twins from M_{21}^c and all second twins from the group M_1^d. The group M_{22} includes all twins from M_{22}^c and all second twins from the group M_2^d. The number of individuals in the groups M_{11}, M_{12}, M_{21}, and M_{22} is equal to $n_{11} = 2K_{11}^c + K_1^d$, $n_{12} = 2K_{12}^c + K_2^d$, $n_{21} = 2K_{21}^c + K_1^d$, and $n_{22} = 2K_{22}^c + K_2^d$, respectively. It holds that $n_{11} + n_{12} + n_{21} + n_{22} = 2n$. Finally we define $\bar{Y}_j^{pq}$, $p, q = 1, 2$ from the formula

$$\bar{Y}_j^{pq} = \frac{\sum\limits_{k \in M_{pq}^c} (Y_{jk,1} + Y_{jk,2}) + \sum\limits_{k \in M_q^d} Y_{jk,p}}{n_{pq}}.$$

The value of S_{2j}^0 is an estimate of the standard deviation of the numerator of Z_{2j}

$$\bar{Y}_j^{11} + \bar{Y}_j^{12} - \bar{Y}_j^{21} - \bar{Y}_j^{22}$$

and can be found in (Begun, 2006)[**B7**].

Articles included in the thesis

B1 Begun, A., Iachine, I. A. and Yashin, A. (2000) Genetic nature of individual frailty: comparison of two approaches. *Twin Research,* ***3***, pp.51-57.

B2 Begun, A., Desjardins, B., Iachine, I. A. and Yashin, A. I. (2000) Multivariate Frailty Model with a Major Gene: Application to Genealogical Data. *Medical Infobahn for Europe: Proceedings of Mie2000 and GMDS2000 (Studies in Health Technology and Informatics),* ***v.77***, pp.412-417.

B3 Begun, A. (2007) A Modification of the Relative Risk Model with Heterogeneity Component for Detecting Genes Contributing to Longevity. *Annals of Human Genetics,* ***72***, pp.111-114.

B4 Begun, A. (2009) Detecting Genes Contributing to Longevity using twin data. *Human Genomics,* ***4(2)***, pp.73-78.

B5 Begun, A. and Yashin, A. I. (2004) Genetic Markers Data in Survival Studies of Twins: The Results of a Simulation Study. *Twin Research and Human Genetics,* ***8***, pp.34-38.

B6 Begun, A. (2013) Joint analysis of bivariate competing risks survival times and genetic markers data. *Journal of Human Genetics,* ***58(10)***, pp.694-699.

B7 Begun, A. (2006) Robust method for detecting differential gene expression in twin studies. *Bioinformatics,* ***22(23)***, pp.2905-2909.

B8 Begun, A. (2008) Power estimation of the t test for detecting differential gene expression. *Functional Integrative Genomics,* ***8***, pp.109-113.

B9 Begun, A. (2006) Age regularities of the mortality of multicellular organisms. *Physica A,* ***360***, pp.401-421.

References

Abernethy, J.D. (1998) Gompertzian mortality originates in the winding-down of the mitotic clock. *J. Theor. Biol., **192,*** pp.419-435.

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993) Statistical Models Based on Counting Processes. New York: Springer Verlag.

Begun, A., Iachine, I. A. and Yashin, A. (2000a) Genetic nature of individual frailty: comparison of two approaches. *Twin Research*, **3**, pp.51-57.

Begun, A., Desjardins, B., Iachine, I. A. and Yashin, A. (2000b) Multivariate Frailty Model with a Major Gene: Application to Genealogical Data. *Medical Infobahn for Europe: Proceedings of Mie2000 and GMDS2000 (Studies in Health Technology and Informatics)*, **v.77**, pp.412-417.

Begun, A. and Yashin, A. I. (2004) Genetic Markers Data in Survival Studies of Twins: The Results of a Simulation Study. *Twin Research and Human Genetics*, **8**, pp.34-38.

Begun, A. (2006) Age regularities of the mortality of multicellular organisms. *Physica A, **360***, pp.401-421.

Begun, A. (2006) Robust method for detecting differential gene expression in twin studies. *Bioinformatics*, **22(23)**, pp.2905-2909.

Begun, A. (2007) A Modification of the Relative Risk Model with Heterogeneity Component for Detecting Genes Contributing to Longevity. *Annals of Human Genetics*, **72**, pp.111-114.

Begun, A. (2008) Power estimation of the t test for detecting differential gene expression. *Functional Integrative Genomics*, **8**, pp.109-113.

Begun, A. (2009) Detecting Genes Contributing to Longevity using twin data. *Human Gennomics*, **4(2)**, pp.73-78.

Begun, A. (2013) Joint analysis of bivariate competing risks survival times and genetic markers data. *Journal of Human Genetics, **58(10)***, pp.694-699.

Bouchard, G. (1989) Population Studies and Genetic Epidemiology in Northeast Quebec. *Canadian Studies in Population. **16(1)***, pp.61-86.

Cox, D.R., Oakes, D. (1984) Analysis of Survival Data. *Chapman and Hall.*

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, A.I., Bonafe, M. and Franchesci, C. (1998) Gene/longevity association studies at

four autosomal loci (REN, THO, PARP, SOD2). *Eur J Hum Genet* ***6***, pp.534-541.

Efron, B., Tibshirani, R., Goss, V. and Chu, G. (2000) Microarrays and their use in a comparative experiment. *Technical Report* ***37B/213****, Statistical Department, Stanford University.*

Efron, B., Tibshirani, R., Storey, G. D. and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.,* ***96,*** pp.1151-1160.

Elbers, C. and Ridder, G. (1982) True and spurious duration dependence: The identifiability of the proportional hazard model. *Review of Economic Studies,* ***49***, pp.403-409.

Garasto, S., Rose, G., Derango, F., Berardelli, M., Corsonello, A., Feraco, F., Mari, V., Maletta, R., Bruni, A., Franchesci, C., Carotenuto, L. and De Benedictis, G. (2003) The study of APOA1, APOC3, and APOA4 Variability in Healthy Aging People Reveals Another Paradox in the Oldest Old Subjects. *Ann Hum Genet* ***67***, 54-62.

association studies at four autosomal loci (REN, THO, PARP, SOD2). *Eur J Hum Genet* ***6***, pp.534-541.

Gjessing, H. K., Aalen, O. O. and Hjort, N. L. (2003) Frailty models based on Levy processes. *Advances in Applied Probability* ***35(2)***, pp.532-550.

Gorfine, M. and Hsu, L. (2011) Frailty-based Competing Risks Model for Multivariate Survival Data. *Biometrics* ***67***, pp.415-426.

Hauge, M.(1981) The Danish twin register. In S.A.Mednik, A.E.Baert, and B.P.Backmann (Eds.), *Prospective Longitudinal Research, An Empirical Basis for the Primary Prevention of Physiological Disorders*, pp. 218-221. London: Oxford University Press.

Honoré, B. E. (1993) Identification results for duration models with multiple spells. *Review of Economic Studies* ***60***, pp.241-246.

Hougaard, P. (1986) Survival models for heterogeneous populations derived from stable distributions. *Biometrika* ***73***, pp.387-396.

Iachine, I. A.(1995) Parameter Estimation in the Bivariate Correlated Frailty Model with observed Covariates via EM-Algorithm. *Working Paper Series: Population Studies of Aging*, **16**, CHS, Odense University.

Iachine, I. A.(2002) The Use of Twins and Family Survival Data in the Population Studies of Aging: Statistical Methods Based on Multivariate Survival Models. PhD Theis, Monograph 8, Department of Statistics and Demography, University of Southern Denmark.

Kalbfleisch, J. D. and Prentice, R. L. (1980) The Statistical Analysis of Failure Time Data. *John Wiley & Sons.*

Kruglyak, L. and Lander, E. S. (1995) Complete multipoint sib pair analysis of quantitative and qualitative traits. *American Journal of Human Genetics,* ***57***, pp.439-454.

Lander, E. S. and Green, P. (1987) Construction of multilocus genetic maps in human. *Proceedings of the National Academy of Science,* ***84,*** pp.2363-2367

Laskiewicz, A., Szymczak, S.Z. and Cebrat, S. (2003) The oldest old and the population heterogeneity. *Int. G. Mod. Phys. C,* ***14(10),*** pp.1355-1362.

Manton, K. G. (1982) Changing concepts of morbidity and mortality in the elderly population. *Milbank Memorial Fund Q. Health Soc.* ***60(2),*** pp.183-244.

McLachalan, G. L. and Basford, K. E. (1988) Mixture Models: Inference and Applicatiopn to Clustering. *New York: Marcel Dekker.*

Morton, N. E. (1955) Sequential tests for the detection of linkage. *American Journal of Human Genetics,* ***7***, pp.277-318.

Neale, M.C. and Cardon, L.R. (1992) Methodology for Genetic Studies of Twins and Families. Nato ASCI Series D: Behavioural and Social Sciences. Kluwer Academic Press.

Ott, J. (1991) Analysis of human genetic linkage. London: The Johzn Hopkins University Press.

Pan, W., Lin, J. and Le, C. (2002) How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biology,* ***3,***, pp.1-11.

Pan, W. (2003) On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics,* ***19,***, pp.1333-1339. Smyth, G. K. (2004) Linear modelsand empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.,* ***3,*** pp.1-26.

Strehler, B. L. and Mildvan, A. S. (1960) General theory of mortality and aging. A stochastic model relates observations on aging, physiologic decline, mortality, and radiation. *Science,* ***132***, pp.14-21.

Tan, Q., Barthum, L., Christiansen, L., De Benedictis, G., Dahlgaard, J., Frizner, N., Vach, W., Vaupel, J.W., Yashin, A.I., Christensen, K. and Kruse, T.A. (2003) Logistic Regression Models for Polymorphic and Antagonistic pleiotropic Gene Action on Human Aging and Longevity. *Ann Hum Genet* ***67***, 598-607.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001) Significant analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA,* ***98,*** pp.5116-5121.

Vaupel, J. W., Manton, K. G. and Stallard, E. (1979) The impact of heterogeneity in

individual frailty on the dynamics of mortality. *Demography,* ***16(3),*** pp.439-454.

Vaupel, J. W. and Yashin, A. I. (1985) Heterogenety's ruses: Some surprising effects of selection on population dynamics. *The American Statistician,* ***39,*** pp.176-185.

Weir, B. S. (1996) Genetic Data Analysis II. Methods for Discrete Population Genetic Data. Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts.

Wienke, A. (2011) Frailty models in survival analysis. CRC Press.

Yashin, A. I. and Iachine, I. A. (1995a) Genetic analysis of durations: Correlated frailty model applied to survival of Danish twins. *Genetic Epidemiology* ***12***, pp.529-538.

Yashin, A. I. and Iachine, I. A. (1995b) How long can human live? Lower bound for biological limit of human longevity calculated from Danish twin data using correlated frailty model. *Mechanisms of Ageing and Development,* ***80***, pp.147-169.

Yashin, A. I. and Iachine, I. A. (1997) How frailty models can be used in the analysis of mortality and longevity limits. *Demography,* ***34***, pp.31-48.

Yashin, A. I. and Iachine, I. A. (1999) What difference does the dependence between durations make? Insights for population studies of aging. *Life Time Data Analysis* ***5***, pp.5-22.

Yashin, A. I, De Benedictis, G., Vaupel, J. W., Tan, Q. et al. (1999) Genes, demography, and life span: The contribution of demographic data in genetic studies on aging and longevity. *Am. J. Hum. Genet.,* ***65***, pp.1178-1193.

Yashin, A. I., Begun, A., Boiko, S. I., Ukraintseva, S. V. and Oeppen, J. (2001) The new trends in survival improvement require a revision of traditional gerontological concepts. *Experimental Gerontology,* ***37***, pp.157-167.

Yashin, A. I., Begun, A., Boiko, S. I., Ukraintseva, S.V. and Oeppen, J. (2002) New age patterns of survival improvement in Sweden: do they characterize changes in individual aging? *Mechanisms of Ageing and Development,* ***123***, pp.637-647.

Yashin, A. I., Arbeev, K. G. and Ukraintseva, S. V. (2007) The accuracy of statisticalestimates in genetic studies of aging can be significantly improved. *Biogerontology* ***8***, pp.243-255.

Die hier vorgelegte Dissertation habe ich eigenständig und ohne unerlaubte Hilfe angefertigt. Die Dissertation wurde in der vorgelegten oder in ähnlicher Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den 04.07.2014

Alexander Begun