ANALYSIS OF CLUSTERED EVENT TIMES WITH RIGHT-CENSORING

by

Lu Wang

B.Sc. in Mathematics, Harbin Institute of Technology, 2009

M.Sc. in Mathematics, Memorial University of Newfoundland, 2011

A Project submitted in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2013

APPROVAL

Name: Lu Wang

Degree: MASTER OF SCIENCE

Title of Project: Analysis of Clustered Event Times with Right-Censoring

Examining Committee: Dr. Tim Swartz
Chair

Dr. X. Joan Hu, Supervisor

Dr. Michelle Zhou, Committee member

Dr. Richard Lockhard, Internal examiner

Date Approved: _____

Partial Copyright Licence

Abstract

Motivated by an infectious disease study at the BC Centre for Disease Control, this project is concerned with clustered event times where the observation is subject to right-censoring, and the cluster size is random. We formulate the dependence of the event times within each cluster with a copula model, and assume a parametric survival model for the margins. Inference on the model parameters are made via MLE (maximum likelihood estimation). In addition, we explore patterns of the cluster sizes and their association with the individuals who define the clusters. The motivating infectious disease study is used throughout this project to illustrate the research.

Keywords: clustered event times, copula model, right-censoring, survival analysis

Acknowledgments

I would like to express my heartfelt gratitude to my supervisor, Professor X. Joan Hu, for her expert guidance, great kindness, constant patience, encouragement and support during the two years of my master program in statistics at Simon Fraser University. Having trained me to become a statistician, she gave me precious advice that will help me enormously throughout my career life.

I am also deeply indebted to Professors Ian Bercovitz, Charmaine Dean, Robin Insley, Richard Lockhart, Thomas Loughin, Brad McNeney, Tim Swartz, Carl Schwarz, Michelle Zhou and all the other professors in the department of Statistics and Actuarial Science for their concern and academic support during my program.

Additionally, I want to thank Charlene Bradbury, Kelly Jay, Sadika Jungic and all the other staff in the department for their excellent help in these two years.

I sincerely acknowledge the School of Graduate Studies and the Department of Statistics and Actuarial Science for the financial support during my master program.

I am much obliged to my friends and fellow students Audrey Beliveau, Sherry Chen, Jack Davis, Joslin Goh, Tianyu Guan, Ruth Joy, Dilinuer Kuebran, Zhenhua Lin, Rachel Lipson, Megan McCorquodale, Nate Payne, Harsha Perera, Abdollah Safari, Maria Santiago, Biljana Stojkova, Fei Wang, Huijing Wang, Qian Wang, Tingting Wen, Vicky Weng, Yi Xiong, Yuanyu Yang, Kasra Yousefi, Rose Yu, Annie Yu and Sabrina Zhang for their encouragement, help and company.

Special thanks go to my family for their endless love and support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Tuberculosis (TB), one of the leading causes of disease and death worldwide, is spread through the air when people with an active TB infection cough, sneeze, or transmit their saliva through the air. Most of the infections result from an asymptomatic, latent infection, which may eventually develop to an active disease. Due to the infectivity of TB, contact tracing is an essential step for TB control programs. Screening for TB can be significantly improved by using efficient standards to identify contacts who are at risk of exposure to active TB patients[12]. To verify this perception, it is necessary to evaluate the association of the TB development among the contacts with potential risk factors.

The BC Centre for Disease Control (BCCDC) conducted an investigation aiming to study the association of the time to TB and latent TB infection (LTBI) with a list of potential risk factors based on the information from the identified individuals in the Greater Vancouver area who had contacts with active infectious TB patients. A total of 7921 people were identified as TB contacts from the BC provincial TB registry. The times to TB on site of the TB contacts were collected up to October 2003. This, together with the staggered study entries, resulted in the observation of the times to TB since the initial contacts subject to a non-informative right-censoring[2].

Motivated by the TB investigation of BCCDC, Cook, Hu and Swartz (2011) explore TB inference under the Cox proportional hazards model with right-censored event times, with covariates missing not at random (MNAR). They propose an approach derived from likelihood estimation utilizing supplementary information. Their approach is based on

the assumption that all the study subjects (the TB contacts) are independent. However, as mentioned in their paper, the TB contacts in the study are naturally clustered according to their TB source cases (the active TB patients). The TB contacts of the same TB source case are likely to be correlated.

This consideration led us to explore clustered event times with right-censoring. We assume that the distribution of the event times within each cluster follows a copula model, allowing the cluster size to be random. The marginal distributions of the event times conditional on potential risk factors are specified into a parametric survival model. The modelling accommodates dependence of individuals within each cluster, and enables an evaluation of the dependence through estimating a parameter. At the same time, significant factors of risk to individual TB development can be identified.

We organize the rest of the project as follows. In Chapter 2, we present some preliminary analyses of the TB data from the BCCDC study. We introduce notation and modelling, and then present an inference procedure and its implementation with the TB data in Chapter 3. Chapter 4 provides final remarks on the analyses and future studies. All numerical analyses in this project were conducted using R 3.0.1.

Chapter 2

Preliminary analysis of the TB data

2.1 Description of the TB data

Information on 17 variables was collected by the TB study. A complete list of these covariates is given in [11]. There are five variables related only to the source cases: the smear test result of the source, ID of the TB cluster of the source, size of the TB cluster of the source, genotype cluster-status of the source (source type) and death date of the source. The other 12 variables are related to the contacts: the ppd convert, gender, age at the diagnosis of the source case, HIV status, drug abuse, indicator of contact death, country of birth, type and number of contacts to the source case, ppd status at baseline of the contact, bcg history and treatment of Latent TB infection.

Some of the variable entries are missing, lightly or heavily. We summarize the missing information in Table 2.1

covariate	drug abuse	HIV status	bcg history	ppd convert	ppd status
No. of missing	7746	7745	3439	741	741
percentage	98%	98%	43%	9%	9%
covariate	age	type	gender	country of birth	
No. of missing	63	56	40	18	
percentage	0.8%	0.7%	0.5%	0.2%	

Table 2.1: Summary of missing data

For illustrative purposes, we conduct analyses with a shorter list of covariates, which are the source type, the smear test result of the source, gender of the contact, age of the contact at the diagnosis of the source case and type of the contacts. Table 2.1 shows that the missing percentages of the variables are relatively small. Thus we assume that the covariate data are missing completely at random in the analyses.

There are 7770 TB contacts with complete covariate information in the study, associated with 559 source cases. The contacts to the same source case are considered to be in one cluster. Thus there are 559 clusters with 72 different cluster sizes, and 43 of the clusters have contacts whose times to TB development are observed. There are at most 5 contacts who have developed active TB in each cluster, i.e. for whom the times to TB development are not censored.

No. of observed event times	1	2	3	4	5	Total
No. of clusters	33	4	3	2	1	43

Table 2.2: Cluster sizes

The detailed number of contacts for each cluster is shown in Appendix A, Table $B.1$. Table 2.3 presents the categorical covariates we use.

covariate	category	code	total	observed event(percentage)
source type	negative	0	6685	40(0.6%)
	positive	1	1085	23(2%)
contact smear test	negative	0	674	8(1%)
	indeterminate	1	1870	12(0.6%)
	positive	2	5226	43(0.8%)
contact gender	female	0	4273	29(0.7%)
	male	1	3497	34(1%)
contact level	casual	0	4004	11(0.3%)
	non-household	1	2376	15(0.6%)
	household	2	1390	37(2.7%)

Table 2.3: Categorical covariates

2.2 Preliminary analyses

We first conduct a generalized linear regression analysis of the counts of active TB with the covariates related to the source case. The analysis units are TB patients (sources), the response variable is the number of TB contacts developing TB associated with a source out of all associated contacts of the source. The independent variables are covariates source type and smear test result associated with sources. Source type has two levels 0 (negative) and 1 (positive). Smear test result has three levels, so we set two dummy variables smear1 and smear2 where smear1=1, smear2=0 means that the test result is indeterminate; smear1=0, smear2=1 means that the test result is positive; smear1=smear2=0 means that the result is negative. The analysis is conducted using the R function glm and the results are shown in Table 2.4

parameter	estimate	std.err	z-value	p-value	odds ratio
(intercept)	-4.63	0.36	-12.77	<0.0001	0.01
source type	1.36	0.27	5.10	<0.0001	3.91
smear1	-0.89	0.47	-1.92	0.05	0.41
smear2	-0.46	0.39	-1.19	0.23	0.63

Table 2.4: Summary of results under Binomial logistic regression

The small p-value (< 0.0001) of source type, indicates a significant association of the TB development. Indeterminate smear result (p-value=0.05) seems to be negatively associated with the development of TB disease, while positive smear result (p-value=0.23) seems not. We can see that, adjusted for smear level, contacts whose source cases have positive source type are about four times more likely to develop an active TB disease than the contacts whose source cases have negative source type. Given source type, contacts whose source cases have a positive (or indeterminate) smear test result are about 1.60 times (or 2.40 times) less likely to develop TB than the contacts whose source cases have negative smear test results. We conduct the wald test for smear test result and the p-value turns out to be 0.15, which indicates that the smear test result may not have a significant effect on the TB development of the TB contacts in the presence of the type of source.

We then explore the relationship of size of cluster with source type and smear test result of the source case. Figure 2.1 shows the histograms and densities (the density

estimates were produced using the R function density) of the size of cluster and logarithm of the size of cluster.

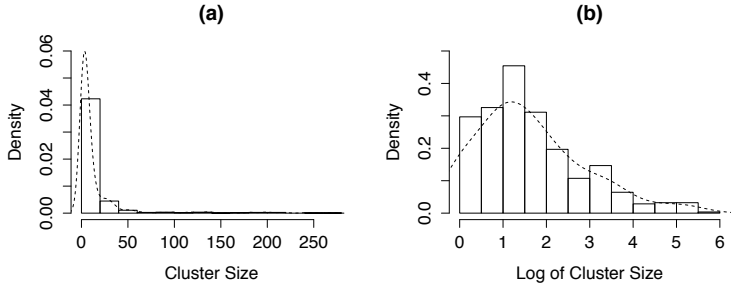

Figure 2.1: (a) Histograms of cluster seize and (b)Histogram of log of cluster size

We start with a Poisson regression analysis of the counts of the clusters, of which the outcome is presented in Table 2.5. We can see that all the risk factors appear significant. To accommodate a potential over-dispersion in the counts, we then fit a quasi-Poisson regression model for the count data, and the results are also shown in Table 2.5. The dispersion parameter seems quite large (61.55 with s.e.=3.71), indicating a large over-dispersion, and the Poisson model is not an appropriate choice. In the quasi-Poisson regression model, the coefficient estimates are all the same as the ones under the Poisson model, but the standard errors are likely more appropriate. According to the estimated coefficients, when smear result is positive (or indeterminate), the mean cluster size is 3.80 (or 1.30) times of the mean cluster size with negative smear result. The mean cluster size with positive source type is 0.77 times the mean cluster size with negative source type. Source type(p-value=0.32) and indeterminate smear test result (p-value=0.46) seem to lack significance in a quasi-Poisson model.

The estimated standard error of the dispersion parameter is calculated by $\widehat{se} = \sqrt{\frac{2n}{df^2}\widehat{\varphi}}$, where n is the sample size, df is the residual degree of freedom and $\widehat{\varphi}$ is the estimated dispersion parameter, and assuming that the dispersion is not large. Therefore the result may not be very accurate in our case since the dispersion is relatively large.

(1) Poisson model	parameter	estimate	std.err	z-value	p-value
	(intercept)	1.87	0.04	48.49	<0.0001
	source type	-0.26	0.03	-7.88	<0.0001
	smear1	0.26	0.04	5.75	<0.0001
	smear2	1.33	0.04	32.37	<0.0001
(2) quasi-Poisson model	parameter	estimate	std.err	t-value	p-value
full model	(intercept)	1.87	0.30	6.18	<0.0001
	source type	-0.26	0.26	-1.00	0.32
	smear1	0.26	0.35	0.73	0.46
	smear2	1.33	0.32	4.13	<0.0001
	dispersion	61.55	3.70		
reduced model(1)	(intercept)	2.65	0.11	25.12	<0.0001
	source type	-0.13	0.28	-0.45	0.65
	dispersion	74.44	4.46		
reduced model(2)	(intercept)	1.85	0.30	6.08	<0.0001
	smear1	0.25	0.35	0.69	0.49
	smear2	1.30	0.32	4.04	<0.0001
	dispersion	62.36	3.74		

Table 2.5: Summary of results under Poisson and quasi-Poisson models

We also conduct an ordinary linear regression analysis with the logarithm of the size of cluster. Consistent results are obtained regarding the significance of the risk factors; see Table 2.6 for the analysis outcome.

parameter	estimate	std.err	t-value	p-value
(intercept)	1.22	0.11	10.67	<0.0001
source type	-0.06	0.14	-0.43	0.67
smear1	0.09	0.14	0.68	0.50
smear2	0.95	0.14	6.91	<0.0001

Table 2.6: Summary of results of the linear regression analysis of the logarithm of the size of cluster

Chapter 3

Statistical inference based on the TB data

3.1 Notation and modelling

Suppose that event time T conditional on the covariates $\underline{Z}$ has the conditional hazard function $h(t|\underline{z}) = h_0(t)e^{\beta_0+\underline{\beta}\underline{z}}$, where $\underline{\beta} = (\beta_1, ..., \beta_m)$ and m is the total number of covariates. We specify $h_0(t) = h_0(t;\alpha)$ belonging to the Weibull family, say, $h_0(t;\alpha) = \alpha t^{\alpha-1}$, $\alpha > 0$. The cumulative conditional hazard function is

$$H(t|\underline{z}) = \int_0^t h_0(s;\alpha)e^{\beta_0+\underline{\beta}\underline{z}}ds = \left[\int_0^t h_0(s;\alpha)ds\right] e^{\beta_0+\underline{\beta}\underline{z}} = t^\alpha e^{\beta_0+\underline{\beta}\underline{z}}, \tag{3.1}$$

and the survivor function is $S(t|\underline{z}) = e^{-H(t|\underline{z})}$.

Let $C(V_1, ..., V_k;\gamma)$, $k \in \mathbb{Z}^+$ be a copula function with the dependence parameter γ. The Archimedean copula with the Clayton generator $\psi(u) = (1+u)^{-\frac{1}{\gamma}}$ and the generator inverse $\psi^{-1}(u) = u^{-\gamma} - 1$ is then

$$C(V_1, ..., V_k;\gamma) = (1 - k + \sum_{j=1}^{k} V_j{}^{-\gamma})^{-\frac{1}{\gamma}}. \tag{3.2}$$

Consider the joint conditional survival function of event times $T_1, ..., T_k$ as

$$C(S(t_1|\underline{z}_1), ..., S(t_k|\underline{z}_k);\gamma) = (1 - k + \sum_{j=1}^{k} e^{\gamma t_j^\alpha e^{\beta_0+\underline{\beta}\underline{z}_j}})^{-\frac{1}{\gamma}}, \tag{3.3}$$

denoted by $G(t_1, ..., t_k|\underline{z}_1, ..., \underline{z}_k)$.

Let $V_1 = S(t_1|\underline{z}_1), ..., V_k = S(t_k|\underline{z}_k)$ and $p \in \mathbb{Z}$, $p \in [0, k]$, then

$$\frac{\partial^p G(t_1, ..., t_k|\underline{z}_1, ..., \underline{z}_k)}{\partial t_1 \cdots \partial t_p} = \frac{\partial^p C(V_1, ..., V_k; \gamma)}{\partial V_1 \cdots \partial V_p} \frac{\partial V_1}{\partial t_1} \cdots \frac{\partial V_p}{\partial t_p}. \tag{3.4}$$

We have

$$\frac{\partial V_j}{\partial t_j} = -\alpha e^{\beta_0 + \underline{\beta}\underline{z}_j} t_j^{\alpha-1} e^{-t_j^\alpha e^{\beta_0 + \underline{\beta}\underline{z}_j}} = -h(t_j|\underline{z}_j) S(t_j|\underline{z}_j), \tag{3.5}$$

and

$$\begin{aligned} \prod_{j=1}^{p} (\frac{\partial V_j}{\partial t_j}) &= \prod_{j=1}^{p} \left[-\alpha t_j^{\alpha-1} e^{\beta_0 + \underline{\beta}\underline{z}_j} e^{-t_j^\alpha e^{\beta_0 + \underline{\beta}\underline{z}_j}} \right] \\ &= (-1)^p \prod_{j=1}^{p} \left[\alpha t_j^{\alpha-1} e^{\beta_0 + \underline{\beta}\underline{z}_j} e^{-t_j^\alpha e^{\beta_0 + \underline{\beta}\underline{z}_j}} \right]. \end{aligned} \tag{3.6}$$

Also, we have

$$\frac{\partial^p C(V_1, ..., V_k; \gamma)}{\partial V_p \cdots \partial V_1} = (1 - k + \sum_{j=1}^{k} e^{\gamma t_j^\alpha e^{\beta_0 + \underline{\beta}\underline{z}_j}})^{-\frac{1}{\gamma} - p} \prod_{j=1}^{p} \{[1 + (j-1)\gamma] \, (e^{-t_j^\alpha e^{\beta_0 + \underline{\beta}\underline{z}_j}})^{-\gamma - 1}\}. \tag{3.7}$$

Hence

$$\begin{aligned} &\frac{\partial^p G(t_1, ..., t_k|\underline{z}_1, ..., \underline{z}_k)}{\partial t_1 \cdots \partial t_k} = \frac{\partial^p C(V_1, ..., V_k; \gamma)}{\partial V_1 \cdots \partial V_p} \prod_{j=1}^{p} \frac{\partial V_j}{\partial t_j} \\ &= (-1)^p (1 - k + \sum_{j=1}^{k} e^{\gamma t_j^\alpha e^{\beta_0 + \underline{\beta}\underline{z}_j}})^{-\frac{1}{\gamma} - p} \prod_{j=1}^{p} \{[1 + (j-1)\gamma] \, \alpha t_j^{\alpha-1} e^{\beta_0 + \underline{\beta}\underline{z}_j} (e^{\gamma t_j^\alpha e^{\beta_0 + \underline{\beta}\underline{z}_j}})\}. \end{aligned} \tag{3.8}$$

Equation (3.8) will be used to calculate the likelihood function later in Section 3.2.

3.2 Likelihood function

Let $i = 1, ..., n$ be the indices of n independent clusters, and $j = 1, ..., K_i$ be the indices of the subjects in the ith cluster.

Assume K_i is a random variable related only to the ith cluster. We model $K_i \sim Poisson\ (e^{\theta_0 + \underline{\theta}' \underline{w}_i})$, where $\underline{\theta}$ is the regression parameter vector and $\underline{W}_i$ denotes the covariate vector related to the ith cluster.

Further, let T_{ij} be the event time of the jth subject of the ith cluster, and C_{ij} be the censoring time. Let $\underline{Z}_{ij} = (\underline{W}_i, \underline{X}_{ij}, K_i)$, where $\underline{X}_{ij}$ denotes the covariate components related to the jth subject in the ith cluster. We assume that T_{ij} is independent of the censoring time C_{ij} conditional on the covariates $\underline{Z}_{ij}$. We consider that the event times in one cluster are dependent on each other and assume that clusters are independent of each other.

The available data from the ith cluster are

$$[\underline{U}_i, \underline{\delta}_i, \underline{Z}_i], \quad i = 1, 2, \cdots, n \tag{3.9}$$

where $\underline{U}_i = (U_{ij}, j = 1, ..., K_i)$, $\underline{\delta}_i = (\delta_{ij}, j = 1, ..., K_i)$, $\underline{Z}_i = (\underline{Z}_{ij}, j = 1, ..., K_i)$,

$$U_{ij} = T_{ij} \wedge C_{ij} = \begin{cases} T_{ij}, & \text{if } T_{ij} \leqslant C_{ij} \\ C_{ij}, & \text{if } T_{ij} > C_{ij} \end{cases} \text{ and } \delta_{ij} = \begin{cases} 1, & \text{if } T_{ij} \leqslant C_{ij} \\ 0, & \text{if } T_{ij} > C_{ij} \end{cases}. \tag{3.10}$$

Using the generic notation for densities or mass probabilities, the information contributed by the ith cluster to the likelihood function with the available data is

$$\left[\underline{U}_i, \underline{\delta}_i | \underline{W}_i, \underline{X}_{ij}, K_i\right] \left[K_i | \underline{W}_i\right] \left[\underline{W}_i, \underline{X}_{ij}\right] \propto \left[\underline{U}_i, \underline{\delta}_i | \underline{W}_i, \underline{X}_{ij}, K_i\right] \left[K_i | \underline{W}_i\right]. \tag{3.11}$$

Without loss of generality, suppose the event times of the first p_i subjects of the ith cluster are not censored. Then

$$\begin{aligned} &P(U_{i1} = u_{i1}, ..., U_{iK_i} = u_{iK_i}, \delta_{i1} = ... = \delta_{ip_i} = 1, \delta_{ip_i+1} = ... = \delta_{iK_i} = 0 | \underline{Z}_i) \\ &\quad \propto (-1)^{p_i} \frac{\partial^{p_i} S(u_{i1}, ..., u_{iK_i} | \underline{Z}_i)}{\partial u_{i1} ... \partial u_{ip_i}} \\ &\quad = (-1)^{p_i} \frac{\partial^{p_i} C(S(u_{i1}|Z_{i1}), ..., S(u_{iK_i}|Z_{iK_i}); \gamma)}{\partial u_{i1} ... \partial u_{ip_i}} \\ &\quad = (-1)^{p_i} \frac{\partial^{p_i} G(u_{i1}, ..., u_{iK_i} | \underline{z}_i)}{\partial u_{i1} \cdots \partial u_{ip_i}}, \ \text{ if } \ p_i > 0. \end{aligned} \tag{3.12}$$

When $p_i = 0$, it means that no partial derivatives are calculated.

Let $U = \{\underline{U_i}, i = 1, \cdots, n\}$, $\delta = \{\underline{\delta_i}, i = 1, \cdots, n\}$, $Z = \{\underline{Z_i}, i = 1, \cdots, n\}$, $K = \{K_i, i = 1, \cdots, n\}$ and $W = \{\underline{W_i}, i = 1, \cdots, n\}$. The likelihood function is

$$L(\alpha, \beta_0, \underline{\beta}, \gamma, \theta_0, \underline{\theta}; U, \delta, Z) = L(\alpha, \beta_0, \underline{\beta}, \gamma; U, \delta|Z) \times L(\theta_0, \underline{\theta}; K|W), \tag{3.13}$$

where

$$\begin{aligned} L(\alpha, \beta_0, \underline{\beta}, \gamma; U, \delta|Z) &\propto \prod_{i=1}^{n} \frac{\partial^{p_i} G(u_{i1}, ..., u_{iK_i}|\underline{z}_i)}{\partial u_{i1} \cdots \partial u_{ip_i}} \\ &\propto \prod_{i=1}^{n} \{(1 - K_i + \sum_{j=1}^{K_i} e^{\gamma u_{ij}^{\alpha} e^{\beta_0 + \underline{\beta}' \underline{z}_{ij}}})^{-\frac{1}{\gamma} - p_i} \\ &\times \prod_{j=1}^{p_i} [1 + (j-1)\gamma] \, \alpha u_{ij}^{\alpha - 1} e^{\beta_0 + \underline{\beta}' \underline{z}_{ij}} (e^{\gamma u_{ij}^{\alpha} e^{\beta_0 + \underline{\beta}' \underline{z}_{ij}}})\}, \end{aligned} \tag{3.14}$$

and

$$L(\theta_0, \underline{\theta}; K|W) \propto \prod_{i=1}^{n} e^{K_i(\theta_0 + \underline{\theta}' \underline{w}_i)} \cdot e^{-e^{\theta_0 + \underline{\theta}' \underline{w}_i}}. \tag{3.15}$$

Then the log-likelihood function of $\alpha, \beta_0, \underline{\beta}$ and γ is

$$\begin{aligned} \ell(\alpha, \beta_0, \underline{\beta}, \gamma) = \sum_{i=1}^{n} \{ &-(\frac{1}{\gamma} + p_i) \log(1 - K_i + \sum_{j=1}^{K_i} e^{\gamma u_{ij}^{\alpha} e^{\beta_0 + \underline{\beta}' \underline{z}_{ij}}}) + \\ &\sum_{j=1}^{p_i} \left[\log(1 + (j-1)\gamma) + \log \alpha + (\alpha - 1) \log u_{ij} + \beta_0 + \underline{\beta}' \underline{z}_{ij} + \gamma u_{ij}^{\alpha} e^{\beta_0 + \underline{\beta}' \underline{z}_{ij}} \right] \}, \end{aligned} \tag{3.16}$$

and the log-likelihood function of θ_0 and $\underline{\theta}$ is

$$\ell(\theta_0, \underline{\theta}) = \sum_{i=1}^{n} \left[K_i(\theta_0 + \underline{\theta}' \underline{w}_i) - e^{\theta_0 + \underline{\theta}' \underline{w}_i} \right]. \tag{3.17}$$

Considering the variables we chose for the analysis, $\underline{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8)$ are the coefficients of cluster size, source type, smear1, smear2, gender, age, type1 and type2. Type1 and type2 are dummy variables where type1=0, type2=0 means that the contact type is casual; Considering the non-casual levels, type1=1,type2=0 means non-household; type1=0,type2=1 means household. And $\underline{\theta} = (\theta_1, \theta_2, \theta_3)'$ are the coefficients of source type, smear1 and smear2 in the Poisson regression model.

We estimate the parameters of the model by the maximum likelihood approach (MLE) using the R function optim. Results are shown in Section 3.3.

3.3 Analysis results

For the parameters θ_0, θ_1, θ_2 and θ_3, maximum likelihood estimation results are shown in Table 3.1. The Wald tests are conducted based on the sandwich standard errors. Results are in agreement with the ones obtained by the glm(quasi-Poisson fitting) function in R.

parameter	estimate	std.error	sandw.std.error	p-value
θ_0	1.87	0.01	0.10	<0.0001
θ_1	-0.26	0.03	0.21	0.21
θ_2	0.26	0.02	0.18	0.14
θ_3	1.33	0.01	0.12	<0.0001

Table 3.1: Summary of results of estimation of θ_0, θ_1, θ_2 and θ_3

We consider the 6 cases of the model specification in Table 3.2 ($*$ denotes the parameters to be estimated). When we set $\gamma = 0$, the event times within each cluster are assumed to be independent. When we set $\alpha = 1$, the hazard rate does not depend on the event times. When $\underline{\beta} = 0$, TB development is not related to the covariates. We can then do model comparisons among these nested models. The estimation results are shown in Table 3.3 and Table 3.4. The p-values are calculated based on sandwich standard errors.

case	γ	α	β_0	$\underline{\beta}$
A1	0	$*$	$*$	0
B1	$*$	$*$	$*$	0
A2	0	1	$*$	$*$
B2	$*$	1	$*$	$*$
A3	0	$*$	$*$	$*$
B3	$*$	$*$	$*$	$*$

Table 3.2: 6 cases of the model specification

A1, A2 and A3 are cases in which we assume the event times are independent within each cluster. B1, B2 and B3 are cases in which the event times in each cluster are correlated.

From Table 3.3, we can see that excluding the risk factors, the times to TB development show a significant positive relationship with the estimated dependence parameter γ=6.94(sandwich std.error=2.58). Plus α=0.28(sandwich std.error=0.03) in case A1 and

α=0.29(sandwich std.error=0.03) in case B1 indicate that the estimated hazard rate decreases while the event time increases. In addition Figure $C.1$ (Appendix 3) reveals that the estimated hazard rates of the clustered case are slightly higher than the ones of the independent case.

parameter	estimate	std.err	sandw.std.err	p-value
A1 case				
log-likelihood	-408.27			
$\gamma \equiv 0$				
α^*	0.28	0.04	0.03	<0.0001
β_0	-5.36	0.14	0.19	<0.0001
B1 case				
log-likelihood	-374.77			
γ	6.94	2.15	2.58	0.007
α^*	0.29	0.04	0.03	<0.0001
β_0	-4.87	0.21	0.23	<0.0001

* The null hypothesis of the tests associated with α is H_0: $\alpha = 1$

Table 3.3: Summary of results of A1 and B1

The analysis results of cases A2, B2, A3 and B3 are presented in Table 3.4. The outcomes of all the four cases are generally in agreement with respect to effects of the risk factors except the cluster size. They show that the source type, age at contact and level of contact are statistically significant risk factors, however, the smear test result and gender are not. The analysis suggests that a positive source type may result in a higher risk of TB development, and that younger contacts are at a higher risk of developing TB than older ones. In addition, we can see that a higher contact level foretells a higher risk of developing TB. Considering the cluster size, in cases A2 and A3 where the event times within each cluster are assumed to be independent, this variable has a relatively significant effect on the TB development. The contacts in one cluster with a smaller cluster size have a higher risk of developing TB. In cases B2 and B3, where the event times within each cluster are assumed to be correlated, this variable has less significance. The estimates of γ, α, β_0 are similar to the ones in cases A1 and A2, which indicates that the event times within each cluster still have a strong positive relationship with each other and the risk of developing TB decreases while the event time increases.

parameter	estimate	std.err	sandw.std.err	p-value	parameter	estimate	std.err	sandw.std.err	p-value
A2 case					A3 case				
log-likelihood	-431.13				log-likelihood	-351.77			
$\alpha \equiv 1$					α^*	0.29	0.04	0.03	<0.0001
β_0	-7.28	0.57	0.80	<0.0001	β_0	-5.89	0.57	0.79	<0.0001
β_1	-1.47	0.73	0.72	0.04	β_1	-1.46	0.73	0.70	0.04
β_2	1.54	0.27	0.35	<0.0001	β_2	1.53	0.27	0.33	<0.0001
β_3	-0.58	0.46	0.61	0.34	β_3	-0.55	0.46	0.60	0.36
β_4	0.41	0.40	0.53	0.44	β_4	0.42	0.40	0.52	0.42
β_5	0.17	0.26	0.28	0.56	β_5	0.17	0.26	0.28	0.53
β_6	-2.91	0.82	0.94	0.002	β_6	-2.89	0.82	0.93	0.002
β_7	0.78	0.40	0.48	0.10	β_7	0.75	0.40	0.47	0.11
β_8	2.07	0.39	0.54	0.0001	β_8	2.03	0.39	0.53	0.0001
B2 case					B3 case				
log-likelihood	-414.57				log-likelihood	-336.42			
$\alpha \equiv 1$					α^*	0.29	0.04	0.03	<0.0001
γ	3.51	1.23	1.57	0.03	γ	3.62	1.32	1.69	0.03
β_0	-7.29	0.62	0.89	<0.0001	β_0	-5.95	0.62	0.85	<0.0001
β_1	-1.66	1.13	1.07	0.12	β_1	-1.63	1.14	1.08	0.13
β_2	1.50	0.36	0.35	<0.0001	β_2	1.49	0.37	0.35	<0.0001
β_3	-0.81	0.54	0.68	0.23	β_3	-0.77	0.54	0.65	0.23
β_4	0.28	0.47	0.65	0.67	β_4	0.33	0.47	0.62	0.59
β_5	0.24	0.25	0.24	0.32	β_5	0.24	0.25	0.24	0.32
β_6	-2.76	0.82	0.94	0.003	β_6	-2.74	0.82	0.94	0.004
β_7	0.91	0.43	0.48	0.06	β_7	0.90	0.43	0.48	0.06
β_8	2.22	0.41	0.51	<0.0001	β_8	2.19	0.41	0.50	<0.0001

* The null hypothesis of the tests associated with α is H_0: $\alpha = 1$

Table 3.4: Summary of results of A2, B2, A3 and B3

Since the parametric survival model we use is a special case of the Cox proportional hazards model, we estimate the coefficients of the covariates using the Cox proportional hazards model with the assumption that the event times are independent of each other. The analysis is conducted using the R package survival. Table 3.5 presents the results of the estimation. The results are similar to the ones of A3 except that the contact types show more significant effects on the TB development in the Cox fit. Comparing the Cox fit and A3, it shows that the parametric survival model we use is appropriate.

parameter	estimate	std.err	sandw.std.err	p-value
β_1	-1.46	0.72	0.70	0.04
β_2	1.52	0.27	0.25	<0.0001
β_3	-0.55	0.46	0.44	0.21
β_4	0.43	0.40	0.39	0.27
β_5	0.17	0.26	0.26	0.51
β_6	-2.89	0.82	0.89	0.001
β_7	0.74	0.40	0.41	0.07
β_8	2.03	0.39	0.44	<0.0001

Table 3.5: Summary of results of the Cox fit

To present the analysis graphically, we set age and cluster size equal to the averages, then we only need to consider combinations of the other variables. There are in total 36 combinations. We split the 36 cases into two groups according to the two source types. When the status is positive, we have 18 combinations of the other variables; when the status is negative, we also have the same 18 combinations of the other variables. The combinations are shown in Table 3.6. The plots of the hazard functions of each combination for cases A2, A3, B2 and B3 are shown in Appendix C.

combination	type1	type2	smear1	smear2	gender
1	1	0	1	0	1
2	1	0	0	1	1
3	1	0	0	0	1
4	1	0	1	0	0
5	1	0	0	1	0
6	1	0	0	0	0
7	0	1	1	0	1
8	0	1	0	1	1
9	0	1	0	0	1
10	0	1	1	0	0
11	0	1	0	1	0
12	0	1	0	0	0
13	0	0	1	0	1
14	0	0	0	1	1
15	0	0	0	0	1
16	0	0	1	0	0
17	0	0	0	1	0
18	0	0	0	0	0

Table 3.6: 18 combinations of the variables

Figure $C.2$, Figure $C.3$ and Figure $C.4$ show that for both independent and clustered event times, a positive source type in a higher hazard rate, however, gender does not have a significant effect on the hazard rates. The first 6 combinations are the ones with non-household contact type, the middle 6 combinations are the ones with household contact type and the last 6 combinations are the ones with casual contact type. We can see clearly that a higher contact level indicates a higher risk of developing TB. These are in agreement with the previous analysis results. In Figure $C.5$, Figure $C.8$ and Figure $C.9$ show quite similar results as Figure $C.2$. In addition, Figure $C.10$ and Figure $C.11$ show that the marginal survival functions of Cox fit and A3 are quite similar, which indicates again that the parametric survival model we use is appropriate.

Chapter 4

Discussion

This project considers clustered event times with right-censoring and random cluster size. We assume that the cluster size follows a Poisson model. The Clayton copula model is adapted to analyze the clustered event times conditional on the cluster size. A parametric survival model is used for the marginal distribution of the event times. The modelling provides an alternative to address the concern about the independent assumption on the event times in the analysis of the same TB data by Cook, Hu and Swartz (2011). The approach potentially has a broad application.

There are a few issues to investigate further. We conducted the analysis assuming that all the covariates are missing completely at random. However, for example, the HIV status in the original data set is heavily missing which is likely not missing at random (MNAR). We may adapt the approach of Cook, Hu and Swartz (2011) to address the MNAR and include the HIV status as one of the potential risk factors. Also, we consider a parametric survival model. It is worth extending the approach to situations with semi-parametric models, such as the Cox proportional hazards model.

Some investigators deal with correlated event times by a frailty model, assuming the times depend with each other through an unobservable random variable. It will be of interest to compare the TB data analysis by our approach with this alternative.

Bibliography

[1] Chen X. and Tsyrennikov, V. (2006). Efficient Estimation of Semiparametric Multivariate Copula Models. *Journal of the American Statistical Association*, 101(475), 1228-1240.

[2] Cook, V.J., Hu, X.J. and Swartz, T.B. (2011). Cox regression with covariates missing not at random. *Statistics in Biosciences*, 3(2), 208-222.

[3] Georges, P., Lamy, A.G., Nicolas, E., Quibel, G. and Roncalli, T. (2001). Multivariate Survival Modelling: A Unified Approach with Copulas. URL: http://ssrn.com/abstract=1032559 or http://dx.doi.org/10.2139/ssrn.1032559 [May 28, 2001].

[4] Gregoriou, G.N. and Pascalau, R. (2012). A joint survival analysis of hedge funds and funds of funds using copulas. *Managerial Finance*, 38(1), 82-100.

[5] Goethals, K., Janssen P. and Duchateau, L. (2008). Frailty models and copulas: similarities and differences. *Journal of Applied Statistics*, 35(9), 1071-1079.

[6] McNeil, A.J. andNeslehova, J. (2007). Multivariate Archimedean Copulas. URL: http://www−1.ms.ut.ee/tartu07/presentations/mcneil.pdf

[7] Schmidt, T. (2006). Coping with Copulas. URL: http://www.math.uni−leipzig.de/ tschmidt/TSchmidt_Copulas.pdf

[8] Shih, J.H. and Louis T.A. (1995). Inferences on the Association Parameter in Copula Models for Bivariate Survival Data. *Biometrics*, 51(4), 1384-1399.

[9] Trivedi, P.K. and Zimmer, D.K. (2005). Copula Modeling: An Introduction for Practitioners. *Foundations and Trends in Econometrics*, 1(1), 1-111.

[10] Wang W. (2003). Estimating the Association Parameter for Copula Models under Dependent Censoring. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65(1), 257-273.

[11] Canadian Tuberculosis Standards 6th edn (2007) Edited by Long, R Canadian Lung Association, Canadian Thoracic Society and Tuberculosis Prevention and Control Centre for Infectious Disease Prevention and Control, Health Canada.

[12] http://en.wikipedia.org/wiki/Tuberculosis

Appendix A

Partial derivatives of the log-likelihood function

We have each partial derivative of the log-likelihood function like below:

$$\begin{aligned}\frac{\partial \ell(\alpha,\beta_0,\underline{\beta},\gamma)}{\partial \alpha} =& \sum_{i=1}^{n}\left[-\frac{(1+\gamma p_i)\sum_{j=1}^{K_i}\log u_{ij}\cdot u_{ij}^{\alpha}e^{\beta_0+\underline{\beta}\underline{z}_{ij}}e^{\gamma u_{ij}^{\alpha}e^{\beta_0+\underline{\beta}\underline{z}_{ij}}}}{1-K_i+\sum_{j=1}^{K_i}e^{\gamma u_{ij}^{\alpha}e^{\beta_0+\underline{\beta}\underline{z}_{ij}}}}\right] \\ &+\sum_{i=1}^{n}\sum_{j=1}^{p_i}\log u_{ij}(1+\gamma u_{ij}^{\alpha}e^{\beta_0+\underline{\beta}\underline{z}_{ij}})+\frac{1}{\alpha}\sum_{i=1}^{n}p_i\end{aligned} \tag{A.1}$$

$$\begin{aligned}\frac{\partial \ell(\alpha,\beta_0,\underline{\beta},\gamma)}{\partial \gamma} =& \sum_{i=1}^{n}\left[\frac{(1+\gamma p_i)\sum_{j=1}^{K_i}u_{ij}^{\alpha}e^{\beta_0+\underline{\beta}\underline{z}_{ij}+\gamma u_{ij}^{\alpha}e^{\beta_0+\underline{\beta}\underline{z}_{ij}}}}{1-K_i+\sum_{j=1}^{K_i}e^{\gamma u_{ij}^{\alpha}e^{\beta_0+\underline{\beta}\underline{z}_{ij}}}}\right]+\sum_{i=1}^{n}\sum_{j=1}^{p_i}(1+\gamma u_{ij}^{\alpha}e^{\beta_0+\underline{\beta}\underline{z}_{ij}}) \\ &-\sum_{i=1}^{n}\sum_{j=1}^{p_i}\frac{1}{1+(j-1)\gamma}+\frac{1}{\gamma}\sum_{i=1}^{n}\log(1-K_i+\sum_{j=1}^{K_i}e^{\gamma u_{ij}^{\alpha}e^{\beta_0+\underline{\beta}\underline{z}_{ij}}})\end{aligned} \tag{A.2}$$

$$\begin{aligned}\frac{\partial \ell(\alpha,\beta_0,\underline{\beta},\gamma)}{\partial \beta_0} =& -\sum_{i=1}^{n}\left[\frac{(1+\gamma p_i)\sum_{j=1}^{K_i}u_{ij}^{\alpha}e^{\beta_0+\underline{\beta}\underline{z}_{ij}+\gamma u_{ij}^{\alpha}e^{\beta_0+\underline{\beta}\underline{z}_{ij}}}}{1-K_i+\sum_{j=1}^{K_i}e^{\gamma u_{ij}^{\alpha}e^{\beta_0+\underline{\beta}\underline{z}_{ij}}}}\right] \\ &+\sum_{i=1}^{n}\sum_{j=1}^{p_i}(1+\gamma u_{ij}^{\alpha}e^{\beta_0+\underline{\beta}\underline{z}_{ij}})\end{aligned} \tag{A.3}$$

Suppose $\underline{\beta}=(\beta_1,\beta_2,\beta_3)$, where β_1,β_2,β_3 are the coefficients of w_i, z_{ij} and K_i.

Then

$$\frac{\partial \ell(\alpha,\beta_0,\underline{\beta},\gamma)}{\partial \beta_1} = -\sum_{i=1}^{n}\left[\frac{(1+\gamma p_i)\sum_{j=1}^{K_i} w_i \cdot u_{ij}^{\alpha} e^{\beta_0+\underline{\beta}\underline{z}_{ij}+\gamma u_{ij}^{\alpha} e^{\beta_0+\underline{\beta}\underline{z}_{ij}}}}{1-K_i+\sum_{j=1}^{K_i} e^{\gamma u_{ij}^{\alpha} e^{\beta_0+\underline{\beta}\underline{z}_{ij}}}}\right] + \sum_{i=1}^{n}\sum_{j=1}^{p_i} w_i(1+\gamma u_{ij}^{\alpha} e^{\beta_0+\underline{\beta}\underline{z}_{ij}}) \tag{A.4}$$

$$\frac{\partial \ell(\alpha,\beta_0,\underline{\beta},\gamma)}{\partial \beta_2} = -\sum_{i=1}^{n}\left[\frac{(1+\gamma p_i)\sum_{j=1}^{K_i} x_{ij} \cdot u_{ij}^{\alpha} e^{\beta_0+\underline{\beta}\underline{z}_{ij}+\gamma u_{ij}^{\alpha} e^{\beta_0+\underline{\beta}\underline{z}_{ij}}}}{1-K_i+\sum_{j=1}^{K_i} e^{\gamma u_{ij}^{\alpha} e^{\beta_0+\underline{\beta}\underline{z}_{ij}}}}\right] + \sum_{i=1}^{n}\sum_{j=1}^{p_i} x_{ij}(1+\gamma u_{ij}^{\alpha} e^{\beta_0+\underline{\beta}\underline{z}_{ij}}) \tag{A.5}$$

$$\frac{\partial \ell(\alpha,\beta_0,\underline{\beta},\gamma)}{\partial \beta_3} = -\sum_{i=1}^{n}\left[\frac{(1+\gamma p_i)\sum_{j=1}^{K_i} K_i \cdot u_{ij}^{\alpha} e^{\beta_0+\underline{\beta}\underline{z}_{ij}+\gamma u_{ij}^{\alpha} e^{\beta_0+\underline{\beta}\underline{z}_{ij}}}}{1-K_i+\sum_{j=1}^{K_i} e^{\gamma u_{ij}^{\alpha} e^{\beta_0+\underline{\beta}\underline{z}_{ij}}}}\right] + \sum_{i=1}^{n}\sum_{j=1}^{p_i} K_i(1+\gamma u_{ij}^{\alpha} e^{\beta_0+\underline{\beta}\underline{z}_{ij}}) \tag{A.6}$$

$$\frac{\partial \ell(\theta_0,\underline{\theta})}{\partial \theta_0} = \sum_{i=1}^{n}\left[K_i - e^{\theta_0+\underline{\theta}\underline{w}_i}\right] \tag{A.7}$$

Suppose $\underline{\theta} = (\theta_1, \theta_2)$, where θ_1, θ_2 are the coefficients of $\underline{w}_i = (w_{i1}, w_{2i})$.

$$\frac{\partial \ell(\theta_0,\underline{\theta})}{\partial \theta_1} = \sum_{i=1}^{n}\left[K_i w_{i1} - e^{\theta_0+\underline{\theta}\underline{w}_i} w_{i1}\right] \tag{A.8}$$

$$\frac{\partial \ell(\theta_0,\underline{\theta})}{\partial \theta_2} = \sum_{i=1}^{n}\left[K_i w_{i2} - e^{\theta_0+\underline{\theta}\underline{w}_i} w_{i2}\right] \tag{A.9}$$

When $\underline{\beta} = \underline{0}$, we have $S(u_{ij}|\underline{z}_{ij}) = e^{-u_{ij}^{\alpha} e^{\beta_0}}$. There is no change in the likelihood function of $\theta_0, \underline{\theta}$, but we have the likelihood function of α, β_0, γ like below:

$$L(\alpha,\beta_0,\gamma) \propto \prod_{i=1}^{n}\{(1-K_i+\sum_{j=1}^{K_i} e^{\gamma u_{ij}^{\alpha} e^{\beta_0}})^{-\frac{1}{\gamma}-p_i} \prod_{j=1}^{p_i}[1+(j-1)\gamma]\,\alpha u_{ij}^{\alpha-1} e^{\beta_0}(e^{\gamma u_{ij}^{\alpha} e^{\beta_0}})\} \tag{A.10}$$

Then the log-likelihood function of α, β_0, γ is

$$\begin{aligned}\ell(\alpha,\beta_0,\gamma) = \sum_{i=1}^{n}\{-(\frac{1}{\gamma}+p_i)\log(1-K_i+\sum_{j=1}^{K_i}e^{\gamma u_{ij}^{\alpha}e^{\beta_0}})+ \\ \sum_{j=1}^{p_i}\Big[\log(1+(j-1)\gamma)+\log\alpha+(\alpha-1)\log u_{ij}+\beta_0+\gamma u_{ij}^{\alpha}e^{\beta_0}\Big]\},\end{aligned} \tag{A.11}$$

Then the estimation equations of α, γ, β_0 are

$$\begin{aligned}\frac{\partial\ell(\alpha,\beta_0,\gamma)}{\partial\alpha} =& \sum_{i=1}^{n}\left[-\frac{(1+\gamma p_i)\sum_{j=1}^{K_i}\log u_{ij}\cdot u_{ij}^{\alpha}e^{\beta_0}e^{\gamma u_{ij}^{\alpha}e^{\beta_0}}}{1-K_i+\sum_{j=1}^{K_i}e^{\gamma u_{ij}^{\alpha}e^{\beta_0}}}\right] \\ &+\sum_{i=1}^{n}\sum_{j=1}^{p_i}\log u_{ij}(1+\gamma u_{ij}^{\alpha}e^{\beta_0})+\frac{1}{\alpha}\sum_{i=1}^{n}p_i\end{aligned} \tag{A.12}$$

$$\begin{aligned}\frac{\partial\ell(\alpha,\beta_0,\gamma)}{\partial\gamma} =& \sum_{i=1}^{n}\left[\frac{(1+\gamma p_i)\sum_{j=1}^{K_i}u_{ij}^{\alpha}e^{\beta_0+\gamma u_{ij}^{\alpha}e^{\beta_0}}}{1-K_i+\sum_{j=1}^{K_i}e^{\gamma u_{ij}^{\alpha}e^{\beta_0}}}\right]+\sum_{i=1}^{n}\sum_{j=1}^{p_i}(1+\gamma u_{ij}^{\alpha}e^{\beta_0}) \\ &-\sum_{i=1}^{n}\sum_{j=1}^{p_i}\frac{1}{1+(j-1)\gamma}+\frac{1}{\gamma}\sum_{i=1}^{n}\log(1-K_i+\sum_{j=1}^{K_i}e^{\gamma u_{ij}^{\alpha}e^{\beta_0}})\end{aligned} \tag{A.13}$$

$$\frac{\partial\ell(\alpha,\beta_0,\gamma)}{\partial\beta_0} = -\sum_{i=1}^{n}\left[\frac{(1+\gamma p_i)\sum_{j=1}^{K_i}u_{ij}^{\alpha}e^{\beta_0+\gamma u_{ij}^{\alpha}e^{\beta_0}}}{1-K_i+\sum_{j=1}^{K_i}e^{\gamma u_{ij}^{\alpha}e^{\beta_0}}}\right]+\sum_{i=1}^{n}\sum_{j=1}^{p_i}(1+\gamma u_{ij}^{\alpha}e^{\beta_0}) \tag{A.14}$$

When $\alpha = 1$, $S(u_{ij}|\underline{z}_{ij}) = e^{-u_{ij}e^{\beta_0+\underline{\beta}\underline{z}_{ij}}}$, then there is no change in the likelihood function of $\theta_0, \underline{\theta}$, but we have the likelihood function of $\gamma, \beta_0, \underline{\beta}$ like below:

$$L(\gamma,\beta_0,\underline{\beta})\propto\prod_{i=1}^{n}\{(1-K_i+\sum_{j=1}^{K_i}e^{\gamma u_{ij}e^{\beta_0+\underline{\beta}\underline{z}_{ij}}})^{-\frac{1}{\gamma}-p_i}\prod_{j=1}^{p_i}[1+(j-1)\gamma]\,e^{\beta_0+\underline{\beta}\underline{z}_{ij}}(e^{\gamma u_{ij}e^{\beta_0+\underline{\beta}\underline{z}_{ij}}})\}. \tag{A.15}$$

Then the log-likelihood function of $\gamma, \beta_0, \underline{\beta}$ is

$$\begin{aligned}\ell(\gamma,\beta_0,\underline{\beta}) = \sum_{i=1}^{n}\{-(\frac{1}{\gamma}+p_i)\log(1-K_i+\sum_{j=1}^{K_i}e^{\gamma u_{ij}e^{\beta_0+\underline{\beta}\underline{z}_{ij}}})+ \\ \sum_{j=1}^{p_i}\Big[\log(1+(j-1)\gamma)+\beta_0+\underline{\beta}\underline{z}_{ij}+\gamma u_{ij}e^{\beta_0+\underline{\beta}\underline{z}_{ij}}\Big]\}.\end{aligned} \tag{A.16}$$

Thus the estimation equations are

$$\frac{\partial \ell(\beta_0, \underline{\beta}, \gamma)}{\partial \gamma} = \sum_{i=1}^{n} \left[\frac{(1+\gamma p_i) \sum_{j=1}^{K_i} u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij} + \gamma u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij}}}}{1 - K_i + \sum_{j=1}^{K_i} e^{\gamma u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij}}}} \right] + \sum_{i=1}^{n} \sum_{j=1}^{p_i} (1 + \gamma u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij}}) - \sum_{i=1}^{n} \sum_{j=1}^{p_i} \frac{1}{1 + (j-1)\gamma} + \frac{1}{\gamma} \sum_{i=1}^{n} \log(1 - K_i + \sum_{j=1}^{K_i} e^{\gamma u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij}}}) \tag{A.17}$$

$$\frac{\partial \ell(\beta_0, \underline{\beta}, \gamma)}{\partial \beta_0} = -\sum_{i=1}^{n} \left[\frac{(1+\gamma p_i) \sum_{j=1}^{K_i} u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij} + \gamma u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij}}}}{1 - K_i + \sum_{j=1}^{K_i} e^{\gamma u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij}}}} \right] + \sum_{i=1}^{n} \sum_{j=1}^{p_i} (1 + \gamma u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij}}) \tag{A.18}$$

Suppose $\underline{\beta} = (\beta_1, \beta_2, \beta_3)$, where $\beta_1, \beta_2, \beta_3$ are the coefficients of w_i, z_{ij} and K_i. Then

$$\frac{\partial \ell(\beta_0, \underline{\beta}, \gamma)}{\partial \beta_1} = -\sum_{i=1}^{n} \left[\frac{(1+\gamma p_i) \sum_{j=1}^{K_i} w_i \cdot u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij} + \gamma u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij}}}}{1 - K_i + \sum_{j=1}^{K_i} e^{\gamma u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij}}}} \right] + \sum_{i=1}^{n} \sum_{j=1}^{p_i} w_i (1 + \gamma u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij}}) \tag{A.19}$$

$$\frac{\partial \ell(\beta_0, \underline{\beta}, \gamma)}{\partial \beta_2} = -\sum_{i=1}^{n} \left[\frac{(1+\gamma p_i) \sum_{j=1}^{K_i} x_{ij} \cdot u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij} + \gamma u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij}}}}{1 - K_i + \sum_{j=1}^{K_i} e^{\gamma u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij}}}} \right] + \sum_{i=1}^{n} \sum_{j=1}^{p_i} x_{ij} (1 + \gamma u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij}}) \tag{A.20}$$

$$\frac{\partial \ell(\beta_0, \underline{\beta}, \gamma)}{\partial \beta_3} = -\sum_{i=1}^{n} \left[\frac{(1+\gamma p_i) \sum_{j=1}^{K_i} K_i \cdot u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij} + \gamma u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij}}}}{1 - K_i + \sum_{j=1}^{K_i} e^{\gamma u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij}}}} \right] + \sum_{i=1}^{n} \sum_{j=1}^{p_i} K_i (1 + \gamma u_{ij} e^{\beta_0 + \underline{\beta} \underline{z}_{ij}}) \tag{A.21}$$

Appendix B

Cluster sizes

Cluster size	1	2	3	4	5	6	7	8	9	10	11	12
No. of Clusters	83	91	75	52	34	33	20	13	12	10	11	9
Cluster size	13	14	15	16	17	18	19	20	21	22	23	24
No. of Clusters	7	7	3	6	3	1	2	1	2	3	2	7
Cluster size	25	26	27	28	29	30	31	32	33	34	35	37
No. of Clusters	4	4	1	3	2	5	2	1	5	3	1	2
Cluster size	38	39	41	42	46	49	50	51	54	55	56	58
No. of Clusters	1	2	1	1	2	1	1	2	1	1	1	1
Cluster size	62	65	74	85	89	99	100	115	116	121	132	136
No. of Clusters	1	1	1	1	1	1	1	1	1	1	1	1
Cluster size	137	144	163	180	182	193	200	201	212	220	241	269
No. of Clusters	1	1	1	1	1	1	1	1	1	1	1	1

Table B.1: Number of clusters for each cluster size

Appendix C

Plots of analysis

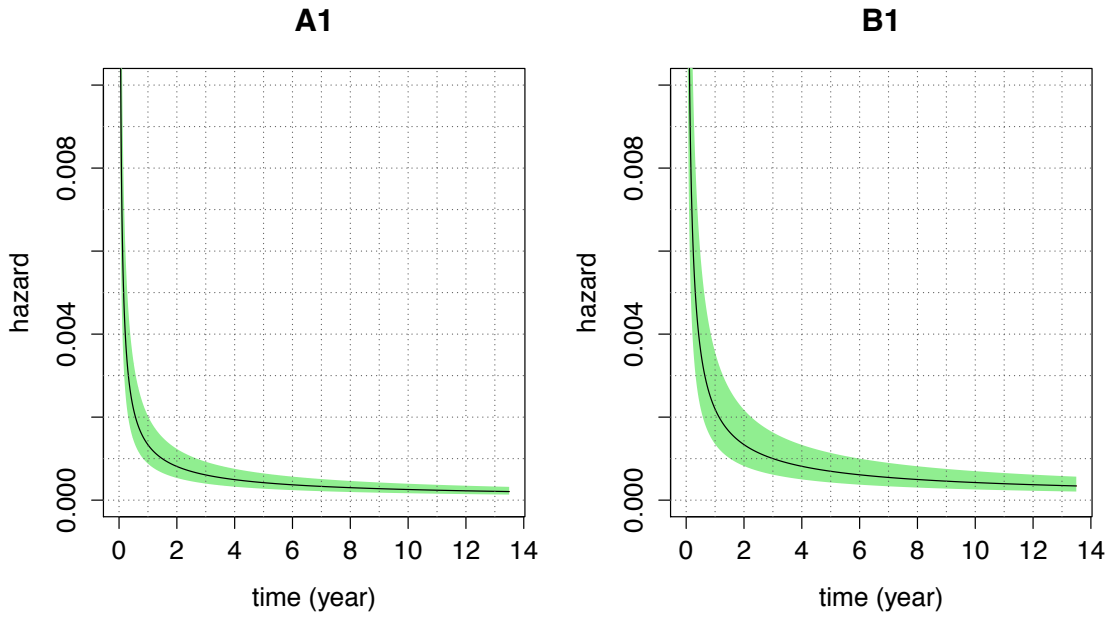

Figure C.1: Marginal hazard functions of A1 and B1 with 95 % CIs

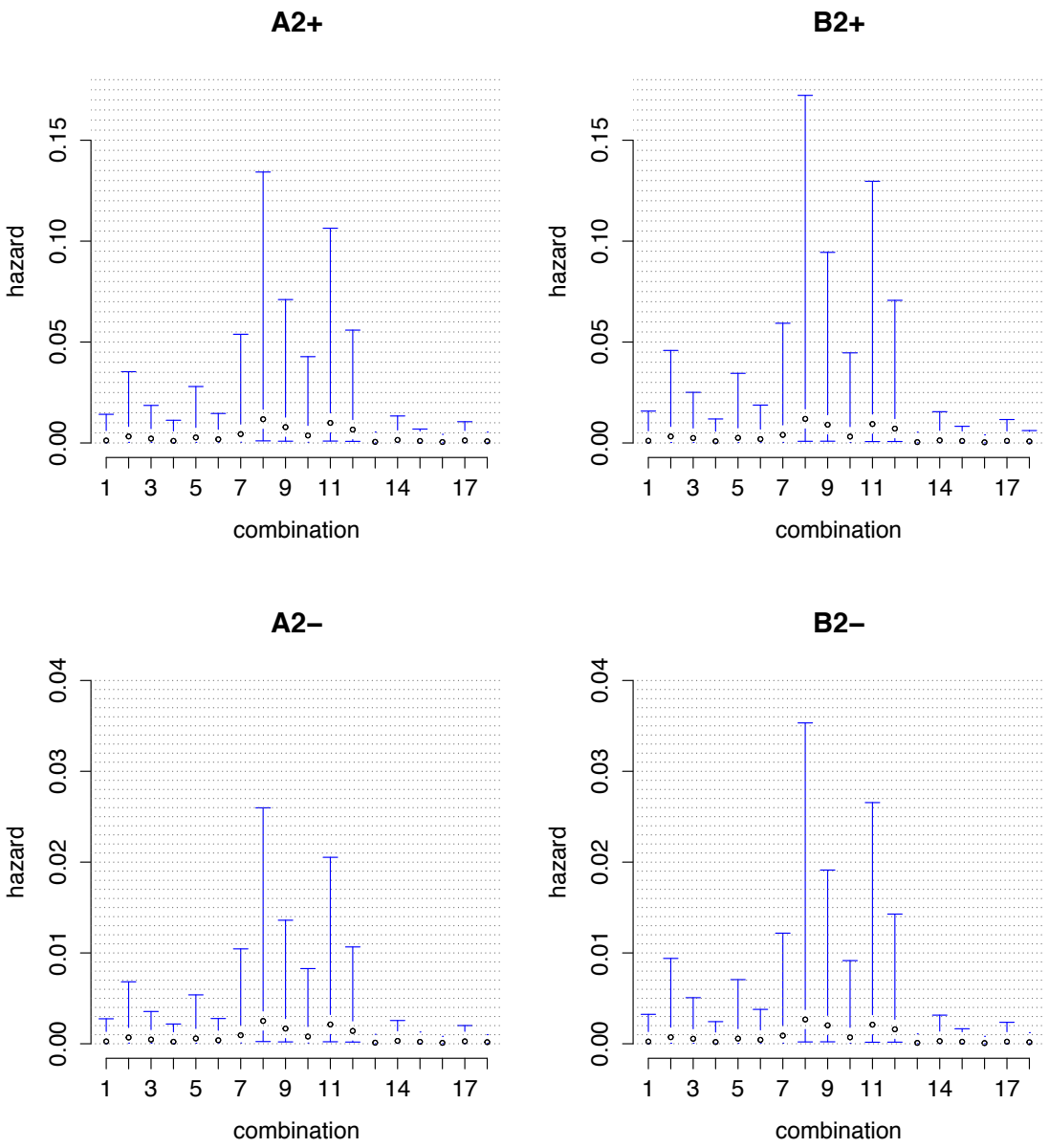

Figure C.2: Marginal hazard functions with 95 % CIs of A2 and B2 with source type positive and negative

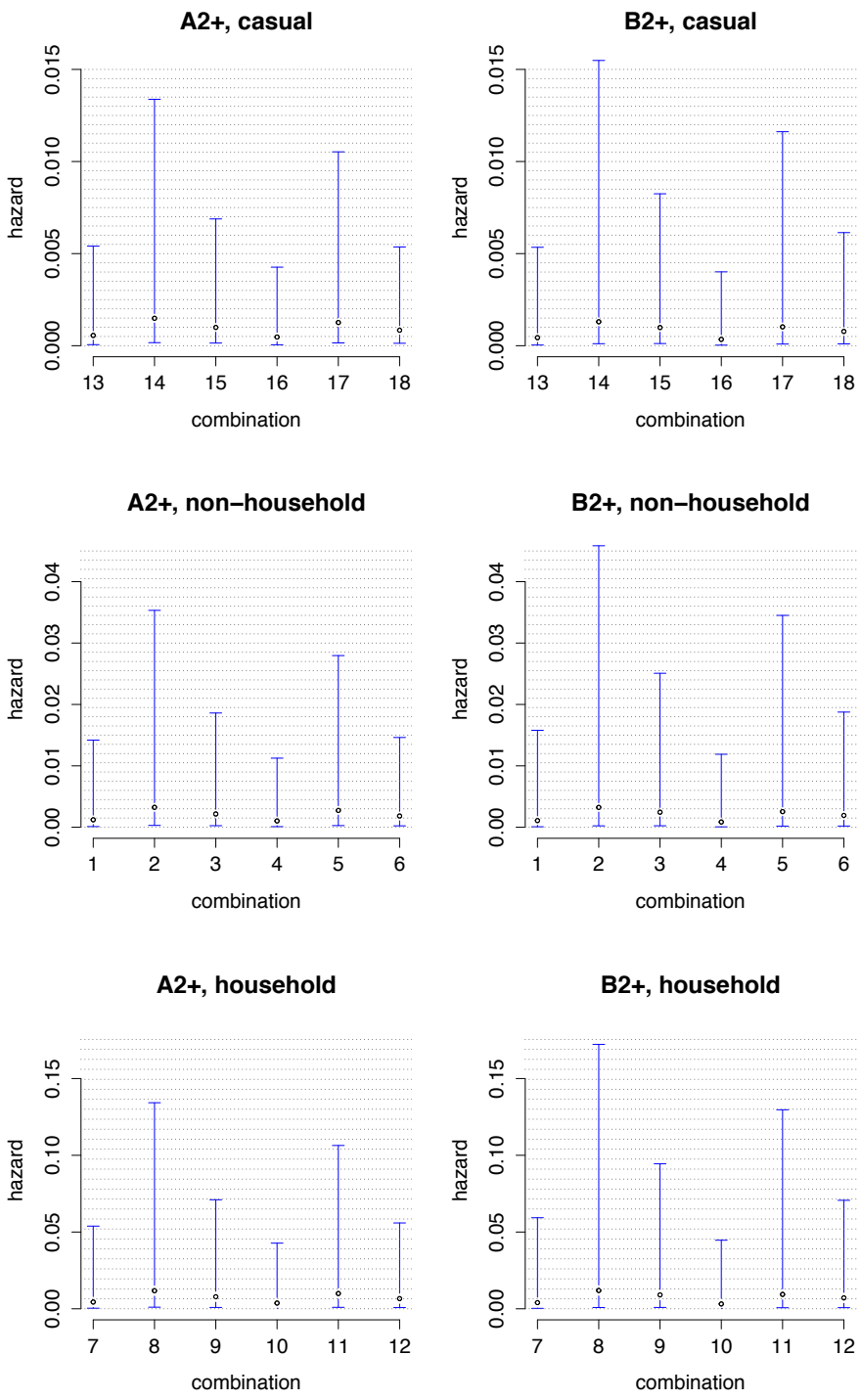

Figure C.3: Marginal hazard functions with 95 % CIs of A2 and B2 with source type positive and three different contact types

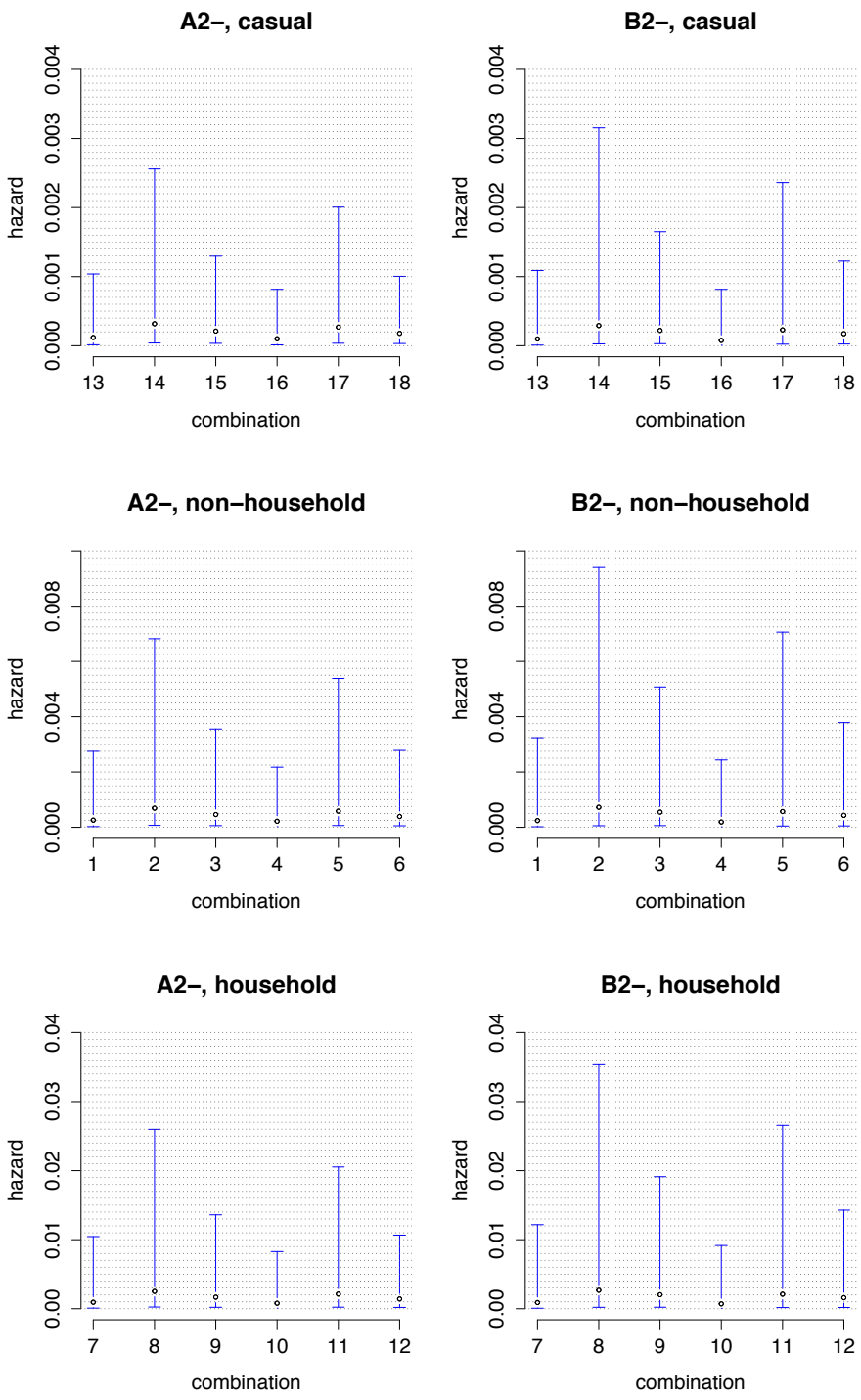

Figure C.4: Marginal hazard functions with 95 % CIs of A2 and B2 with source type negative and three different contact types

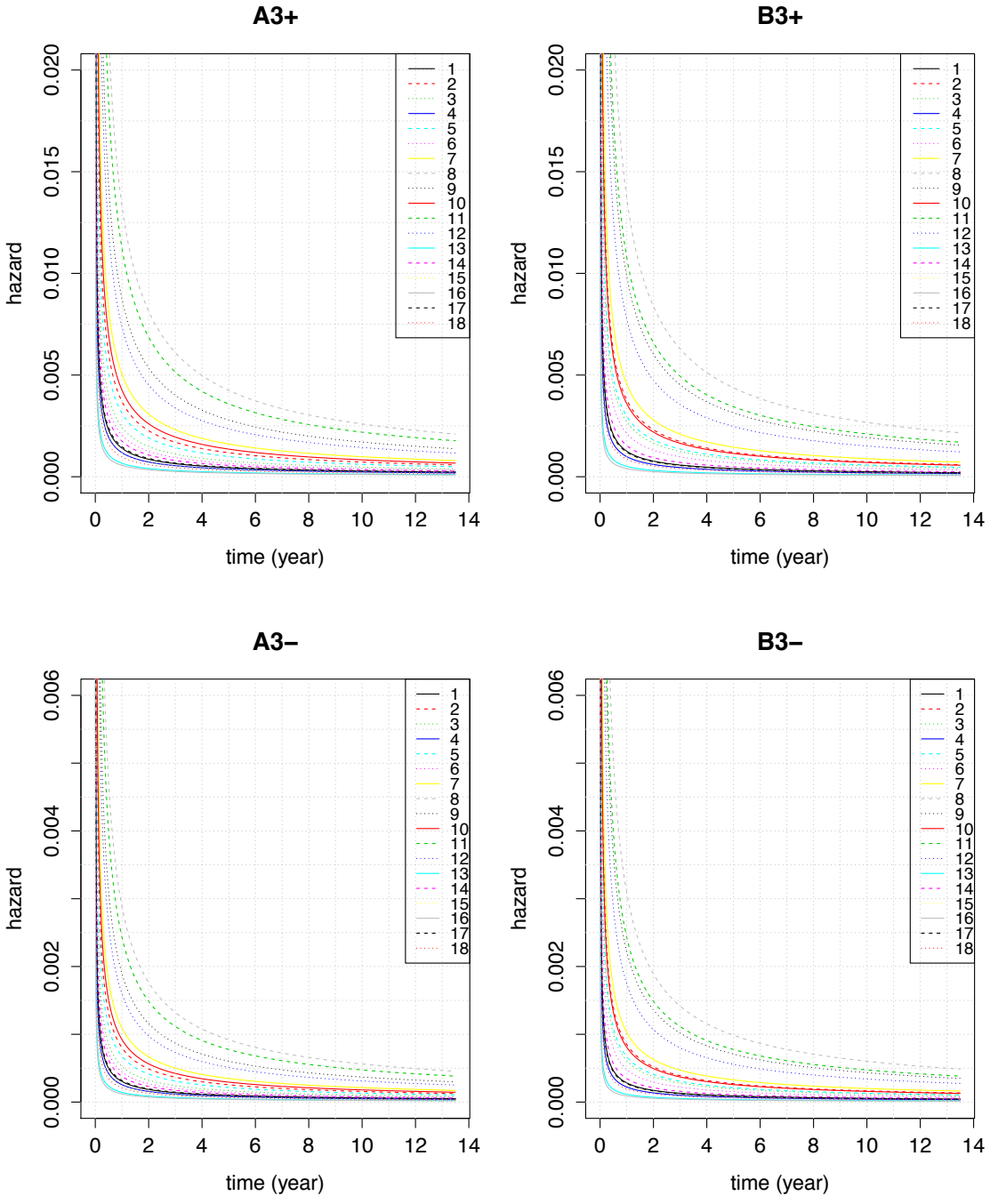

Figure C.5: Marginal hazard functions of A3 and B3 with source type positive and negative

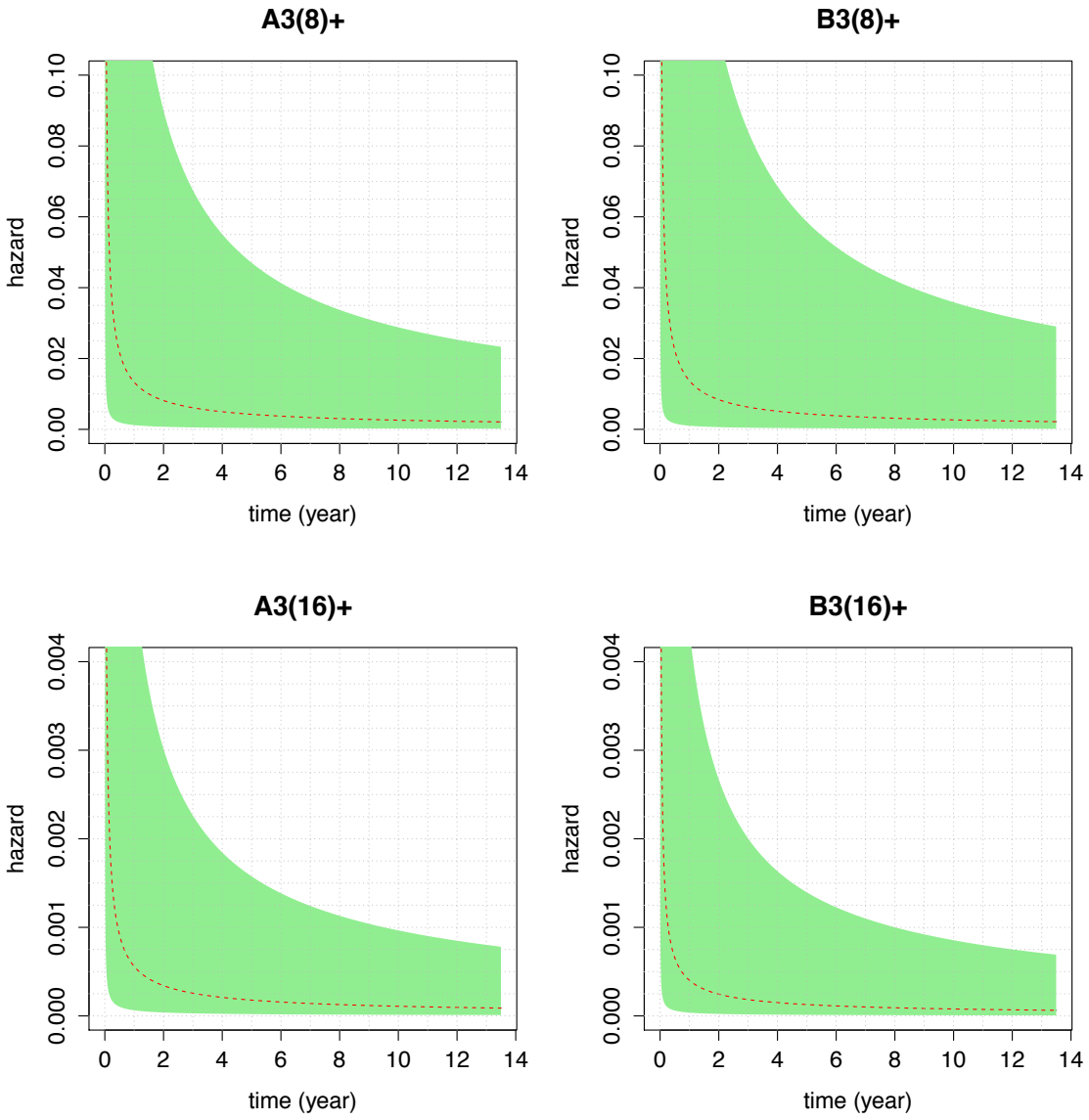

Figure C.6: Marginal hazard functions with 95% CIs of A3(8,16) and B3(8,16) with source type positive

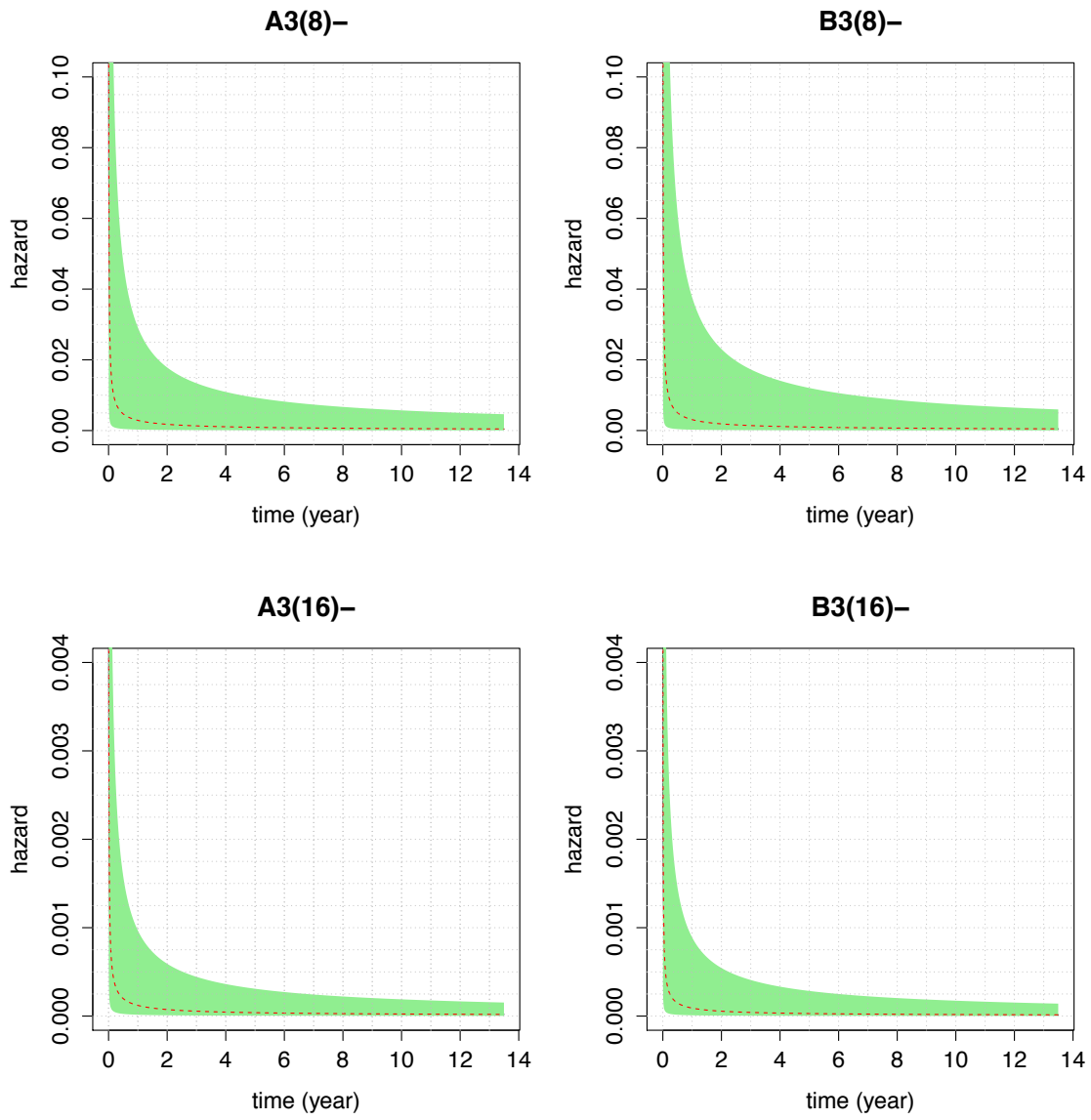

Figure C.7: Marginal hazard functions with 95% CIs of A3(8,16) and B3(8,16) with source type negative

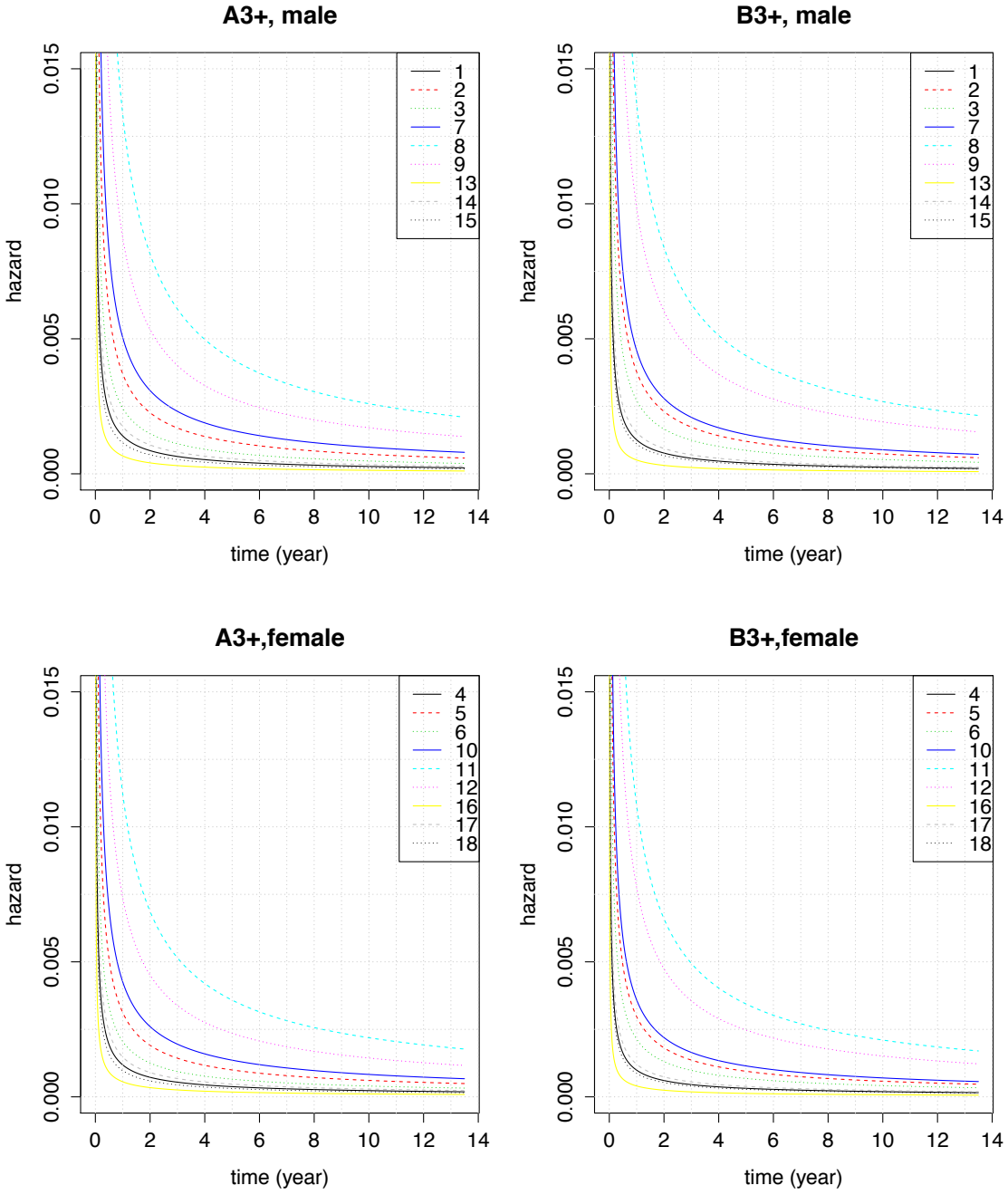

Figure C.8: Marginal hazard functions of A3 and B3 with source type positive and different gender

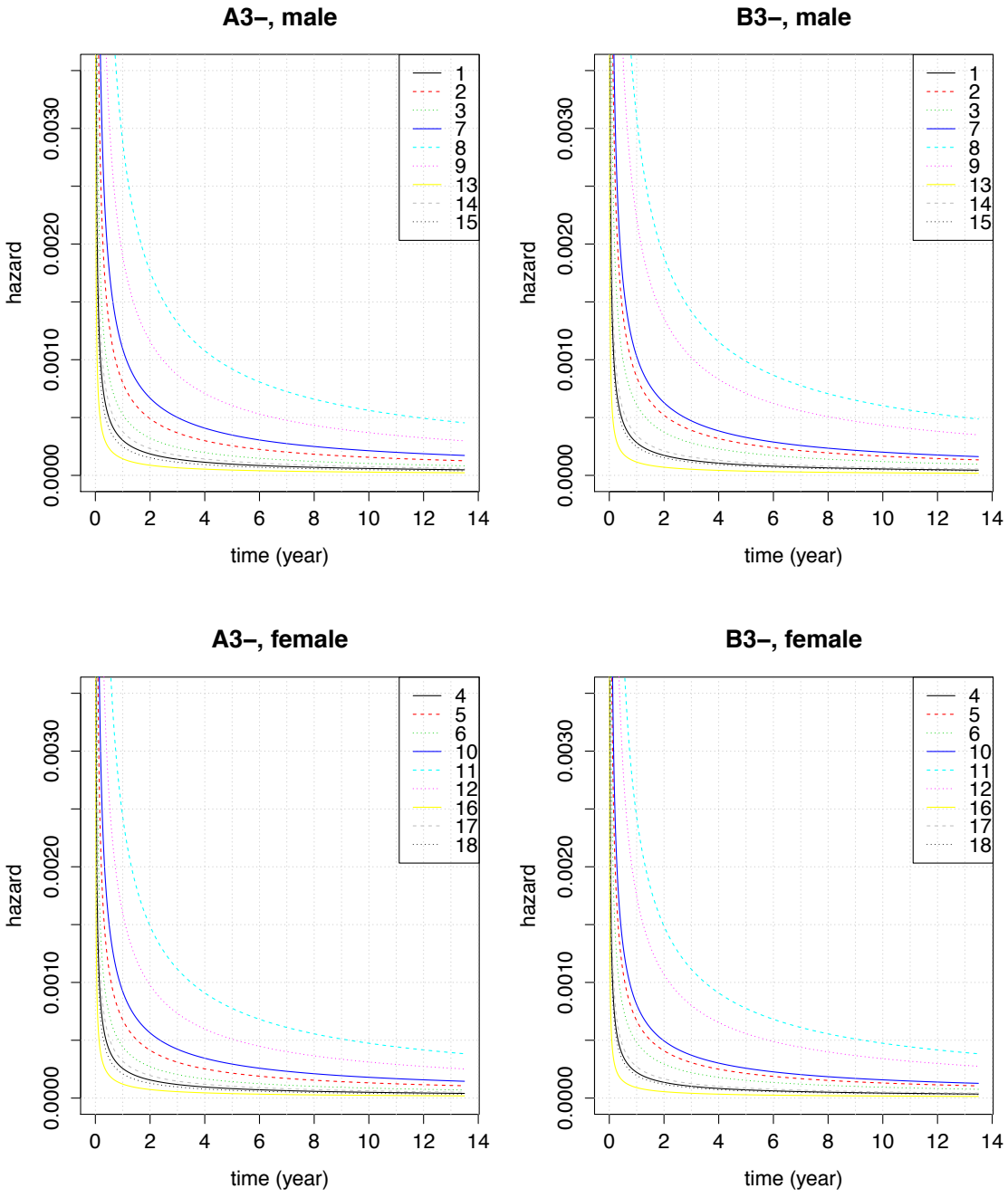

Figure C.9: Marginal hazard functions of A3 and B3 with source type negative and different gender

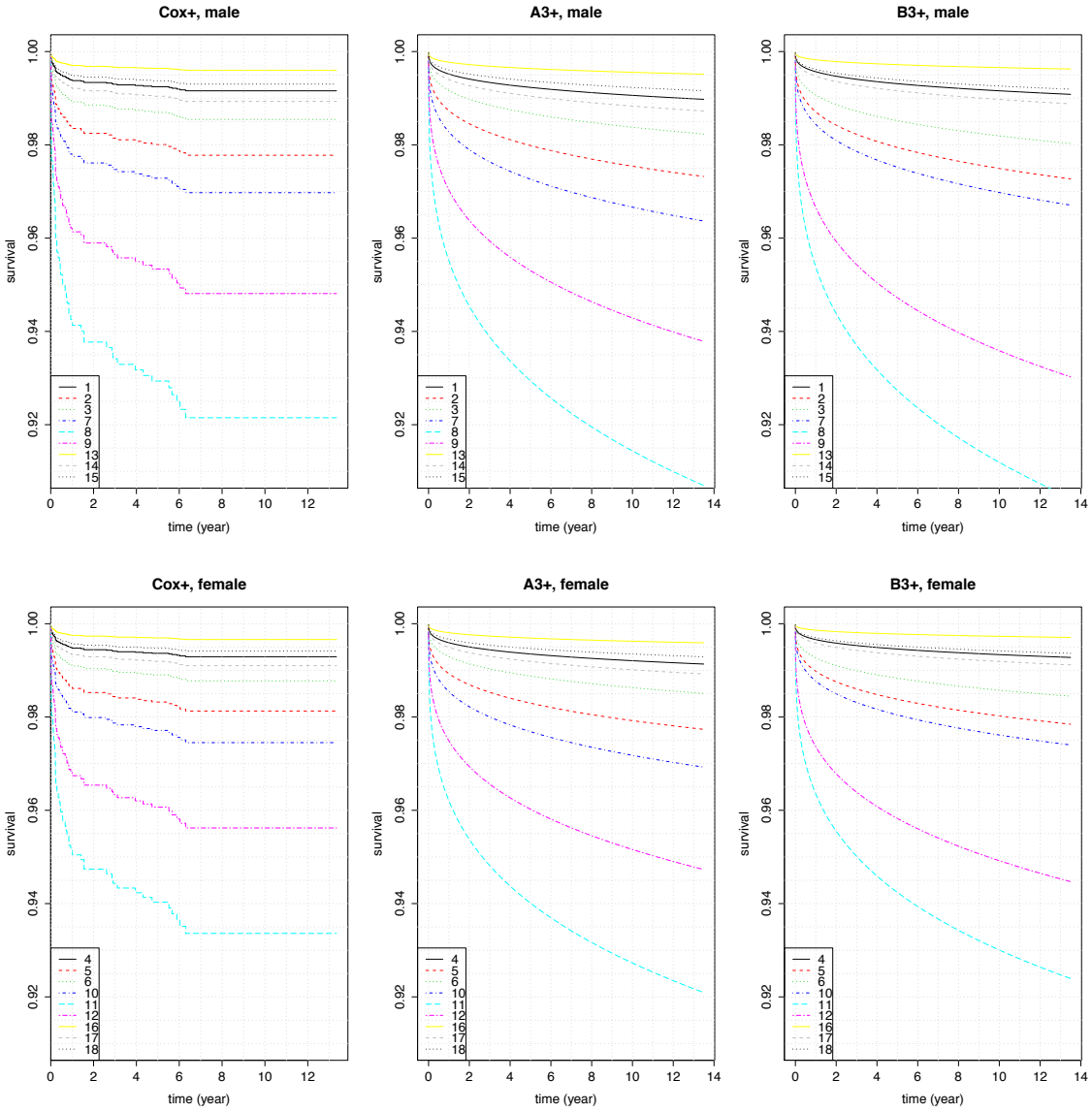

Figure C.10: Marginal survival functions of the Cox proportional hazards model fit, A3 and B3 with source type positive and different gender

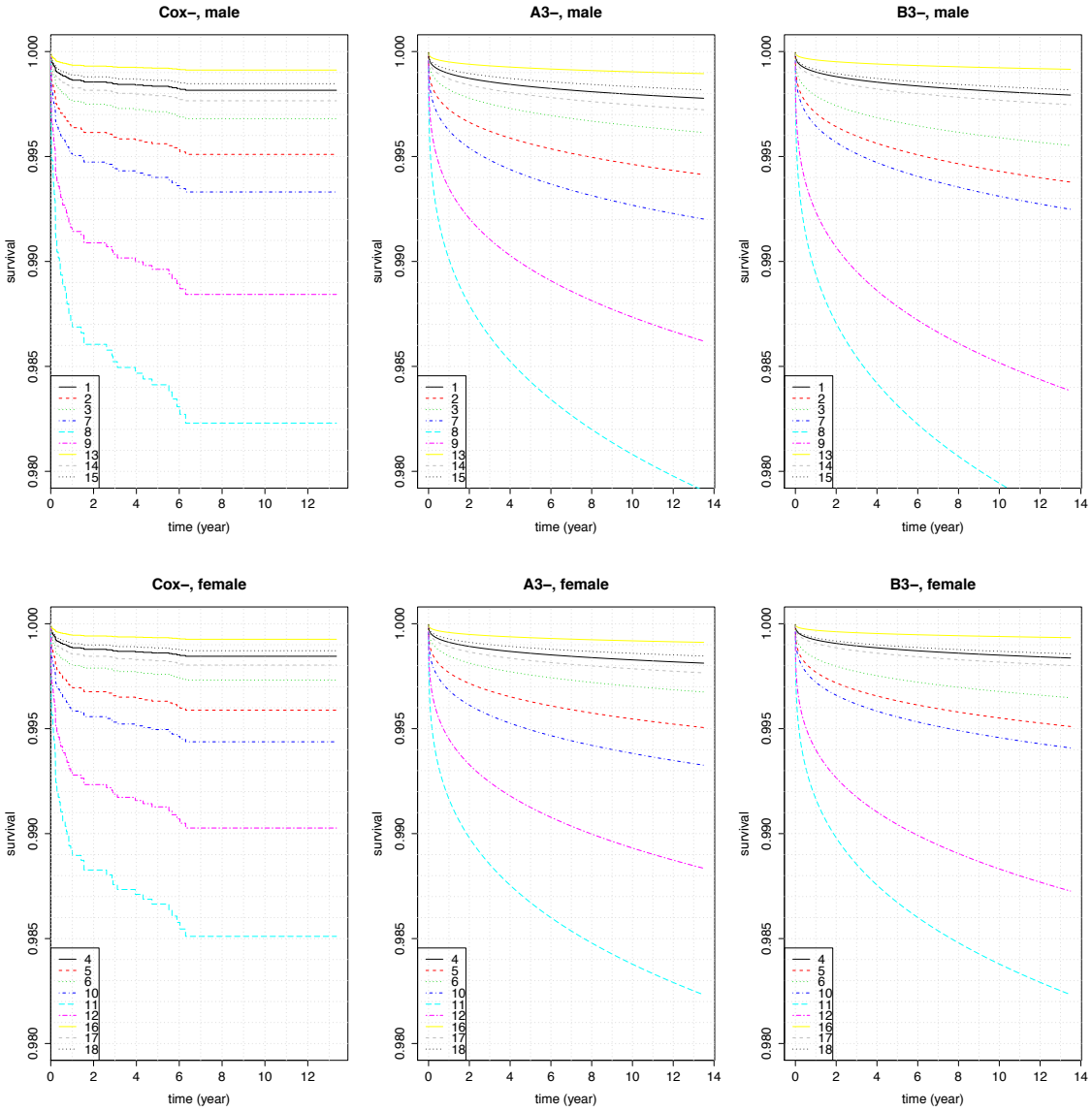

Figure C.11: Marginal survival functions of the Cox proportional hazards model fit, A3 and B3 with source type negative and different gender