Heriot-Watt University

Models for Income Protection Insurance Incorporating Cause of Sickness

Sing Yee Ling

January 10, 2009

Submitted for the degree of
Doctor of Philosophy
on completion of research in the
Department of Actuarial Mathematics & Statistics,
School of Mathematical and Computer Sciences.

I hereby declare that the work presented in this thesis was carried out by myself at Heriot-Watt University, Edinburgh, except where due acknowledgement is made, and has not been submitted for any other degree.

Sing Yee Ling (Candidate)

Professor Howard R. Waters (Supervisor)

Professor A. David Wilkie (Supervisor)

Date

Abstract

The Continuous Mortality Investigation (CMI) of the Institute of Actuaries and the Faculty of Actuaries in the UK established, in CMI Report 12 (1991), a multiple state model consisting of three states (Healthy, Sick and Dead) for the analysis of Income Protection Insurance (IPI) data. The transition intensities between states, estimated using a set of homogeneous male IPI data from 1975-78, are also presented in this report. Based on these estimated transition intensities, premium and reserve in respect of IPI business can be calculated. By using this model, in which there is only one Sick state to represent all causes of sickness, a whole portfolio of claims, regardless of their cause of sickness, will be subject to the same termination assumption. With cause of sickness as an important source of heterogeneity among IPI claimants, Cordeiro (1998, 2002) further developed this model so that it can be used to analyse IPI data by cause of sickness and obtained approximations to the cause-specific transition intensities defined in this new model. The main application of obtaining cause-specific termination assumptions is in the area of reserving more reliably for a portfolio of claims consisting of different causes of sickness.

In this thesis, we present methods and results for the estimation of the recovery and mortality intensities from sick by cause of sickness using IPI data provided by the CMI. There are 70 possible causes of sickness. The recovery intensity model for each cause of sickness assumes a multiplicative structure and is estimated in a structured manner with the use of the Cox model (Cox, 1972) and generalised linear models (GLM). The mortality intensity from sick is modelled using an additive relative survival model in which the excess mortality as a result of being sick is measured relative to the mortality intensity for a standard population. Finally, two applications of the recovery and mortality intensities from sick by cause of sickness are presented.

Acknowledgements

First of all, I wish to express my deepest appreciation and gratitude to my supervisors Professor Howard Waters and Professor David Wilkie for their dedication, guidance, encouragement and patience throughout this study. I have benefited greatly from their wealth of experience and insights in the course of bringing my research to fruition. They have my utmost admiration and I feel very honoured to be their student.

I also wish to thank the Continuous Mortality Investigation of the Institute of Actuaries and the Faculty of Actuaries which, through my supervisors, has provided me with the set of data used in this thesis.

I express my utmost appreciation to my family for their unconditional love and support all this time. Their unwavering encouragement gives me the strength to achieve my dreams and for that I am indebted to them.

I am also very grateful to all my friends for their wonderful friendship and for making my stay in Edinburgh most enjoyable. I express my special appreciation to Mannyee Leong and Yeewien Liew for their psychological aid despite the physical distance; Asad Nazir for delighting me with pleasant meals; and all my fellow PhD students in the department for making my time here a memorable one. Thanks and appreciations are also due to Xiaobin Lin, Kenny Mcivor and Xingchen Wang.

Contents

List of Figures

List of Tables

Introduction

Income protection insurance (IPI) is a class of long-term insurance that provides an income while the insured is unable to work due to sickness or disability for a period longer than the deferred period specified in the policy. It is known as disability insurance in the US, disability income insurance in Australia and was formerly known as permanent health insurance in the UK.

As with other insurance contracts, the insurer needs to assess the future expected cash outflows associated with a portfolio of IPI contracts so that a suitable premium can be set. These cash outflows will be dominated by benefits payable to insured lives who subsequently become unable to work as a result of sickness or disability. To make sure that the insurer can meet its future obligation while remaining solvent, the actuary is required to assess the magnitude and timing of these uncertain future cash flows and to recommend a suitable premium rate that will meet these future benefit payments to ensure the profitablity of the business. The largest and most uncertain future cash flow in the management of IPI buiness is the payment of the insured benefit which can vary considerably in terms of the timing of the claim and the duration of the claim. To assist the actuary in the assessment of the future liabilities in relation to IPI business, a number of morbidity tables, based on data collected from life offices, have been produced in various countries. The tables published outside of the UK include the 1985 Commissioners Individual Disability A Table (CIDA) from US and IAD 89-93 Standard Table from Australia.

In the UK, the Institute of Actuaries and Faculty of Actuaries set up the Continuous Mortality Investigation (CMI) to carry out research into mortality and morbidity experience using data collected from UK life insurers. In particular, the Income Protection (IP) Sub-committee of the CMI is devoted to analysis of IPI data. In the early

years of their investigation, the methodology adopted by the IP Sub-committee was to compare the actual weeks of sickness to the expected weeks of sickness calculated on the basis of the Manchester Unity table. The Manchester Unity table (1893-197) is a morbidity table constructed based on the sickness experience in England in the late nineteenth century.

The CMI, in CMI Report 12 (1991), established a new methodology for the analysis of IPI data in the form of a multiple state model with three states: Healthy, Sick and Dead. This model is defined in terms of transition intensities between the states: sickness inception intensity (i.e. transition from Healthy state to Sick state), recovery intensity (i.e. transition from Sick state to Healthy state), mortality intensity from sick (i.e. transition from Sick state to Dead state) and mortality intensity from healthy (i.e. transition from Healthy state to Dead state). In order to make this model operational for pricing and reserving purposes, these transition intensities have to be estimated. For this purpose, CMI Report 12 (1991) presented graduation formulae for the sickness inception intensity, recovery intensity and mortality intensity from sick based on a set of homogeneous male IPI data from 1975-1978. This set of graduation intensities, known as SM1975-78, is used as a yardstick against which the sickness experience for subsequent years is compared using the actual/expected ratios. The results of such analyses were published in a series of CMI reports.

In the three-state model introduced in CMI Report 12 (1991), there is only one Sick state which the IPI policyholder enters when he becomes sick, regardless of his cause of sickness. In recognition of the cause of sickness or disability as an important observable source of heterogeneity among IPI claimants, Cordeiro (1998, 2002) presented a new multiple state model for the analysis of IPI data by cause of sickness or disability, in which the Sick state in the model defined in the previous paragraph is replaced by n different Sick states, one for each cause of sickness. This new model, which consists of $n + 2$ states, is also defined in terms of transition intensities which are n sickness inception intensities (i.e. transitions from Healthy to each of the n Sick states), n recovery intensities (i.e. transitions from each of the n Sick states to Healthy state), n mortality intensities from sick (i.e. transitions from each of the n Sick states to Dead state) and mortality intensity from healthy (i.e. transition from

Healthy state to Dead state). Constrained by a much less detailed set of IPI data which makes it impossible to obtain graduations for these transition intensities in a similar manner to the graduations obtained in CMI Report 12, Cordeiro (1998) devised a way to derive continuous functions which can be taken as approximations to these transition intensities using IPI male data from 1975 to 1978 classified by 18 sickness categories. The motivation behind this approximation is that these transition intensities, possibly for some cause of sickness, may share the same feature or shape as the corresponding graduated intensities proposed in CMI Report 12 but having a different level to the latter, which is verified by conducting tests of hypotheses based on the distributions of average sickness durations. In connection with this new multiple state model, Cordeiro (1998) also defined the basic probabilities which are required for the calculation of quantities concerning IPI business and presented formulae for the basic probabilities. The estimated transition intensities are then fed into these probabilities which are evaluated efficiently using numerial algorithms.

A later attempt to analyse IPI data by cause of sickness was carried out by the CMI Cause of Disability Working Party which was set up by the CMI IP Sub-committee in 2004 to analyse the IP sickness experience and possibly obtain graduations of termination experience by cause of sickness using a huge amount of IPI per policy data by cause of sickness from 1975 to 2002. In their first published report, CMI Working Paper 23 (2006), they presented an initial analysis of the claim inceptions and claim terminations (recovery and death) by major cause group using the actual/expected ratios on the basis of SM1975-78. The benefit of analysing IP sickness experience was also discussed in this working paper. It is reckoned that the area of most potential benefit to IPI practitioners is the ability to reserve more reliably for claims in payment. By using cause-specific termination assumptions, the resulting aggregate reserve for a whole portfolio of claims will be more reliable than one which is calculated using a single aggregate termination assumption. This is because if the termination assumption is cause-specific, the aggregate reserve can take into account the changing mix of underlying causes in the portfolio and any cause-specific medical advancement in the future. In addition to that, the study of IPI claims by cause is useful to the underwriting and claim control processes because it gives a better indi-

cation of the relative importance of the various causes to claims costs and a better understanding of the average length of claim by cause. The underwriting and claim control procedures adopted by life offices can have an impact on the profitability of their IPI business (Sanders and Silby, 1986).

The main purpose of this thesis is to estimate the recovery intensity and mortality intensity by cause of sickness as defined in Cordeiro (1998). This thesis can be viewed as an extension to the work in Cordeiro (1998) and CMI Working Paper 23 (2006). To achieve our objective, we were provided by the CMI with a large amount of IPI claim data, each of which contains various data fields describing the attributes of the IPI policy and the insured person as well as the cause of disability that gives rise to the claim. This thesis will highlight the role that a number of survival model techniques have to play in the analysis of IPI data. In particular, the Cox proportional hazards model, generalised linear models and relative survival models will be considered and their suitability for morbidity modelling in respect of IPI will be discussed. The structure of the thesis is as follows:

In Chapter 1 we describe IPI and present the multiple state model as proposed in CMI Report 12 (1991). Then, we present a multiple state model which enables the analysis of IPI claims by cause of sickness as proposed in Cordeiro (1998). Finally, we review the actuarial literature and CMI reports concerning the analysis of sickness experience among UK IPI claimants, which will place the research in Chapter 3 in an appropriate context.

In Chapter 2 we describe in detail the data set which is used for the graduation of the recovery intensity and mortality intensity from sick by cause of sickness. This set of IPI data comprises claim records for which payments have been made from 1975 to 2002 inclusive. We also describe the information included for each IPI claim and present the classification of the 70 possible causes of sickness into 12 sickness categories to enable meaningful comparison of results at the exploratory analysis stage.

In Chapter 3 we present the estimation of the recovery intensity model for each cause of sickness. The recovery intensity model takes on a multiplicative structure consisting of two components. The first component is the baseline intensity which is a function of sickness duration alone while the second component is the relative

risk which describes the multiplicative effects of covariates. The parameterisation for each component is carried out separately. The relative risk component is estimated separately from the baseline intensity by using the Cox model (Cox, 1972). Implicit in the Cox model is that covariate effects remain constant for all sickness durations (i.e the proportional hazards assumption). This assumption will be assessed by using a graphical diagnostic check based on the Schonefeld residuals (Schoenfeld, 1982). In the case that the covariate effects are duration-dependent, the Cox model is extended to include a duration-varying coefficient by expressing it as the constant coefficient of a suitably defined duration-dependent covariate. To parameterise the baseline intensity, we assume a piecewise constant structure and graduate the estimates by using a mathematical formula. Finally, with a fully parameterised recovery intensity model, all the parameters are estimated together using maximum likelihood estimation. The fitted model then undergoes a series of test to assess its goodness of fit. This chapter concludes with the presentation of the estimated recovery intensity models for a few causes of sickness with the remainder in Appendix B.

In Chapter 4 we propose a mortality intensity from sick model which is the sum of two different components. The first component is the 'base' mortality intensity which is derived from a standard population while the second component is the excess mortality intensity which can be interpreted as the mortality in excess of that experienced by a standard population as a result of being sick. In the case of IPI claimants, we regard the UK assured lives population as a reasonably comparable group from which the base mortality intensity can be obtained. This chapter is dedicated to the modelling of the 'base' mortality intensity, in which we present a structured approach to the estimation of the 'base' mortality intensity, separately for males and females, in the GLM framework by using the UK assured lives data set.

In Chapter 5 we estimate the 'excess mortality intensity', which is the other constituent component in the mortality intensity from sick model as proposed in Chapter 4. We show that by having a multiplicative structure for the excess mortality intensity model, all the parameters in the model can be estimated by using a GLM with Poisson error structure and a specially constructed link function. As in the recovery intensity model, the parameterisation of the excess mortality intensity is carred out

in a structured manner. We also describe the steps involved in transforming the IPI per policy data from the CMI so that it can be used for GLM analysis. Given the low number of deaths in the large majority of causes of sickness, we classify the 70 causes of sickness into 15 sickness groups which are then further grouped into 5 different sickness categories according to the shape of their mortality curves. Finally, we present the estimated excess mortality model for each of the 5 sickness categories.

In Chapter 6, we show the application of using the estimated recovery intensities and mortality intensities from sick by cause of sickness presented in Chapters 3 – 5 in the following two aspects:

(i) Calculating the expected present values of annuities by cause of sickness.

(ii) Deriving the aggregate recovery intensity and mortality intensity from sick.

Finally, in Chapter 7, we present our contributions and ideas for further research.

Chapter 1

Background

1.1 Introduction

In this chapter we give some background to the work presented in this thesis. In Section 1.2 we describe in general terms the nature of Income Protection Insurance (IPI). In Section 1.3 we present the multiple state model as proposed in CMI Report 12 (1991) to analyse IPI claims data. A generalisation of this multiple state model which enables analyses of IPI claims by cause of sickness is presented in Section 1.4. Finally, in Section 1.5, we present the results published in CMI reports and the actuarial literature which will serve as a useful introduction to Chapter 3.

1.2 Income Protection Insurance

Income protection insurance (IPI), formerly known as permanent health insurance in the UK, is a class of long-term insurance that provides an income while the insured is unable to work due to sickness or disability. Once an IPI policy is effected, it cannot be cancelled by the insurer, other than in very exceptional circumstances specified in the policy.

There are two main types of IPI policy: individual policies and group policies. In this thesis we will focus on the claim experience in relation to individual IPI policies. The basic features of an IPI policy are illustrated in Figure 1.1. The symbols 'H','S' and 'DP' in this diagram represent 'healthy', 'sick' and 'deferred period', respectively.

Under an individual IPI policy, the insurer is obliged to pay the policyholder income during periods of disability longer than the deferred period. In other words, a policyholder has to remain sick for at least as long as the deferred period of his policy in order to make a claim and receive benefit. The claim payments will stop once the claimant recovers from the sickness/disability. However, if the same sickness recur within a very short period of recovery, it will usually be treated as claim revival by the insurance company with the deferred period requirement being waived so as to encourage earlier recovery. In exchange for these benefits, the policyholder has to pay premium from the time he/she effects the policy until the end of the policy term or to the retirement age of 60 for women and 65 for men, except when the policyholder is in receipt of benefit. The common deferred periods in the UK market are 1 week, 4 weeks, 13 weeks, 26 weeks and 52 weeks. In this thesis, they are represented by the symbols DP1, DP4, DP13, DP26 and DP52, respectively.

Both benefit and premium can take many forms. The benefit received may be in the form of a regular income, increasing income linked to inflation or income that includes an investment element on a unit-linked or with-profit basis. In all cases, the benefit received is usually less than a set percentage of the claimant's previous income so as to provide an incentive for the claimant to return to work. The premium paid may be renewable or reviewable after a period of a few years or guaranteed for the policy term.

While the basic features of various individual IPI policy issued by different companies are generally the same, the individual IPI product is a complex and highly variable product because the specific policy conditions can vary considerably between insurers. These policy conditions include the definition of disability that triggers benefit payment, exclusion of certain cause of claims, the size of benefit relative to the insured's pre-disability income and age at entry of the policyholder. The varying policy conditions adopted by each insurer that write IPI business in UK can be found in Kluwer's Income Protection Insurance (2001).

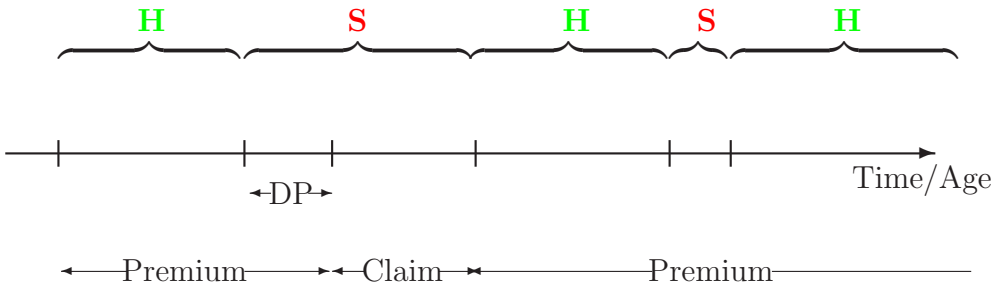


Figure 1.1: Policy design of IPI.

1.3 A Multiple State Model for IPI

The CMI introduced a semi-Markov model for pricing and reserving in respect of IPI business in CMI Report 12 (1991). This model consists of three states: Healthy (H), Sick (S) and Dead (D). A diagrammatic representation of this three-state model is shown in Figure 1.2 and will be used to give an intuitive explanation of this model.

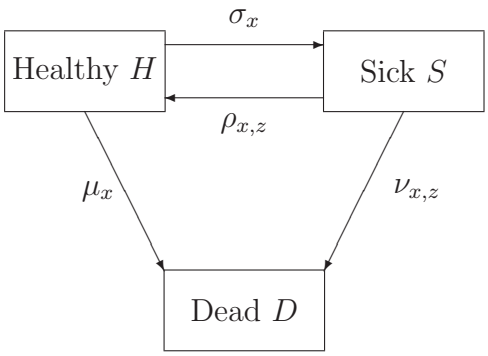


Figure 1.2: A multiple state model for IPI in which a policyholder may move between these three states, with death as an absorbing state.

Once an IPI policy is effected, at which time the policyholder is supposed to be healthy, he enters state H. From this state, he may transfer at any future time either to state S (i.e. he becomes sick) or to state D (i.e. he dies). The transition intensities associated with these two transitions are denoted σ_x and μ_x, respectively. Both these transition intensities are dependent only on x, the policyholder's attained age.

Once the policyholder is in state S, he may transfer either back to state H (i.e. he recovers) or to state D (i.e. he dies). The transition intensities in connection with these transitions are denoted by $\rho_{x,z}$ and $\nu_{x,z}$, respectively. Both these transition intensities depend on both x, the policyholder's attained age, and on z, the duration

of his current sickness (no account is taken of previous sickness). Both states H and S are transitive while state D is absorbing.

The movement of a policyholder in this multiple states model can be described by a pair of continuous time stochastic processes

$$\{Y(x), Z(x)\} \quad x > 0 \quad \text{and} \quad y > 0 \tag{1.1}$$

where $Y(x)$ denotes the state in which the policyholder is at age x and $Z(x)$ denotes the duration of his sojourn so far in the current state $Y(x)$. In formal terms, $Z(x)$ is defined as follows:

$$Z(x) = \max\{t : t \leq x \text{ and } Y(x-h) = Y(x) \ \forall \ h : \ 0 \leq h \leq t\}. \tag{1.2}$$

$Y(x)$ can takes any of the three values H, S and D while $Z(x)$ takes value in the set $[0,\infty)$.

The joint process (1.1) is assumed to be a Markov process so that the future of the process depends only on the values of $Y(x)$ and $Z(x)$ and not on any information prior to age x. This means that if a policyholder has just become sick, his transition probability into either state H or state D takes no account of his prior sickness history. On the other hand, the transition probability from state H to state S is the same for a healthy policyholder who has just effected his policy as for a policyholder of the same age who is healthy but has just recovered from a long sickness. We make this strong assumption because it is not possible to infer any prior sickness history (e.g. whether or not the person has just recovered from the same or another sickness) about the claimant from the available data to enable the fitting of a more realistic model.

1.4 A Multiple State Model for IPI by Cause of Sickness

In the three-state semi-Markov model presented in Figure 1.2, there is only one sick state to represent all possible causes of sickness which the policyholder enters whenever he becomes sick. This model assumes that regardless of the cause of sickness, all IPI

claimants are subject to the same recovery and mortality intensities.

Cordeiro (1998, 2002) developed a new multiple state model which enables the analysis of IPI claims by cause of disability. Figure 1.3 gives a diagrammatic representation of this new multiple state model by cause of disability. There are $n+2$ states in this new multiple state model. In addition to states H and D, there are n sick states $(S_1, S_2, \ldots, S_n)$, each of which represents a specific cause of sickness, in place of a single sick state to represent all of them.

As in the three-state semi-Markov model in Section 1.3, a policyholder enters state H when his policy is effected. From there, he can transfer at any future time to state D or to one of the n sick states depending on his disability or sickness. The transition intensities associated with these transitions are respectively denoted by $\sigma(i)_x$: $H \rightarrow S_i$ $(i = 1, 2, \ldots, n)$ and $\mu_x : H \rightarrow D$. Both $\sigma(i)_x$ and μ_x depend only on x, the policyholder's attained age.

From the sick state S_i $(i = 1, 2, \ldots, n)$, a person can return to either state H (i.e. recover) or state D (i.e. die) at any future time. The transition intensities associated with these transitions are respectively denoted by $\rho(i)_{x,z}$: $S_i \rightarrow H$ $(i = 1, 2, \ldots, n)$ and $\nu(i)_{x,z} : S_i \rightarrow D$. Both $\rho(i)_{x,z}$ and $\nu(i)_{x,z}$ depend on the policyholder's attained age x, and the duration of his sickness z.

It should be noted that in this new multiple state model, transitions between the n sickness states, which represent different cause of disability are not allowed. We assume that a person who falls sick from a specific sickness cannot develop another disease until after he recovers (returns to state H) from his original sickness. This is despite the fact that it is entirely possible, for example, for a person who suffers from diabetes to develop stroke without having to recover from diabetes first. If an IP claimant is affected by another illness while he is claiming under the original sickness, it is common practice for insurance companies to continue with the payment of benefit under the original cause of disability. The nature of the data gives us no information about any transition between the sickness states and thus, these transition intensities cannot be modelled.

As in Section 1.3, the movement of a policyholder in this new multiple state model can also be described by a pair continuous time stochastic processes

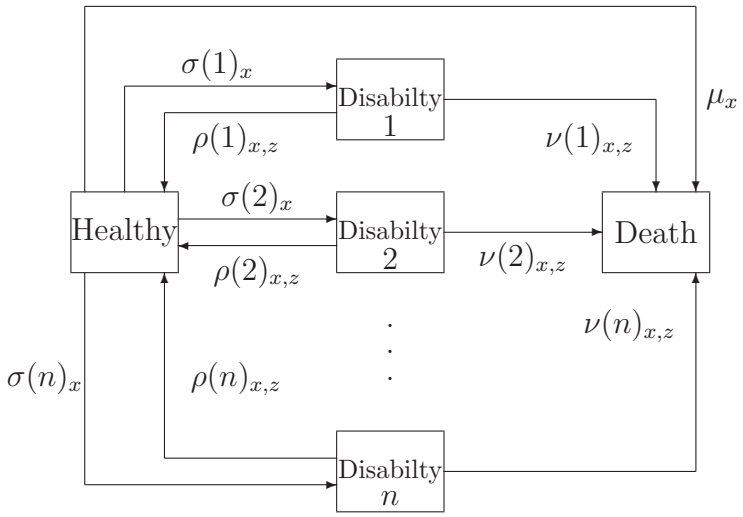


Figure 1.3: A multiple state model for the analysis of IPI claims by cause of disability.

$$\{Y(x), Z(x)\} \quad x > 0 \tag{1.3}$$

where $Y(x)$ is the state in which the policyholder is at age x and $Z(x)$ denotes the duration of the sojourn so far in the current state $Y(x)$. In this new multi-state model, $Y(x)$ takes values in the set $H, S_1, S_2, \ldots, S_n, D$ while $Z(x)$, as defined by Equation (1.2), takes values in $[0,\infty]$. As in Section 1.3, the joint process (1.3) is assumed to be a Markov process.

1.5 A Review of Past Results

The main focus of this thesis is the graduation of recovery intensities and mortality intensities from sick by cause of sickness associated with the multiple state model by cause of sickness as presented in Figure 1.3. There are a number of papers that deal with statistical and actuarial problems in related fields, such as long-term care and Continuing Care Retirement Communities. In particular, Jones (1992) used a Markov process to analyse US statistical data concerning long-term care and Jones (1995) presented a multi-state stochastic model for analyzing continuing care retirement community populations.

Despite having a large database of information relating to cause of sickness, the CMI has not carried out such an analysis and only limited use has been made of this dataset to date. Nevertheless, the CMI has, in the past, produced graduated recovery intensities and mortality intensities from sick using aggregate data from all causes of sickness. For example, CMI Report 12 (1991) presented both the graduation of the recovery intensity and mortality intensity from sick in Part B and the graduation of the sickness inception intensity in Part C. These graduated intensities, derived from a set of homogenous IPI data from 1975-78, are collectively referred to as SM1975-78 and form the basis for comparison for subsequent quadrenniums. In addition to SM1975-78, CMI Working Paper 5 (2004) presented recovery and mortality intensities from sick based on more recent sickness experience by using IPI data during the period 1991-1998. This set of intensities is referred to as IPM91-98.

As a prelude to Chapter 3 which centres on estimating the recovery intensity by cause of sickness, we examine, in this Section, the modelling techniques used in the graduation of recovery intensities as presented in both the CMI reports mentioned above as well as their relevant important findings. Subsequently, in search of the evidence of year trend in the sickness experience, we examine results on the basis of SM1975-78 for the sickness experience for each quadrennium in the period 1975 – 2002 as published in a series of CMI Reports. We also present findings from Renshaw and Haberman (2000) who studied year trends in the sickness experience. Finally, we summarise the main findings of these past analyses.

CMI Report 12 (1991)

In association with the three-state semi-Markov Model for IPI business (see Section 1.3), CMI Report 12 (1991), in Part B, presented the graduation of the recovery intensity $S \rightarrow H$ using male 'Standard' data for 1975-78. The idea of 'Standard' data, first introduced in CMI Report 7 (1987), refers to a more homogeneous subset of the total data that consists of policies issued in the UK, policies without occupational rating or known health impairment and with a regular benefit payment.

The general form of the graduation formula was developed by investigating the age and durational effects. At the exploratory analysis phase, a multiplicative model

of durational factors and age factors was fitted for each deferred period. For DP1, a plot of the log of the durational factors for up to one year against the square root of duration produced a graph that was approximately linear and decreasing. Apart from 4-week 'run-in' periods of lower recovery intensities immediately after the end of the deferred period, a similar linearity was found for DP4, DP13 and DP26 and there was no significant difference between deferred periods. The linear trend, however, is not continued for sickness durations exceeding one year. The 'run-in' phenomenon was discussed at some length in CMI Report 12, Part B, Section 3.3. It is generally regarded as being caused by people who do not submit a claim when their recoveries are imminent at the end of their deferred period. To assess the variation of recovery intensity by age, the age factors for each deferred period were plotted against age. It transpired that there is a broadly linear relationship of the factors with age with no great disparity between the deferred periods. Apart from the need to make special adjustment for the 'run-in' periods, the data from all deferred periods was combined to recalculate the age and durational factors. A detailed diagnostic check revealed that such a simple multiplicative model fails to fit the data sufficiently well. Further piecewise adjustment was made to account for the change in the slope of the log linear variation of recovery intensity with duration after one year and the linear dependency of age with $\sqrt{z}$ with a steeper negative slope before four weeks of sickness duration. For practical reasons, the recovery intensities depend only on attained age after five years of sickness (i.e $z > 5$).

The complete graduation formula for the recovery intensity in SM1975-78 is as follows:

$$\rho_{y+z,z} = r\ \{a + b(1 + q(4 - wz)_+)\sqrt{Z}(Y - 50)\}\ e^{-c\sqrt{Z}} \tag{1.4}$$

where Y is a function of the exact age y (in years) at sickness inception while Z is a function of sickness duration z in years, such that

$$Y = \begin{cases} y & \text{for} \quad z \leq 5 \\ y + z - 5 & \text{for} \quad z > 5 \end{cases} \qquad Z = \begin{cases} z & \text{for} \quad z \leq 1 \\ 1 + s(z-1) & \text{for} \quad 1 < z \leq 5 \\ 1 + 4s & \text{for} \quad z > 5 \end{cases}$$

and r is defined as

$$r = \begin{cases} 1 & \text{for DP1} \\ \min\{(p + wz - d)(1-p)/4, 1\} & \text{for DP4, DP13, DP26} \end{cases}$$

where w= 52.18 (assuming that there are 52.18 weeks in a year), d is the deferred period measured in weeks, and

$$a = 51.057202 \qquad b = -2.687089 \qquad c = 4.914441$$
$$p = 0.205111 \qquad q = 1.419428 \qquad s = 0.362456$$

Figure 1.4 provides a visual comparison of the graduated recovery intensities for a male aged 40 at sickness inception for the different deferred periods as a function of sickness duration. This figure shows the 'run-in' periods for DP4, DP13 and DP26, the change of slope before four weeks of sickness duration and after one year of sickness duration.

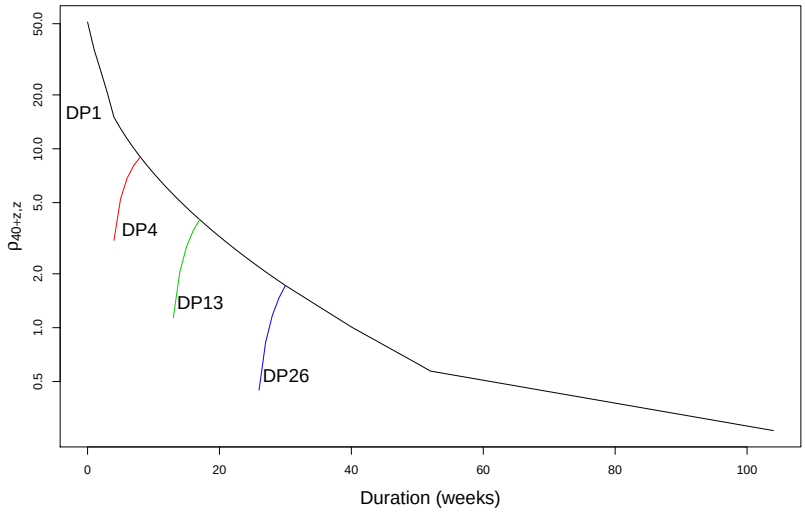


Figure 1.4: The recovery intensities for a male IPI claimant aged 40 at sickness inception according to SM1975-78.

CMI Working Paper 5 (2004)

In addition to SM1975-78, a set of graduated recovery intensities based on more recent sickness experience was published in CMI Working Paper 5 (2004). The data used in this graduation is from IP males, occupational class 1 from 1991 to 1998 inclusive. Despite the data spanning eight years, the calendar year effect was not investigated because preliminary investigation by the author revealed no clear time trend. The combined data from each year was used for the graduation. In common with SM1975-78, the recovery intensity after five years of sickness is dependent only on attained age and similar exploratory analysis techniques to those adopted in CMI Report 12 (1991) were used. While there are broad similarities in the recovery pattern between this graduation and that of SM1975-78, there are a few distinctive differences too. The resulting graduation formula is an additive log linear function of age y at sickness inception, duration z and deferred period d given by

$$\log \rho(d, y, z) = s_d + g_z + q_z + r_z + f_{yz} + h_{yz}.$$

In view of the rather elaborate terms comprised in each of the components above, the full graduation formula is presented in Appendix D. In essence, the meanings of the various components in the above model are as follows:

(i) s_d consists of factors that account for the differences in the level of recovery intensity between different deferred periods.

(ii) g_z consists of piecewise linear terms of the transformed duration variable, $t(z)$, such that $t(z) = w/(1 + 0.025w)$ where $w = 365z/7$.

(iii) q_z is a piece-wise linear function of $t(z)$ aimed to adjust for the change in slope during $8 < w \leq 16$ in DP4.

(iv) r_z consists of piece-wise linear terms that account for the increasing linear trend during the four weeks of run-in period noticeable in DP4 and DP13.

(v) f_{yz} is a cubic function of age y with the coefficient of the linear age term varying negatively with $t(z)$.

(vi) h_{yz} is a function of age y and duration z applicable only to sickness durations less than four weeks.

To give a visual impression of features (i) to (iv), the graduated recovery intensities for a male aged 40 at sickness inception, computed separately for each deferred period, are shown in Figure 1.5.

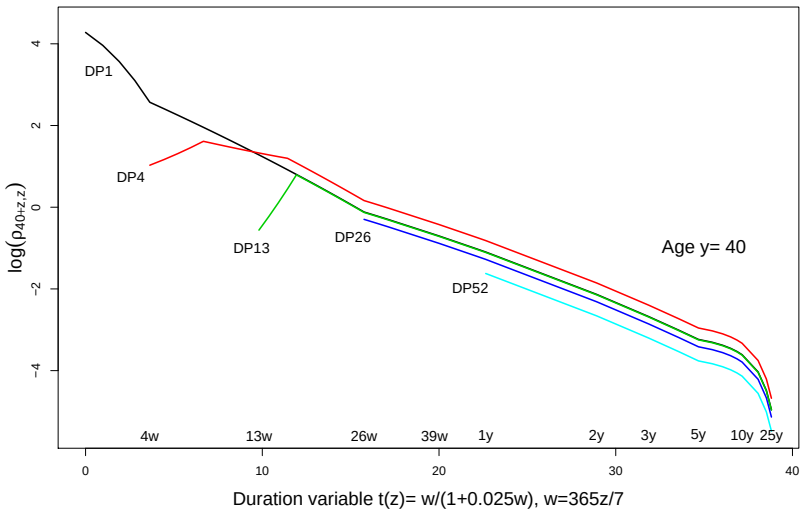


Figure 1.5: The graduated recovery intensities for a male IPI claimant aged 40 for separate deferred periods.

This set of graduated intensities is referred to as IPM91-98 and is used as a comparison basis for other males and females occupational classes from the 1991-1998 data. The results of such an analysis are reported in CMI Working Paper 7 (2004).

Other CMI Reports

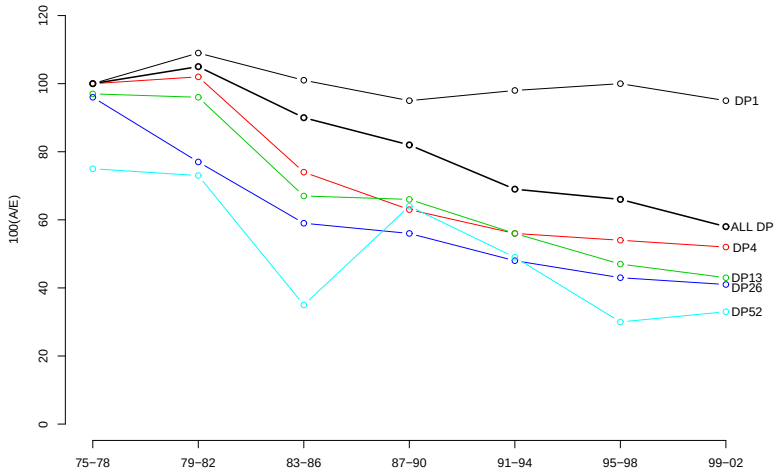
The standard table SM1975-78 has been used by the CMI to assess future sickness experience. The methods used by the CMI to analyse IPI claims experience are based on a comparison of actual number (A) of recoveries and deaths versus expected (E) on the basis of SM1975-78. The detailed methodology was set out in CMI Report 15 (1996).

The IPI data is typically grouped by quadrennia before such comparisons are carried out. For the years 1975-1990, the analysis is based on 'Standard' data; for years since 1991, it is based on the 'Standard*' data. This new subset of the total data called 'Standard*' data, introduced in CMI Report 18 (2000), is designed to make use of the 'occupational class' information collected since 1991. This set of data is created by using the same criteria as for the 'Standard' data but ignores the content of the 'occupational rating' field. A description of the data fields included in each IPI data can be found in Section 2.2. The 'Standard*' data therefore forms a larger subset of the total data than the 'Standard' data. The results are then published in a series of CMI Reports, i.e. CMI Report 15 (1996) for the 1975-78, 1979-82, 1983-86 and 1987-90 experience, CMI Report 18 (2000) for the 1991-94 experience, CMI Report 20 (2001) for the 1995-98 experience and CMI Report 22 (2005) for the 1999-02 experience.

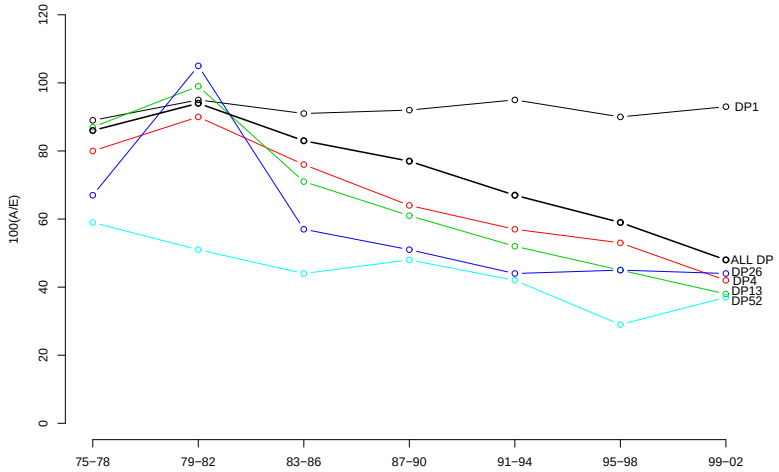
Using the A/E values obtained from these CMI Reports, Figure 1.6 shows $100A/E$ for the number of recoveries by deferred period and quadrennium over 1975-2002 for all sickness durations combined, separately for males and females. Figure 1.7 depicts the $100A/E$ for the number of recoveries by sickness duration and quadrennium over 1975-2002 for all deferred periods combined, separately for males and females. From these graphs, we see that

(i) There is a strong declining trend in both males and females recovery intensities over the period 1975-2002 for all deferred periods apart from DP1.

(ii) The sickness experiences for the first 3-4 weeks, in both males and females, seem to depart from the general falling trend observed in sickness durations greater than 3 weeks. No explanation is given by the IP Sub-committee for this feature.

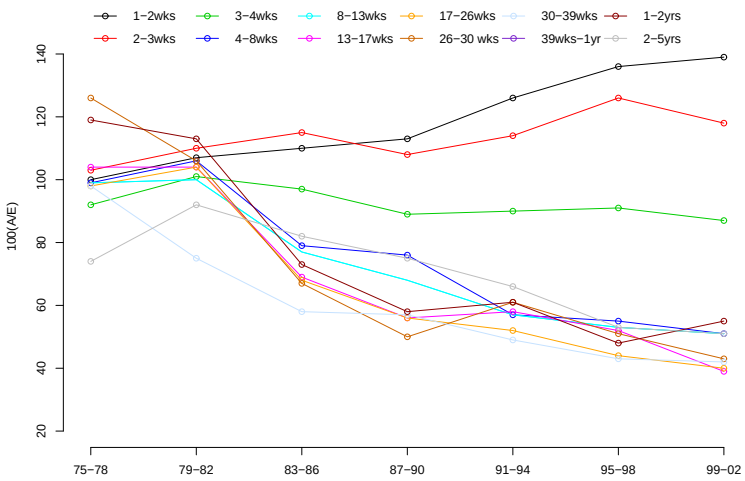


1.6.1: Male

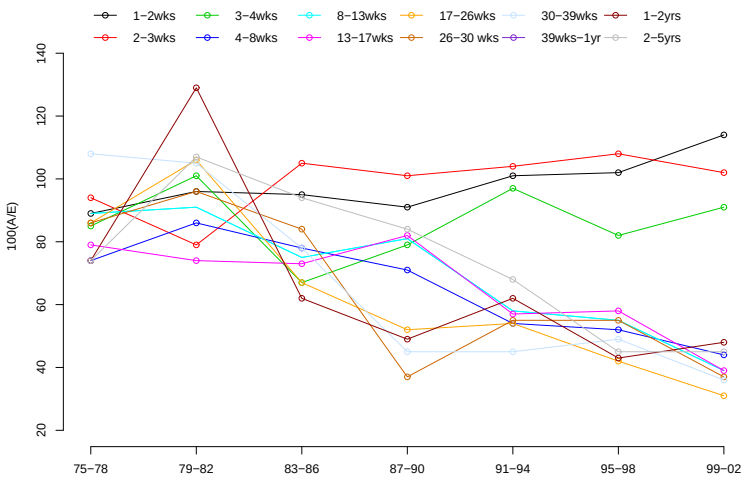


1.6.2: Female

Figure 1.6: Values of $100A/E$ (all durations combined) by deferred periods and quadrennium over 1975-2002. Standard data for 1975-1990; Standard* data, all occupational classes for 1991-2002. Expected values based on SM1975-78. All values obtained from CMIR 15, 18, 20 and 22.



1.7.1: Male



1.7.2: Female

Figure 1.7: Values of $100A/E$ (all DP combined) by sickness durations and quadrennium over 1975-2002. Standard data for 1975-1990; Standard* data, all occupational classes for 1991-2002. Expected values based on SM1975-78. All values obtained from CMIR 15, 18, 20 and 22.

Renshaw and Haberman (2000)

In both CMI Report 12 (1991) and CMI Working Paper 5 (2004), no time trend is incorporated in the recovery intensity model due to the relatively short period of calendar years that the data covers. Both Figures 1.6 and 1.7 show evidence of a general declining trend in the sickness experience from 1975 to 2002. The results from Renshaw and Haberman (2000) lend support to this phenomenon.

Renshaw and Haberman (2000) studied the presence of any significant time trend in the sickness experience of UK IPI claimants for males and females separately by using IPI data spanning 20 calendar years from 1975 to 1994 inclusive. To do so, they make use of the results in an earlier paper, Renshaw and Haberman (1995), which used a GLM based approach to model the transition intensities in the IPI multiple-state model separately for each deferred period by using IPI male 'Standard' data for 1975-78, the same set of data used in the graduation of the recovery intensity in CMI Report 12 (1991). The underlying method in Renshaw and Haberman (2000) is to use the model structure for each deferred period found in Renshaw and Haberman (1995) as their starting point and let the parameters in each model vary by every single calendar year from 1975 to 1994.

To illustrate how their method works, we focus on the recovery intensity for males DP1 as an example. Using the same notation as in Renshaw and Haberman (2000), we take $\rho_{y+z,z}$ as the recovery intensity for a person aged y (in years) at sickness inception and has been sick for duration z (in years). The model structure for males DP1 in Renshaw and Haberman (1995) is given by

$$\log(\rho_{y+z,z}) = \beta_0 + \beta_1\sqrt{z} + \beta_2 z + \beta_3 y + \beta_4 y\sqrt{z} + \beta_5 yz$$

In the search for a time trend, the parameters in the above model are to vary by every single calendar year from 1975 to 1994. Each set of these year-dependent coefficients is tested for statistical significance and is plotted against calendar year t to examine the trend of dependency. The recovery intensity as a log-linear function of age y, duration z and year t for males DP1 in Renshaw and Haberman (2000) is

$$\log(\rho_{y+z,z,t}) = (\beta_0+\beta_6 t)+(\beta_1+\beta_7 t)\sqrt{z}+(\beta_2+\beta_8 t)z+\beta_3 y+\beta_4 y\sqrt{z}+\beta_5 yz+\beta_9 y^2+\beta_{10} y^3$$

where

$\beta_0 = 2.76175$ $\beta_1 = -9.9429 \times 10^{-1}$ $\beta_2 = 3.2281 \times 10^{-2}$ $\beta_3 = 1.6973 \times 10^{-1}$

$\beta_4 = 5.6928 \times 10^{-3}$ $\beta_5 = -3.4555 \times 10^{-4}$ $\beta_6 = 4.2984 \times 10^{-2}$ $\beta_7 = -2.633 \times 10^{-2}$

$\beta_8 = 1.3632 \times 10^{-3}$ $\beta_9 = -4.4399 \times 10^{-3}$ $\beta_{10} = 2.9982 \times 10^{-5}$

Renshaw and Haberman (2000) interpreted this model from three perspectives, which include the following:

(i) The log recovery intensity varies linearly with year for fixed z. The coefficient of year is duration dependent and is given by $4.2984 \times 10^{-2} - 2.633 \times 10^{-2}\sqrt{z} + 1.3632 \times 10^{-3} z$. This implies that the recovery intensities have increased over the years for sickness durations less than three weeks and over 307 weeks, i.e. a positive year coefficient. For other sickness durations, the recovery intensity decreases over the years investigated, i.e. a negative year coefficient. This conclusion is generally consistent with the observations in Figure 1.7.

(ii) The log recovery intensity can be viewed as a quadratic function in $\sqrt{z}$. The coefficient of $(\sqrt{z})^2$ is $3.2281 \times 10^{-2} + 1.3632 \times 10^{-3} t - 3.4555 \times 10^{-4} y$ and is positive for all y and t values within the domain of data. Thus, it is a convex function with the minimum turning point lying beyond the domain of data.

(iii) The log recovery intensity is a cubic function in age y. The coefficient of the linear age y term is a quadratic function of $\sqrt{z}$ while the coefficients for y^2 and y^3 are constant terms.

The recovery intensity for males and other deferred periods are investigated separately in the same manner. For each deferred period, the following model structure is adopted:

$$\log(\rho_{y+z,z,t}) = \beta_0 + \beta_6 t + \beta_1 y + \beta_2 z + \beta_3 (z - z_0)_+ + \beta_4 (z - z_1)_+ + \beta_5 y (z - z_0)_+$$

for DP4, DP13, DP26 with $\beta_3 = \beta_5 = \beta_6 = 0$ for DP52. The value for z_0 is set equal to the length of the respective deferred period plus four weeks in view of the the 4 weeks of 'run-in' period reported in CMI Report 12 (1991). The authors found that this feature persists throughout the 20 investigation years. The location of z_1 is found by varying its locations in the model and examining the resulting deviance profile. The value of z_1 is 60.88 weeks for DP4 and 65.22 weeks for both DP13 and DP26. The coefficient for β_6 is negative for all deferred period, indicating a drop in recovery intensity over the years investigated.

Conclusions

After reviewing the results from the above four sources, we know that

(i) The year effect does not stay constant for all sickness durations. For sickness duration less than three weeks, the recovery intensity increases over the years but it decreases over the years thereafter (see Figures 1.6 and 1.7 and Renshaw and Haberman (2000)).

(ii) The age effect is duration-dependent (see CMI Report 12 (1991), CMI Working Paper 5 (2004) and Renshaw and Haberman (2000)).

(iii) The existence of a lower but linearly increasing recovery intensity during the 4 weeks of 'run-in' period that occur after the end of DP4, DP13 and DP26 respectively (CMI Report 12, 1991; see Figure 1.4). In CMI Working Paper 5 (2004), such a phenomenon is not observed in DP26 (see Figure 1.5).

Chapter 2

Data

2.1 Introduction

We wish to develop models for income protection insurance by cause of sickness. For this purpose, we were provided by the CMI with a set of IPI data, comprising claim records for which payments have been made during each investigation year from 1975 to 2002 inclusive. This claim data was contributed to the CMI for analysis by UK life insurance companies that write IPI business. In Section 2.2 we present the information included in each IPI claim data. For the purpose of presenting the IPI data split by 70 causes of sickness in a way that meaningful comparisons and contrasts can be made when exploratory analysis of the IPI data is conducted in Section 2.4, we present in Section 2.3 the classification of these causes of sickness into 12 sickness categories after reviewing various such groupings in the IPI literature.

2.2 Structure of the Data

Each IPI claim record contains various data fields which contain information about the claimant, attributes of the IPI policy and the nature of the claim. The coding of the data fields is set out in CMI Report 2 (1976). From 1991, the CMI started to collect 'occupational class' information by asking the life offices to submit their own occupational class code which is then converted by the CMI to its equivalent CMI occupational class code. The information included in each claim record is as follows:

(a) Sex: male and female are coded as 1 and 2, respectively

(b) Deferred period: 1, 4, 13, 26 and 52 weeks and are coded as 1, 2, 3, 4 and 5, respectively

(c) Occupational rating: no rating, rated and more rated are coded as 0, 1 and 2, respectively

(d) Age last birthday at sickness inception, taking values from 17 to 69

(e) Calendar year of the claim payment, taking values from 1975 to 2002

(f) Occupational class: The four occupational classes employed by the CMI are translated into codes 1, 2, 3 and 4, respectively. These occupational classess are described in CMI Report 18 (2000) and they are reproduced here as follows:

Class 1 Professional, managerial, executive, administrative and clerical classes not engaged in manual labour

Class 2 Master craftsmen and tradesmen engaged in management and supervision; skilled operatives engaged in light manual work in non-hazardous occupations

Class 3 Skilled operatives engaged in manual work in non-hazardous occupations

Class 4 Skilled and semi-skilled operatives engaged in heavy manual work or subject to special hazard

(g) Cause of claim: There are 70 possible causes of sickness and they are coded as 1-70 according to the Abbreviated List C in the Eighth Revision of the *Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death*. In this thesis, this list will be referred to as 'ICD8'. A copy of this list is produced in Appendix A.

(h) Month and year of birth

(i) Date of sickness inception

(j) Date of payment commencement

(k) Mode of payment commencement: continuation of claim payment from preceding year, new claim, new claim after interruption, revival and benefit change are coded as 0, 1, 2, 3 and 4, respectively. New claim after interruption refers to new claim that starts after certain interruptions during the deferred period while revival refers to claim which is revived due to recurrence of the same sickness within a very short period of recovery. In the investigation of recovery and mortaliy intensities by cause of sickness in the subsequent chapters, no account is taken of the mode of payment commencement

(l) Date of payment cessation

(m) Mode of payment cessation: recovery, death, expiry and continuation of claim payment in succeeding year are coded as 0, 1, 2, 3 and 4, respectively

As noted above, the CMI did not collect 'occupational class' code information until 1990. However, for IPI data collected before 1990, we find values of 0, 1 or blank in the 'occupational class' field. The blank is not the same as 0 and is changed to 5 instead. Despite the collection of occupational class information, the field for 'occupational rating' is still being used after 1990. The majority of claim records with occupational class = 2, 3 or 4 have occupational rating = 1. However, there are some claim records with occupational class = 1 but with occupational rating = 1. Thus, for the purpose of using the data set for the entire 28 year-period from 1975 to 2002 in our analysis, we reckon that the best available interpretation is to regard those with occupational classes = 0, 1 and 5 and with occupational rating = 1 as "non-rated" and everything else (i.e. occupational class= 0, 1, 5 with occupational rating = 1 or 2 and also occupational class = 2, 3, 4 regardless of occupational rating) as "rated".

Using the information described above, we are able to calculate the following quantities for each claim record:

(i) Sickness duration at the start of the claim payment. This is given by the difference between the date of payment commencement and the date of sickness inception. Note that a claimant's sickness duration at the start of the claim

payment should be equal to or longer than the deferred period specified in the claimant's policy.

(ii) Sickness duration at the end of the claim payment. This is given by the difference between the date of payment cessation and the date of sickness inception.

(iii) Exact age at the start of claim payment. With only month and year of birth given, we will assume that 15th is the day of birth. The exact age of the claimant at the start of claim payment is given by the difference between the date of payment commencement and the date of birth.

Due to the data being collected and recorded on a yearly basis, an individual who is still sick at the end of each investigation year will have his/her observation right-censored and will probably (but not always owing to the lack of homegeneity in the offices that contribute data from year to year) enter into our study in the subsequent investigation year with 'continuation of claim payment' as the mode of claim commencement. Hence, until an IPI claimant experiences either recovery or death, he or she may have multiple claim records, one for each investigation year visited until the IPI policy expires.

The IPI claim data submitted by life offices undergoes screening and is scrutinised for possible errors of coding by the CMI. There is also a procedure in place to identify and remove duplicate claims that arise as a result of an individual having more than one IPI policy. Despite such scrutiny, during the preliminary investigation of the data, we detected and removed erroneous data as a result of

1. New claims that start and end on the same day.

2. New claims for which the sickness duration at the date of payment commencement is less than the pre-specified deferred period. CMI Working Paper 6 (2004) reveals a small percentage of new claims that start a few days before or after the pre-specified deferred period. For example, some claim payments belonging to DP4 policies only start at 30 or 31 days (1 month) after sickness inception. In view of this, an allowance of $\pm$ 3 days from the pre-specified deferred period is used to identify genuine new claims.

3. Claims with other modes of commencement for which the sickness duration at the date of payment commencement is less than the pre-specified deferred period.

Table 2.1 gives a summary of data removed due to the above three reasons. Nevertheless, it is impossible to ensure that the data is flawless because there may be unidentifiable errors such as coding mistakes by the administrative personnel in some life offices when preparing the claim data for submission to the CMI.

Table 2.1: A summary of erroneous IPI claim records removed from analysis.

	DP1	DP4	DP13	DP26	DP52
Intial number of claim records	63,584	60,722	51,900	42,912	17,541
Cause of removal					
1	153	329	423	135	117
2	76	937	1500	829	369
3	-	203	303	235	170
Number of claim records removed	229	1,469	2,226	1,199	656
(%)	0.36%	2.42%	4.29%	2.79%	3.74%
Final number of claim records	63,355	59,253	49,674	41,713	16,885

2.3 Grouping of Causes of Sickness

Since this set of IPI claim data will be used to obtain recovery intensity and mortality intensity from sick by cause of sickness in the following chapters, we will describe the data in greater detail. There are a total of 70 causes of sickness. Such a large number of causes of sickness does not lend itself readily for presentation. As many of the individual causes of sickness have a small amount of data, we wish to group these 70 causes of sickness into fewer sickness categories. For this purpose, we review the various groupings of causes of sickness by ICD8 code in the context of IPI. CMI Report 8 (1986) decided to group the causes of sickness into 14 sickness categories purely on medical grounds using IPI claim data from 1975-78. In their grouping, causes of sickness regarded to be of little significance are amalgamated together while any cause of sickness that accounts for at least 5% of the total experience was included as a specific category. Cordeiro (1998) analysed the IPI claim data classified by CMI Report 8 (1986)'s grouping for the same 4-year period and, for computational convenience, decided to further categorise them into the following five classes based on their levels of recovery intensity approximated by statistical properties.

Class I	Very high recovery intensities
Class II	High recovery intensities
Class III	Medium recovery intensities
Class IV	Low recovery intensities
Class V	Very low recovery intensities and very high mortality from sick intensities

In CMI Working Paper 23 (2006), the 70 causes of sickness are grouped into 11 sickness categories in such a way that the number of terminations (recovery and death) in each category is sufficiently large and no causes of sickness that show dissimilar termination patterns are put in the same category. To achieve the latter aim, the actual number of terminations (A) for each cause of sickness is compared to that expected (E) under SM1975-78 by using the A/E ratio. This analysis is conducted by using combined data from 1991-2002, both sexes, DP4-DP52 and occupational class 1. The classification of 70 causes of sickness into different sickness categories adopted by these three different investigations into IPI claim experience is shown in Table 2.2.

Table 2.2: The different grouping of 70 causes of sickness (represented by their ICD8 code) by CMI Report 8 (1986), CMI Working Paper 23 (2006) and Cordeiro (1998)

ICD8	CMI Report 8 (1986)	CMI Working Paper 23 (2006)	Cordeiro (1998)
01-19	Other infective	Infections & acute respiratory	Class II
20	Neoplasms	Neoplasms	Class V
21	Neoplasms	Neoplasms	Class II
22-26	Endocrine & Metabolic	All other known causes	Class IV
27	Mental	Mental illness	Class IV
28-30	All Others	Nervous system & sensory organs	Class III
31	Nervous	Nervous system & sensory organs	Class III
32-33	All others	Circulatory	Class III
34-38	Circulatory	Circulatory	Class III
39-41	Acute respiratory	Infections & acute respiratory	Class I
42-45	Chronic respiratory	All other known causes	Class II
46	All others	All other known causes	Class III
47-51	Digestive	Digestive (non-infectious)	Class III
52-55	Genito-urinary	Genito-urinary	Class II
56-58	All others	All other known causes	Class III
59	All others	Infections & acute respiratory	Class III
60	All others	All other known causes	Class III
61	Musculoskeletal	Arthritis	Class IV
62	Musculoskeletal	Musculoskeletal	Class III
63-65	All others	All other known causes	Class III
66	Road traffic accident injuries	Injuries	Class III
67-70	Other injuries	Injuries	Class III

With reference to the different groupings shown in Table 2.2, we eventually decided to classify the 70 causes of sickness (as represented by their ICD8 codes) into 12 sickness categories as set out in Table 2.3. Note that the grouping we adopted is very similar to that used by CMI Working Paper 23 (2006). We include in Table 2.3, for each sickness category, the total number of claim inceptions during the investigation years, the total exposed to risk of claim termination (recovery or death) calculated in days and the total number of recoveries and deaths. The total exposed to risk gives the total IPI claims duration during the period under review and therefore includes claims which were already in force at the outset of the period of investigation. The largest five sickness categories in terms of number of inceptions are 'G10 Musculoskeletal', 'G11 Injuries', 'G7 Respiratory', 'G6 Circulatory' and 'G4 Mental Illness'. In terms of exposed to risk, 'G5 Nervous system & sensory organs' replaced 'G7 Respiratory' in

these top five sickness categories. With the exception of 'G2 Neoplasms', the numbers of recoveries for other sickness cateogories are a lot higher than the number of deaths. The number of deaths from 'G2 Neoplasms' constitues about half of the total number of deaths.

Table 2.3: The grouping of the 70 causes of sickness into 12 sickness categories.

Sickness Category	ICD8 code	Number of inceptions	Exposed to risk (days)	Number of Recoveries	Deaths
G1 Infections & acute respiratory	01–19	6,355	527,121	4,895	39
G2 Neoplasms	20, 21	4,499	2,188,366	2,165	1,765
G3 Endocrine & Metabolic	22–26	868	640,812	596	48
G4 Mental illness	27	8,512	8,616,813	5,280	207
G5 Nervous system & sensory organs	28–31	4,473	3,795,765	2,861	194
G6 Circulatory	32–38	9,360	7,773,247	6,163	549
G7 Respiratory	39–45	14,041	955,585	10,440	103
G8 Digestive (non-infectious)	47–51	6,115	1,016,999	5,752	113
G9 Genito-urinary	52–55	2,986	509,259	2,603	56
G10 Musculoskeletal	61, 62	17,200	9,519,873	13,125	139
G11 Injuries	66–70	15,636	4,234,648	13,749	85
G12 All other known causes	46, 56–60, 63–65	6,542	527,121	4,895	39
All Sickness Categories		96,587	40,305,609	72,524	3,337

2.4 Exploratory Analysis of Data

We wish to investigate the distribution of duration from the onset of sickness (i.e. including the deferred period) for IPI claims which ended due to recovery or death of the claimant during the period of investigation. For this purpose, we use a box plot to convey visually the important aspects of a distribution through its five-number summaries: the smallest observation (not considered as an outlier), lower quartile, median, upper quartile, and the largest observation (not considered as an outlier). These five numbers are represented by five horizontal lines, arranged from bottom to top, in a box plot. The smallest and largest observations not considered as an outlier are no more than 1.5 times the interquartile range away from the lower and upper quartiles respectively. Those observations that lie beyond these two observations are considered as outliers and are represented by dots.

Figure 2.1 shows the box plots of duration since onset of sickness until recovery on a logarithmic scale by sickness category for each deferred period. Note that these box plots only include sickness durations for individuals who recover during the period of investigation and have ignored the presence of right-censored durations. Therefore, these box plots should not be construed to indicate the true distribution of the sickness duration until recovery. They are meant as descriptive statistics and are not relied upon in the estimation of recovery intensities in Chapter 3. The number of recoveries upon which each box plot is constructed is indicated at the bottom of the box plot. These box plots show that the sickness duration until recovery is heavily skewed to the right, in which case, the median is a better measure than the mean. For comparison purposes, we present in Figure 2.2 the medians (represented by lines) by sickness category for each deferred period. The numbers underneath the lines are the number of recoveries used in the derivation of their respective median. This graph shows that the median increases with longer deferred periods for all sickness categories. In both DP1 and DP4, where most of the recoveries are concentrated, 'G1 Infections & acute respiratory' and 'G7 Respiratory' have lower medians than the others while 'G2 Neoplasms', 'G4 Mental illness' and 'G6 Circulatory' have higher medians.

Due to the low number of deaths in most of the sickness categories, we will use

combined data from all deferred periods to construct a box plot of duration since onset of sickness until death. These box plots, presented on a logarithmic scale in Figure 2.3, show that the sickness duration until death is skewed to the right. The medians for 'G4 Mental illness', 'G5 Nervous system & sensory organs' and 'G10 Musculoskeletal' rank higher than the others.

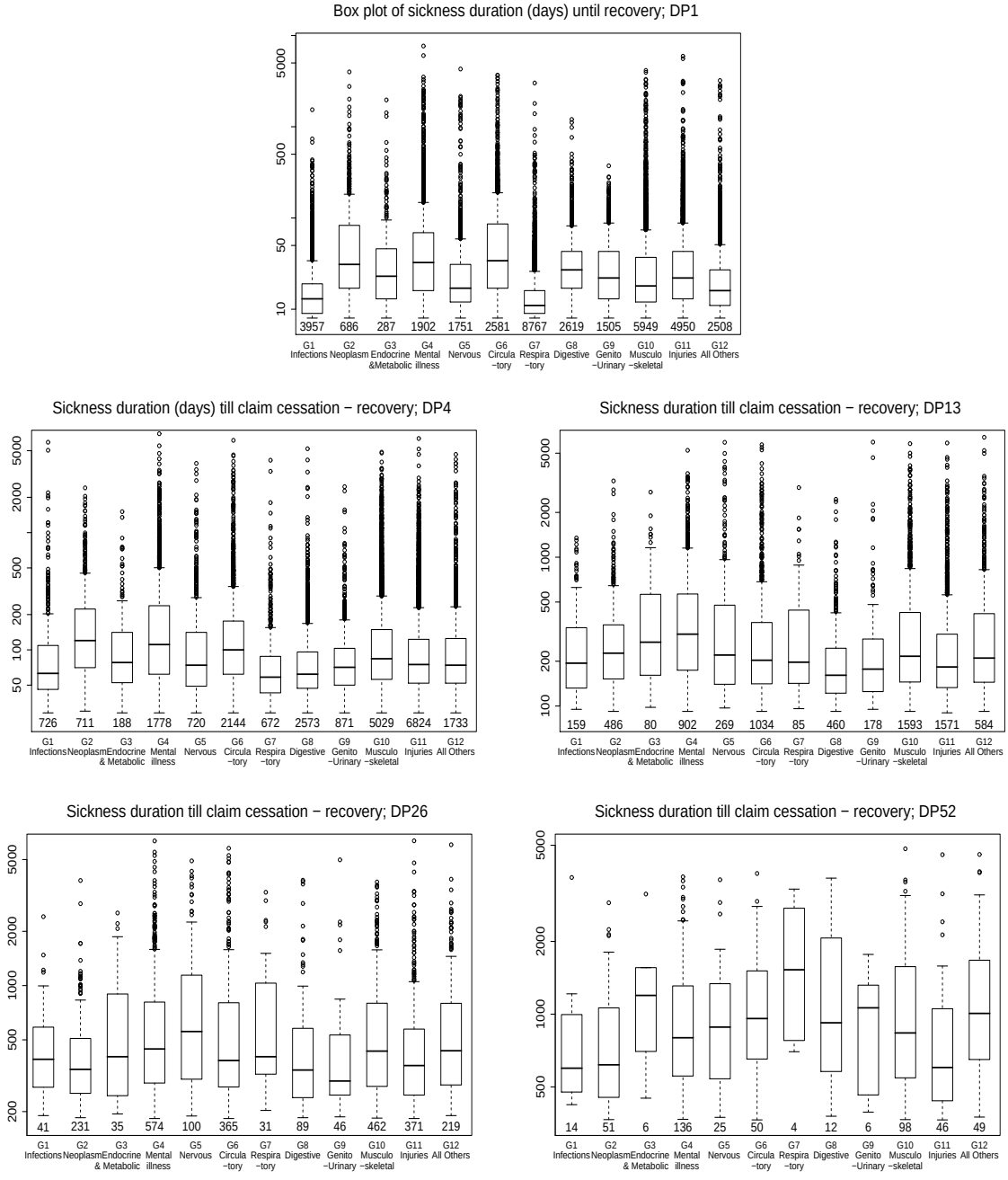


Figure 2.1: The box plots of sickness duration until recovery by sickness category for each deferred period.

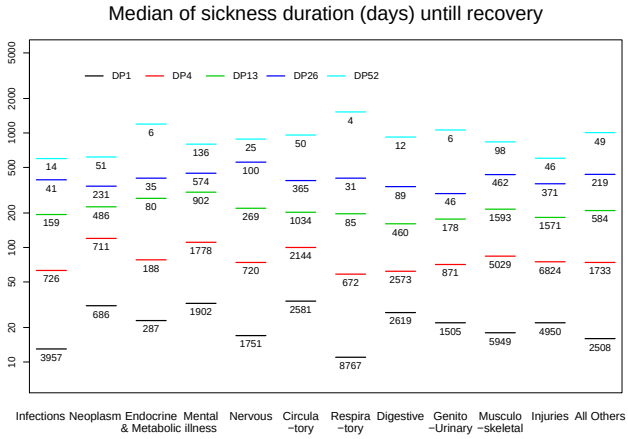


Figure 2.2: Median of sickness duration until recovery by sickness category for each deferred period.

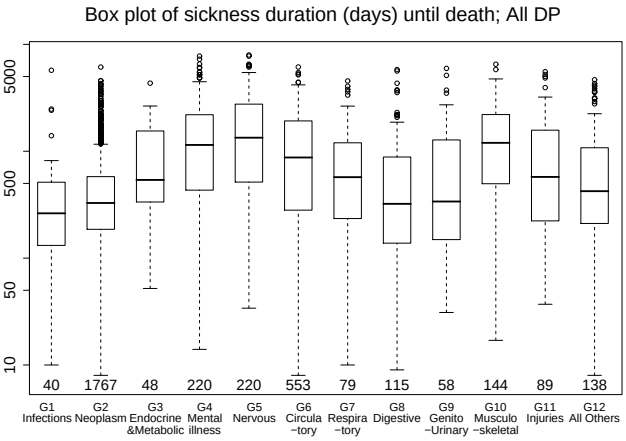


Figure 2.3: Box plots of sickness duration until death by sickness category.

We show in Table 2.4 the number of recoveries expressed as percentages of the total for each sickness duration interval by sickness category. In terms of the percentage of the total recoveries that occur within the first eight weeks of sickness, it is 83% and 92% for 'G1 Infections & acute respiratory' and 'G7 Respiratory', respectively, between 55% and 63% for 'G5 Nervous system & sensory organs', 'G8 Digestive', 'G9 Genito-urinary' and 'G12 All other known causes', between 45% and 49% for 'G3 Endocrine & Metabolic', 'G10 Musculoskeletal' and 'G11 Injuries', 33% for 'G6 Circulatory', 31% for 'G4 Mental illness' and 26% for 'G2 Neoplasms'. Both 'G2 Neoplasms' and 'G4 Mental Illness', with 18% and 24%, respectively, have the largest percentage of their total number of recoveries remaining after one year.

Table 2.5 shows the number of deaths expressed as percentages of the total for each sickness duration interval by sickness category. In contrast to the percentage of recoveries in Table 2.4, a large percentage of deaths occur at longer sickness durations for all sickness categories. At 33.3%, 'G1 Infections & acute respiratory' has the least percentage of its total deaths remaining after one year, followed by 'G8 Digestive', 'G2 Neoplasms' and 'G9 Genito-urinary'. The other eight sickness categories have more than 50% of their total deaths remaining after one year.

Tables 2.6 and 2.8 show, respectively, the median of sickness duration until recovery and until death for each quadrennium by sickness category. The numbers of recoveries and deaths used to derive these medians are presented in Table 2.7 and 2.9, respectively. Apart from 'G1 Infections & acute respiratory', 'G5 Nervous system & sensory organs', 'G7 Respiratory' and 'G8 Digestive', the median of sickness duration until recovery generally increases over the quadrennia. For 'G2 Neoplasms' and 'G6 Circulatory', the two sickness categories which contain a significant number of deaths, the median of sickness duration until death increases over the quadrennia.

Tables 2.10–2.19 show the exposed to risk of claim termination (recovery or death) as a percentage of the total for each age group by sex and deferred period. For males, 'G4 Mental illness', 'G6 Circulatory' and 'G10 Musculoskeletal', in general, assume greater importance than the other sickness categories. In particular, the exposed to risk for 'G6 Circulatory' as a percentage of the total increases with age for males across all deferred periods. The exposed to risk for 'G11 Injuries' dominates younger

age groups. For females, both 'G4 Mental illness' and 'G10 Musculoskeletal' are the two most important sickness categories in terms of exposed to risk. The exposed to risk for 'G1 Infections & acute respiratory' is concentrated in the youngest age group of DP1 for both sexes.

Tables 2.20–2.21 show for each sickness category the total exposed to risk of claim termination (recovery or death), the number of recoveries and deaths as well as percentages of these totals by sex and rating indicator. We also present in each table the median of sickness duration in days until recovery and deaths using all data and subdivisions of it by sex and rating indicator. We note that

(i) Males' exposed to risk is higher than females' for all sickness categories.

(ii) With the exception of 'G11 Injuries', the exposed to risk for 'non-rated' is higher than 'rated' for the other sickness categories

(iii) The median of sickness duration until recovery is higher in 'rated' than in 'non-rated' for all sickness categories.

(iv) The median of sickness duration until recovery is noticeably higher in females than in males for 'G2 Neoplasms', 'G4 Mental Illness', 'G3 Endocrine & Metabolic', 'G9 Genito-urinary' and 'G12 All other known causes'. Apart from 'G6 Circulatory', where the median for males is higher than females, the median is broadly similar for both sexes for the remaining six sickness categories.

Figure 2.4 shows for each sickness category, the exposed to risk of claim termination (recovery or death) as a proportion of the total over each successive quadrennium in 1975-2002. For all sickness categories, the percentage of exposed to risk spent in the first sickness duration interval and last sickness duration interval decreases and increases, respectively, over the quadrennia.

Table 2.4: Percentages of recoveries for each sickness duration interval by sickness category.

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
1-2wks	48.09	6.24	14.09	6.63	24.64	7.43	61.30	8.59	18.09	16.31	10.61	21.64
2-4wks	22.47	8.59	14.60	9.39	18.94	11.36	22.09	16.67	16.75	14.13	11.14	15.90
4-8wks	12.50	10.81	18.46	15.47	19.19	14.67	8.86	31.87	25.12	18.20	24.14	18.07
8-13wks	6.23	11.59	12.58	11.95	10.38	14.31	3.23	18.12	17.52	14.43	18.15	12.27
13-26wks	5.78	20.88	14.77	16.65	10.52	24.53	2.31	15.68	13.64	16.72	19.27	14.47
26-39wks	1.78	15.70	8.72	9.85	4.19	10.03	0.67	4.16	4.15	6.32	6.97	5.13
39-1yr	0.92	8.55	4.03	6.48	2.69	4.85	0.45	1.63	1.61	3.61	3.01	3.13
1-2yrs	1.33	11.59	6.71	12.44	4.23	6.96	0.62	2.29	1.96	5.92	4.27	4.78
2-5yrs	0.80	5.31	4.70	8.43	3.46	4.11	0.34	0.70	0.92	3.41	1.96	3.24
5-12yrs	0.10	0.74	1.34	2.71	1.75	1.75	0.12	0.30	0.23	0.95	0.49	1.38
Total number of recoveries	4,895	2,165	596	5,280	2,861	6,163	10,440	5,752	2,603	13,125	13,749	4,895

Table 2.5: Percentages of deaths for each sickness duration interval by sickness category.

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
1-2wks	2.56	0.17	0.00	0.97	0.00	0.91	1.94	0.88	0.00	0.00	0.00	0.73
2-4wks	0.00	0.34	0.00	0.97	0.00	0.91	2.91	3.54	0.00	0.72	0.00	2.92
4-8wks	5.13	2.49	2.08	0.97	1.03	4.01	1.94	7.96	14.29	2.88	3.53	4.38
8-13wks	5.13	4.87	2.08	1.93	1.55	3.10	4.85	5.31	1.79	3.60	7.06	5.11
13-26wks	28.21	16.15	8.33	6.76	6.19	8.56	5.83	15.04	16.07	5.76	8.24	10.22
26-39wks	12.82	17.62	6.25	3.86	5.15	6.92	12.62	13.27	12.50	5.04	12.94	13.14
39-1yr	12.82	13.03	12.50	5.80	3.09	4.74	13.59	11.50	7.14	5.04	12.94	8.76
1-2yrs	17.95	28.61	31.25	18.36	19.59	16.58	18.45	15.04	14.29	12.95	12.94	21.90
2-5yrs	10.26	13.82	20.83	26.57	29.38	28.60	21.36	18.58	17.86	32.37	23.53	21.17
5-12yrs	5.13	2.89	16.67	33.82	34.02	25.68	16.50	8.85	16.07	31.65	18.82	11.68
Total number of recoveries	39	1,765	48	207	194	549	103	113	56	139	85	39

Table 2.6: Median of sickness duration until recovery (days) for each quadrennium by sickness category.

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
75-78	15.0	73.0	46.0	62.5	39.0	79.0	13.0	44.0	37.0	44.5	47.0	31.5
79-82	15.0	73.0	65.0	85.5	39.0	84.0	13.0	47.0	35.0	47.0	51.0	34.0
83-86	15.0	103.0	61.0	89.0	41.0	92.0	13.0	48.0	42.0	59.0	58.0	32.0
87-90	15.0	96.5	55.5	82.5	38.0	93.5	12.0	55.0	48.0	59.0	62.0	56.0
91-94	14.0	154.0	68.0	127.0	43.0	115.0	11.0	59.0	55.0	64.0	76.0	70.5
95-98	14.0	199.0	85.0	190.0	22.0	111.0	11.0	48.0	46.0	63.0	78.0	88.0
99-02	13.0	265.0	106.5	252.5	31.0	125.0	11.0	52.0	60.5	90.0	90.0	90.0

Table 2.7: Number of recoveries for each quadrennium by sickness category.

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
75-78	861	212	85	594	350	874	2026	1101	382	1226	1902	734
79-82	766	231	77	518	388	887	1544	1014	343	1429	1780	711
83-86	1004	294	106	596	425	1022	1961	1014	444	1853	2506	1092
87-90	917	292	102	761	475	990	1806	939	458	2327	2991	915
91-94	691	364	81	862	460	997	1379	757	411	2610	2136	757
95-98	365	323	77	936	409	724	1045	478	344	1925	1272	435
99-02	293	449	68	1025	358	680	679	450	224	1761	1175	449

Table 2.8: Median of sickness duration until death (days) for each quadrennium by sickness category.

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
75-78	107.0	214.0	307.0	184.0	491.0	592.0	228.0	79.0	49.0	53.0	140.0	194.5
79-82	181.0	220.5	521.0	1390.0	914.0	596.0	411.5	268.0	250.0	324.0	306.0	437.0
83-86	217.0	249.0	860.0	1312.0	1411.0	685.5	337.0	189.5	333.0	629.0	561.0	130.5
87-90	274.0	293.0	447.5	492.0	1130.0	881.0	274.0	468.0	748.0	655.0	791.0	409.0
91-94	447.0	328.0	501.0	1011.0	842.5	1020.0	580.0	522.0	196.0	1161.5	246.0	353.0
95-98	209.5	397.0	369.0	1272.0	1326.5	1373.0	2004.0	203.5	459.0	1338.0	839.0	1140.5
99-02	322.0	434.0	1338.5	1051.0	1616.5	1101.0	685.0	971.5	1595.0	2051.5	1574.0	1654.5

Table 2.9: Number of deaths for each quadrennium by sickness category.

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
75-78	4	127	3	16	21	63	15	12	5	7	4	16
79-82	6	146	7	19	22	61	14	25	10	9	7	11
83-86	5	193	7	17	35	89	13	14	4	9	16	17
87-90	9	281	8	37	28	113	22	20	8	17	20	27
91-94	5	327	13	57	38	82	25	20	12	32	18	31
95-98	8	340	6	38	45	82	9	10	15	31	12	18
99-02	3	353	4	36	31	63	5	14	4	39	12	18

Table 2.10: Exposed to risk of claim termination (recovery or death) as percentage of totals for each age group: Males - DP1.

Sickness Category	17-29	30-39	40-49	50-59	60-65	All ages
G1 Infections	10.7	6.4	2.3	1.5	1.3	2.3
G2 Neoplasms	2.0	1.2	3.6	4.2	6.2	3.9
G3 Endocrine & Metabolic	2.0	0.7	0.7	0.8	2.1	0.9
G4 Mental Illness	11.0	19.7	25.3	18.6	8.1	19.2
G5 Nervous	4.4	12.3	9.0	7.4	10.1	8.5
G6 Circulatory	0.8	6.1	17.0	28.6	34.3	23.5
G7 Respiratory	11.8	8.3	4.4	3.7	6.0	4.7
G8 Digestive	3.5	3.0	2.9	3.0	4.4	3.2
G9 Genito-Urinary	0.7	1.0	1.2	1.6	2.5	1.5
G10 Musculoskeletal	17.7	19.7	21.4	22.1	16.9	21.0
G11 Injuries	25.6	16.5	8.1	4.6	5.4	7.2
G12 All others	9.8	5.2	4.1	3.8	2.6	4.0
Total exposed to risk (days)	100,557	487,956	1,299,354	2,629,099	539,422	5,056,388

Table 2.11: Exposed to risk of claim termination (recovery or death) as percentage of totals for each age group: Male - DP4.

Sickness Category	17-29	30-39	40-49	50-59	60-65	All ages
G1 Infections	1.8	2.3	1.0	0.6	1.0	1.2
G2 Neoplasms	1.7	3.3	4.0	5.6	8.2	4.5
G3 Endocrine & Metabolic	0.4	0.5	0.9	1.8	1.5	1.2
G4 Mental Illness	12.0	14.2	15.0	14.1	8.7	14.0
G5 Nervous	5.0	8.9	8.3	6.8	7.2	7.6
G6 Circulatory	2.0	6.5	14.7	26.8	34.2	17.7
G7 Respiratory	2.1	1.7	2.6	2.8	5.8	2.6
G8 Digestive	4.0	2.7	3.9	2.6	3.3	3.2
G9 Genito-Urinary	0.6	0.8	1.0	1.3	1.2	1.0
G10 Musculoskeletal	22.8	27.7	26.0	25.2	19.3	25.5
G11 Injuries	39.8	24.8	17.0	7.5	5.2	15.8
G12 All others	7.9	6.6	5.6	4.9	4.4	5.6
Total exposed to risk (days)	572,618	1,958,324	3,035,909	3,581,860	477,553	9,626,264

Table 2.12: Exposed to risk of claim termination (recovery or death) as percentage of totals for each age group: Male - DP13.

Sickness Category	17-29	30-39	40-49	50-59	60-65	All ages
G1 Infections	0.9	1.2	1.3	0.6	1.2	1.0
G2 Neoplasms	2.1	3.6	5.1	5.6	10.4	5.2
G3 Endocrine & Metabolic	2.6	1.3	1.5	1.6	1.9	1.6
G4 Mental Illness	11.8	19.1	18.1	16.7	6.6	16.8
G5 Nervous	4.2	11.2	10.6	9.1	8.7	9.8
G6 Circulatory	4.4	7.0	15.3	26.8	33.6	19.3
G7 Respiratory	1.1	1.5	1.6	2.4	4.2	2.0
G8 Digestive	2.4	2.0	2.5	2.4	1.7	2.3
G9 Genito-Urinary	0.9	1.1	0.9	0.9	1.2	0.9
G10 Musculoskeletal	24.2	23.8	23.3	22.0	23.5	22.9
G11 Injuries	35.3	20.1	13.4	6.5	3.0	11.8
G12 All others	10.1	8.1	6.5	5.5	4.0	6.3
Total exposed to risk (days)	348,802	1,726,361	3,453,957	4,176,970	589,271	10,295,361

Table 2.13: Exposed to risk of claim termination (recovery or death) as percentage of totals for each age group: Male - DP26.

Sickness Category	17-29	30-39	40-49	50-59	60-65	All ages
G1 Infections	4.1	2.3	1.2	1.2	0.3	1.3
G2 Neoplasms	4.5	6.0	3.5	3.9	9.9	4.4
G3 Endocrine & Metabolic	2.4	2.5	1.3	1.6	2.5	1.7
G4 Mental Illness	26.4	21.9	29.0	21.4	7.6	23.3
G5 Nervous	10.6	13.8	11.5	9.3	10.8	10.6
G6 Circulatory	2.7	5.3	17.1	30.5	35.0	23.2
G7 Respiratory	0.0	1.1	2.1	2.4	5.7	2.3
G8 Digestive	1.8	1.7	2.3	1.6	1.7	1.8
G9 Genito-Urinary	0.7	1.1	0.5	1.0	0.5	0.8
G10 Musculoskeletal	16.8	19.2	16.2	17.8	19.2	17.5
G11 Injuries	18.6	16.7	7.8	4.2	2.2	6.8
G12 All others	11.4	8.5	7.5	5.0	4.7	6.3
Total exposed to risk (days)	145,293	920,185	2,948,554	4,283,308	515,025	8,812,365

Table 2.14: Exposed to risk of claim termination (recovery or death) as percentage of totals for each age group: Male - DP52.

Sickness Category	17-29	30-39	40-49	50-59	60-65	All ages
G1 Infections	0.9	0.4	1.2	0.7	0.5	0.8
G2 Neoplasms	3.0	3.1	4.4	3.5	9.2	4.0
G3 Endocrine & Metabolic	0.0	2.6	1.0	1.0	2.6	1.2
G4 Mental Illness	22.1	27.4	35.4	25.8	8.3	28.2
G5 Nervous	3.5	10.6	12.3	8.6	7.9	9.9
G6 Circulatory	0.6	4.9	10.4	27.8	34.0	19.7
G7 Respiratory	4.0	1.7	0.8	1.3	4.8	1.4
G8 Digestive	0.0	1.2	1.5	1.6	2.5	1.5
G9 Genito-Urinary	0.0	0.5	0.2	0.8	1.3	0.6
G10 Musculoskeletal	17.3	28.0	16.2	19.1	19.7	18.9
G11 Injuries	31.8	8.0	7.8	3.6	2.3	5.9
G12 All others	16.9	11.6	8.9	6.2	6.9	7.8
Total exposed to risk (days)	72,736	299,603	1,211,592	1,800,952	178,182	3,563,065

Table 2.15: Exposed to risk of claim termination (recovery or death) as percentage of totals for each age group: Female - DP1.

Sickness Category	17-29	30-39	40-49	50-59	60-65	All ages
G1 Infections	11.8	4.4	2.0	2.0	0.4	3.3
G2 Neoplasms	2.6	3.4	7.0	5.5	6.0	5.3
G3 Endocrine & Metabolic	1.1	2.0	0.8	3.9	0.0	2.1
G4 Mental Illness	10.5	10.5	13.1	14.8	19.8	13.1
G5 Nervous	1.7	4.5	6.2	11.4	4.2	7.2
G6 Circulatory	0.4	12.2	3.6	7.9	10.7	6.8
G7 Respiratory	9.7	9.2	5.0	3.1	3.2	5.6
G8 Digestive	4.7	2.2	5.0	1.5	0.0	3.1
G9 Genito-Urinary	2.0	6.5	9.3	4.1	0.7	6.1
G10 Musculoskeletal	20.3	12.6	29.7	37.8	16.8	27.6
G11 Injuries	19.9	19.0	10.8	5.6	26.4	12.0
G12 All others	15.5	13.4	7.5	2.3	11.9	7.8
Total exposed to risk (days)	36,379	90,894	147,130	144,377	11,973	430,753

Table 2.16: Exposed to risk of claim termination (recovery or death) as percentage of totals for each age group: Female - DP4.

Sickness Category	17-29	30-39	40-49	50-59	60-65	All ages
G1 Infections	1.5	2.4	1.4	1.2	0.0	1.6
G2 Neoplasms	5.3	3.6	9.4	10.5	8.8	7.6
G3 Endocrine & Metabolic	0.8	1.8	1.4	0.8	0.0	1.3
G4 Mental Illness	22.4	20.2	20.4	13.7	17.4	19.3
G5 Nervous	0.9	10.8	7.7	7.8	7.3	7.7
G6 Circulatory	2.0	6.0	6.4	3.5	25.5	5.3
G7 Respiratory	1.2	0.8	0.8	2.2	0.1	1.1
G8 Digestive	3.1	4.2	1.5	1.2	10.5	2.4
G9 Genito-Urinary	5.1	4.2	5.2	3.8	6.0	4.6
G10 Musculoskeletal	24.8	21.1	26.2	29.6	8.0	25.2
G11 Injuries	21.2	12.8	11.3	9.4	15.7	12.7
G12 All others	11.6	12.0	8.3	16.3	0.6	11.3
Total exposed to risk (days)	1,707,59	345,058	517,799	263,949	9,391	1,306,956

Table 2.17: Exposed to risk of claim termination (recovery or death) as percentage of totals for each age group: Female - DP13.

Sickness Category	17-29	30-39	40-49	50-59	60-65	All ages
G1 Infections	1.8	2.2	0.8	1.0	0.0	1.3
G2 Neoplasms	2.3	7.3	11.1	10.3	5.2	9.1
G3 Endocrine & Metabolic	2.1	4.0	1.5	2.4	0.0	2.4
G4 Mental Illness	26.0	25.1	27.0	17.2	0.0	23.9
G5 Nervous	8.3	9.6	8.0	6.7	0.0	8.1
G6 Circulatory	6.6	2.8	4.2	9.9	13.7	5.5
G7 Respiratory	3.6	1.7	2.3	1.8	0.3	2.1
G8 Digestive	3.3	1.6	2.1	1.0	12.4	1.8
G9 Genito-Urinary	1.9	2.0	1.9	3.5	0.0	2.3
G10 Musculoskeletal	18.9	19.9	24.0	30.8	68.1	24.3
G11 Injuries	10.3	10.7	7.6	6.8	0.3	8.4
G12 All others	14.9	13.1	9.4	8.6	0.0	10.7
Total exposed to risk (days)	163,422	436,809	683,209	416,496	8,695	1,708,631

Table 2.18: Exposed to risk of claim termination (recovery or death) as percentage of totals for each age group: Female - DP26.

Sickness Category	17-29	30-39	40-49	50-59	60-65	All ages
G1 Infections	0.9	3.2	0.8	1.0	0.0	1.4
G2 Neoplasms	4.6	5.8	6.9	8.8	8.6	7.1
G3 Endocrine & Metabolic	5.4	2.6	2.0	2.0	1.5	2.3
G4 Mental Illness	31.7	31.0	33.3	23.8	21.1	29.7
G5 Nervous	5.4	8.2	9.4	9.9	13.3	9.1
G6 Circulatory	3.9	3.7	5.7	8.2	17.2	6.0
G7 Respiratory	0.6	0.5	0.6	2.2	7.3	1.1
G8 Digestive	0.0	2.5	2.5	1.6	0.0	2.1
G9 Genito-Urinary	1.0	0.7	2.3	1.3	0.0	1.6
G10 Musculoskeletal	24.9	22.6	23.0	29.3	30.8	25.0
G11 Injuries	7.1	5.0	4.6	4.2	0.3	4.7
G12 All others	14.4	14.4	9.0	7.6	0.0	10.0
Total exposed to risk (days)	124,062	490,698	1,008,985	727,423	8,303	2,359,471

Table 2.19: Exposed to risk of claim termination (recovery or death) as percentage of totals for each age group: Female - DP52.

Sickness Category	17-29	30-39	40-49	50-59	60-65	All ages
G1 Infections	0.0	2.6	1.4	0.4	0.0	1.2
G2 Neoplasms	11.9	4.1	5.9	6.2	29.9	5.9
G3 Endocrine & Metabolic	7.1	1.1	2.8	0.9	0.0	2.1
G4 Mental Illness	22.9	35.0	36.2	27.2	0.0	32.6
G5 Nervous	1.4	6.7	8.1	9.4	0.0	7.9
G6 Circulatory	11.4	4.3	4.6	9.5	0.0	6.4
G7 Respiratory	0.0	0.4	0.9	2.5	24.7	1.3
G8 Digestive	2.4	1.1	2.6	0.7	0.0	1.7
G9 Genito-Urinary	0.6	1.2	0.8	1.2	0.0	1.0
G10 Musculoskeletal	15.7	19.6	20.9	31.0	45.4	23.6
G11 Injuries	11.5	10.3	4.1	3.4	0.0	5.4
G12 All others	15.2	13.6	11.9	7.5	0.0	11.0
Total exposed to risk (days)	56,234	236,005	573,450	394,718	1,065	1,261,472

Table 2.20: Exposed to risk of claim termination (recovery or death), number of recoveries and deaths, median of sickness duration until recovery and death for sickness categories G1 – G6.

	Exposed to risk (days)	Number recovery	Number death	Median of sickness duration (days) recovery	Median of sickness duration (days) death
		G1 Infections & acute respiratory			
Total	527,121	4,895	39	15.0	257.0
Male	81.47	85.58	100.00	15.0	257.0
Female	18.53	14.42	0.00	16.0	NA
Non-rated	72.25	91.62	74.36	14.0	274.0
Rated	27.75	8.38	25.64	65.5	195.5
		G2 Neoplasms			
Total	2,188,366	2,165	1,765	139.0	328.0
Male	76.56	73.44	86.40	128.5	306.0
Female	23.44	26.56	13.60	181.0	437.0
Non-rated	70.68	79.17	74.16	118.5	328.0
Rated	29.32	20.83	25.84	202.0	330.0
		G3 Endocrine & Metabolic			
Total	640,812	596	48	63.0	539.0
Male	77.52	79.87	85.42	56.5	521.0
Female	22.48	20.13	14.58	111.5	818.0
Non-rated	67.76	76.68	91.67	48.0	539.0
Rated	32.24	23.32	8.33	103.0	1147.0
		G4 Mental Illness			
Total	8,616,813	5,280	207	119.0	1061.0
Male	79.20	80.64	89.37	104.0	1063.0
Female	20.80	19.36	10.63	212.5	1004.5
Non-rated	79.45	80.30	88.89	100.0	1072.0
Rated	20.55	19.70	11.11	194.0	695.0
		G5 Nervous			
Total	3,795,765	2,861	194	35.0	1094.0
Male	85.25	87.14	94.33	35.5	1127.0
Female	14.75	12.86	5.67	34.5	949.0
Non-rated	72.92	82.17	78.87	26.0	964.0
Rated	27.08	17.83	21.13	97.0	1449.0
		G6 Circulatory			
Total	7,773,247	6,163	549	96.0	843.0
Male	94.92	95.31	98.18	98.0	843.0
Female	5.08	4.69	1.82	62.0	786.0
Non-rated	75.35	78.26	80.51	87.0	881.0
Rated	24.65	21.74	19.49	136.0	762.0

Table 2.21: Exposed to risk of claim termination (recovery or death), number of recoveries and deaths, median of sickness duration until recovery and death for sickness categories G7 – G12.

	Exposed to risk (days)	Number recovery	Number death	Median of sickness duration (days) recovery	Median of sickness duration (days) death
			G7 Respiratory		
Total	955,585	10,440	103	12.0	509.0
Male	88.97	86.59	93.20	12.0	520.0
Female	11.03	13.41	6.80	13.0	396.0
Non-rated	65.94	92.28	73.79	12.0	513.0
Rated	34.06	7.72	26.21	47.0	396.0
			G8 Digestive		
Total	1,016,999	5,752	113	49	309.0
Male	86.29	92.65	97.35	49	294.0
Female	13.71	7.35	2.65	54	539.0
Non-rated	59.83	68.29	66.37	39	329.0
Rated	40.17	31.71	33.63	75	300.0
			G9 Genito-Urinary		
Total	509,259	2,603	56	45.0	304.5
Male	67.77	67.61	91.07	35.0	301.0
Female	32.23	32.39	8.93	69.0	308.0
Non-rated	69.35	85.25	64.29	39.0	259.0
Rated	30.65	14.75	35.71	88.5	355.0
			G10 Musculoskeletal		
Total	9,519,873	13,125	139	59.0	1179.0
Male	82.25	88.58	90.65	58.0	1192.0
Female	17.75	11.42	9.35	69.0	828.0
Non-rated	52.19	65.67	66.19	34.0	1161.5
Rated	47.81	34.33	33.81	115.0	1205.0
			G11 Injuries		
Total	4,234,648	13,749	85	62.0	519.0
Male	87.86	90.89	97.65	62.0	519.0
Female	12.14	9.11	2.35	65.0	1419.5
Non-rated	42.49	55.07	60.00	42.0	764.0
Rated	57.51	44.93	40.00	89.0	363.5
			G12 All others		
Total	2,857,232	5,087	137	47.0	409.0
Male	74.77	79.91	90.51	42.0	390.5
Female	25.23	20.09	9.49	77.0	1101.0
Non-rated	64.48	72.03	71.53	28.0	415.5
Rated	35.52	27.97	28.47	90.0	409.0

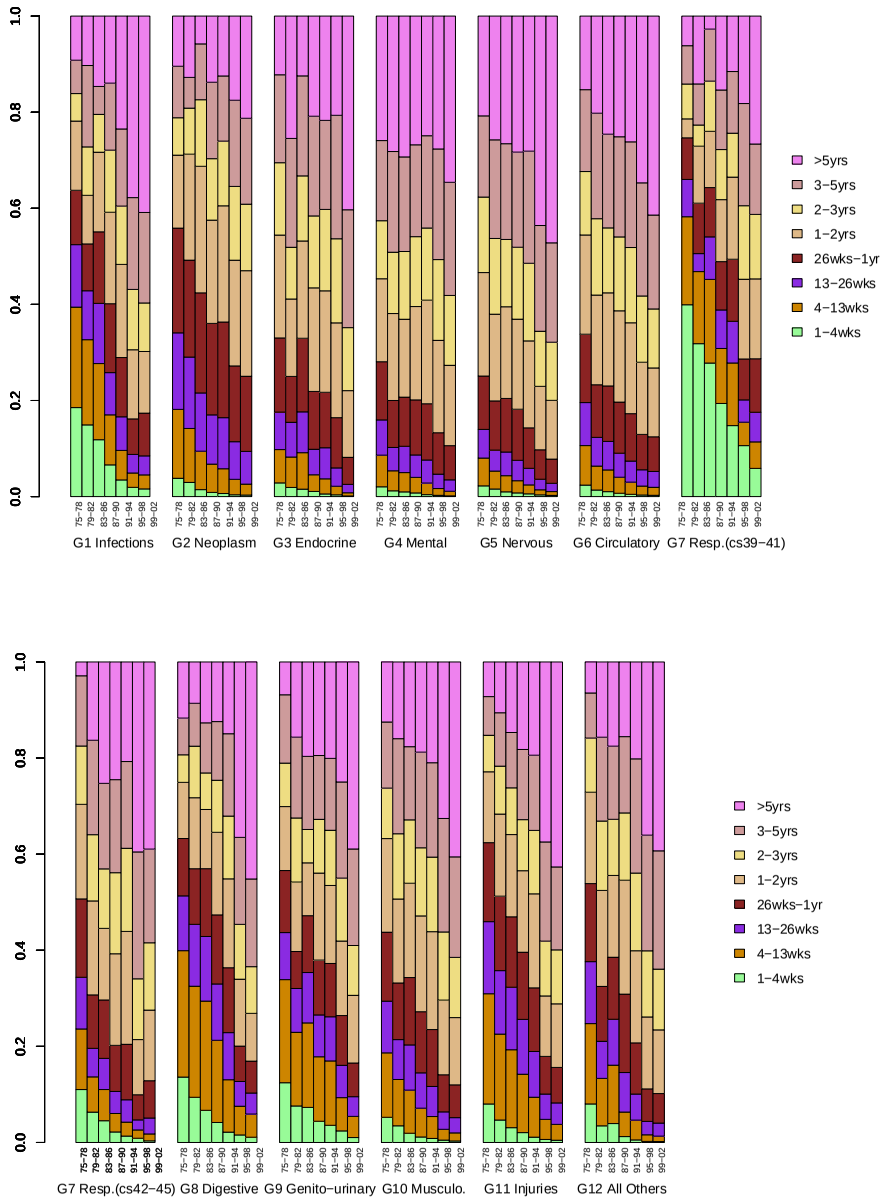


Figure 2.4: Barplots showing the exposed to risk as proportion of totals for each sickness duration interval by quadrennium and sickness category.

Chapter 3

Modelling the Recovery Intensity

3.1 Introduction

The purpose of this chapter is to present methods and results fot the estimation of the recovery intensity model by cause of sickness using the IPI claim data by ICD8 cause as described in Chapter 2. In terms of substantial work done on the analysis of IPI claim experience by cause of sickness, CMI Working Paper 23 (2006) examined the variation of sickness experience by quadrennium, deferred period and occupational class for each sickness category, separately for both sexes, using IPI 'standard*' data from 1991–2002. The sickness categories employed by CMI Working Paper 23 (2006) and the causes of sickness (represented by their ICD8 code) which constitute each sickness category are presented in Table 2.2. This one-way analysis of claim experience is conducted by comparing the ratio between the actual number (A) of recoveries and deaths and that expected (E) under SM1975–78. The key conclusions from this report are as follows:

(i) There is a wide variation in both recovery and mortality from sick experience by sickness category. The $100A/E$ values for 'Infections & acute respiratory', 'Digestive' and 'Genito-urinary' are relatively higher than the rest of the sickness categories.

(ii) The $100A/E$ values for recoveries for each sickness category and both sexes decline over the quadrennia.

(iii) There is a general trend of slight reduction in $100A/E$ values for recoveries with increasing deferred period in each sickness category with the exception that $100A/E$ values for recoveries increase with longer deferred period for 'Neoplasms'.

(iv) There is little variation in both recovery and mortality from sick experience by occupational class in all sickness categories with the exception that the occupational class 1 $100A/E$ values for recoveries for 'Musculoskeletal' and 'Injuries' are slightly higher than those of the other occupational classes.

The marginal analyses presented in CMI Working Paper 23 (2006) do not shed light on how the covariates jointly relate to the recovery and mortality intensities from sick. We therefore require a regression-type model to incorporate the dependence of recovery intensity on covariates. There are examples of regression-type models in insurance modelling. In respect of long-term care insurance, Czabo and Rudolph (2002) studied the effect of the covariates on the transition intensities between possible states by using the Cox regression model (Cox, 1972). Jones (1997) also presented the Cox model approach and demonstrated the methodology in analysing continuing care retirement community (CCRC) data. Pitt (2007) used a mixture parametric regression model that takes into account the probability of being totally and permanently disabled in the modelling of claim duration for IPI policyholders.

We let the recovery intensity model for a specific cause of sickness take on a multiplicative structure consisting of a baseline intensity and a relative risk component. The baseline intensity, denoted by $\rho_0(z)$, is a function of sickness duration z alone. The relative risk component, represented by $\exp(\mathbf{x}\boldsymbol{\beta}(z))$, describes the multiplicative covariate effects of sex, age, deferred period, rating indicator and calendar year (denoted by covariate vector $\mathbf{x}$), where $\boldsymbol{\beta}(z)$ represents the vector of duration-varying coefficients of $\mathbf{x}$. We make such an allowance because when aggregate data from all causes of sickness is analysed, there is evidence that some of the covariate effects are duration-dependent (see Section 1.5). Thus, the recovery intensity for a particular

cause of sickness, allowing for the possible duration-varying effects of covariates, can be written as

$$\rho(z, \mathbf{x}) = \rho_0(z) \exp(\mathbf{x}\boldsymbol{\beta}(z)) \quad . \tag{3.1}$$

The parameterisation of the recovery intensity model is split into three stages. The first stage, presented in Section 3.2 involves estimating the relative risk, a task that includes

(i) Assuming a proportional hazards (PH) model for the recovery intensity (i.e. $\boldsymbol{\beta}(z) = \boldsymbol{\beta}$) and selecting which covariates to include in the relative risk.

(ii) Testing the PH assumption for the covariates.

(iii) Relaxing the PH assumption, if necessary, by modelling the duration-varying effect of the covariates.

The second stage is about finding a suitable parametric formula for the baseline intensity while the last stage involves estimating all the parameters in the recovery intensity model by using maximum likelihood estimation. These last two stages are described in Sections 3.3 and 3.4, respectively. The goodness-of-fit of the estimated model is assessed by using methods presented in Section 3.5. Finally, the estimated recovery intensity models for several causes of sickness are presented in Section 3.6, with the remainders given in Appendix B.

3.2 Parameterisation of the Relative Risk

The Cox PH model

As a starting point, the covariate effects are assumed to stay constant for all sickness durations z. Let $\boldsymbol{\beta}$ denote the vector of duration-fixed regression coefficients. We therefore have a PH model for the recovery intensity given by

$$\rho(z, \mathbf{x}) = \rho_0(z) \exp(\mathbf{x}\boldsymbol{\beta}) \tag{3.2}$$

The Cox PH model (Cox, 1972) enables the estimation of $\boldsymbol{\beta}$ without making any assumption about the functional form of $\rho_0(z)$ through the use of the partial likelihood, $L_p(\boldsymbol{\beta})$, as given by

$$L_p(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\mathbf{x}_{(j)}\boldsymbol{\beta})}{\sum_{l \in R(z_{(j)})} \exp(\mathbf{x}_l \boldsymbol{\beta})} \tag{3.3}$$

where $\{z_{(j)}\}$ is the set of unique event times, sorted in ascending order such that $z_{(1)} < z_{(2)} < \cdots < z_{(r)}$, $R(z_{(j)})$ denotes the set of individuals at risk at event time $z_{(j)}$, $\mathbf{x}_{(j)}$ denotes the covariate vector for the individual who experiences the event at time $z_{(j)}$ and $\mathbf{x}_l$ denotes the covariate vector of individuals belonging to each risk set.
The log partial likelihood obtained by taking the log on both sides of Equation (3.3) is given by $l_p(\boldsymbol{\beta})$, where

$$l_p(\boldsymbol{\beta}) = \log L_p(\boldsymbol{\beta}) = \sum_{j=1}^{r} \left\{ \mathbf{x}_j \boldsymbol{\beta} - \log \left(\sum_{l \in R(z_{(j)})} \exp(\mathbf{x}_l \boldsymbol{\beta}) \right) \right\} \quad . \tag{3.4}$$

While the partial likelihood does not use all the information available from the data (information between event times is discarded), it has been shown to maintain the properties of a full likelihood (Andersen and Gill, 1982). The estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, found by maximising either $L_p(\boldsymbol{\beta})$ or $l_p(\boldsymbol{\beta})$ is asymptotically normal, consistent, efficient and unbiased.

The partial likelihood in Equation (3.3) is developed by treating time as continuous and assuming that there are no ties in the event times. The recovery time in the IPI claim data is measured in days and it is very common to find more than one recovery at any event time. The exact partial likelihood for ties in event time is based on a discrete time process where events do happen at exactly the same time. For each risk set $R(z_{(j)})$, let d_j be the number of tied recoveries at event time $z_{(j)}$ and let $D(z_{(j)})$ denote the set of d_j tied recoveries. Let $Q(z_{(j)})$ be the set of all possible subsets of d_j individuals which can be selected from $R(z_{(j)})$. The exact partial likelihood is given by

$$L_p(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\prod_{k \in D(z_{(j)})} \exp(\mathbf{x}_k \boldsymbol{\beta})}{\sum_{l \in Q(z_{(j)})} \exp(\mathbf{x}_l \boldsymbol{\beta})} \quad . \tag{3.5}$$

This method involves numerous permutations of the possible risk set at each tied event time and can be very time consuming to calculate if there is a large number of events at each tied event time. For example, if there were 20 tied recoveries at a particular recovery time from a pool of 60 individuals in the risk set, the sum in the denominator would be over all $\binom{60}{20}$ subsets, which is computationally prohibitive to calculate.

We will use Efron's approximation (1977) to the exact partial likelihood. Efron's approximation treats time as continuous and ties happen because of imprecise measurement. To explain the intuition behind Efron's approximation, suppose that there are four individuals in a particular risk set with a tied recovery time happening to individuals 1 and 4. The risk set for the first recovery will contain all four individuals. If the first recovery were to happen to individual 1, then the risk set for the second recovery would consist of individuals 2, 3 and 4. On the other hand, if individual 4 were to recover first, the risk set for the second recovery would consist of individuals 1, 2 and 3. Since both cases are equally likely to happen, each of the two possible risk sets has a probability of 0.5 of being the second risk set. Efron's approximate partial likelihood is given by

$$L_p(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\prod_{k \in D(z_{(j)})} \exp(\mathbf{x}_k \boldsymbol{\beta})}{\prod_{g=1}^{d_j} \left[\sum_{l \in R(z_{(j)})} \exp(\mathbf{x}_l \boldsymbol{\beta}) - \frac{g-1}{d_j} \sum_{l \in D(z_{(j)})} \exp(\mathbf{x}_l \boldsymbol{\beta}) \right]} \quad . \tag{3.6}$$

Selection of covariates

The information provided in each individual claim record is presented in Section 2.2. Table 3.1 gives a description of the information which can be included as explanatory variables in the regression model for the recovery intensity.

Table 3.1: The potential explanatory variables for the recovery intensity model.

Predictor	Description
Age	Age at sickness inception (17-69)
Year	Attained calendar year (1975-2002)
Rating Indicator	Indicator variable for rated; rated=1, non-rated=0
Sex	Indicator variable for female; female=1, male=0
Deferred Period	Possible values are deferred period of 1 week (DP1), 4 weeks (DP4), 13 weeks (DP13), 26 weeks (DP26) and 52 weeks (DP52)

Both age and year are continuous covariates and, for computational stability, their values are scaled and are represented by x_{age} and x_{year} respectively, where

$$x_{\text{age}} = (\text{age} - 43)/26, \quad x_{\text{year}} = (\text{year} - 1988)/13$$

so that x_{age} ranges from -1 to 1 when age ranges from 17 to 69 and x_{year} ranges from -1 to 1.076923 (≈ 1) when year ranges from 1975 to 2002.

The age and calendar year effects are modelled by using Chebycheff polynomials. The Chebycheff polynomial $C_n(x_{\text{age}})$ of degree n is generated by the following recurrence relation

$$C_0(x_{\text{age}}) = 1, \quad C_1(x_{\text{age}}) = x_{\text{age}}, \quad C_{n+1}(x_{\text{age}}) = 2(x_{\text{age}})C_n(x_{\text{age}}) - C_{n-1}(x_{\text{age}}) \qquad \text{for} \quad n \geq 1$$

so that by denoting $x_{\text{agei}} = C_i(x_{\text{age}})$, $i = 2, 3, 4$

$$x_{\text{age0}} = 1, \quad x_{\text{age1}} = x_{\text{age}}, \quad x_{\text{age2}} = 2x_{\text{age}}^2 - 1, \quad x_{\text{age3}} = 4x_{\text{age}}^3 - 3x_{\text{age}}, \quad x_{\text{age4}} = 8x_{\text{age}}^4 - 8x_{\text{age}}^2 + 1,$$

are approximately an orthogonal basis (see Forfar *et al*, 1988). The Chebycheff polynomials for x_{year} are constructed and defined in the same manner.

The discrete covariates sex, rated and deferred period are coded in the following way:

$$x_{\text{sex}} = \begin{cases} 1 & \text{for female} \\ 0 & \text{for male} \end{cases} \qquad x_{\text{rated}} = \begin{cases} 1 & \text{for rated} \\ 0 & \text{for non-rated} \end{cases}$$

$$x_{\text{dp4}} = \begin{cases} 1 & \text{for DP4} \\ 0 & \text{for other deferred periods} \end{cases} \qquad x_{\text{dp13}} = \begin{cases} 1 & \text{for DP13} \\ 0 & \text{for other deferred periods} \end{cases}$$

$$x_{\text{dp26}} = \begin{cases} 1 & \text{for DP26} \\ 0 & \text{for other deferred periods} \end{cases} \qquad x_{\text{dp52}} = \begin{cases} 1 & \text{for DP52} \\ 0 & \text{for other deferred periods} \end{cases}$$

Let S be the set comprising the covariates in Table 3.1, the Chebycheff polynomials of age and year up to degree four and all possible interaction terms. To determine which covariates in S to include in the Cox model, we rely on the Akaike Information Criterion (AIC) (Akaike, 1974) defined by

$$\text{AIC} = -2\log(L_p(\hat{\boldsymbol{\beta}})) + 2n \tag{3.7}$$

where $L_p(\hat{\boldsymbol{\beta}})$ is the partial likelihood evaluated at the estimated coefficient vector $\hat{\boldsymbol{\beta}}$ and n is the number of parameters in the model. The AIC values from models fitted with varying numbers and combinations of covariates from S are compared and the model that gives the lowest AIC value is selected. This model selection procedure is carried out by using the 'stepAIC' function in the 'R' statistical package which enables various models to be explored and their AIC values compared in an automated manner. It is should be noted that selection of covariates based on AIC may lead to inclusion of non-significant or marginally significantly covariates.

Testing of PH assumptions

The key assumption in using the Cox model is that the PH assumptions for all the covariates are valid. Suppose that a Cox regression model is fitted with n covariates and let $\hat{\boldsymbol{\beta}}$ be the vector of estimated regression coefficients. The Schoenfeld residual (Schoenfeld, 1982) is useful for identifying violation of proportionality assumption. It compares the observed and expected covariate value and is calculated for each covariate at observed event times. The Schoenfeld residual, $r_{s_{ji}}$, is defined as the jth covariate value for the individual who experiences the event at time z_i, $x_j^{(i)}$, minus its expected value at that time, $\hat{a}_j^{(i)}$. This expected value is a weighted average of the covariate, with the weight given by the likelihood of risk for each individual in the risk set at the event time. Hence

$$r_{s_{ji}} = x_j^{(i)} - \hat{a}_j^{(i)} \tag{3.8}$$

where

$$\hat{a}_j^{(i)} = \frac{\sum_{l \in R(z_i)} x_{jl} \ \exp(\mathbf{x}_l \hat{\boldsymbol{\beta}})}{\sum_{l \in R(z_i)} \exp(\mathbf{x}_l \hat{\boldsymbol{\beta}})} \quad . \tag{3.9}$$

The Schoenfeld residulas assess the relative magnitude of an individual's covariate value in comparison to what we expect given his or her event time. If the proportionality assumption holds, the Schoenfeld residuals will be unrelated to time and a plot of the Schoenfeld residuals versus observed event time should reveal no consistent trend.

Let $\mathbf{r}_{s_i} = (r_{s_{1i}}, r_{s_{2i}}, \ldots, r_{s_{ni}})^T$ denote the vector of Schoenfeld residuals. The scaled Schoenfeld residuals as proposed by Therneau and Grambsch (1994), $r^*_{s_{ji}}$, are the improved version of the original Schoenfeld residuals and are the components of the vector

$$\mathbf{r}^*_{s_i} = d_r \text{var}(\hat{\boldsymbol{\beta}}) \mathbf{r}_{s_i} \tag{3.10}$$

where d_r is the total number of recoveries and $\text{var}(\hat{\boldsymbol{\beta}})$ is the variance-covariance matrix of the parameter estimates in the fitted Cox regression model. Therneau and

Grambsch (1994) show that

$$E(r^*_{s_{ji}}) + \hat{\beta}_j \approx \beta_j(z_i) \tag{3.11}$$

where $\hat{\beta}_j$ is the Cox' estimate for β_j and $\beta_j(z)$ is the duration-varying coefficient.

By plotting $r^*_{s_{ji}} + \hat{\beta}_j$ against event time z_i or some function of event time $g(z_i)$, we obtain an approximation to the functional form of $\beta_j(z)$. Interpretation of such a plot is facilitated by using smoothing splines on the residuals. If the PH assumption is valid, such a smoothed plot should reveal a horizontal line that suggests that the coefficient of x_j stays constant for all durations. Otherwise, the examination of such a plot will reveal the form of deviation from the PH assumption. There exists a formal test, developed by Grambsch and Therneau (1994), to detect a linear relationship between the scaled Schoenfeld residuals and $g(z)$. The test statistic used in this test is shown to be analogous to the standard test of assessing the correlation between two variables and has an asymptotic χ^2_1 distribution when the PH assumption is true. However, this test may fail to detect a non-linear association between these residuals and $g(z)$. It is therefore recommended to examine graphically the smoothed plot of scaled Schoenfeld residuals. Hess (1995) advocates the smoothed plots of scaled Schoenfeld residuals following a review of eight graphical methods of assessing the PH assumption on three different data sets.

Estimation of coefficients using GLM

To estimate the duration-varying coefficients, the functional form of duration dependency of the coefficient must be specified parametrically. Suppose that the coefficient of covariate x varies with duration z and is included in the model as $x\beta(z)$ where $\beta(z) = \beta_0 + \beta_1 f(z)$. The functional form of $f(z)$ can be specified by examining the smoothed plot of the scaled Schoenfeld residuals. The term $x\beta(z)$ can also be written as $x\beta_0 + x(z)\beta_1$ where $x(z) = xf(z)$ is a duration-dependent variable. Thus, estimating $\beta(z)$ involves estimating the coefficients of x and $xf(z)$. The Cox model, developed assuming PH (i.e. constant coefficients for the covariates), can be extended to estimate a duration-varying coefficient by expressing it as the constant coefficient of a suitably defined duration-dependent covariate.

As the partial likelihood in Equation (3.3) requires the relative risk of every individual in the risk set at each ordered event time to be included in the denominator, the duration-dependent covariate has to be evaluated at every event time over the follow-up time of each individual claim record. Thus, a set of pseudo-observations designed to accommodate this duration-varying covariate has to be generated for each individual claim record. Depending on the size of the original data set and the form of the duration-dependent covariates used, such an exercise can potentially increase the size of the data set to the extent that estimation of parameters is no longer computationally viable.

To reduce data storage and computational time, we can group the data according to their covariate pattern (distinct combinations of the covariate values) and estimate the parameters by using the grouped data version of the Cox partial likelihood. From the original individual claim records, the data is cross classified by sex (2 levels: male and female), deferred period (5 levels: 1, 4, 13, 26 and 52 weeks), occupational rating (2 levels: rated and non-rated), age last birthday at sickness inception (49 levels: 20 to 69), calendar year (28 levels: 1975 to 2002) and sickness duration partitioned into 149 discrete intervals of

- single days from 7 to 133 days (or 19 weeks) of sickness (i.e. intervals of 7–8 days, 8–9 days, ..., 132–133 days).
- single weeks from 19 weeks to 30 weeks of sickness (i.e. intervals of 19–20 weeks, 20–21 weeks, ..., 29–30 weeks) followed by intervals of 30–39 weeks and 39 weeks–1 year.
- single years from 1 year to 8 years of sickness (i.e intervals of 1–2 years, 2–3 years , ..., 7–8 years) followed by 8–12 years and 12–16 years.

We refer the above partition of sickness duration as partitioning system A to differentiate it from other types of partition found in later sections. The exposed-to-risk and the number of recoveries for each distinct combination of covariate pattern, indexed by l, and sickness duration interval, indexed by m, are calculated and are denoted by r_{lm} and d_{lm} respectively. Let z_m be the mid-point of the mth sickness

duration interval. In grouped data, the duration-dependent covariate $x(z_m) = xf(z_m)$ is treated as an interaction term between the covariate x and $f(z_m)$. With duration-varying coefficients expressed in terms of duration-dependent covariates, the grouped data version of Cox's partial likelihood, $L_g(\boldsymbol{\beta})$, is given by

$$L_g(\boldsymbol{\beta}) = \prod_{l,m} \left(\frac{\exp(\mathbf{x}_l(z_m)\boldsymbol{\beta}_g)}{\sum_l r_{lm} \exp(\mathbf{x}_l(z_m)\boldsymbol{\beta}_g)} \right)^{d_{lm}} \quad . \tag{3.12}$$

Although the use of grouped data will result in the loss of efficiency in parameter estimation due to discretisation of continuous covariates, the reduction in computational time outweighs this drawback. Breslow (1985) compared the grouped and continuous Cox model analysis using the Montana smelter workers data set and found that both approaches give similar results. He also commented on the considerable saving in computational time by using the grouped data.

There is a total of 149 sickness duration intervals. Holford (1976) showed that Equation (3.12) can be obtained by assuming that the baseline intensity in each sickness duration interval is constant. To illustrate this, we let the sickness duration interval cut-points be τ_m $(m = 1, \ldots, M)$ and set $\rho_0(z) = \rho_m$ for $m \in [\tau_{m-1}, \tau_m)$. The likelihood of the data is given by

$$L = \prod_{l,m} \exp(-r_{lm}\rho_m \exp(\mathbf{x}_l(z_m)\boldsymbol{\beta}))(\rho_m \exp(\mathbf{x}_l(z_m)\boldsymbol{\beta}))^{d_{lm}} \quad . \tag{3.13}$$

The log likelihood of the data obtained by taking logs on both sides of Equation (3.13) is given by

$$\log L = \sum_{l,m} \{-r_{lm}\rho_m \exp(\mathbf{x}_l(z_m)\boldsymbol{\beta}) + d_{lm} \log(\rho_m) + d_{lm}\mathbf{x}_l(z_m)\boldsymbol{\beta}\} \quad . \tag{3.14}$$

Setting the derivative of Equation (3.14) with respect to ρ_m equal to zero yields

$$\hat{\rho}_m = \frac{\sum_l d_{lm}}{\sum_l r_{lm} \exp(\mathbf{x}_l(z_m)\boldsymbol{\beta})} \tag{3.15}$$

which can be substituted into Equation (3.14) to give

$$\log L(\boldsymbol{\beta}) = \sum_{l,m} d_{lm}(\mathbf{x}_l(z_m)\boldsymbol{\beta}) - \sum_m \left(\sum_l d_{lm} \right) \log \left(\sum_l r_{lm} \exp(\mathbf{x}_l(z_m)\boldsymbol{\beta}) \right) + C \quad (3.16)$$

where C is a constant term. By exponentiating Equation (3.16), we obtain Equation (3.12).

Holford (1980) and Laird and Oliver (1981) pointed out that we can maximise the likelihood in Equation (3.13) by using a Poisson regression model in the GLM. To illustrate this, we assume that the observed number of recoveries d_{lm} follows a Poisson distribution with mean

$$\mu_{lm} = E(d_{lm}) = r_{lm}\rho_m \exp(\mathbf{x}_l(z_m)\boldsymbol{\beta}) \quad . \quad (3.17)$$

The kernel of the likelihood by treating d_{lm} with the above mean coincides with the log likelihood under a piece-wise constant baseline intensity in Equation (3.13). Taking logs on both sides of Equation (3.17), we obtain

$$\log(\mu_{lm}) = \log(r_{lm}) + \log(\rho_m) + \mathbf{x}_l(z_m)\boldsymbol{\beta} \quad . \quad (3.18)$$

As Equation (3.12) is also the maximum likelihood estimator for $\boldsymbol{\beta}$ in a Poisson regression model, we can use a GLM with a Poisson error structure, log link function and $\log(r_{lm})$ as an offset term to estimate $\boldsymbol{\beta}$. However, such estimation requires the piece-wise constant baseline intensities, as represented by binary indicators, to be estimated alongside $\boldsymbol{\beta}$. The sickness duration has been partitioned into 149 intervals (i.e. partitioning system A) when we transform individual claim record data into grouped data. If 149 binary indicators are to be estimated, the computational time will increase dramatically. Since our primary purpose is to estimate $\boldsymbol{\beta}$, we will use only 32 binary indicators to represent the following 32 sickness intervals: 1–2 weeks, 2–3 weeks, 3–4 weeks, ..., 15–16 weeks, 16–18 weeks, 18–20 weeks, 20–23 weeks, 23–26 weeks, 26–30 weeks, 30–39 weeks, 39 weeks – 1 year, 1–2 years, 2–3 years, ..., 7–8 years, 8–12 years, 12–16 years and 16–20 years. We are aware that the reduction in computational time is achieved at the expense of efficiency of parameter estimation.

3.3 Parameterisation of the Baseline Intensity

There are two ways to estimate the piece-wise constant baseline intensity. The more direct or convenient way is to estimate jointly the piece-wise constant baseline intensity and the relative risk by using the Poisson regression approach in a GLM (see Section 3.2). Alternatively, we can obtain the parameter estimator, $\hat{\boldsymbol{\beta}}_g$, by maximising Equation (3.12) and setting $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_g$ in Equation (3.15). Both approaches will produce the same estimate for the piece-wise constant baseline intensity.

However, there is a need to obtain a smooth baseline intensity because the true recovery intensity is assumed to be a reasonably smooth mathematical function so that functions of practical importance calculated from the model will share this property as well.

The piecewise constant baseline intensity can be graduated by using either a parametric or non-parametric method. For non-parametric smoothing methods such as moving weighted average graduation and kernel smoothing, the degree of smoothness is varied by the choice of bandwidth and this often involves an element of subjectivity. We will therefore graduate the piece-wise constant baseline intensities parametrically by using Chebycheff polynomials, such that

$$\rho_0(z, \mathbf{b}) = \exp\left(\sum_{i=0}^{s} b_i C_i(t_k(z))\right)$$

where $\mathbf{b}$ represents the vector of regression coefficients and $C_i(t_k(z))$ denotes the Chebycheff polynomial in $t_k(z)$ of degree i with $t_k(z)$ being a function of duration z defined as $t_k(z) = z/(1+kz)$, so that by letting $y = t_k(z)$

$$C_0(y) = 1, \quad C_1(y) = y, \quad C_2(y) = 2y^2 - 1$$
$$C_3(y) = 4y^3 - 3y, \quad C_4(y) = 8y^4 - 8y^2 + 1.$$

The transformed duration variable $t_k(z)$ is used in the baseline intensity because it will tend towards an upper limit of $1/k$ as z increases without limit so that when the recovery intensity for very long sickness durations is calculated by extrapolating the graduation formula beyond the data range, the results obtained are more likely to be

sensible. This transformed duration variable is also used in CMI Working Paper 15 (2004) due to this property.

We need to explore which degree of Chebycheff polynomial is suitable as the baseline intensity and what value of k should be used in the duration transformation. This involves trying out a range of k values for varying degrees of Chebycheff polynomials and comparing their maximised likelihoods. When the total number of parameters (for both $\boldsymbol{\beta}$ and $\mathbf{b}$) or the size of the data is large, such a procedure can be very time consuming. Since our interest lies in finding the optimal structure for the baseline intensity and its associated k value, we will fix $\boldsymbol{\beta}$ at its estimated value, $\hat{\boldsymbol{\beta}}_g$. By substituting $\rho_m = \rho_0(z_m, \mathbf{b})$ and removing constant terms, Equation (3.14) becomes

$$\log L(\mathbf{b}) = \sum_{l,m} \left\{ -r_{kl}\rho_0(z_m, \mathbf{b}) \exp(\mathbf{x}_l \hat{\boldsymbol{\beta}}_{\boldsymbol{g}}) + d_{kl} \log(\rho_0(z_m, \mathbf{b})) \right\} \quad . \tag{3.19}$$

By putting the Chebycheff polynomial of degree s ($s = 1, \ldots, 6$) in turn as the baseline intensity, we will maximise Equation (3.19) by using a range of k values, say $k = 0.0, 0.1, \ldots, 4.0$, and the value of k that yields the largest log likelihood is selected. The optimal degree of polynomial is selected by examining the values of the log likelihood produced by the selected k for each degree of polynomial. Twice the difference in the log likelihood between two different degrees of polynomial is approximated by a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters used.

3.4 Fully Parameterised Recovery Intensity Model

The general expression for a fully parameterised recovery intensity model is as follows:

$$\begin{aligned}\rho(z,\mathbf{x}) &= \exp(\sum_{i=0}^{n} b_i C_i(t_k(z)))\exp(x_{\text{sex}}\beta_{\text{sex}}(z) + x_{\text{rated}}\beta_{\text{rated}}(z) \\ &\quad + x_{\text{dp4}}\beta_{\text{dp4}}(z) + x_{\text{dp13}}\beta_{\text{dp13}}(z) + x_{\text{dp26}}\beta_{\text{dp26}}(z) + x_{\text{dp52}}\beta_{\text{dp52}}(z) \\ &\quad + \sum_{i=1}^{n} x_{\text{yeari}}\beta_{\text{yeari}}(z) + \sum_{i=1}^{n} x_{\text{agei}}\beta_{\text{agei}}(z) + \mathbf{\Phi})\end{aligned} \tag{3.20}$$

where each subscripted $\beta(z)$ denotes the duration-varying coefficient of its corresponding subscripted covariate x and $\mathbf{\Phi}$ denotes the interaction terms.

The duration-varying coefficient can take many forms. To describe the possible forms taken, we take $\beta_{\text{dp4}}(z)$ as an example. In the case of duration-fixed covariate effect, $\beta_{\text{dp4}}(z) = \alpha_{\text{dp4}}$. In the case of duration-dependent covariate effect, $\beta_{\text{dp4}}(z) = \alpha_{\text{dp4}} + f_{\text{dp4}}(z)$. It transpires that the duration-varying effect can either persist for all durations or only for a certain period of sickness duration. In the latter case, the point at which the duration-varying effect started or ended is known as the "break-point". We use $t_k(z)$ and z to model, respectively, a long period and a short period of duration-varying effect. Therefore, $f_{\text{dp4}}(z)$ consists of one or several of the following components that represent different types of duration dependency.

(i) $\gamma_{\text{dp4}}(\tau_{\text{dp4}} - z)_+$

(ii) $\theta_{\text{dp4}_i} C_i(t_k(z)), \quad i = 1, 2, \ldots$

(iii) $\zeta_{\text{dp4}_i}(C_1(t_k(\tau_{\text{dp4}})) - C_1(t_k(z)))_+^i, \quad i = 1, 2, \ldots$

(iv) $\phi_{\text{dp4}_i}(C_1(t_k(z)) - C_1(t_k(\tau_{\text{dp4}})))_+^i, \quad i = 1, 2, \ldots$

where $\gamma_{\text{dp4}}, \theta_{\text{dp4}_i}, \zeta_{\text{dp4}_i}$ and ϕ_{dp4_i} are regression parameters, τ_{dp4} is a break point and $y_+ = y$ if $y > 0$ and 0 otherwise. Note that $\theta_{\text{dp4}_1} = \theta_{\text{dp4}}$, $\zeta_{\text{dp4}_1} = \zeta_{\text{dp4}}$ and $\phi_{\text{dp4}_1} = \phi_{\text{dp4}}$. In the case of two breakpoints, we will refer to the first one as τ_{dp4_1} and the second one as τ_{dp4_2}. The breakpoint (if any) is placed after examining the variation of the log

hazard ratio with duration as revealed by the smoothed plot of the scaled Schoenfeld residuals. The "partial residual effects" plot, which will be discussed in Section 3.5, is also used to guide the placement of breakpoint.

Since the recovery intensity model in Equation (3.20) can be written as an additive log-linear model, all the parameters can be estimated by using a Poisson regression model with log link function in a GLM.

3.5 Model Assessment

Let $\rho_0(z, \mathbf{b}) \exp(\mathbf{x}\boldsymbol{\beta}(z))$ be a fully estimated recovery intensity model for a specific cause of sickness where $\hat{\mathbf{b}}$ and $\hat{\boldsymbol{\beta}}$ are the maximum likelihood estimates of $\mathbf{b}$ and $\boldsymbol{\beta}$ respectively.

"Partial residual effects" and "residual effects" plots

We wish to check whether the estimated covariate effects, be it duration-fixed or duration-dependent, give a reasonable representation of the actual recovery pattern.

Let the covariate vector $\mathbf{x}$ be partitioned into $\mathbf{x} = (x_1, \mathbf{x}_2)$ where x_1 is a discrete covariate that takes the value 0 or 1. Gray (1990) refers to "residual effects" as the intensity from which all covariate effects have been removed while "partial residual effects" of x_1 means all covariate effects except x_1 have been removed. Gray (1990) proposed using the "residual effects" and "partial residual effects" plots to indicate the approximate form of $\rho_0(z)$ and $\rho_0(z) \exp(x_1\beta_1(z))$, respectively and estimated these intensities by applying kernel-based smoothing to the Breslow (1974) estimator for the cumulative baseline intensity but suggested using piece-wise constant hazards and other smoothing methods as alternatives to kernel-based smoothing.

We will estimate $\rho_0(z) \exp(x_1\beta_1(z))$ using data for which $x_1 = 1$, by partitioning the sickness duration into 64 intervals of

- three days from 7 to 133 days (19 weeks) of sickness (i.e. intervals of 7–10 days, 10–13 days, ..., 127–130 days, 130–133 days).

- single weeks from 19 weeks to 30 weeks of sickness (i.e. intervals of 19–20 weeks, 20–21 weeks, ..., 29–30 weeks) followed by intervals of 30–39 weeks and 39 weeks–1 year.

- single years from 1 year to 8 years of sickness (i.e intervals of 1–2 years, 2–3 years, ..., 7–8 years) followed by 8–12 years and 12–16 years.

Note that the above partition of sickness duration is different to that in partitioning system A and will be referred to as partitioning system B. We assume that the intensity in each interval is constant. By using the notation of grouped data described in Section 3.2, the "partial residual" effect of x_1 in the mth interval ($m = 1, \ldots, 64$), $\rho(z_m, x_1) = \rho_0(z_m) \exp(x_1 \beta_1(z_m))$, is estimated by $\hat{\rho}(z_m, x_1)$ where

$$\hat{\rho}(z_m, x_1) = \frac{\sum_l d_{lm}}{\sum_l r_{lm} \exp(\mathbf{x}_{2l} \hat{\boldsymbol{\beta}}_{\mathbf{2}}(z))} \quad . \tag{3.21}$$

Note that the covariate x_1 is not included in the exponential term in the denominator and therefore the effect of x_1 is not "removed" from the estimate. The"partial residual effects" plot provides a direct way of estimating the intensity. In the case that the log hazard ratio for x_1 varies with duration, the "partial residual effects" plot for x_1 will reveal the functional form of the variation (including the location of any break points) which should be similar to that revealed by the relevent smoothed Schoenfeld residuals plot. By overlaying the estimated smooth intensity, $\rho_0(z, \hat{\mathbf{b}}) \exp(x_1 \hat{\beta}_1(z))$, on the "partial residual effects" plot, we can see whether it fits reasonably well. The "partial residual effects" plot is constructed separately for female data ($x_{\text{sex}} = 1$), rated data ($x_{\text{rated}} = 1$), DP4 data ($x_{\text{dp4}} = 1$), DP13 data ($x_{\text{dp13}} = 1$), DP26 data ($x_{\text{dp26}} = 1$) and DP52 data ($x_{\text{dp52}} = 1$).

The techniques described above are also extended to continuous covariates such as age and year. To do so, age, year and sickness duration are categorised into discrete bands as follows:

(i) Age band (9) : 17–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59, 60–65

(ii) Year band (7) : 1975–1978, 1979–1982, 1983-1986, 1987–1990, 1991–1994, 1995-1998, 1999-2002

(iii) sickness duration band (28) : 1–2, 2–3, 3–4, ..., 10–11, 11–13, 13–15, 15–17, 17–21, 21–25, 25–30, 30–39, 39 weeks–1 year, 1–2, 2–3, ..., 7–8, 8–12, 12–16, 16–20 years

We refer to the above partition of age, year and sickness duration as partitioning system C. For each sickness duration band, we construct a "partial residual effects" plot for age. The estimates are calculated using Equation (3.21) but with the effects of sickness duration (i.e. the baseline intensity) and all other covariate effects apart from age included in the exponential term of the denominator. From these plots, we are able to view the shape of the hazard ratio as a function of the discrete age bands and how this shape changes over the sickness duration bands. To see how well the smooth intensity of age fitted the actual experience, we overlay on the "partial residual effects" plot for the mth sickness duration band ($m = 1, 2, \ldots, 28$) the smooth curve $\exp(\sum_i x_{\text{agei}} \hat{\beta}_{\text{agei}}(z_m))$, where z_m is the mid point of the mth sickness duration band. The "partial residual effects" plot for year is constructed analogously.

The "residual effects" plot, in which all covariate effects are removed, is constructed separately using all the data, males data ($x_{\text{sex}} = 0$), non-rated data ($x_{\text{rated}} = 0$) and DP1 data ($x_{\text{dp4}} = x_{\text{dp13}} = x_{\text{dp26}} = x_{\text{dp52}} = 0$). The estimates are calculated by including all the covariate effects in the denominator of Equation (3.21). The smooth baseline intensity, $\rho_0(z, \hat{\mathbf{b}})$, is then overlaid on these plots.

The confidence interval for the "partial residual effects" and "residual effects" can be obtained on the basis of normal approximation by ensuring that the numbers of recoveries used to calculate these estimates are sufficiently large, say greater than 8 through appropriate merger of discrete bands. Suppose we wish to calculate the confidence interval for the "partial residual effects" of x_1, $\rho(z_m, x_1)$. Let $d_{.m} = \sum_l d_{lm}$ and $R_{.m} = \sum_l r_{lm} \exp(\mathbf{x}_{2l} \hat{\boldsymbol{\beta}}_{\mathbf{2}}(\boldsymbol{z}))$. The 95% confidence interval for $\rho(z_m, x_1)$ is

$$\left(\hat{\rho}(z_m, x_1) - z_\alpha \sqrt{(\tfrac{\hat{\rho}(z_m, x_1)}{R_{.m}})}, \qquad \hat{\rho}(z_m, x_1) + \ z_\alpha \sqrt{(\tfrac{\hat{\rho}(z_m, x_1)}{R_{.m}})} \right) \tag{3.22}$$

where $\alpha = 0.025$ and $z_\alpha = 1.96$. By putting $\hat{\rho}(z_m, x_1) = d_{.m}/R_{.m}$, the interval (3.22) becomes

$$\left(\tfrac{d_{.m} - z_\alpha \sqrt{d_{.m}}}{R_{.m}}, \qquad \tfrac{d_{.m} + z_\alpha \sqrt{d_{.m}}}{R_{.m}} \right) \tag{3.23}$$

A more accurate confidence interval for $\rho(z_m, x_1)$ can be provided by assuming that the number of recoveries in the mth sickness duration band, $d_{.m}$, follows a Poisson distribution (see Forfar *et al*, 1988). If $d_{.m} = 0$, the lower limit is 0; if $d_{.m} > 0$, the lower limit is the unique positive root of the equation

$$\left.\begin{array}{r} \sum_{k=d_{.m}}^{\infty} e^{-\mu_L} \frac{\mu_L^k}{k!} = \alpha \\ \text{or equivalently,} \\ \sum_{k=0}^{d_{.m}-1} e^{-\mu_L} \frac{\mu_L^k}{k!} = 1 - \alpha \end{array}\right\} \quad . \tag{3.24}$$

The upper limit is the unique positive root of the equation

$$\left.\begin{array}{r} \sum_{k=0}^{d_{.m}} e^{-\mu_U} \frac{\mu_U^k}{k!} = \alpha \\ \text{or equivalently,} \\ \sum_{k=d_{.m}+1}^{\infty} e^{-\mu_U} \frac{\mu_U^k}{k!} = 1 - \alpha \end{array}\right\} \tag{3.25}$$

where $\mu_L = \rho_L(z_m, x_1) \sum_l r_{lm} \exp(\mathbf{x}_{2l}\hat{\boldsymbol{\beta}}_{\mathbf{2}}(z))$ and $\mu_U = \rho_U(z_m, x_1) \sum_l r_{lm} \exp(\mathbf{x}_{2l}\hat{\boldsymbol{\beta}}_{\mathbf{2}}(z))$. We will then solve for $\rho_L(z_m, x_1)$ and $\rho_U(z_m, x_1)$ which, respectively, form the lower and upper limits of the 95% confidence interval for $\rho(z_m, x_1)$. Since calculating the confidence interval based on the Poisson assumption is not computationally intensive, we will use it for $d_{.m} \leq 80$ and only use Equation (3.23) for $d_{.m} > 80$. Since the relevant calculations are simple for a computer, we have chosen 80 as the cut-off point even though 10 is usually sufficient for the normal approximation to hold.

Both confidence intervals, based on the Poisson or Normal assumption, are obtained by ignoring the estimation of $\boldsymbol{\beta}$. The true variability of the estimates is therefore higher than that given by these confidence intervals. These confidence intervals are therefore used only as a rough indicator of the magnitude of the variability of the estimates.

χ^2 test

In addition to the "partial residual effects" and "residual effects" plots described above, we wish to conduct a formal goodness-of-fit test on the model. For this pur-

pose, sickness duration, age and year are discretised into 9, 7 and 28 bands as in partitioning system C.

Using the same terminology as in CMI Report 15 (1996), the sickness experience for each combination of covariate levels for sex, rating indictor, deferred period and year band is referred to as a tableau. In each tableau, data are laid out in a two dimensional array with age bands as rows and sickness duration bands as columns. Two such arrays are constructed for each tableau, one containing the actual number of recoveries (A) and the other the expected number of recoveries (E). In each cell (the intersection of each distinct row and column), we will calculate $z = D/\sqrt{E}$, incorporating continuity corrections to allow for the fact that the actual number of recoveries is necessarily an integer, where:

$$D = \begin{cases} A - E - 0.5 & \text{if} \quad 0.5 < A - E \\ 0.0 & \text{if} \quad -0.5 \leq A - E \leq 0.5 \\ A - E + 0.5 & \text{if} \quad A - E < -0.5 \quad . \end{cases}$$

For each z to approximate a normal variate, its expected number of recoveries E has to be greater than a certain number which is usually taken as 8 in CMI Report 15 (1996). Thus, cells with fewer than 8 expected recoveries have to be merged with adjacent cells until at least 8 expected recoveries is obtained. We have a total of $2 \times 2 \times 5 \times 7 = 140$ possible tableaux and the total number of expected recoveries in each tableau has to exceed 8 before grouping of cells within the tableau can be carried out. Details about the merger of tableaux and the subsequent grouping of cells within each tableau can be found in Appendix E. The methodology used to group the cells in a tableau is described elsewhere in Appendix A of CMI Report 15 (1996).

The sum of the squares of the z values, after necessary grouping is carried out, is approximated by a χ^2 distribution with the number of degrees of freedom equal to the number of cells (after grouping) minus the number of parameters fitted in the model.

Two-dimensional plot of deviance residuals

Apart from the χ^2 statistic, the deviance statistic can be used to evaluate the goodness of fit of a generalised linear model (see McCullagh and Nelder, 1989, for details). For the Poisson model, the deviance statistic takes the following form:

$$\text{Deviance} = 2\sum_i \left\{ A_i \log(\frac{A_i}{E_i}) - (A_i - E_i) \right\}.$$

where A_i and E_i are respectively the actual and expected number of recoveries in cell i.

The deviance residual measures the contribution of each cell i to the deviance and is defined by

$$\text{sign}(A-E)\sqrt{2\left(A\log\frac{A}{E} - (A-E)\right)} \tag{3.26}$$

where sign(x) is a function that extracts the sign of x. Unlike in calculating χ^2 statistics, there is no need to merge cells so that the expected number of recoveries in each cell is greater than 8. An example of a two-dimensional plot of deviance residuals is given in Figure 3.1. This plot is constructed with age bands as rows and sickness duration bands as columns for each quadrennium arranged side by side. Positive and negative deviance residuals are shown in red and blue rectangles respectively. Both colours are represented by three different intensities, representing the different ranges of values for the deviance residual. Cells with no exposed-to-risk, and therefore for which deviance residuals cannot be calculated, are depicted in white. For each distinct age band, there are 28 rectangles between two adjacent year bands, representing the 28 sickness duration intervals arranged in increasing order. For a good fit to the data, the positive and negative deviance residuals of different ranges of values should be randomly scattered.

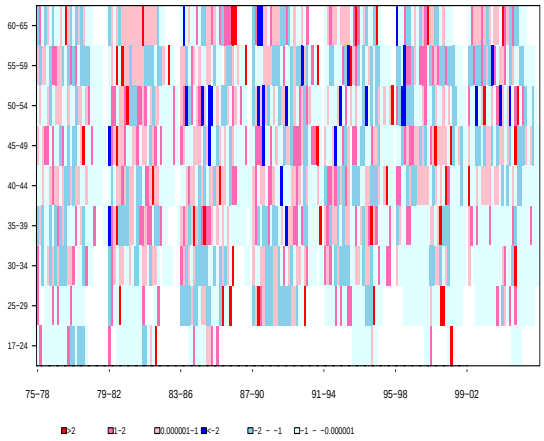


Figure 3.1: A two-dimensional plots of deviance residuals.

3.6 Estimation Results

There is a total of 70 causes of sickness and they are classified into 12 sickness categories as set out in Table 2.3 and reproduced here in Table 3.2.

Table 3.2: Grouping of the 70 causes of sickness into 12 categories.

Name	Sickness Category	ICD8 code for cause of sickness
G1	Infections & acute respiratory	01–19
G2	Neoplasms	20, 21
G3	Endocrine & Metabolic	22–26
G4	Mental Illness	27
G5	Nervous system & sensory organs	28–31
G6	Circulatory	32–38
G7	Respiratory	39–45
G8	Digestive (non-infectious)	47–51
G9	Genito-urinary	52–55
G10	Musculoskeletal	61, 62
G11	Injuries	66–70
G12	All other known causes	46, 56–60, 63–65

For the rest of this chapter, each cause of sickness will be represented by its ICD8 code prefixed by "cs". For example, malignant neoplasm is represented by cs20.

For each sickness category, we wish to check whether its constituent causes of sickness have broadly similar shape or level of recovery intensity. To do so, each cause of sickness is coded by a binary indicator variable that is denoted by the symbol $I_{\text{cs}i}$,

where i is the ICD8 code for the cause of sickness. For example, cs20 is coded by the binary indicator I_{cs20}. These binary indicators, apart from the one representing the reference cause of sickness, are estimated using the Cox model. We then rely on the smoothed plot of the scaled Schoenfeld residuals (see Section 3.2) to establish whether the recovery patterns for the causes of sickness in each sickness category are proportional to each other. Causes of sickness with recovery patterns which are proportional to each other will be modelled together by adding a set of binary indicators representing these causes of sickness into the covariate vector. On the other hand, causes of sickness which exhibit a disparate recovery pattern from the others will be modelled separately.

The stages involved in the estimation of a fully parameterised recovery intensity model for a cause of sickness are described in Sections 3.2, 3.3 and 3.4. In essence, these stages can be represented by three different models: Model I, Model II and Model III, where

(a) Model I is the Cox model in which the covariates are selected and estimated without making any assumption about the baseline intensity. The PH assumption for those covariates selected based on AIC criterion are assessed by examining the smoothed plots of their scaled Schoenfeld residuals (see Section 3.2).

(b) Model II is Model I extended to include duration-varying coefficients of covariates for which the PH assumption is not valid (see Section 3.2).

(c) Model III is Model II but with a parametric baseline intensity (see Section 3.3) and all the parameters are estimated by maximum likelihood estimation (see Section 3.4).

To avoid having a voluminous chapter, we will only present the estimation results from the above three models for causes of sickness in sickness categories G2, G4, G6 and G10. For the remaining sickness categories, only the estimation results from Model III are presented in Appendix B.

3.6.1 G2 Neoplasms

The causes of sickness in G2 Neoplasms alongside their ICD8 code, exposed to risk in units of days and the number of recoveries are presented in Table 3.3.

Table 3.3: The causes of sickness in sickness category G2 Neoplasms.

ICD8	Cause of sickness	Exposed to risk (days)	Recoveries
20	Malignant neoplasms, including neoplasms of lymphatic and haematopoietic	1,919,924	1,621
21	Benign neoplasms and neoplasms of unspecified nature	292,164	544

We first created a binary indicator for cs21, I_{cs21}, and estimated it using the Cox model. The proportionality between cs20 and cs21 is then assessed using the smoothed plots of the scaled Schoenfeld residuals. Due to the discrete nature of the covariate I_{cs21}, the resulting unscaled Schoenfeld residual at event time z_i (see Equation (3.8)) is $r_{s_i} = 1 - \hat{a}^{(i)}$ for $I_{cs21} = 1$ and $r_{s_i} = 0 - \hat{a}^{(i)}$ for $I_{cs21} = 0$, where $\hat{a}^{(i)}$ is given by Equation (3.9). Therefore, the scaled version of these residuals (i.e. the scaled Schoenfeld residuals as given in Equation (3.10)) appear as two horizontal bands of black dots at the top and bottom of Figure 3.2. To facilitate interpretation of these scaled Schoenfeld residuals, we use smoothing spline on the residuals. The solid line in this graph denotes the smoothed scaled Schoenfeld residuals while the broken lines represent a $\pm$ 2-standard-error band around the fit. Cox's estimate for I_{cs21} (i.e. the log hazard ratio between cs21 and cs20) is 0.848 and is represented by the broken red line. As discussed in Section 3.2, the smoothed scaled Schoenfeld residuals indicate the variation of the log hazard ratio with sickness duration. This plot indicates that the recovery intensity for cs21 is higher than cs20 until both intensities converge (i.e. $\beta(z) = 0$) somewhere between 200 to 500 days. This convergence in intensities may be due to the fact that as time goes by there are more neoplasms of unspecified nature left in the cs21 data which are in fact of malignant nature. Given the non-PH between both intensities, we decided to estimate them separately.

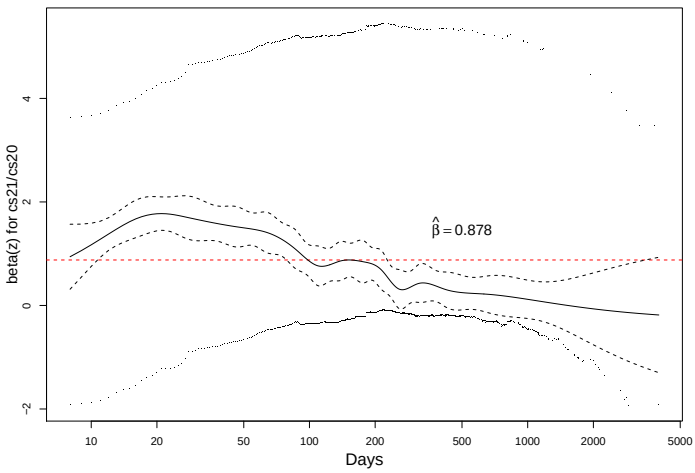


Figure 3.2: The smoothed plot of the Schoenfeld residuals, indicating the log hazard ratio between cs21 and cs20.

Recovery intensity model for cs20

For cs20, the estimation results from intermediate models (i.e. Model I and II) leading up to a fully specified recovery intensity model (i.e. Model III) are presented in Table 3.4. The first column of Table 3.4 shows the notation representing different types of regression coefficients as explained in Section 3.4. The second column of Table 3.4 shows the estimation results from Model I which is a Cox model that assumes that the covariate effects are duration-fixed (i.e. the PH assumption holds). This assumption is tested for each covariate by using the smoothed plot of the Schoenfeld residuals. The covariate effects of x_{year}, x_{dp4} and x_{dp13} are found to be duration-dependent because the smoothed plot of their respective scaled Schoenfeld residuals as presented in Figure 3.3 suggest that

(i) The year effect has a concave shape until some point between 39 weeks to 1 year, from which it starts tailing off till the end, but with a wider confidence interval. The falling trend at the right of the graph, observed also in the other two graphs, is most likely due to the few residuals at the bottom right of the graph. The functional form for $\beta_{\text{year}}(z)$ is therefore specified as

$$\beta_{\text{year}}(z) = \alpha_{\text{year}} + \zeta_{\text{year}_1}(C_1(t_{2.3}(\tau_{\text{year}})) - C_1(t_{2.3}(z)))_+ + \zeta_{\text{year}_2}((C_1(t_{2.3}(\tau_{\text{year}})) - C_1(t_{2.3}(z)))_+)^2$$

where $\tau_{\text{yr}} = 319.5$.

(ii) There is a 'run-in' period for DP4 and DP13, during which the recovery intensity is lower, such that

$$\beta_{\text{dp4}}(z) = \alpha_{\text{dp4}} + \gamma_{\text{dp4}}(\tau_{\text{dp4}} - z)_+$$
$$\beta_{\text{dp13}}(z) = \alpha_{\text{dp13}} + \gamma_{\text{dp13}}(\tau_{\text{dp13}} - z)_+$$

where $\tau_{\text{dp4}} = 58.5$ and $\tau_{\text{dp13}} = 185.5$.

Model II is then obtained by extending Model I to include additional covariates used to model the duration-varying effects of $x_{\text{year}}, x_{\text{dp4}}$ and x_{dp13}. The estimated

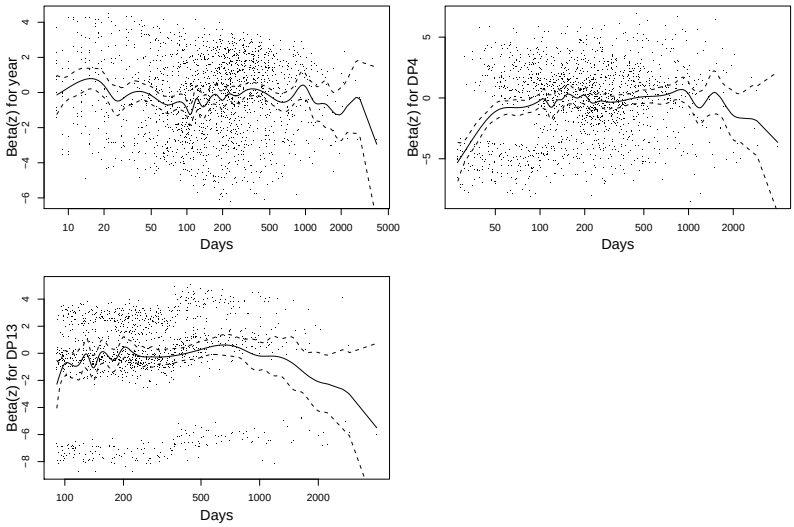


Figure 3.3: The smoothed plots of the scaled Schoenfeld residuals for $x_{\text{year}}, x_{\text{dp4}}$ and x_{dp13} in cs20.

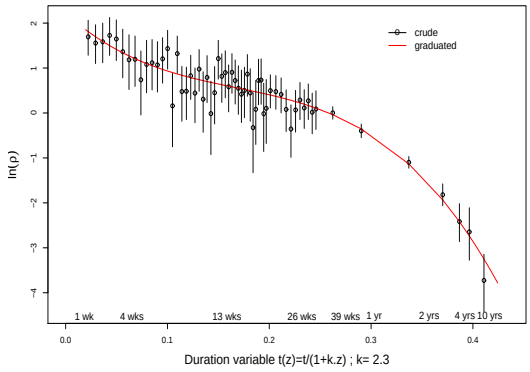
parameters in Model II are presented in column three of Table 3.4. Model II gives the complete parameterisation of the relative risk component and as in Model I, all the covariate effects are estimated without imposing any functional form on the baseline intensity. Lastly, Model III retains all the parameters in Model II but with the baseline intensity described by a $\exp(\sum_{i=0}^{3} b_i C_i(t_{2.3}(z)))$ formula. The estimated parameters in this fully parameterised recovery intensity model are presented in column four of Table 3.4. The standard error for each estimated parameters in these three models is given by the value in bracket.

The goodness-of-fit of Model III is then assessed using a series of techniques presented in Section 3.5. The value of the χ^2 statistic is 118.83. With a total of 125 cells (after grouping) and 17 parameters fitted in the model, the probability value is 0.224 on 108 degrees of freedom, indicating a reasonably good fit to the data. The "residual effects" and "partial residual effects" plots for the discrete covariates are presented in Figure 3.4. The data upon which each plot is based is indicated by the caption of the plot. Figures 3.5 and 3.6 show the "partial residual effects" plots for year and age, respectively, for each sickness duration band. The red curve overlaid on each of the plots in Figures 3.4–3.6 is the estimated smooth intensity according to Model III and all of them have fitted the actual experience reasonably well.

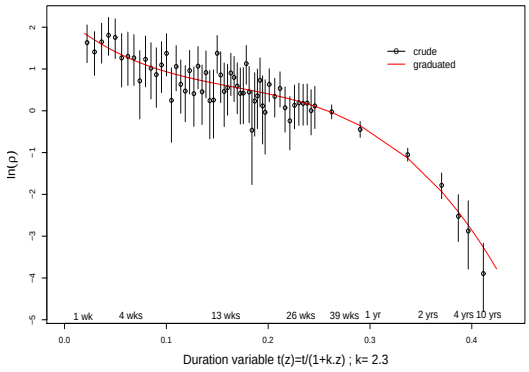
The two-dimensional plots of the deviance residuals for all data as well as its subsets are shown in Figure 3.7. The positive and negative deviance residuals for 'All' and 'Male & Not-rated' data are roughly randomly scattered. For other subsets, the plots are dominated by white and light blue cells because most of the cells have either zero exposed to risk or no recoveries.

Table 3.4: Parameters in the intermediate and final recovery intensity models for cs20.

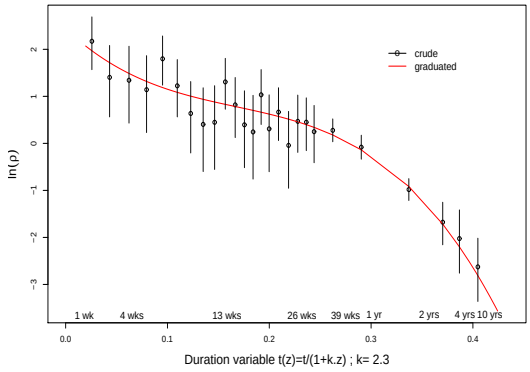
	Model I	Model II	Model III
k			2.3
b_0			50.6322
			(8.4196)
b_1			-162.0055
			(22.1178)
b_2			48.1491
			(8.2888)
b_3			-46.6650
			(6.3935)
α_{sex}	0.2350	0.2356	0.2402
	(0.0618)	(0.0615)	(0.0615)
α_{rated}	-0.2631	-0.2661	-0.2647
	(0.0644)	(0.0646)	(0.06458)
α_{age}	-1.3901	-0.9967	-0.9899
	(0.2201)	(0.0925)	(0.0926)
α_{year}	-0.2352		
	(0.0557)		
τ_{year}		319.5	319.5
ζ_{year_1}		-10.0621	-10.2329
		(1.5267)	(1.4919)
ζ_{year_2}		43.7485	44.0975
		(7.3581)	(7.1534)
α_{dp4}	-0.5241	-0.2498	-0.2776
	(0.0827)	(0.0914)	(0.0879)
τ_{dp4}		58.5	58.5
γ_{dp4}		-0.0856	-0.0746
		(0.0116)	(0.0101)
α_{dp13}	-0.6230	-0.2590	-0.3029
	(0.0860)	(0.1031)	(0.1004)
τ_{dp13}		185.5	185.5
γ_{dp13}		-0.0073	-0.0068
		(0.0022)	(0.0019)
α_{dp26}	-0.8360	-0.4804	-0.5885
	(0.1021)	(0.1137)	(0.1080)
α_{dp52}	-0.7502	-0.4794	-0.4977
	(0.1733)	(0.1789)	(0.1760)
$\alpha_{\text{sex:age}}$	0.3870	0.3340	0.3381
	(0.1676)	(0.1677)	(0.1677)
$\alpha_{\text{dp4:age}}$	0.4202	0.4465	0.4438
	(0.1449)	(0.1451)	(0.1452)



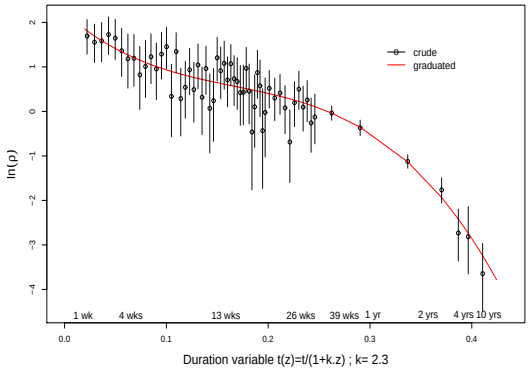
3.4.1: All



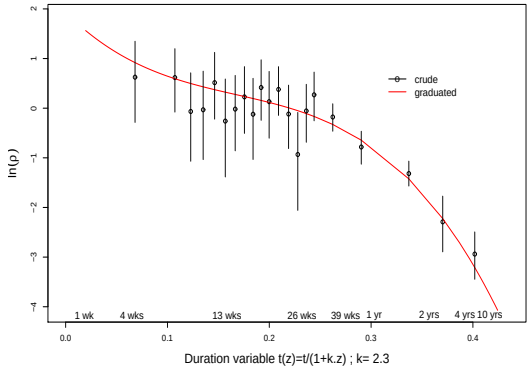
3.4.2: Male



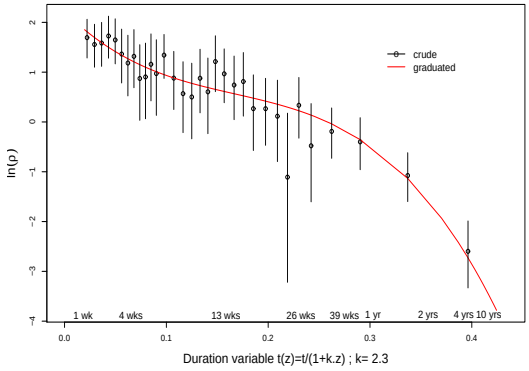
3.4.3: Female



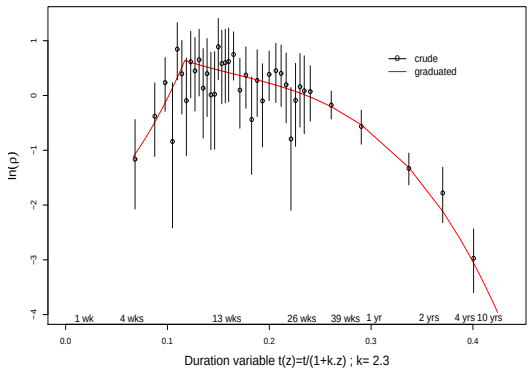
3.4.4: Non-rated



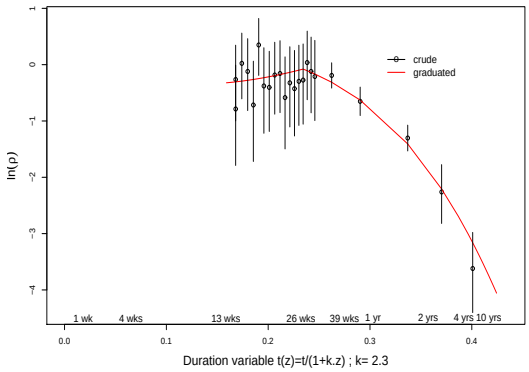
3.4.5: Rated



3.4.6: DP1



3.4.7: DP4



3.4.8: DP13

Figure 3.4: The "residual effects" and "partial residual effects" plots for the discrete covariates – cs20.

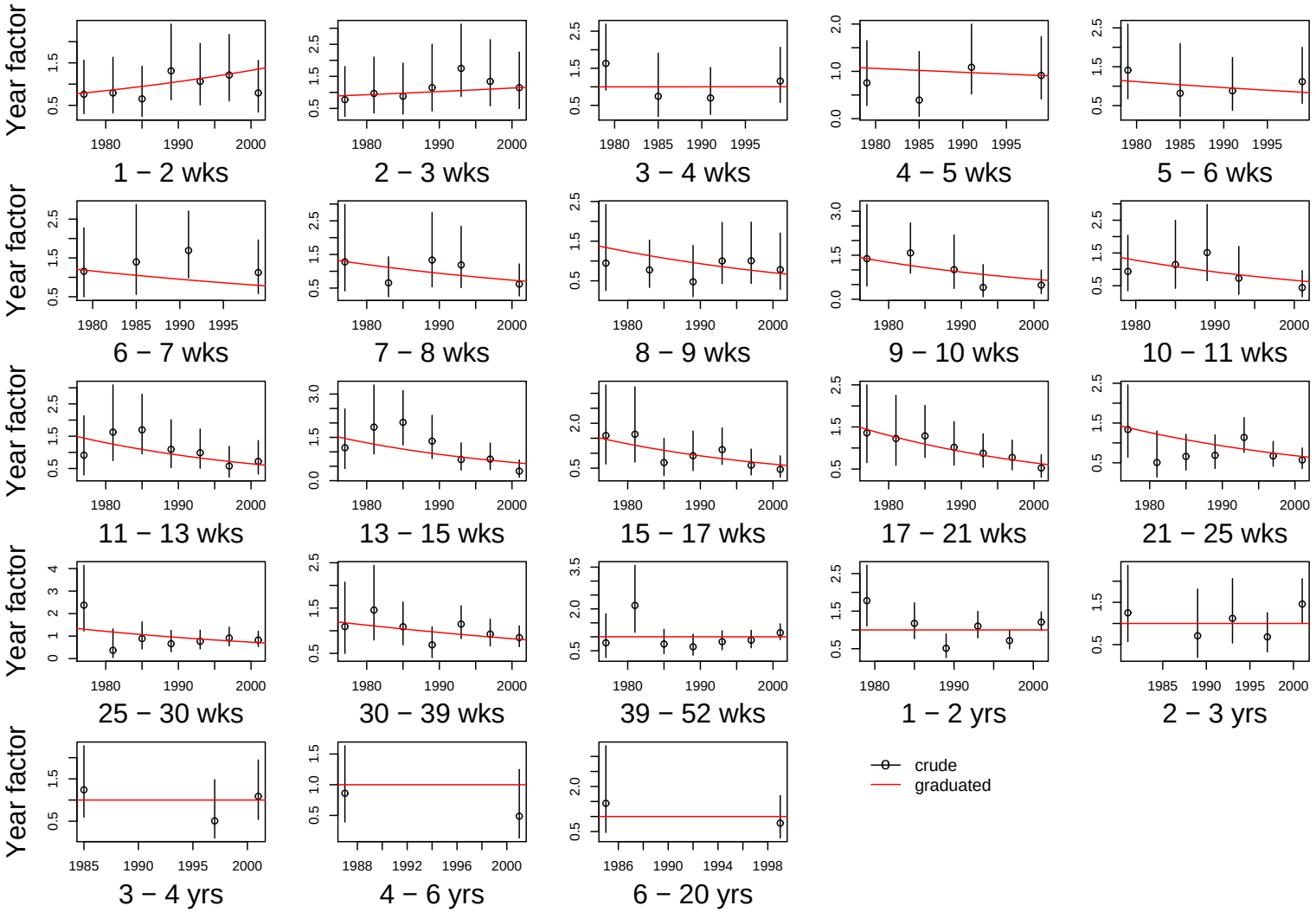


Figure 3.5: The "partial residual effects" plots for year – cs20.

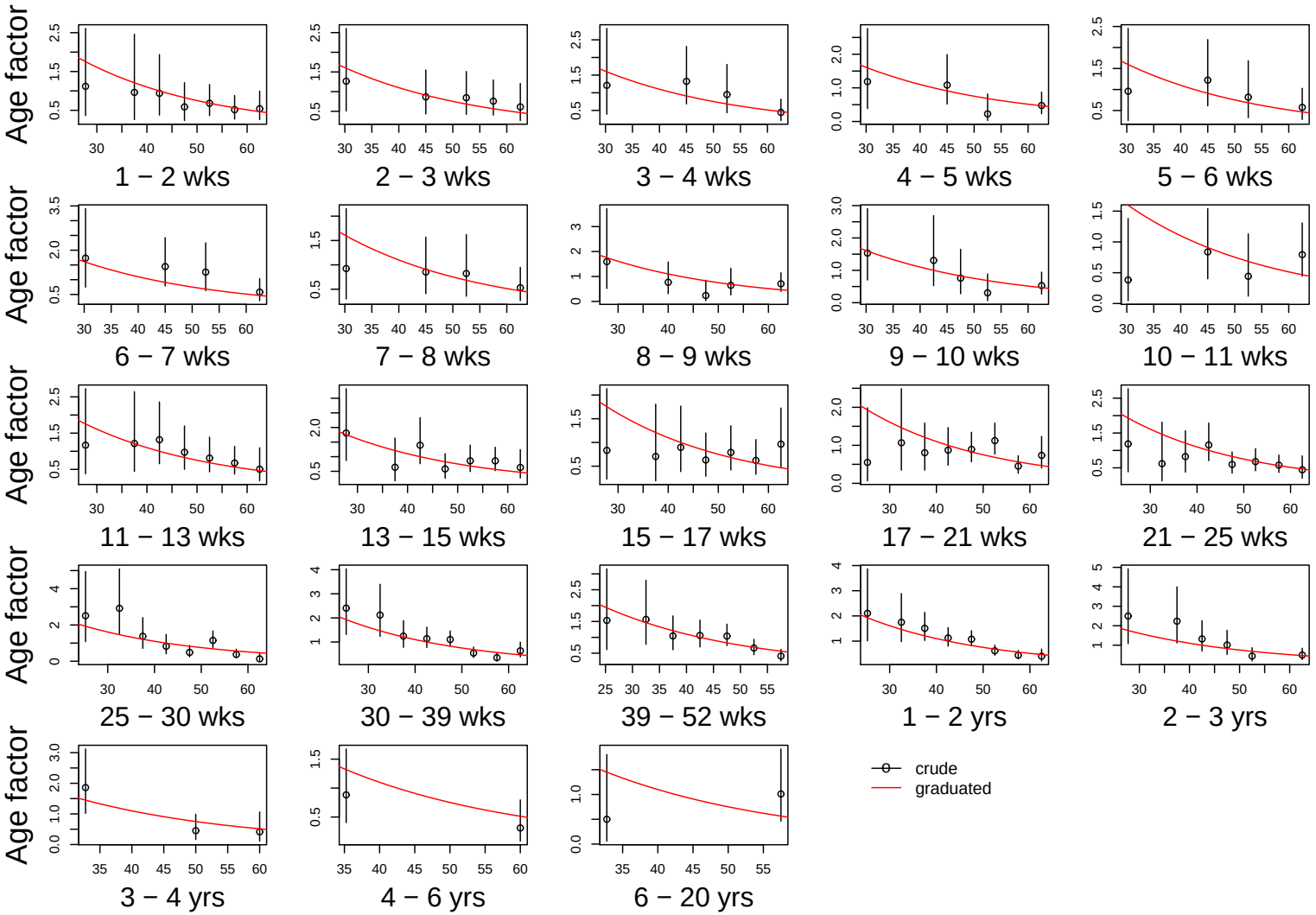


Figure 3.6: The "partial residual effects" plots for age – cs20.

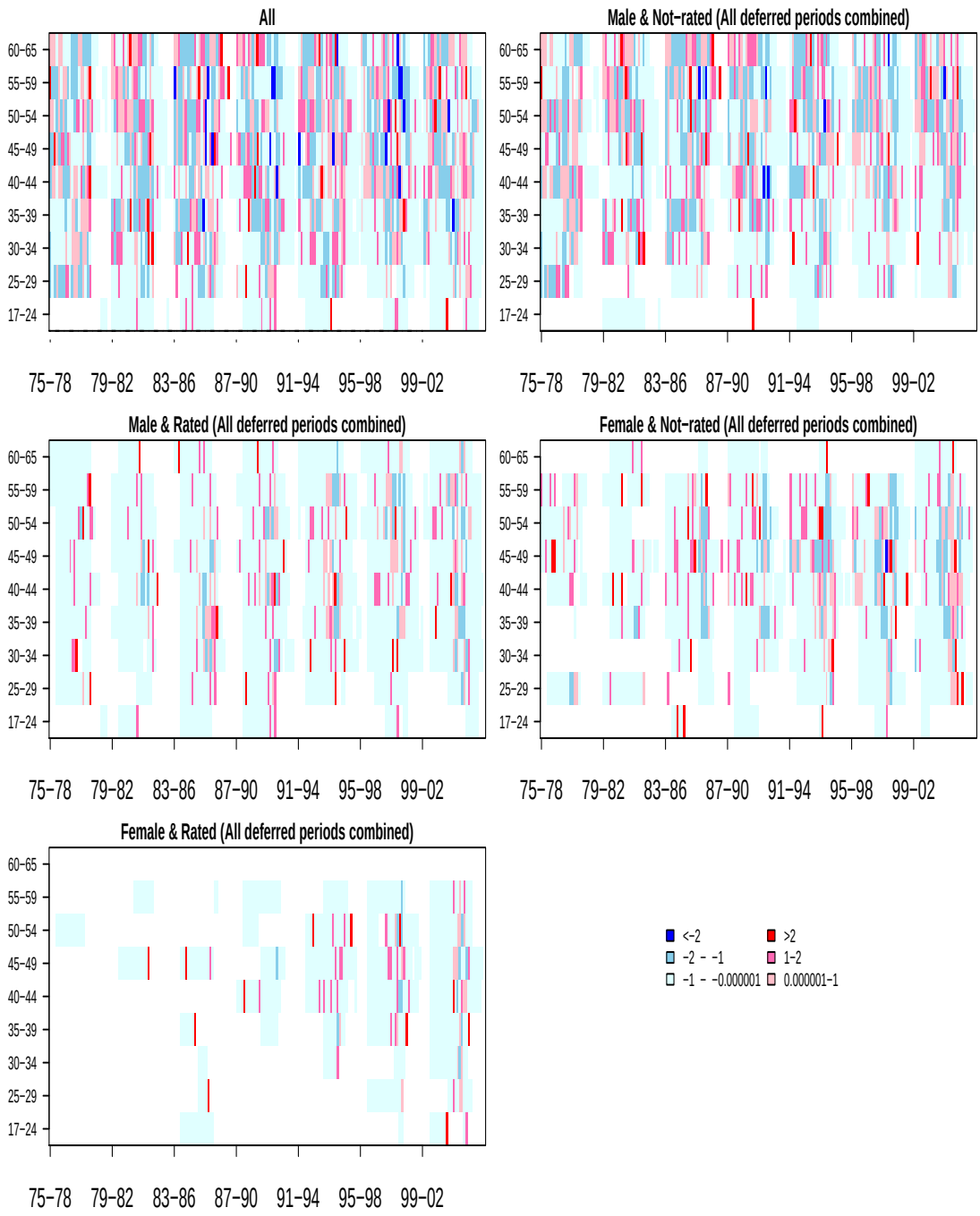


Figure 3.7: Two-dimensional plots of deviance residuals using cs20 data.

Recovery intensity model for cs21

For cs21, Table 3.5 shows the estimated regression coefficients from intermediate models (i.e. Model I and II) leading up to the fully-specified recovery intensity model (i.e. Model III), in columns two to four, respectively. Model I is a Cox model which assumes that the covariate effects are duration-fixed. This assumption is not valid for x_{year}, x_{dp4} and x_{dp13} because the smoothed plot of their respective scaled Schoenfeld residuals in Figure 3.8 suggests that their respective coefficients can be parameterised as

$$\begin{aligned}
\beta_{\text{sex}}(z) &= \alpha_{\text{sex}} + \gamma_{\text{sex}}(\tau_{\text{sex}} - z)_+ \\
\beta_{\text{year}}(z) &= \alpha_{\text{year}} + \zeta_{\text{year}}(C_1(t_{2.3}(\tau_{\text{year}})) - C_1(t_{2.3}(z)))_+ \\
\beta_{\text{dp4}}(z) &= \alpha_{\text{dp4}} + \gamma_{\text{dp4}}(\tau_{\text{dp4}} - z)_+ \\
\beta_{\text{dp13}}(z) &= \alpha_{\text{dp13}} + \gamma_{\text{dp13}}(\tau_{\text{dp13}} - z)_+
\end{aligned}$$

where $\tau_{\text{sex}} = 74.5$, $\tau_{\text{year}} = 126.5$, $\tau_{\text{dp4}} = 52.5$ and $\tau_{\text{dp13}} = 165.5$.

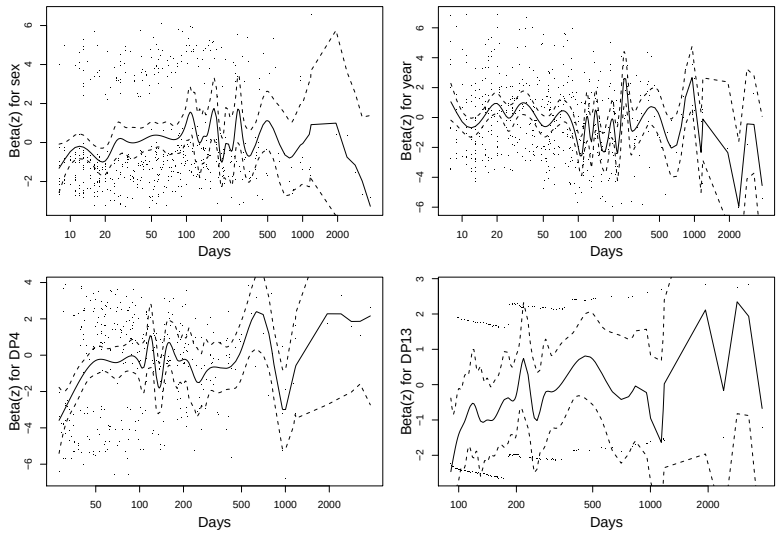


Figure 3.8: The smoothed plots of the scaled Schoenfeld residuals for $x_{\text{sex}}, x_{\text{year}}, x_{\text{dp4}}$ and x_{dp13} for cs21.

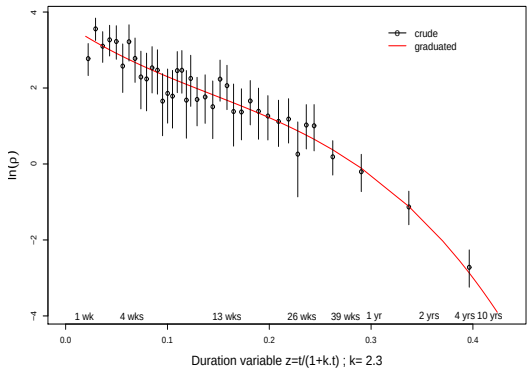
Model II is then obtained by including in Model I the additional covariates created to describe the duration-varying effects of $x_{\text{sex}}, x_{\text{year}}, x_{\text{dp4}}$ and x_{dp13}. Lastly, Model III is the fully parameterised recovery intensity model with the baseline intensity

modelled by $\exp(\sum_{i=0}^{3} b_i C_i(t_{2.3}(z)))$. This final fitted model is then assessed for its goodness-of-fit. The value of the χ^2 statistic is 41.30. With 48 cells and 15 parameters fitted in the model, the probability value is 0.152 on 33 degrees of freedom, indicating a reasonably good fit to the data. The "residual effects" and "partial residual effects" plots for the discrete covariates are presented in Figure 3.9. The data upon which each plot is based is indicated by the caption of the plot. Figures 3.10 and 3.11 show the "partial residual effects" plots for year and age, respectively, for each sickness duration band. We only show the "partial residual effects" plot for age for the first few sickness durations bands because of the small amount of data at longer sickness durations bands. The red curves overlaid on the plots in Figures 3.9–3.11 are the estimated smooth intensities according to Model III and they have represented the actual experience reasonably well.

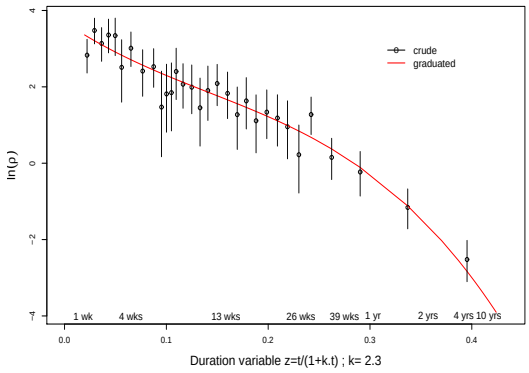
The two-dimensional plots of the deviance residuals for all data as well as its subsets are shown in Figure 3.12. For 'All' and 'Male & Not-rated' data, the cells belonging to longer sickness durations bands are mostly in blue or white because most of the recoveries happen at shorter sickness duration, leaving very few recoveries or exposed-to-risk at longer sickness durations bands. For other subsets, there are a greater number of white and light blue cells because most of the cells have either zero exposed-to-risk or no recoveries.

Table 3.5: Parameters in the intermediate and final recovery intensity models for cs21.

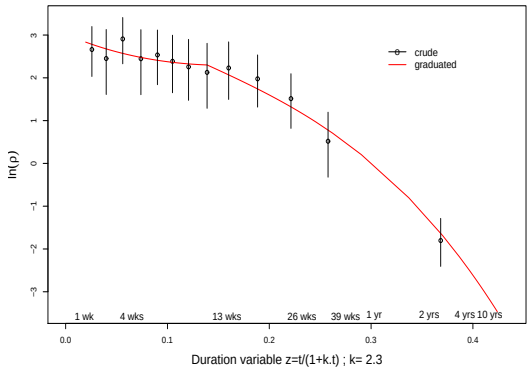
	Model I	Model II	Model III
k			2.3
b_0			16.9710
			(13.4547)
b_1			-75.3337
			(37.2265)
b_2			13.3819
			(13.3147)
b_3			-20.4126
			(11.0738)
α_{sex}	0.0507	0.3743	0.3502
	(0.104)	(0.1387)	(0.1381)
τ_{sex}		74.5	74.5
γ_{sex}		-0.0132	-0.0123
		(0.0039)	(0.0039)
α_{age}	-0.6901	-0.7235	-0.7273
	(0.131)	(0.1318)	(0.1319)
α_{year}	-0.1860	-0.5655	-0.5433
	(0.106)	(0.1616)	(0.1611)
τ_{year}		126.5	126.5
ζ_{year}		5.0649	4.8752
		(1.592)	(1.5980)
α_{dp4}	-0.6424	-0.3106	-0.3993
	(0.135)	(0.1591)	(0.1459)
τ_{dp4}		52.5	52.5
γ_{dp4}		-0.0848	-0.0805
		(0.0180)	(0.0158)
α_{dp13}	-1.0603	-0.3705	-0.4018
	(0.186)	(0.2376)	0.2276
τ_{dp13}		165.5	165.5
γ_{dp13}		-0.0185	-0.0198
		(0.0068)	(0.0057)
α_{dp26}	-0.9282	-0.4638	-0.4635
	(0.267)	(0.2891)	(0.2800)
$\alpha_{\text{sex:age}}$	0.7290	0.7457	0.7415
	(0.288)	(0.2898)	(0.2895)



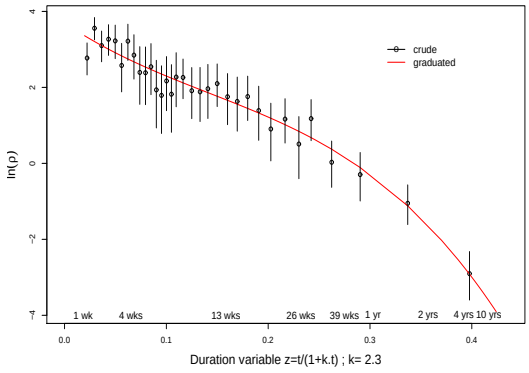
3.9.1: All



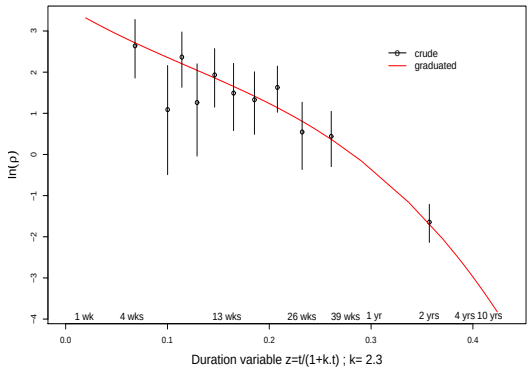
3.9.2: Male



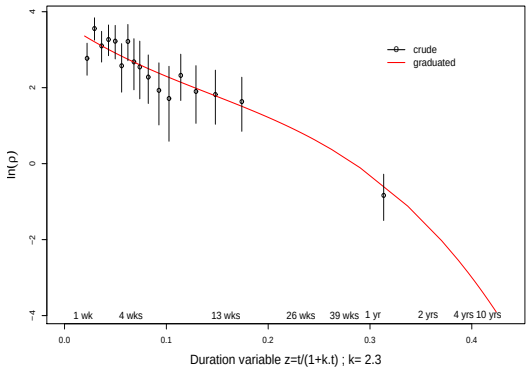
3.9.3: Female



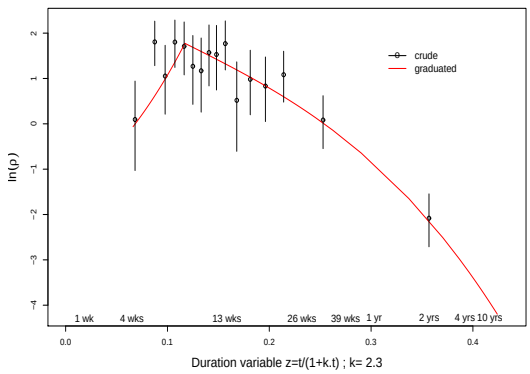
3.9.4: Non-rated



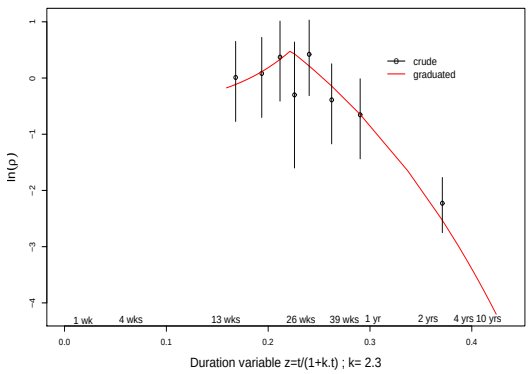
3.9.5: Rated



3.9.6: DP1



3.9.7: DP4



3.9.8: DP13

Figure 3.9: The "residual effects" and "partial residual effects" plots for the discrete covariates – cs21.

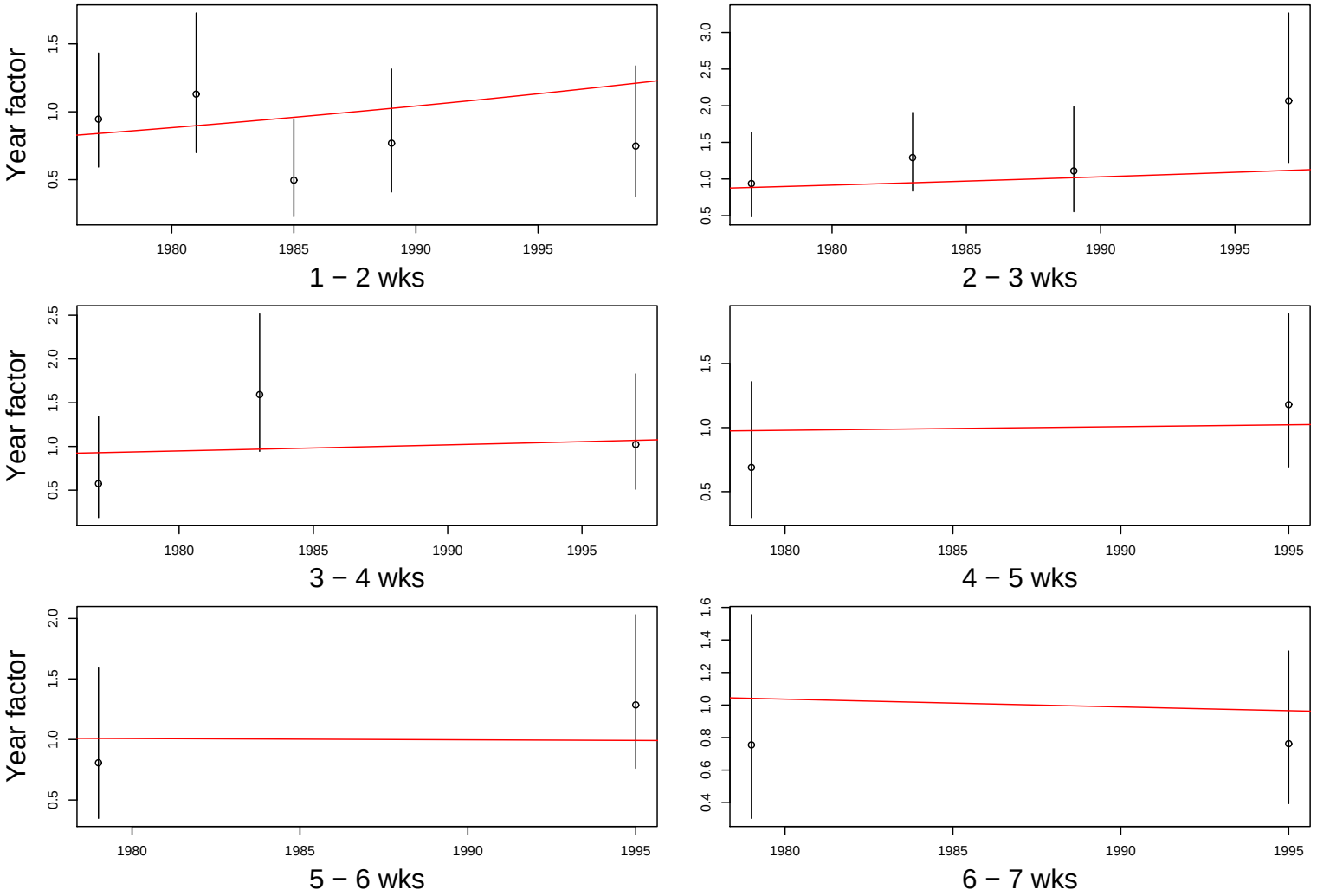


Figure 3.10: The "partial residual effects" plots for year – cs21.

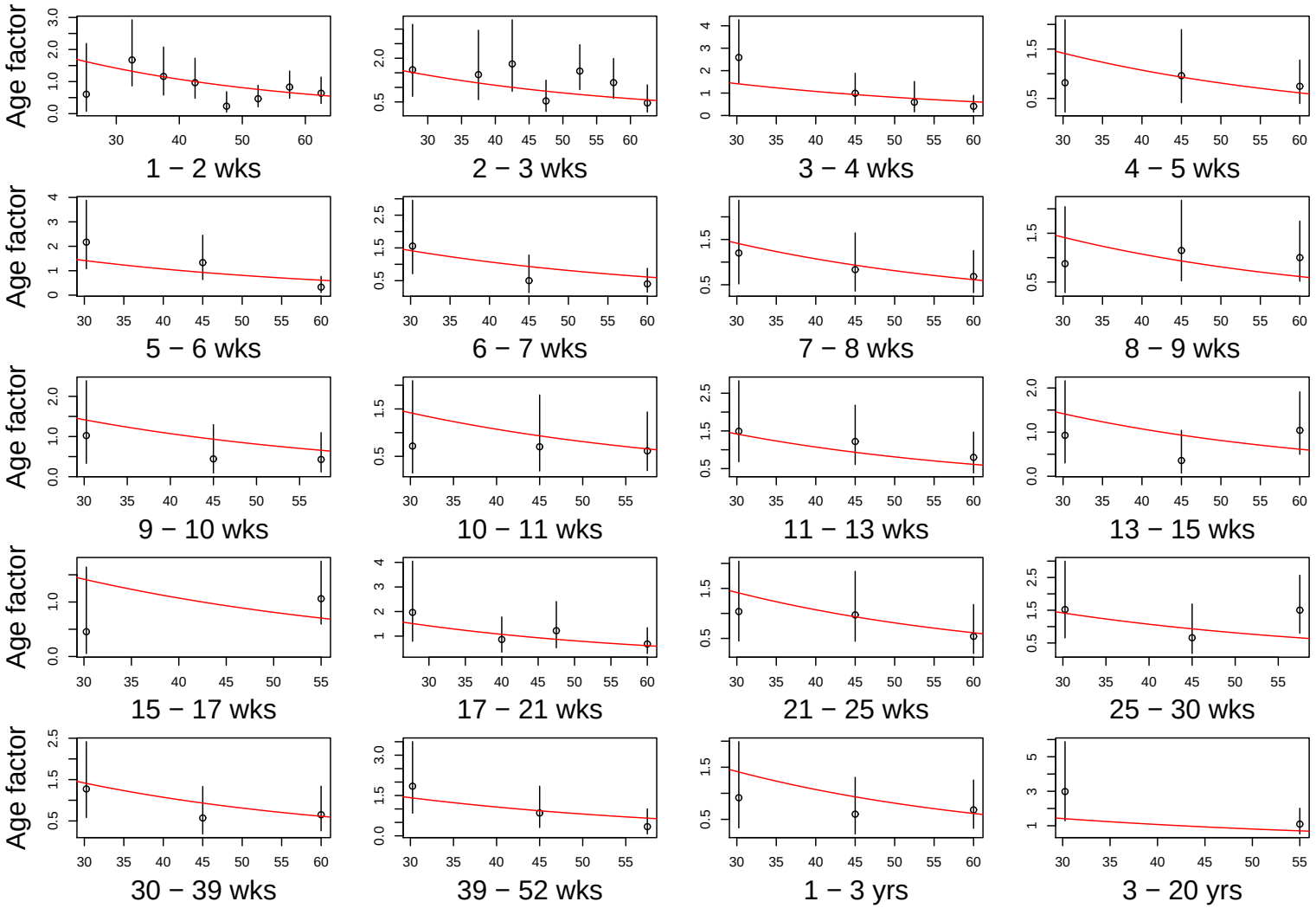


Figure 3.11: The "partial residual effects" plots for age – cs21.

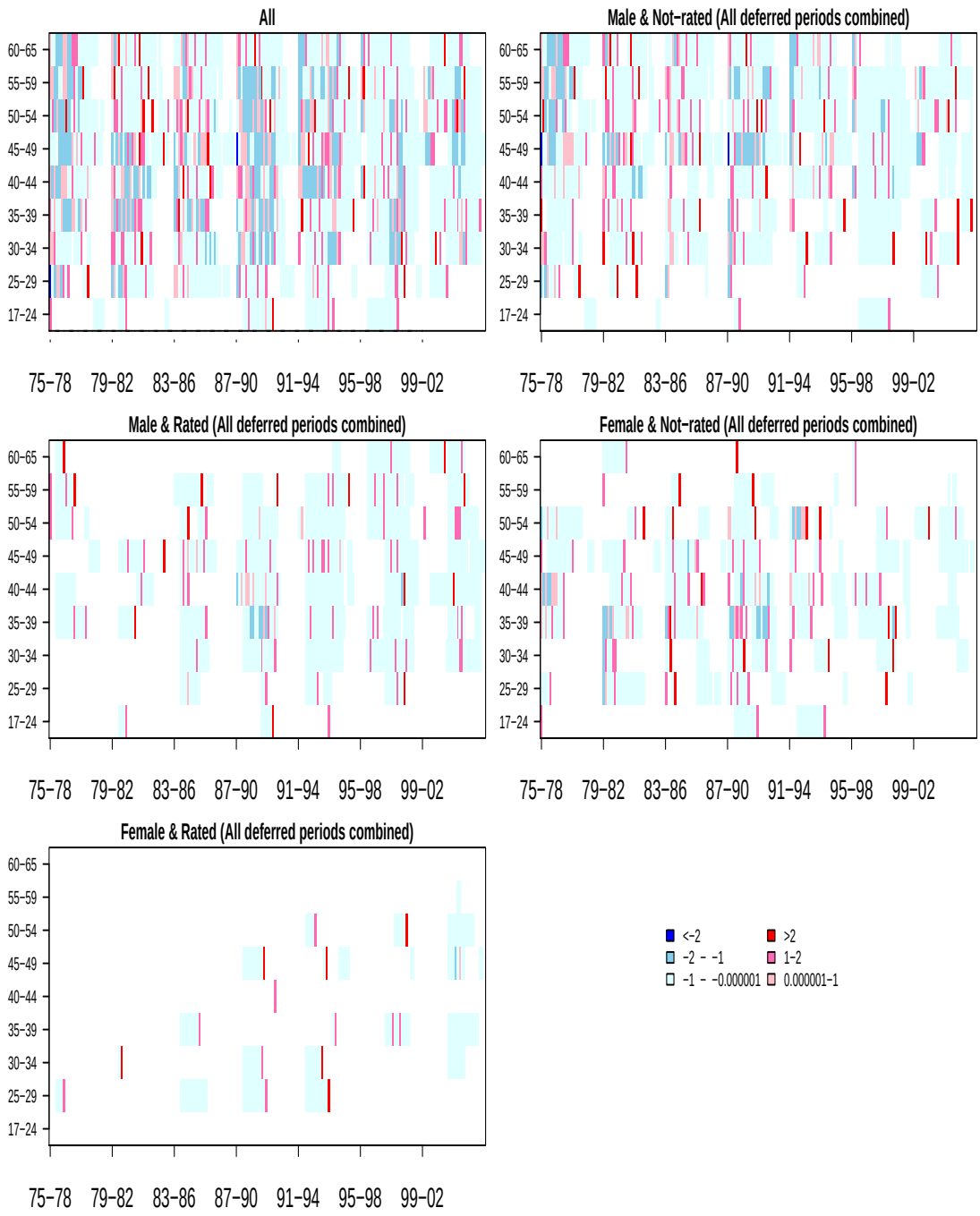


Figure 3.12: Two-dimensional plots of deviance residuals using cs21 data.

3.6.2 G4 Mental Illness

There is only one cause of sickness in G4 Mental Illness. Its ICD8 code, exposed to risk in units of days and the number of recoveries are presented in Table 3.6.

Table 3.6: The cause of sickness in sickness category G4 Mental Illness.

ICD8	Cause of sickness	Exposed to risk (days)	Recoveries
27	Psychoses and non-psychotic mental disorders	8,845,266	5,287

Recovery intensity model for cs27

For cs27, Table 3.7 shows, in columns two to four, the estimated regression coefficient from Models I, II and III, respectively. Model I assumes that all the covariate effects are duration-fixed, an assumption which is not valid for $x_{\text{year}}, x_{\text{age}}, x_{\text{dp4}}, x_{\text{dp13}}$ and x_{dp26} because the smoothed plots of their respective scaled Schoenfeld residuals shown in Figure 3.13 suggest that their respective coefficients can be parameterised as

$$\beta_{\text{year}}(z) = \alpha_{\text{year}} + \gamma_{\text{year}}(\tau_{\text{year}} - z)_+$$

$$\beta_{\text{age}}(z) = \alpha_{\text{age}} + \zeta_{\text{age}_1}(C_1(t_{2.3}(\tau_{\text{age}})) - C_1(t_{2.3}(z)))_+ + \zeta_{\text{age}_2}((C_1(t_{2.3}(\tau_{\text{age}})) - C_1(t_{2.3}(z)))_+)^2$$

$$\beta_{\text{dp4}}(z) = \alpha_{\text{dp4}} + \gamma_{\text{dp4}_1}(\tau_{\text{dp4}_1} - z)_+ + \gamma_{\text{dp4}_2}(\tau_{\text{dp4}_2} - z)_+$$

$$\beta_{\text{dp13}}(z) = \alpha_{\text{dp13}} + \gamma_{\text{dp13}}(\tau_{\text{dp13}} - z)_+$$

$$\beta_{\text{dp26}}(z) = \alpha_{\text{dp26}} + \gamma_{\text{dp26}}(\tau_{\text{dp26}} - z)_+$$

where $\tau_{\text{year}} = 26.5, \tau_{\text{age}} = 106.5, \tau_{\text{dp4}_1} = 45.5, \tau_{\text{dp4}_2} = 94.5, \tau_{\text{dp13}} = 198.5$ and $\tau_{\text{dp26}} = 250.5$.

Model II is Model I but with the inclusion of additional covariates created to describe the duration-varying effects of $x_{\text{year}}, x_{\text{age}}, x_{\text{dp4}}, x_{\text{dp13}}$ and x_{dp26}. Lastly, Model III is the fully parameterised recovery intensity model with the baseline intensity modelled by $\exp(\sum_{i=0}^{4} b_i C_i(t_{2.3}(z)))$ formula. This final fitted model is then assessed for its goodness-of-fit. The value of the χ^2 statistic is 383.3666. With 451 cells and

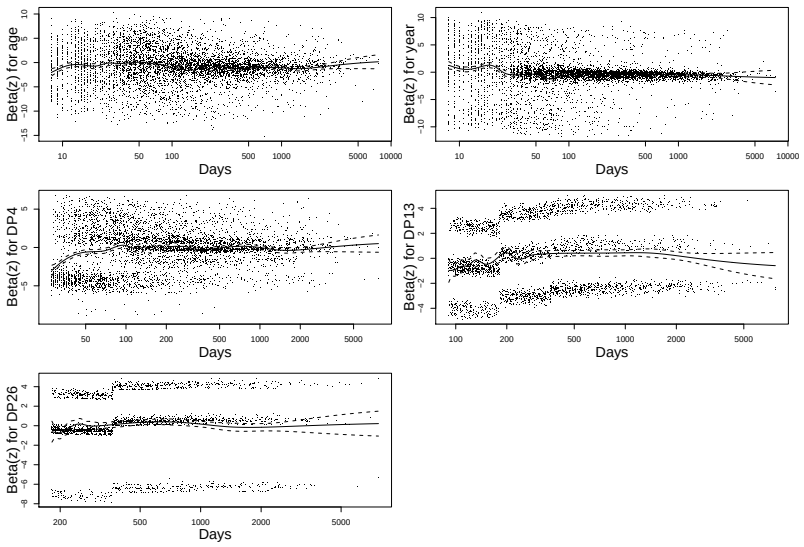


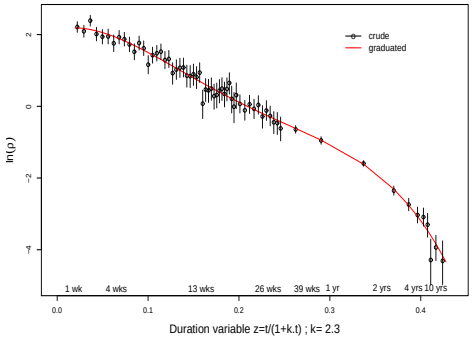
Figure 3.13: The smoothed plots of the scaled Schoenfeld residuals for $x_{\text{year}}, x_{\text{age}}, x_{\text{dp4}}, x_{\text{dp13}}$ and x_{dp26} in cs27.

26 parameters fitted in the model, the probability value is 0.927088 on 425 degrees of freedom, indicating a good fit to the data. The "residual effects" and "partial residual effects" plots for the discrete covariates are presented in Figure 3.14. The data upon which each plot is based is indicated by the caption of the plot. Figures 3.15 and 3.16 show the "partial residual effects" plots for year and age, respectively, for each sickness duration band. The red curves overlaid on the plots in Figures 3.14–3.16 are the estimated smooth intensities according to Model III, all of which have fitted the actual experience well. The two-dimensional plots of the deviance residuals for all data as well as its subsets are shown in Figure 3.17. Apart from 'Female & rated' for which the data is sparse, the positive and negative deviance residuals of different ranges of values are roughly randomly scattered in other subsets of the data.

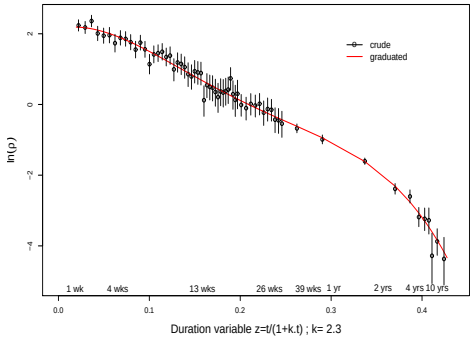
Table 3.7: Parameters in the intermediate and final recovery intensity models for cs27.

	Model I	Model II	Model III
k			2.3
b_0			-418.8920
			(61.9760)
b_1			521.1417
			(84.3994)
b_2			-531.0750
			(77.3801)
b_3			171.7831
			(27.0574)
b_4			-110.0431
			(15.5167)
α_{sex}	0.0338	0.0232	0.0228
	(0.0504)	(0.0505)	(0.0505)
α_{age}	-0.8362	-1.2531	-1.2484
	(0.0466)	(0.06450)	(0.0643)
τ_{age}		106.5	106.5
ζ_{age_1}		25.9934	26.3914
		(2.8529)	(2.8394)
ζ_{age_2}		-158.7982	-162.5144
		(21.3096)	(21.2099)
α_{age2}	-0.2495	-0.2733	-0.2703
	(0.0496)	(0.04990)	(0.0499)
α_{rated}	0.0526	0.0365	0.0375
	(0.0440)	(0.0434)	(0.0434)
α_{year}	-0.4182	-0.5215	-0.5211
	(0.0337)	(0.0362)	(0.0361)
τ_{year}		26.5	26.5
γ_{year}		0.0463	0.0460
		(0.0064)	(0.0064)
α_{year2}	0.2786	0.1219	0.1221
	(0.0876)	(0.0332)	(0.0332)

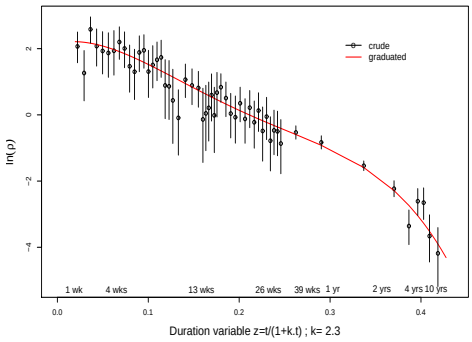
	Model I	Model II	Model III
α_{dp4}	-0.1241	0.1767	0.1729
	(0.0979)	(0.0479)	(0.0451)
τ_{dp4_1}		45.5	45.5
γ_{dp4_1}		-0.0251	-0.0379
		(0.0119)	(0.0103)
τ_{dp4_2}		94.5	94.5
γ_{dp4_2}		-0.0181	-0.0159
		(0.0021)	(0.0018)
α_{dp13}	-0.5820		
	(0.0511)		
τ_{dp13}		198.5	198.5
γ_{dp13}		-0.0090	-0.0094
		(0.0012)	(0.0011)
α_{dp26}	-0.7271	-0.2384	-0.2385
	(0.0601)	(0.0563)	(0.0555)
τ_{dp26}		250.5	250.5
γ_{dp26}		-0.0087	-0.0105
		(0.0036)	(0.0034)
α_{dp52}	-1.0602	-0.6071	-0.5993
	(0.0992)	(0.0951)	(0.0943)
$\alpha_{\text{age:dp26}}$	-0.8273	-0.4547	-0.4554
	(0.1368)	(0.1431)	(0.1430)
$\alpha_{\text{age:dp52}}$	-1.1170	-0.7453	-0.7442
	(0.2866)	(0.2900)	(0.2899)
$\alpha_{\text{age:rated}}$	0.1534	0.2516	0.2489
	(0.1063)	(0.1085)	(0.1085)
$\alpha_{\text{year:rated}}$	-0.1820	-0.1440	-0.1474
	(0.0731)	(0.0738)	(0.0738)
$\alpha_{\text{sex:year2}}$	-0.1623	-0.1647	-0.1631
	(0.0684)	(0.0684)	(0.0684)
$\alpha_{\text{sex:dp4}}$	-0.1893	-0.1845	-0.1820
	(0.0763)	(0.0763)	(0.0763)



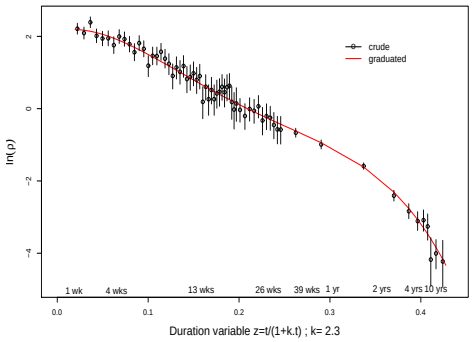
3.14.1: All



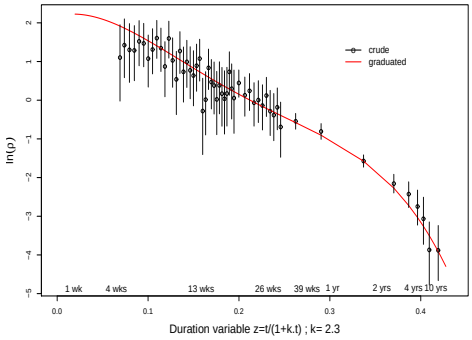
3.14.2: Male



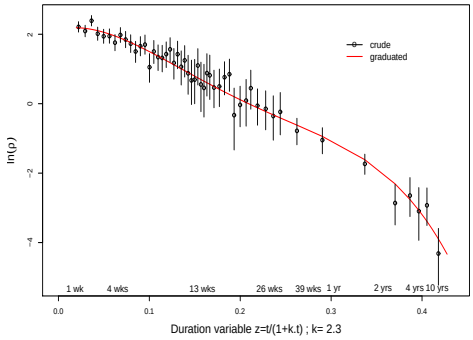
3.14.3: Female



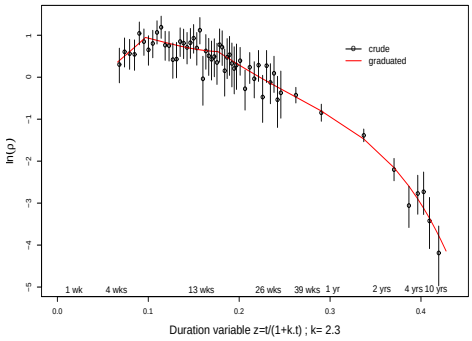
3.14.4: Non-rated



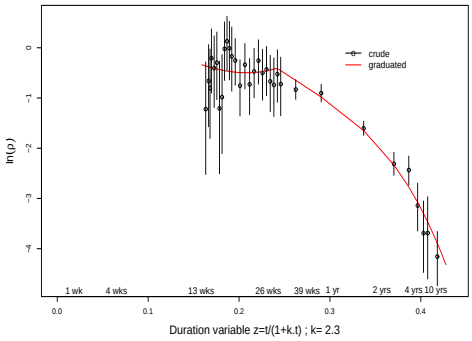
3.14.5: Rated



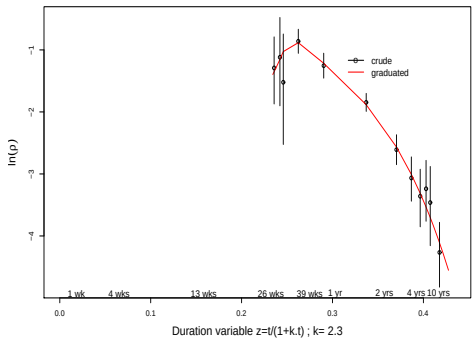
3.14.6: DP1



3.14.7: DP4



3.14.8: DP13



3.14.9: DP26

Figure 3.14: The "residual effects" and "partial residual effects" plots for the discrete covariates – cs27.

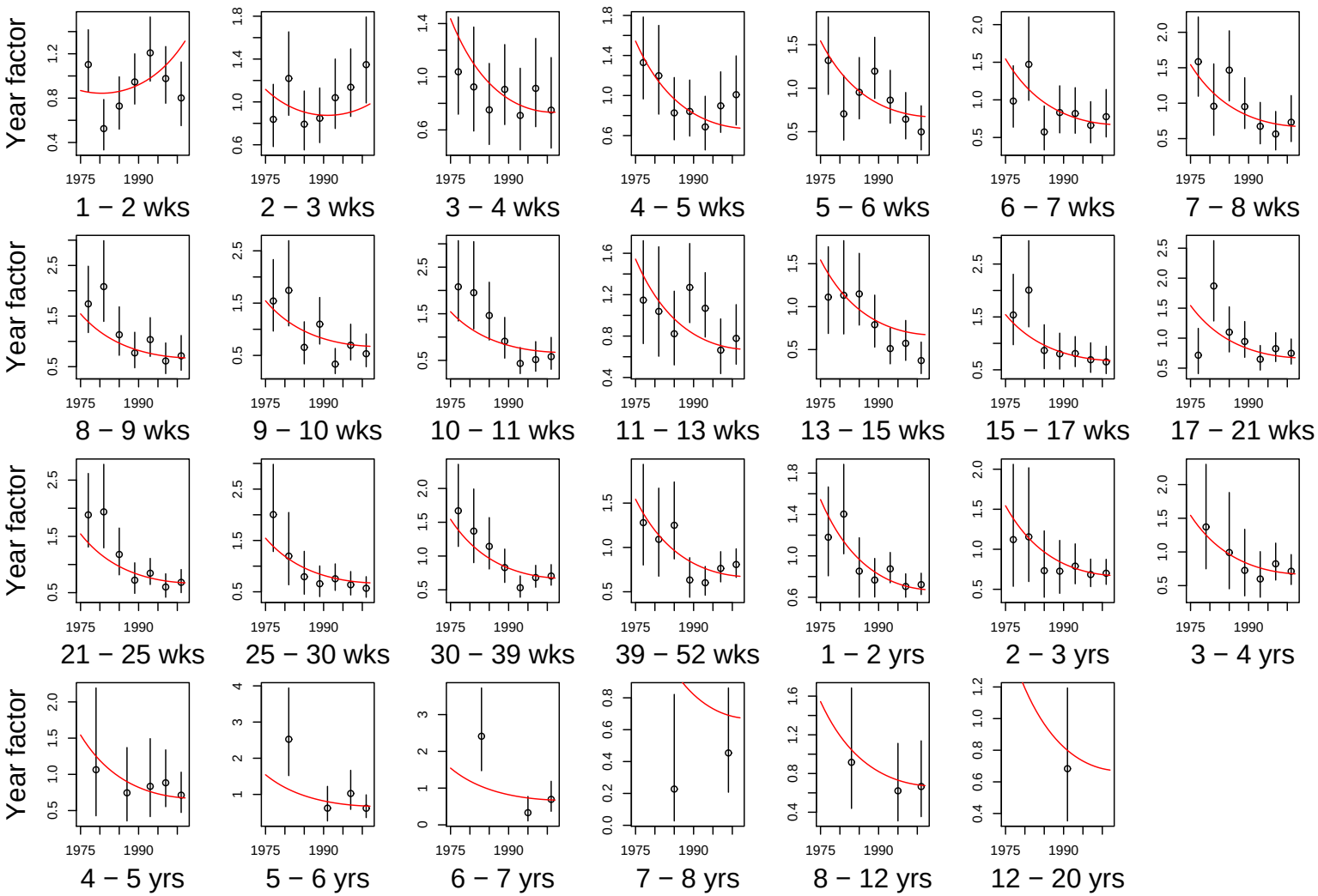


Figure 3.15: The "partial residual effects" plots for year – cs27.

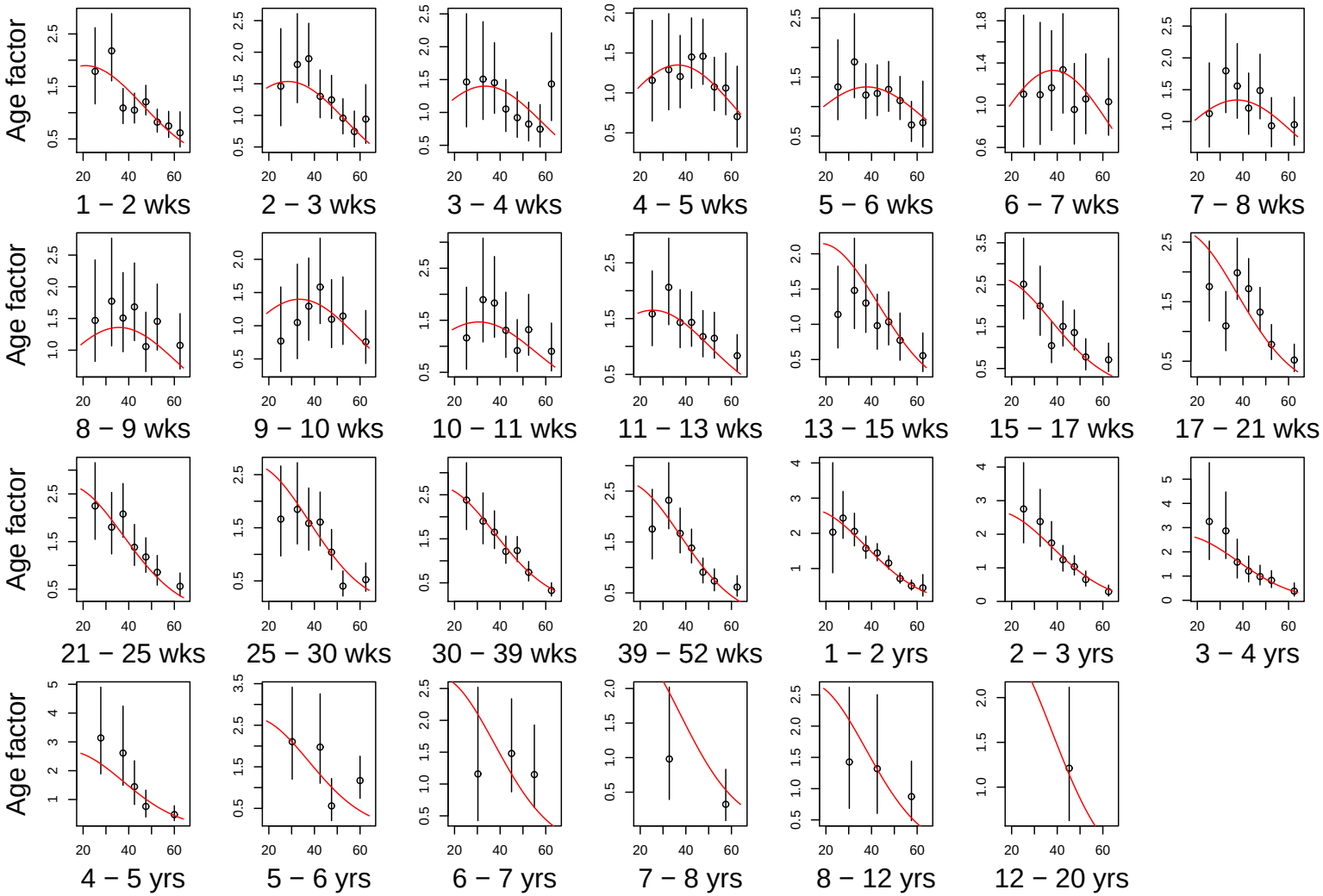


Figure 3.16: The "partial residual effects" plots for age – cs27.

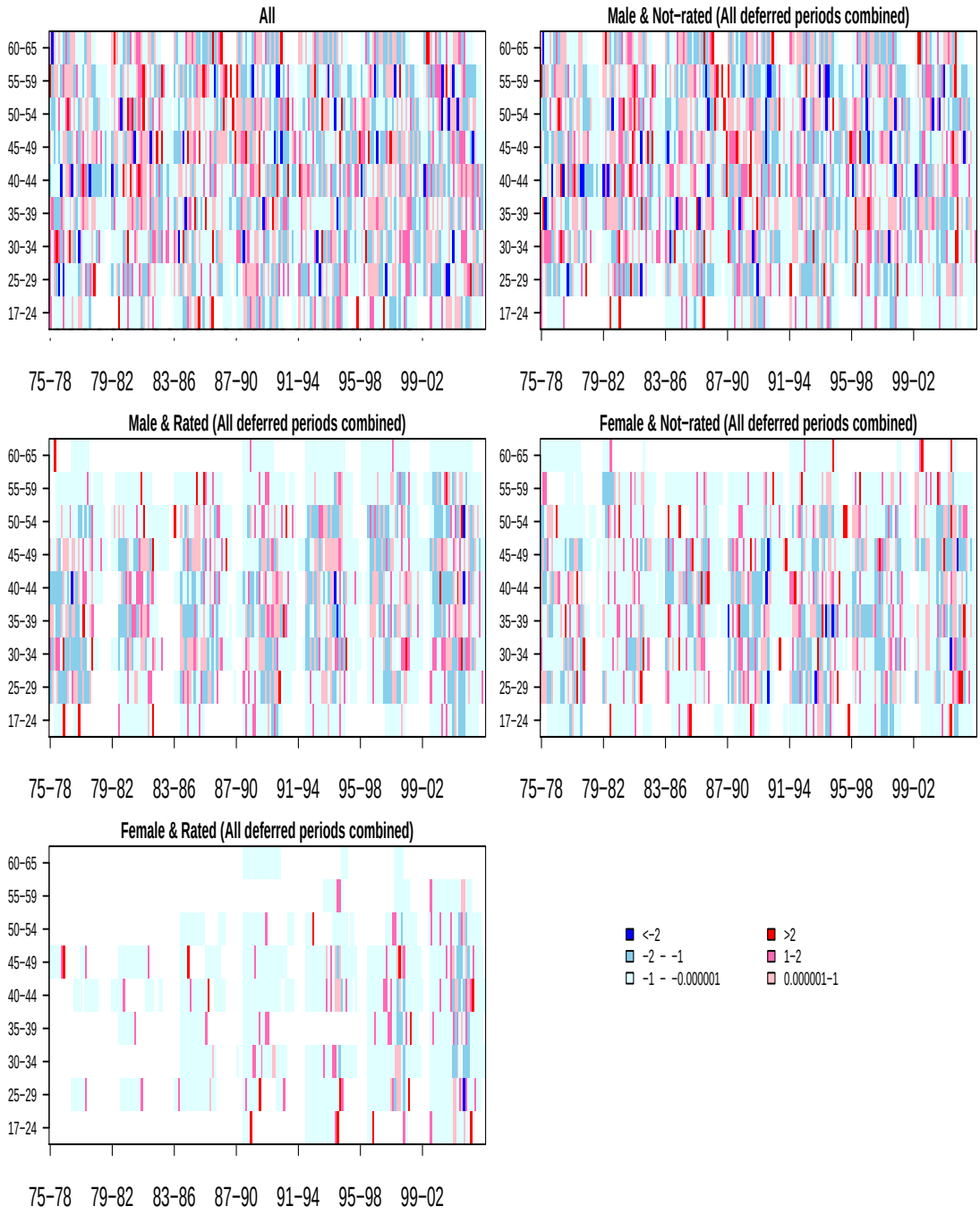


Figure 3.17: Two-dimensional plots of deviance residuals for cs27 data.

3.6.3 G6 Circulatory

The causes of sickness in G6 Circulatory alongside their ICD8 code, exposed to risk in units of days and the number of recoveries are presented in Table 3.8.

Table 3.8: The causes of sickness in sickness category G6 Circulatory.

ICD8	Cause of sickness	Exposed to risk (days)	Recoveries
32	Active rheumatic fever	7,172	11
33	Chronic rheumatic heart disease	49,898	23
34	Hypertensive disease	657,927	566
35	Ischaemic heart disease	4,701,560	3,426
36	Cerebrovasular disease	1,503,893	389
37	Venous thrombosis and embolism	221,987	288
38	Other diseases of circulatory system	824,861	1,470

The reference cause of sickness is cs38. The binary indicators representing the remaining causes of sickness in this sickness category are estimated using the Cox model. The proportionality of the recovery patterns for these causes of sickness are assessed using the smoothed plots of the scaled Schoenfeld residuals which are presented in Figure 3.18. The broken red lines overlaid on this plots are the Cox's estimate. These plots suggest that the log hazard ratio between cs38 and each of cs32, cs33, cs34, cs37 and cs38 stay reasonably constant at all sickness durations. On the other hand, the log hazard ratio between cs35 and cs36 versus cs38 deviates from a horizontal line, suggesting that their recovery patterns are not proportional to cs38. Therefore, we will use a proportional hazard model incorporating cause of sickness as a factor to describe the recovery intensities for cs32, cs33, cs34, cs37 and cs38. The recovery intensity for cs35 and cs36 will be estimated separately.

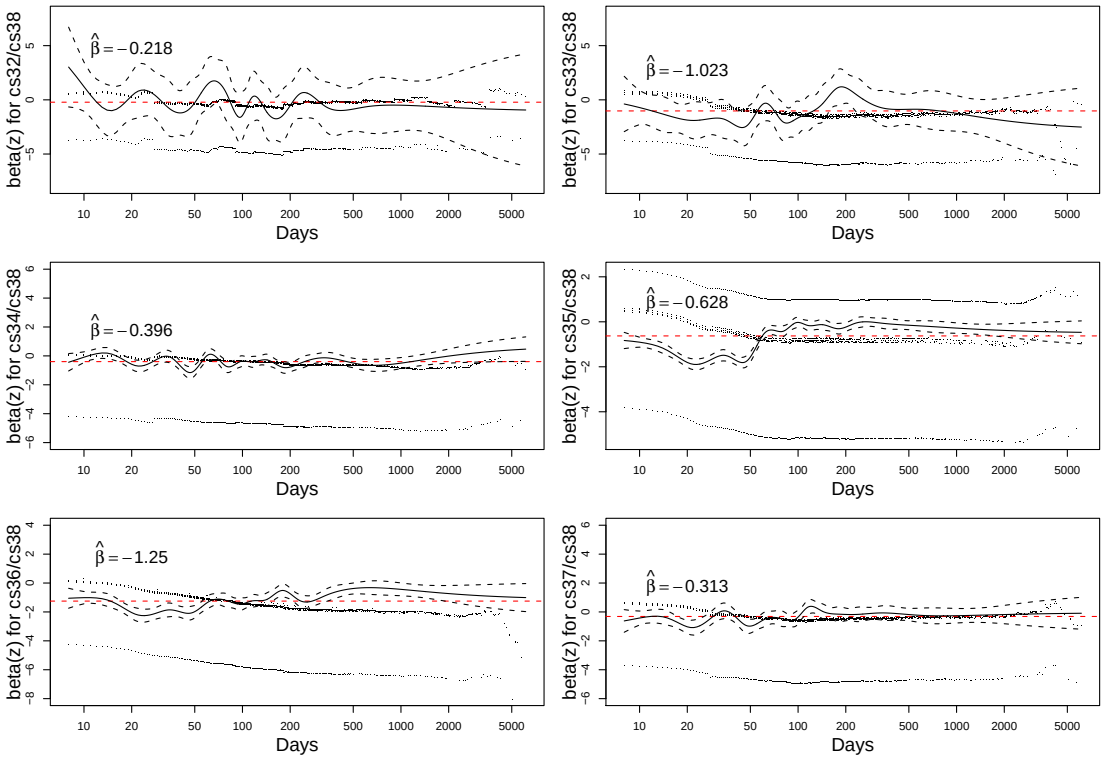


Figure 3.18: The smoothed plot of the Schoenfeld residuals for the causes of sickness in G6.

Recovery intensity model for cs32, cs33, cs34, cs37 and cs38

The recovery intensities for cs32, cs33, cs34, cs37 and cs38 are of different levels but are proportional to each other. We therefore use a proportional hazards model to describe their recovery intensities in which cs38 is the reference cause of sickness while the remaining four causes of sickness are represented by their binary indicators in the model. Table 3.9 shows, in columns two to four, the estimated parameters from Models I, II and III, respectively. Model I assumes that all the covariate effects are duration-fixed, an assumption which is not valid for $x_{\text{year}}, x_{\text{dp4}}$ and x_{dp13} because the smoothed plots of their respective scaled Schoenfeld residuals shown in Figure 3.13 suggests that their respective coefficients can be parameterised as

$$\begin{aligned}
\beta_{\text{year}}(z) &= \alpha_{\text{year}} + \theta_{\text{year}} C_1(t_{6.7}(z)) \\
\beta_{\text{dp4}}(z) &= \alpha_{\text{dp4}} + \gamma_{\text{dp4}_1}(\tau_{\text{dp4}_1} - z)_+ + \gamma_{\text{dp4}_2}(\tau_{\text{dp4}_2} - z)_+ \\
\beta_{\text{dp13}}(z) &= \alpha_{\text{dp13}} + \gamma_{\text{dp13}}(\tau_{\text{dp13}} - z)_+
\end{aligned}$$

where $\tau_{\text{dp4}_1} = 37.5, \tau_{\text{dp4}_2} = 70.5$ and $\tau_{\text{dp13}} = 125.5$.

Model II has all the parameters in Model I as well as additional covariates which are created to describe the duration-varying effects of $x_{\text{year}}, x_{\text{dp4}}$ and x_{dp13}. Lastly, Model III has all the parameters in Model II and a baseline intensity which is described by a $\exp(\sum_{i=0}^{4} b_i C_i(t_{6.7}(z)))$ formula. The final fitted model, Model III, is then assessed for its goodness-of-fit. The value of the χ^2 statistic is 177.6417. With 195 cells and 26 parameters fitted in the model, the probability value is 0.3091 on 169 degrees of freedom, indicating a reasonably good fit to the data.

The "residual effects" and "partial residual effects" plots for the discrete covariates for cs34, cs37 and cs38 are presented in Figures 3.20 – 3.22, respectively. The data upon which each plot is based is indicated by the caption of the plot. These plots are not constructed for cs32 and cs33 because of the small amount of data in these causes of sickness. The "partial residual effects" plots for age and year is only constructed for cs38 due to its reasonably large amount of data and they are presented in Figures 3.23 and 3.24, respectively. The red curves overlaid on the plots in Figures 3.20 – 3.24 are the estimated smooth intensities according to Model III and they fitted

the actual experience reasonably well. The two-dimensional plots of the deviance residuals for cs34, cs37 and cs38 as well as their subsets for are shown in Figures 3.25 – 3.27. In these figures, the positve and negative residuals do not seem to be randomly scattered since cells for longer sickness duration bands are mostly in blue and white. In particular, the cells for younger age range are dominantly white because there is very small exposure to risk for these age ranges.

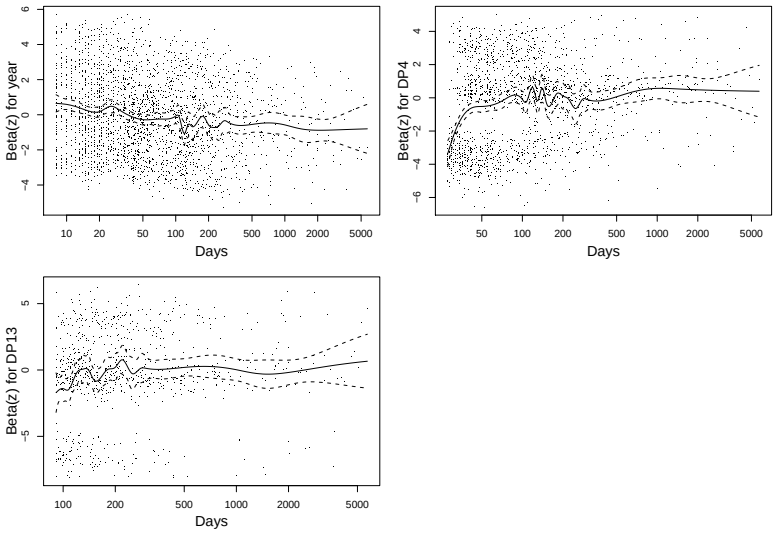
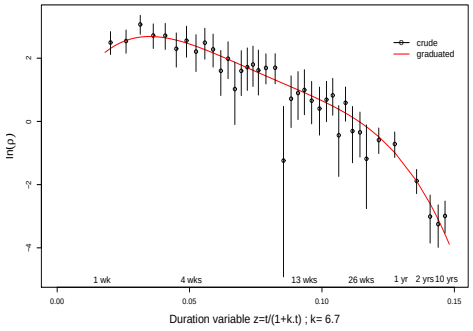


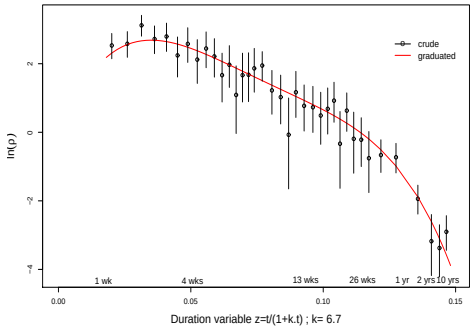
Figure 3.19: The smoothed plots of the scaled Schoenfeld residuals for $x_{\text{year}}, x_{\text{dp4}}$ and x_{dp13} in the recovery intensity model for cs32, cs33, cs34, cs37 and cs38.

Table 3.9: Parameters in the intermediate and final recovery intensity models for cs32, cs33, cs34, cs37 and cs38.

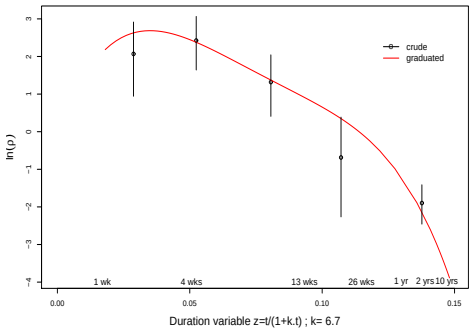
	Model I	Model II	Model III
k			6.7
b_0			-49621.41
			(6521.934)
b_1			29804.68
			(4072.891)
b_2			-65398.31
			(8597.399)
b_3			9869.151
			(1348.994)
b_4			-15776.75
			(2075.904)
I_{cs32}	0.4101	0.5085	0.5272
	(0.3184)	(0.3185)	(0.3184)
I_{cs33}	-0.7315	-0.7352	-0.7275
	(0.2114)	(0.2114)	(0.2114)
I_{cs34}	-0.2115	-0.2070	-0.2093
	(0.0582)	(0.0581)	(0.0581)
I_{cs37}	-0.2560	-0.2791	-0.2787
	(0.0711)	(0.0726)	(0.0725)
α_{age}	-0.6002	-0.6071	-0.6092
	(0.0718)	(0.0720)	(0.0719)
α_{age2}	-0.3072	-0.2817	-0.2881
	(0.0707)	(0.0707)	(0.0707)
α_{rated}	-0.1461	-0.1255	-0.1257
	(0.0603)	(0.0606)	(0.0605)
$\alpha_{\text{rated}} \times I_{\text{cs32}}$	-2.2613	-2.3477	-2.3618
	(1.0507)	(1.0507)	(1.0506)
α_{year}	-0.1054	0.5913	0.5774
	(0.0457)	(0.1005)	(0.1004)
θ_{year}		-9.7122	-9.4564
		(1.2198)	(1.2166)
α_{dp4}	-0.3318	0.2171	0.1347
	(0.0709)	(0.0964)	(0.0896)
$\alpha_{\text{dp4}} \times I_{\text{cs34}}$	-0.2034	-0.2497	-0.2384
	(0.1109)	(0.1114)	(0.1113)
τ_{dp4_1}		37.5	37.5
γ_{dp4_1}		-0.1805	-0.2136
		(0.0456)	(0.0393)
$\gamma_{\text{dp4}_1} \times I_{\text{cs37}}$		0.1331	0.1332
		(0.0644)	(0.0644)
τ_{dp4_2}		70.5	70.5
γ_{dp4_2}		-0.0258	-0.0197
		(0.0051)	(0.0039)
α_{dp13}	-0.7313	-0.1734	-0.2037
	(0.0953)	(0.1130)	(0.1101)
$\alpha_{\text{dp13}} \times I_{\text{cs37}}$	0.3802	0.5064	0.5062
	(0.1736)	(0.1745)	(0.1744)
τ_{dp13}		125.5	125.5
γ_{dp13}		-0.0413	-0.0394
		(0.0107)	(0.0095)
α_{dp26}	-0.6018	-0.1067	-0.1649
	(0.1706)	(0.1787)	(0.1756)
α_{dp52}	-1.1524	-0.6483	-0.7099
	(0.2831)	(0.2883)	(0.2860)
$\alpha_{\text{age:dp26}}$	-0.8772	-1.0545	-1.0881
	(0.3726)	(0.3723)	(0.3691)



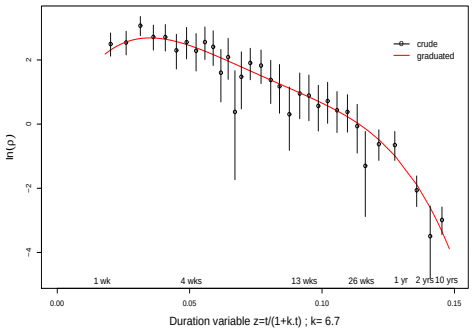
3.20.1: All



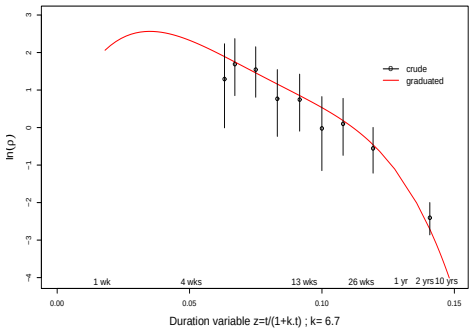
3.20.2: Male



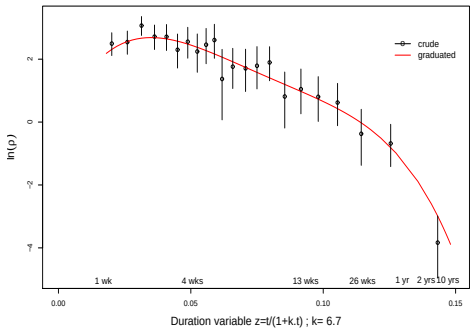
3.20.3: Female



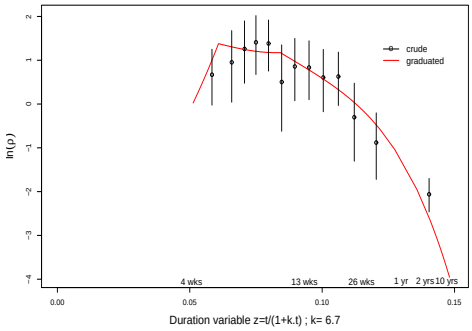
3.20.4: Non-rated



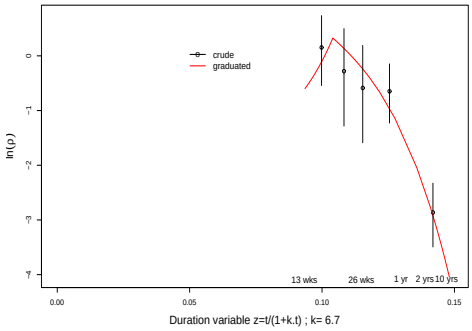
3.20.5: Rated



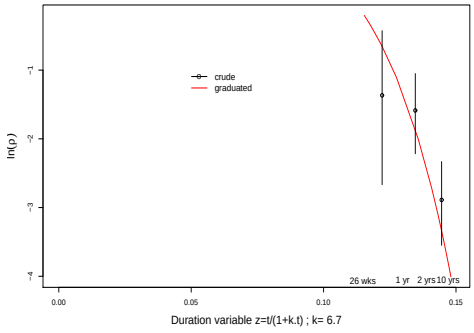
3.20.6: DP1



3.20.7: DP4

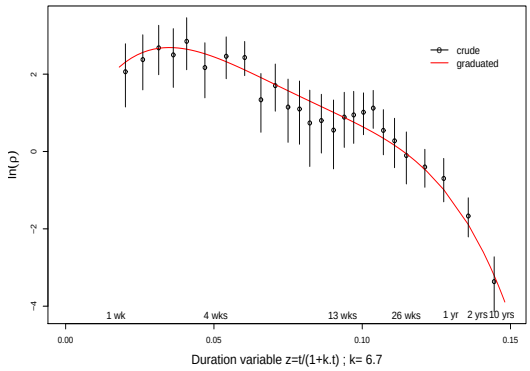


3.20.8: DP13

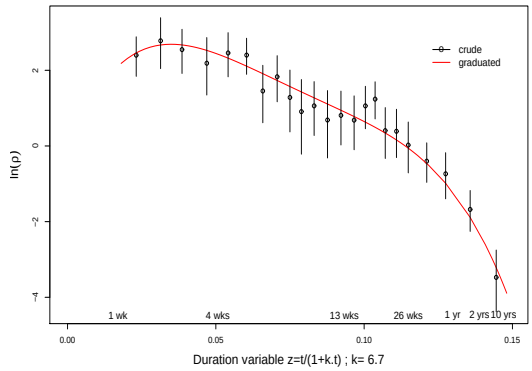


3.20.9: DP26

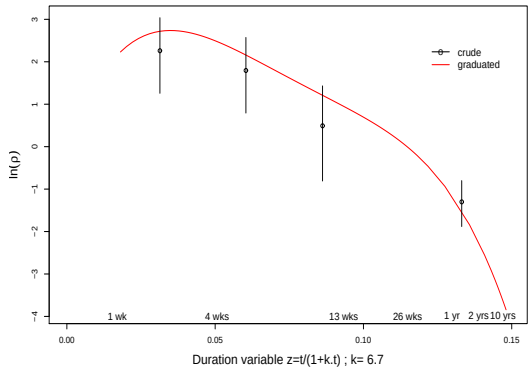
Figure 3.20: The "residual effects" and "partial residual effects" plots for the discrete covariates – cs34.



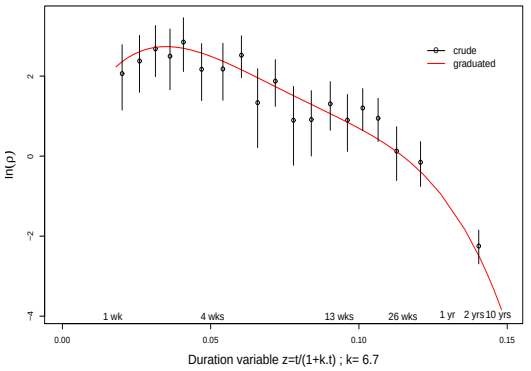
3.21.1: All



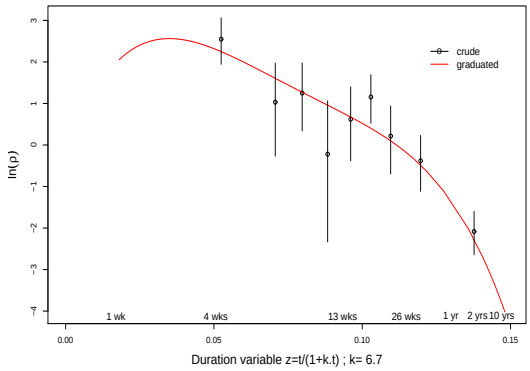
3.21.2: Male



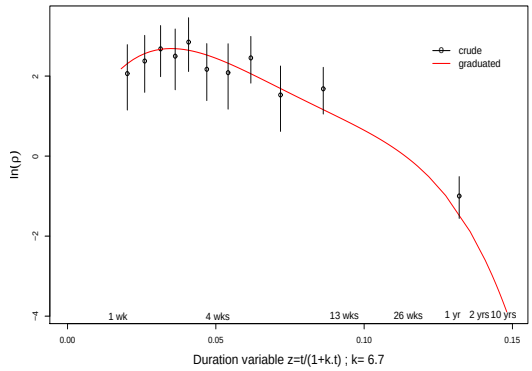
3.21.3: Female



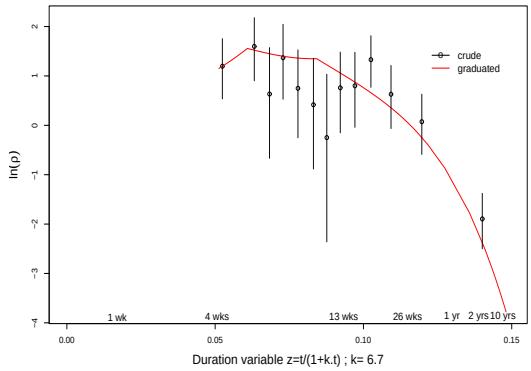
3.21.4: Non-rated



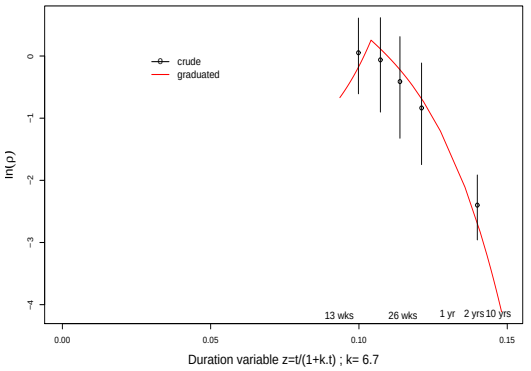
3.21.5: Rated



3.21.6: DP1

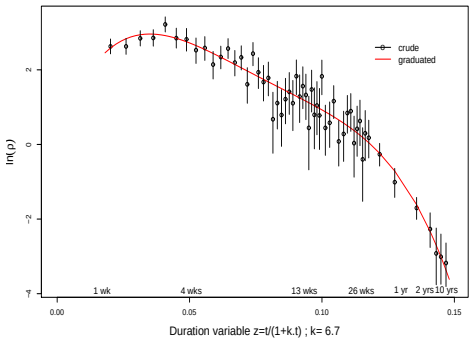


3.21.7: DP4

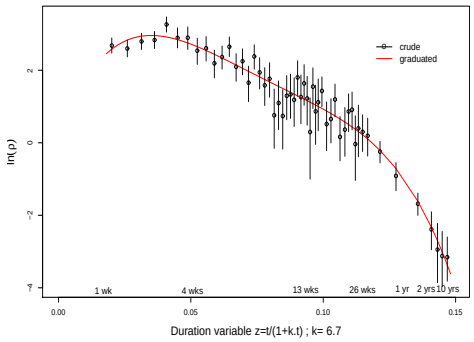


3.21.8: DP13

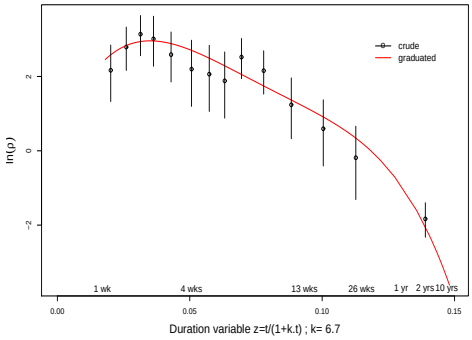
Figure 3.21: The "residual effects" and "partial residual effects" plots for the discrete covariates – cs37.



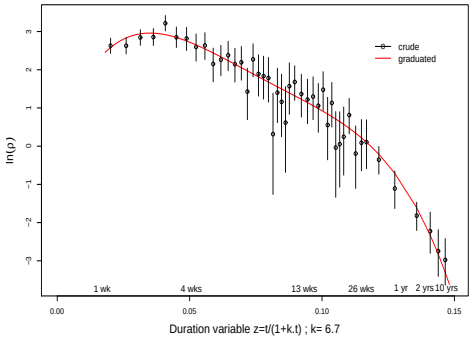
3.22.1: All



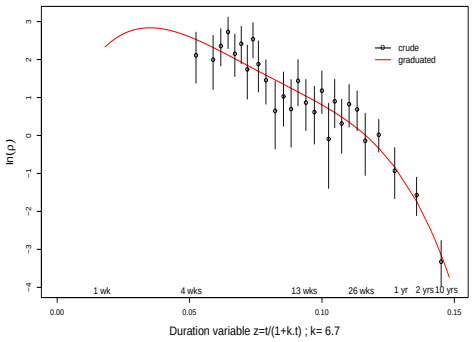
3.22.2: Male



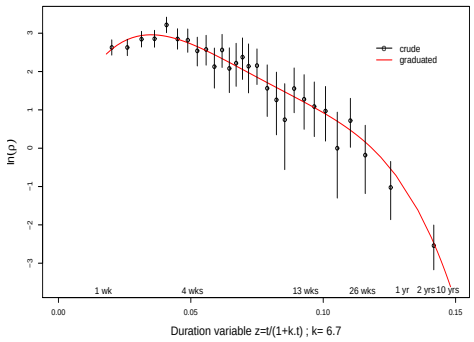
3.22.3: Female



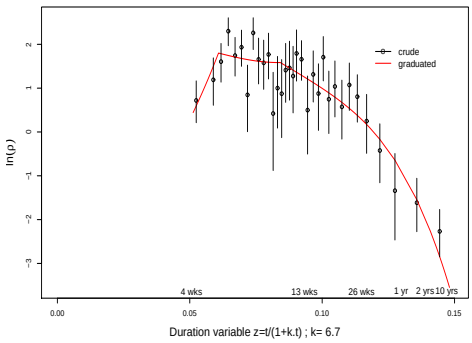
3.22.4: Non-rated



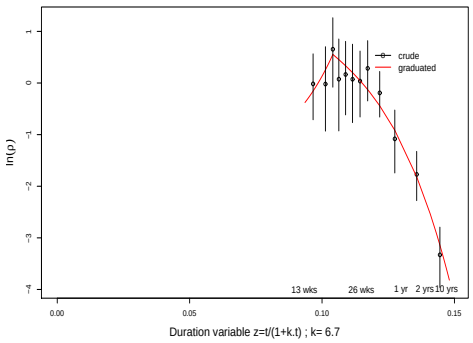
3.22.5: Rated



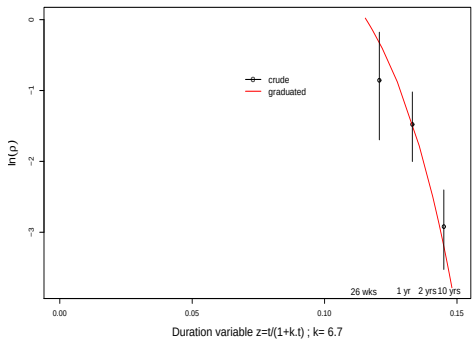
3.22.6: DP1



3.22.7: DP4



3.22.8: DP13



3.22.9: DP26

Figure 3.22: The "residual effects" and "partial residual effects" plots for the discrete covariates – cs38.

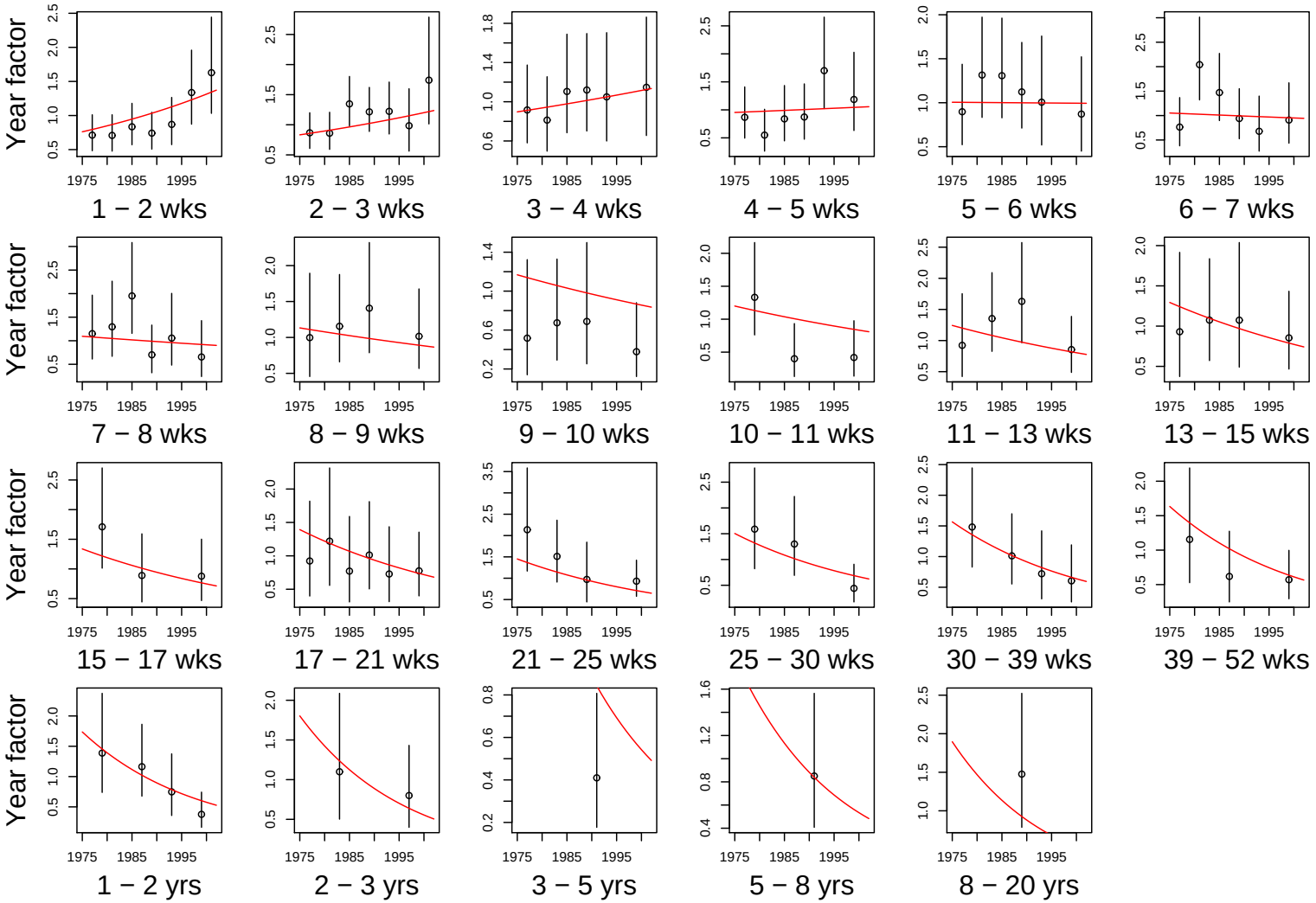


Figure 3.23: The "partial residual effects" plots for year – cs38.

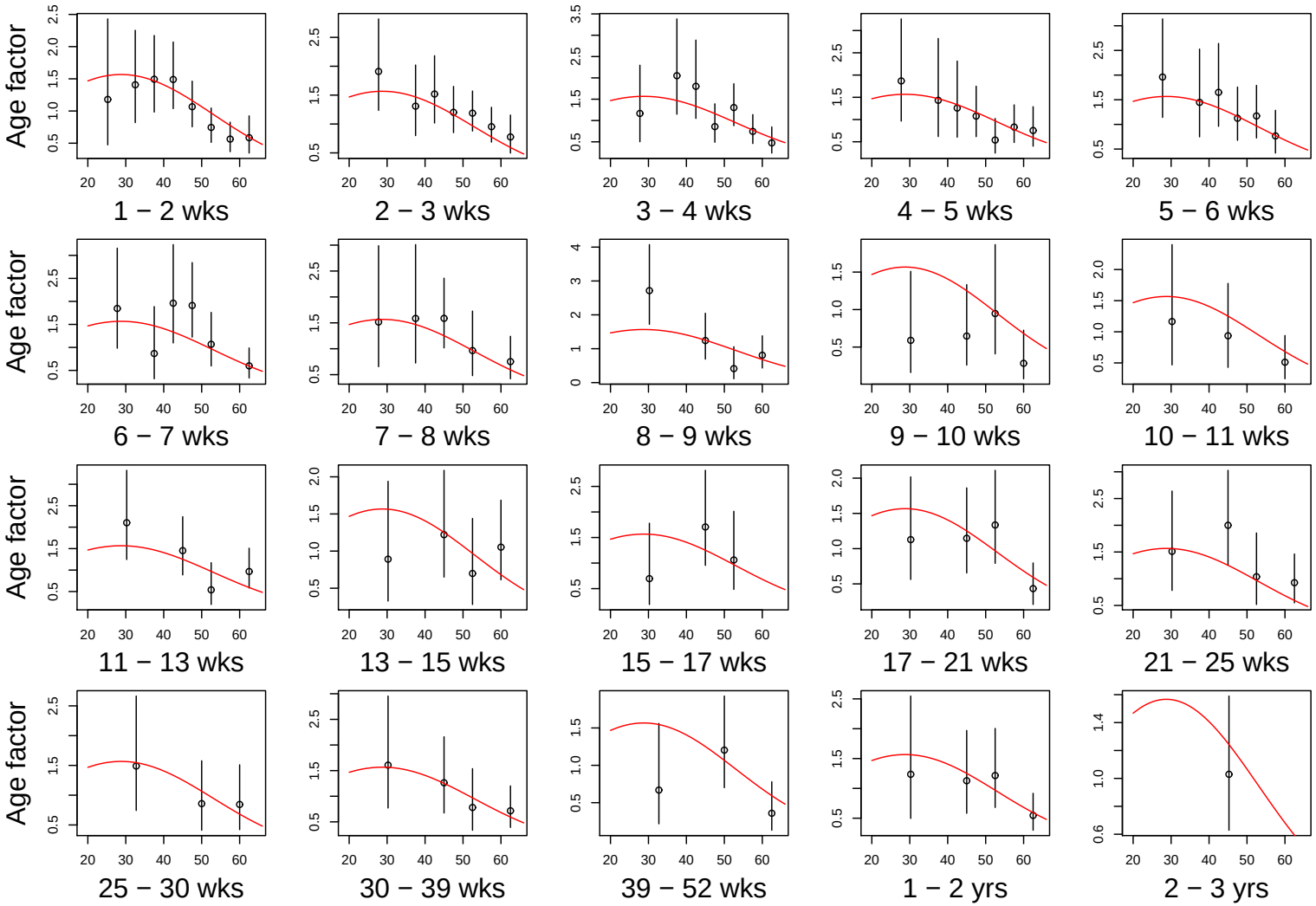


Figure 3.24: The "partial residual effects" plots for age – cs38.

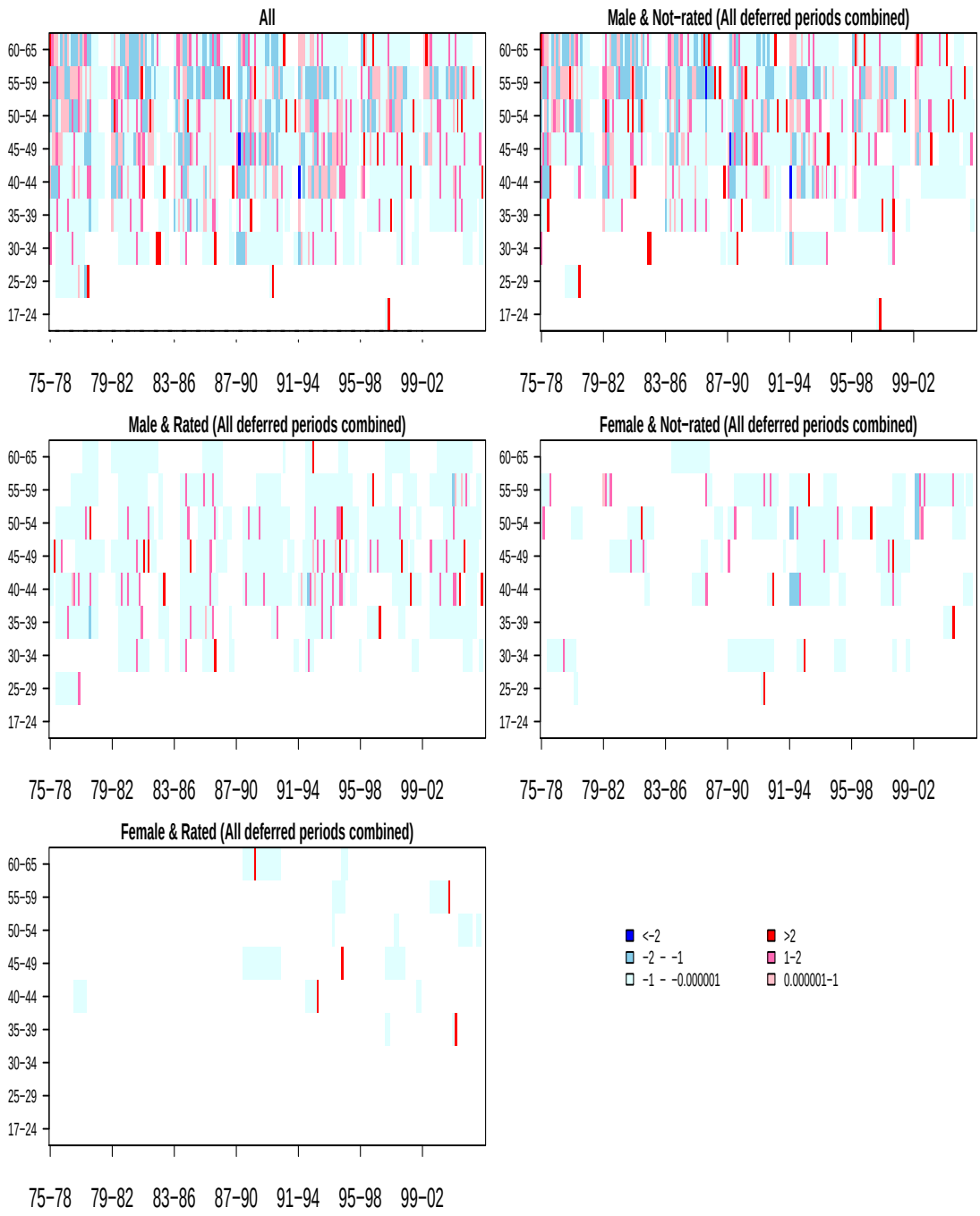


Figure 3.25: Two-dimensional plots of deviance residuals for cs34 data.

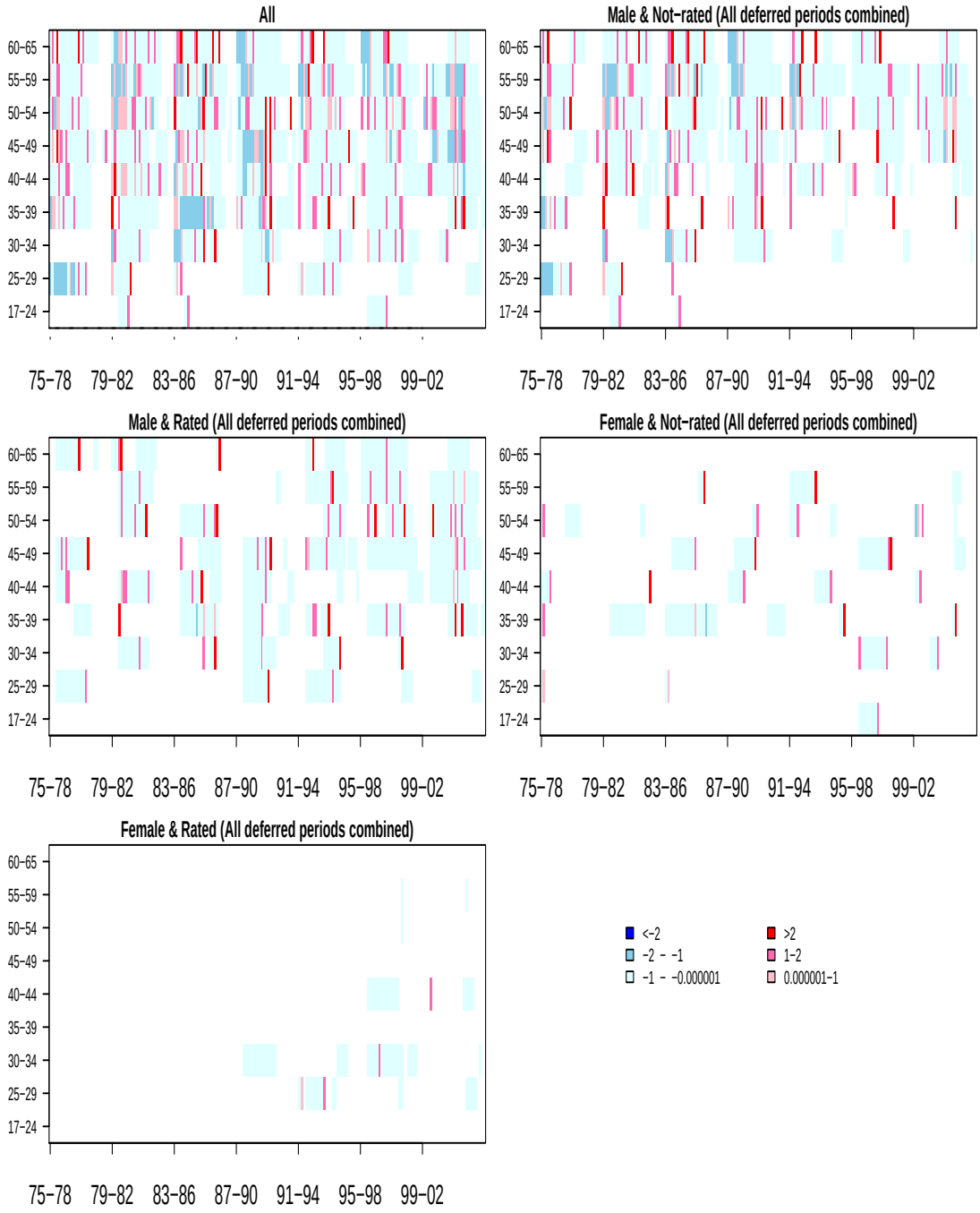


Figure 3.26: Two-dimensional plots of deviance residuals for cs37 data.

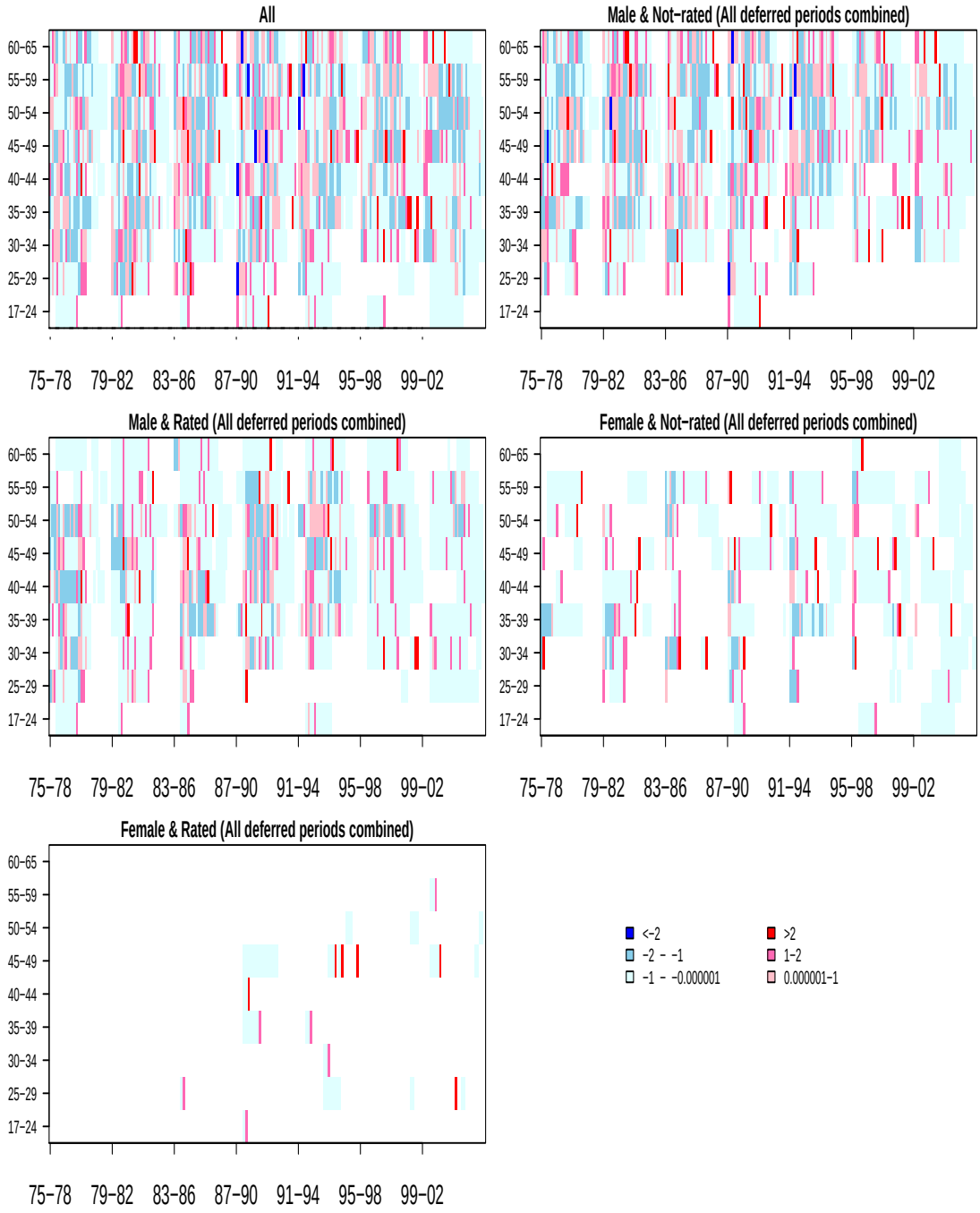


Figure 3.27: Two-dimensional plots of deviance residuals for cs38 data.

Recovery intensity model for cs35

For cs35, Table 3.10 shows the estimated regression coefficients from intermediate models (i.e. Models I and II) leading up to the fully-specified recovery intensity model (i.e. Model III), in columns two to four, respectively. Model I is a Cox model which assumes that all the covariate effects are duration-fixed, an assumption which is not valid for $x_{\text{year}}, x_{\text{dp4}}, x_{\text{dp13}}$ and x_{dp26}. This is because the smoothed plot of their respective scaled Schoenfeld residuals in Figure 3.28 suggests that their respective coefficients can be parameterised as

$$\begin{aligned}
\beta_{\text{year}}(z) &= \alpha_{\text{year}} + \gamma_{\text{year}}(\tau_{\text{year}} - z)_+ \\
\beta_{\text{dp4}}(z) &= \alpha_{\text{dp4}} + \gamma_{\text{dp4}}(\tau_{\text{dp4}} - z)_+ \\
\beta_{\text{dp13}}(z) &= \alpha_{\text{dp13}} + \gamma_{\text{dp13}}(\tau_{\text{dp13}} - z)_+ \\
\beta_{\text{dp26}}(z) &= \alpha_{\text{dp26}} + \gamma_{\text{dp26}}(\tau_{\text{dp26}} - z)_+
\end{aligned}$$

where $\tau_{\text{year}} = 25.5$, $\tau_{\text{dp4}} = 72.5, \tau_{\text{dp13}} = 185.5$ and $\tau_{\text{dp26}} = 265.5$.

Model II is obtained by including in Model I additional covariates created to describe the duration-varying effects of $x_{\text{year}}, x_{\text{dp4}}, x_{\text{dp13}}$ and x_{dp26}. Lastly, Model III is Model II but with a parametric baseline intensity described by a $\exp(\sum_{i=0}^{3} b_i C_i(t_{6.7}(z)))$ formula. The final fitted model, Model III, is then assessed for its goodness-of-fit. The value of the χ^2 statistic is 270.2829. With 289 cells and 18 parameters fitted in the model, the probability value is 0.5009 on 271 degrees of freedom, indicating a good fit to the data. The "residual effects" and "partial residual effects" plots for the discrete covariates are presented in Figure 3.29. The data upon which each plot is based is indicated by the caption of the plot. Figures 3.30 and 3.31 show the "partial residual effects" plots for age and year, respectively, for each sickness duration band. The red curves overlaid on the plots in Figures 3.29–3.31 are the estimated smooth intensities according to Model III and they fitted the actual experience reasonably well. The two-dimensional plots of the deviance residuals for all data as well as its subsets are shown in Figure 3.32. The positve and negative residuals are reasonably randomly scattered. However, the cells for younger age range are

dominantly white because there is very small exposure to risk for these age ranges.

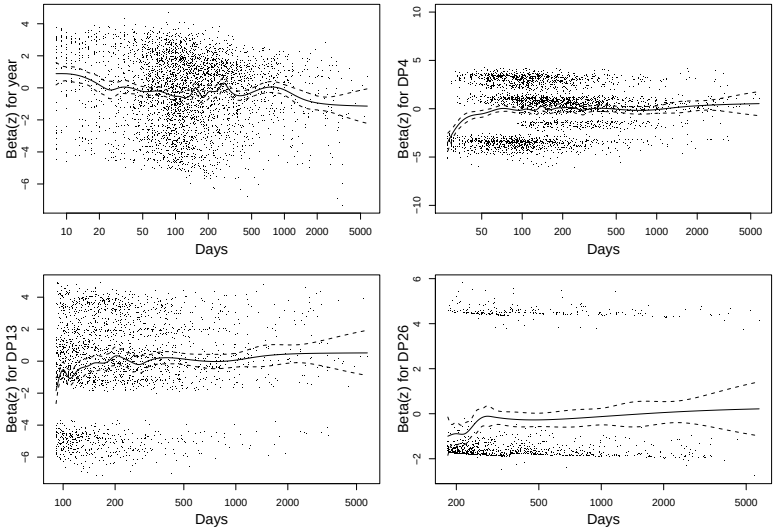
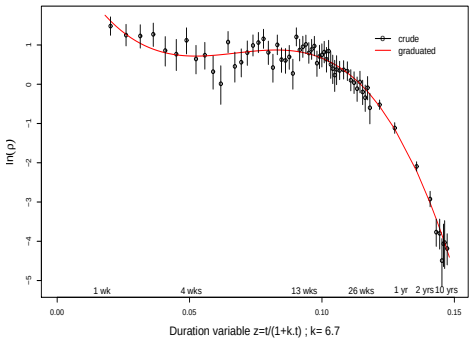


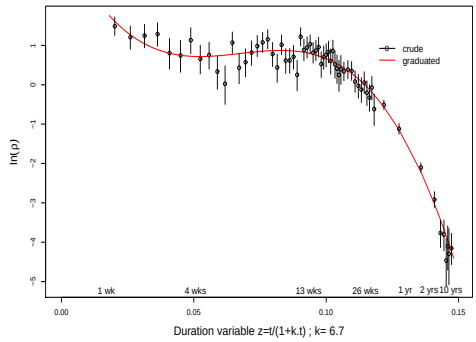
Figure 3.28: The smoothed plots of the scaled Schoenfeld residuals for $x_{\text{year}}, x_{\text{dp4}}, x_{\text{dp13}}$ and x_{dp26} in cs35.

Table 3.10: Parameters in the intermediate and final recovery intensity models for cs35.

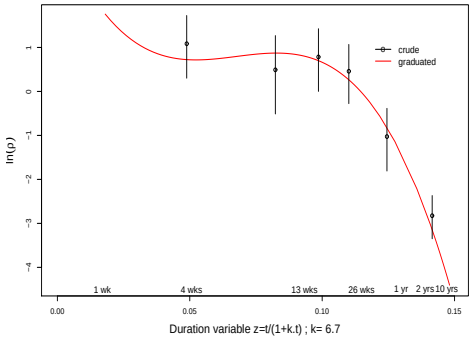
	Model I	Model II	Model III
k			6.7
b_0			1128.687
			(65.498)
b_1			-8475.895
			(388.616)
b_2			1124.992
			(65.292)
b_3			-2777.223
			(126.278)
α_{age}	0.4939	0.4600	0.4602
	(0.3809)	(0.3791)	(0.3787)
α_{age2}	-0.7186	-0.7058	-0.6989
	(0.2076)	(0.2066)	(0.2063)
α_{age3}	0.3526	0.3403	0.3400
	(0.1254)	(0.1250)	(0.1249)
α_{rated}	-0.5699	-0.5482	-0.5489
	(0.1119)	(0.1110)	(0.1109)
α_{year}	-0.2466	-0.3129	-0.3115
	(0.0380)	(0.0392)	(0.0391)
τ_{year}		25.5	25.5
γ_{year}		0.0705	0.0648
		(0.0107)	(0.0103)
α_{year2}	0.0973	0.0944	0.0923
	(0.0387)	(0.0387)	(0.0387)
α_{dp4}	-0.1426		
	(0.0466)		
τ_{dp4}		72.5	72.5
γ_{dp4}		-0.0261	-0.0295
		(0.0036)	(0.0031)
α_{dp13}	-0.3174		
	(0.05596)		
τ_{dp13}		185.5	185.5
γ_{dp13}		-0.0090	-0.0083
		(0.0011)	(0.0011)
α_{dp26}	-0.1988	0.1521	0.1931
	(0.1411)	(0.1440)	(0.1427)
τ_{dp26}		265.5	265.5
γ_{dp26}		-0.0117	-0.0137
		(0.0037)	(0.0035)
α_{dp52}	-0.8835	-0.6690	-0.6343
	(0.1836)	(0.1816)	(0.1806)
$\alpha_{\text{age3:dp26}}$	0.6104	0.6477	0.6477
	(0.1716)	(0.1716)	(0.1707)
$\alpha_{\text{age2:rated}}$	-0.4902	-0.4782	-0.4784
	(0.1437)	(0.1436)	(0.1436)



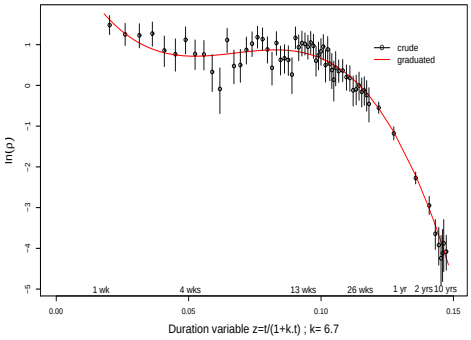
3.29.1: All



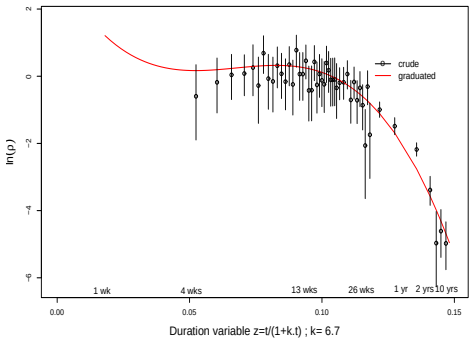
3.29.2: Male



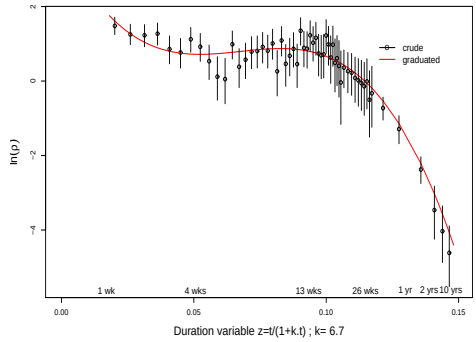
3.29.3: Female



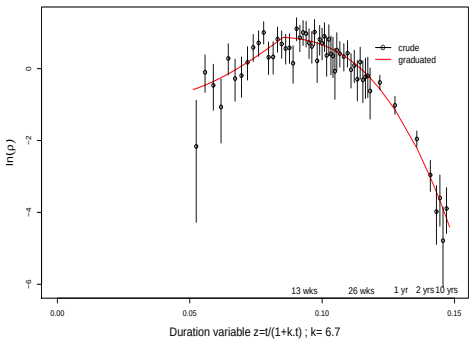
3.29.4: Non-rated



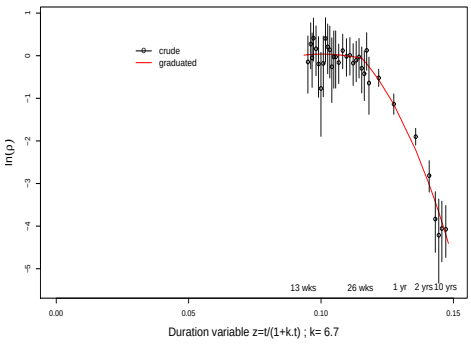
3.29.5: Rated



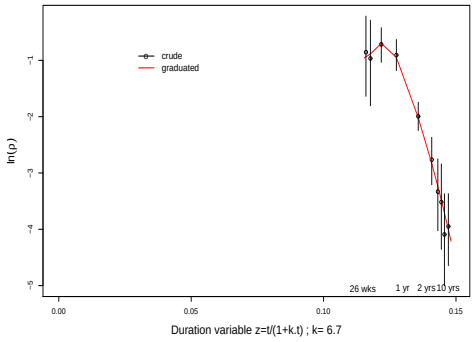
3.29.6: DP1



3.29.7: DP4



3.29.8: DP13



3.29.9: DP26

Figure 3.29: The "residual effects" and "partial residual effects" plots for the discrete covariates – cs35.

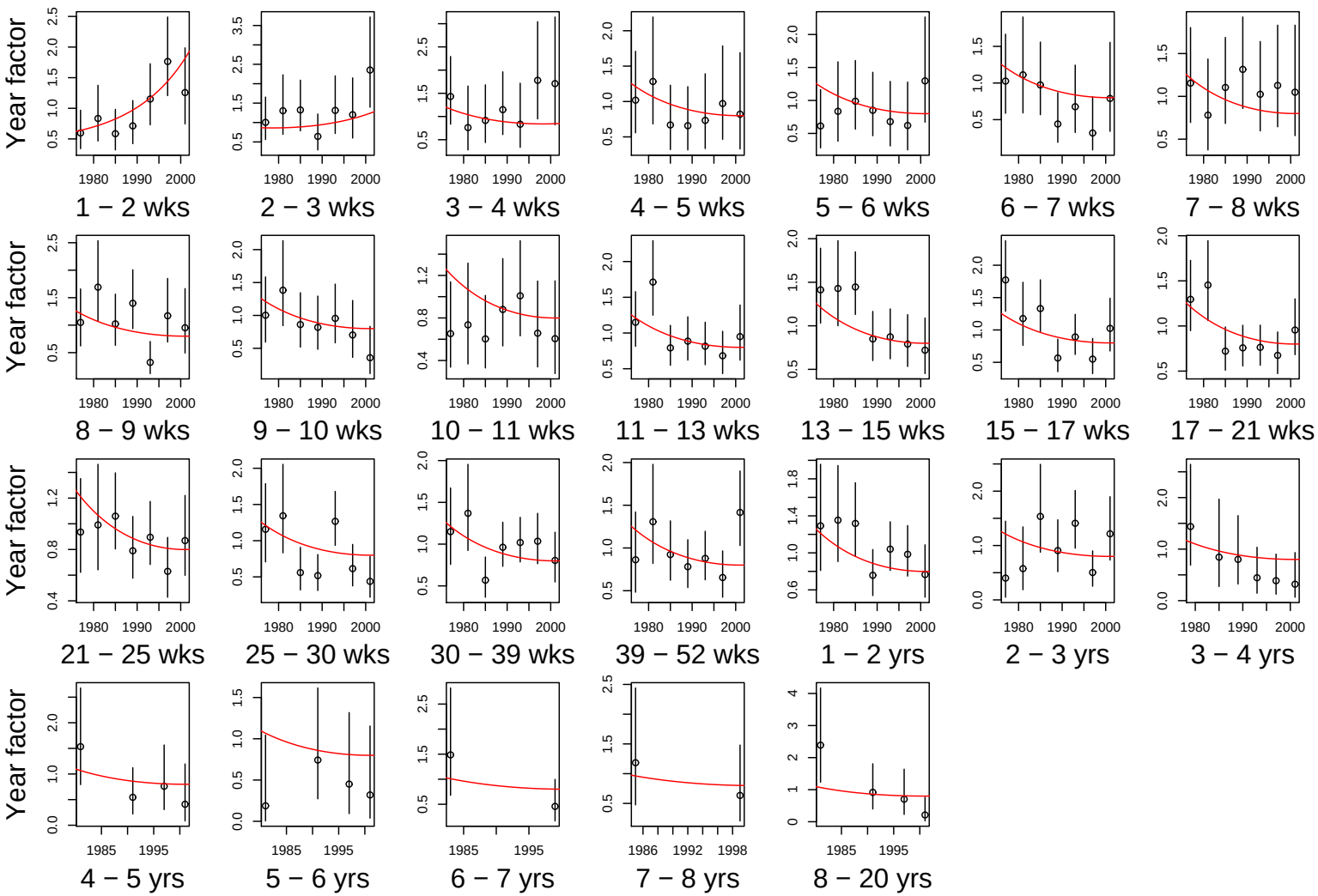


Figure 3.30: The "partial residual effects" plots for year – cs35.

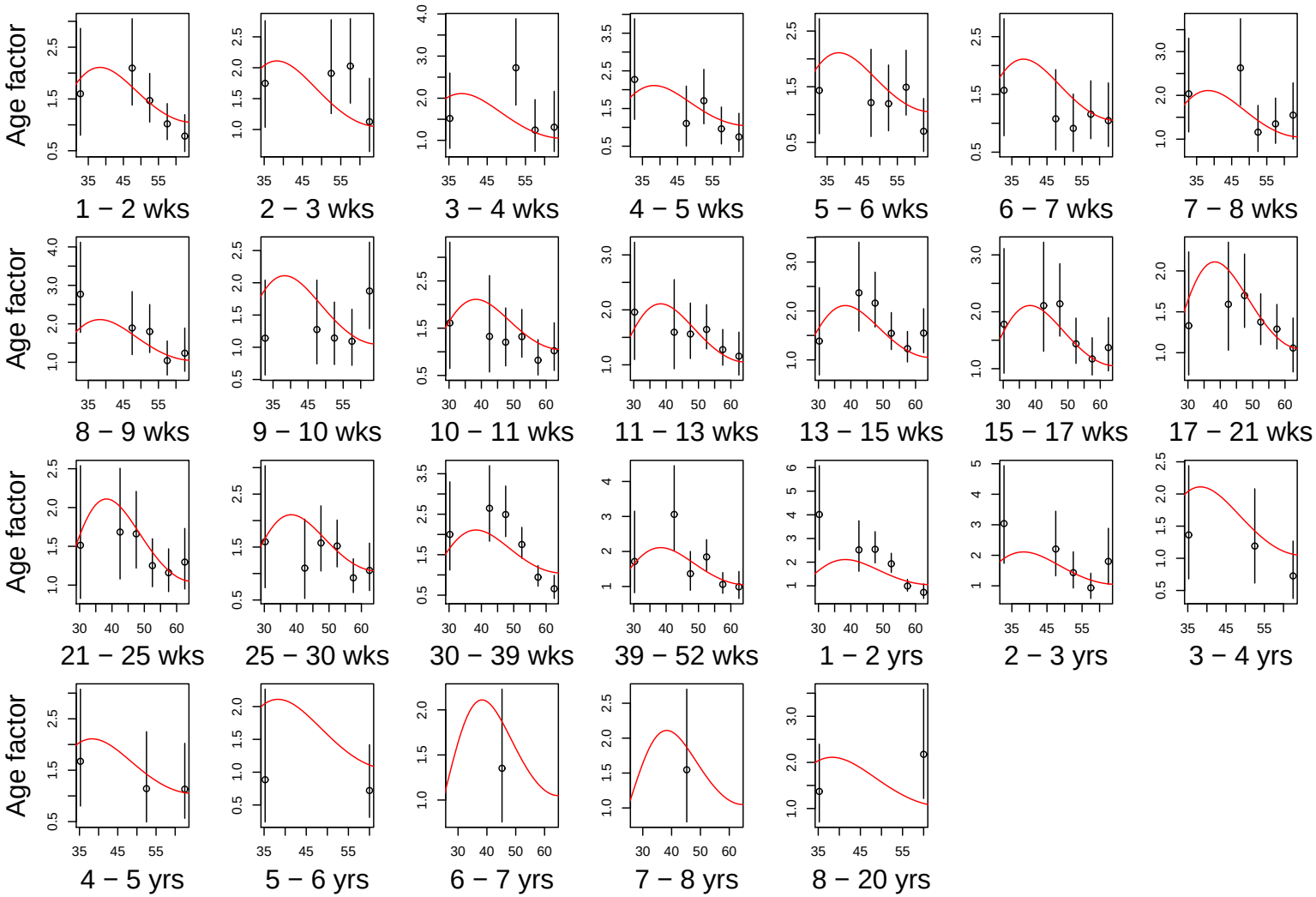


Figure 3.31: The "partial residual effects" plots for age – cs35.

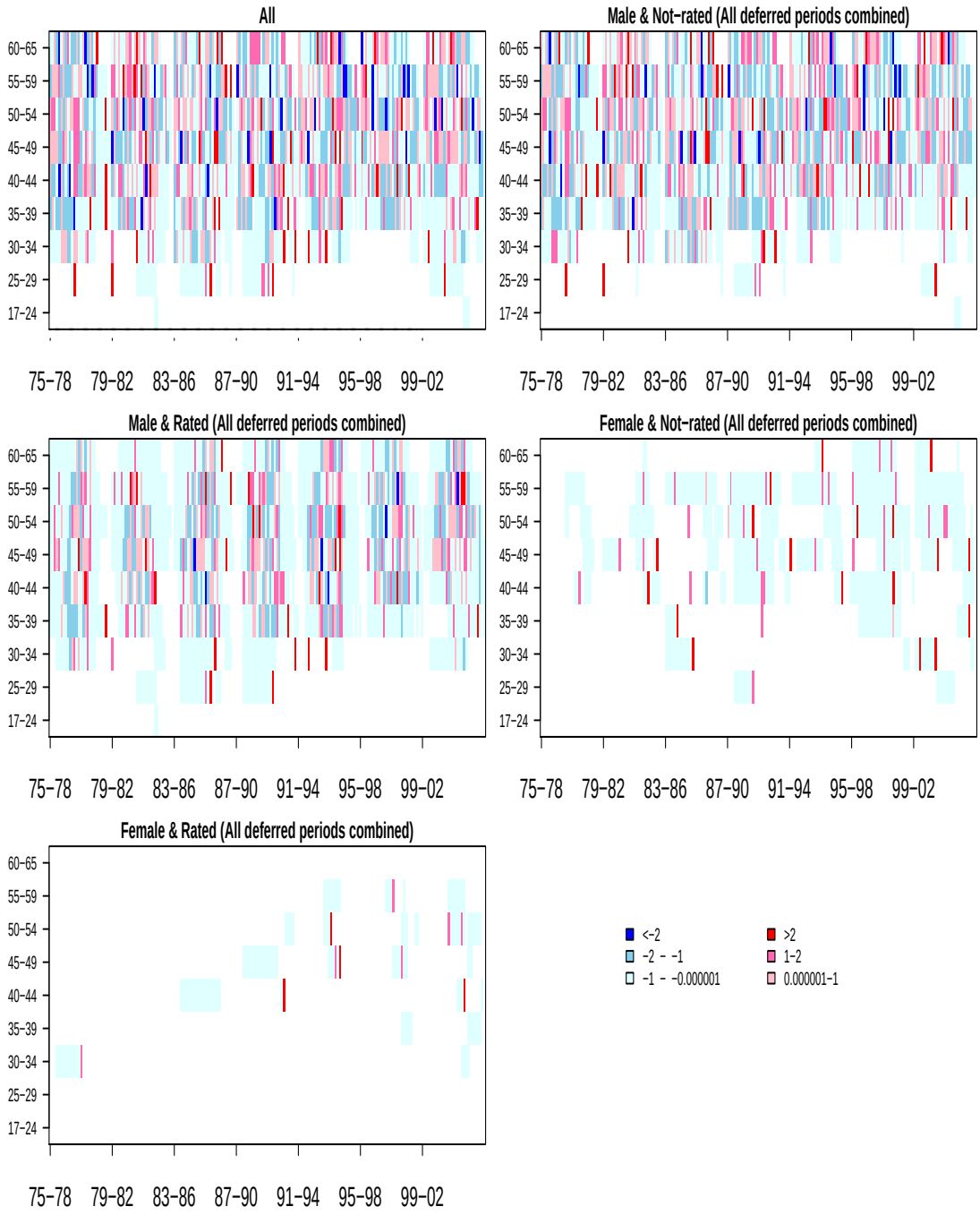


Figure 3.32: Two-dimensional plots of deviance residuals for cs35 data.

Recovery intensity model for cs36

For cs36, Table 3.11 shows, in column two to four, the results from Models I, II and III, respectively. Model I assumes that all the covariate effects are duration-fixed, an assumption which is not valid for x_{dp4} and x_{dp13} are duration-dependent because the smoothed plots of their scaled Schoenfeld residuals shown in Figure 3.33 suggests that their respective coefficients can be parameterised as

$$\beta_{\text{dp4}}(z) = \alpha_{\text{dp4}} + \gamma_{\text{dp4}}(\tau_{\text{dp4}} - z)_+$$
$$\beta_{\text{dp13}}(z) = \alpha_{\text{dp13}} + \gamma_{\text{dp13}}(\tau_{\text{dp13}} - z)_+$$

where $\tau_{\text{dp4}} = 81.5$ and $\tau_{\text{dp13}} = 180.5$.

Model II is Model I but with the inclusion of additional covariates created to describe the duration-varying effects of x_{dp4} and x_{dp13}. Lastly, Model III is the fully parameterised recovery intensity model with the baseline intensity described by a $\exp(\sum_{i=0}^{3} b_i C_i(t_{2.3}(z)))$ formula. This final fitted model is then assessed for its goodness-of-fit. The value of the χ^2 statistic is 19.7367. With 31 cells and 14 parameters fitted in the model, the probability value is 0.2879 on 17 degrees of freedom, indicating a reasonably good fit to the data. The "residual effects" and "partial residual effects" plots for the discrete covariates are presented in Figure 3.34. The red curve overlaid on the plots in Figure 3.34 are the estimated smooth intensities according to Model III and they seem to fit the actual experience rather well. The data upon which each plot is based is indicated by the caption of the plot. We do not construct the "partial residual effects" plots for age and year because the amount of data is not sufficiently large to make such an exercise meaningful. The two-dimensional plots of the deviance residuals for all data as well as its subsets are shown in Figure 3.35. Due to the sparseness of the data, the plots appear mostly in blue and white.

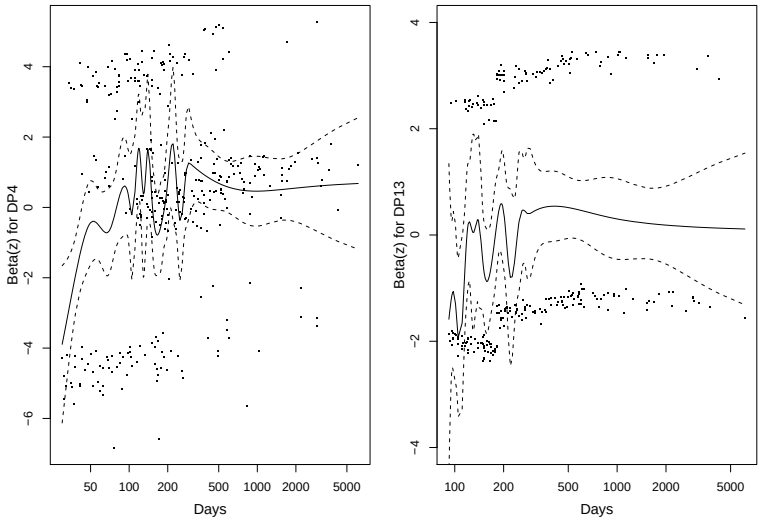
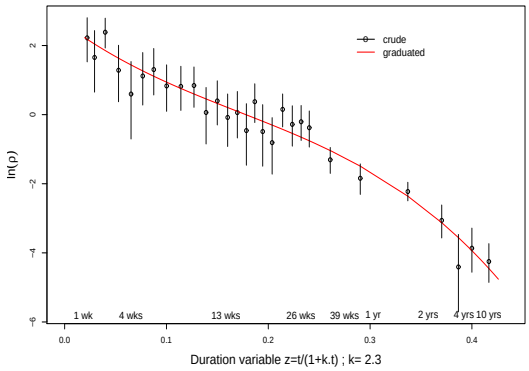


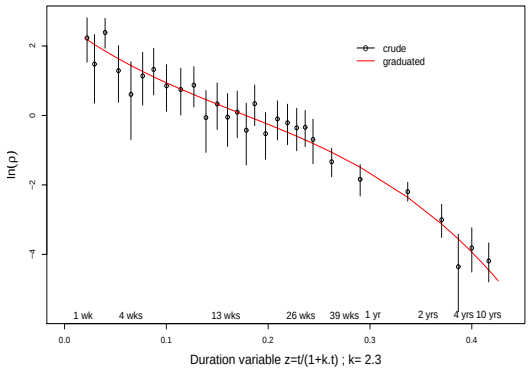
Figure 3.33: The smoothed plots of the scaled Schoenfeld residuals for x_{dp4} and x_{dp13} in cs36.

Table 3.11: Parameters in the intermediate and final recovery intensity models for cs36.

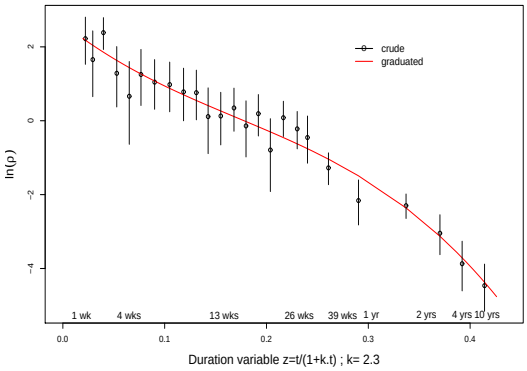
	Model I	Model II	Model III
k			2.3
b_0			32.1920
			(14.8636)
b_1			-107.9258
			(38.2616)
b_2			29.5602
			(14.6404)
b_3			-28.7525
			(11.0018)
α_{sex}	-0.6698	-0.6302	-0.6349
	(0.2785)	(0.2786)	(0.2787)
α_{age}	-0.7812	-0.7916	-0.8054
	(0.1473)	(0.1473)	(0.1465)
α_{rated}	-0.3959	-0.3923	-0.3976
	(0.1741)	(0.1745)	(0.1744)
α_{year}	-0.4774	-0.4608	-0.4584
	(0.1119)	(0.1119)	(0.1119)
α_{dp4}	0.2023	0.3769	0.4191
	(0.1412)	(0.1568)	(0.1487)
τ_{dp4}		81.5	81.5
γ_{dp4}		-0.0302	-0.0338
		(0.0090)	(0.0074)
α_{dp13}	-0.0228	0.2609	0.3185
	(0.1616)	(0.1766)	(0.1743)
τ_{dp13}		180.5	180.5
γ_{dp13}		-0.0140	-0.0141
		(0.0054)	(0.0046)
α_{dp52}	-1.3888	-1.2462	-1.1757
	(0.5904)	(0.5922)	(0.5913)
$\alpha_{\text{rated:dp13}}$	0.6790	0.6688	0.6679
	(0.2751)	(0.2754)	(0.2753)



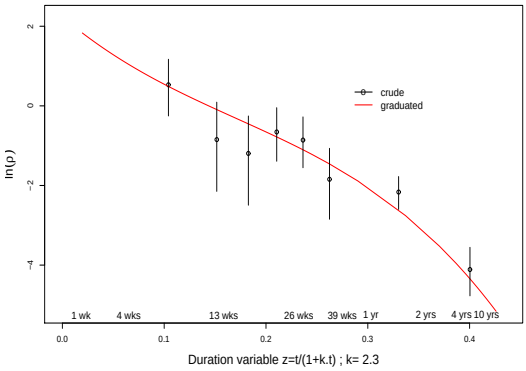
3.34.1: All



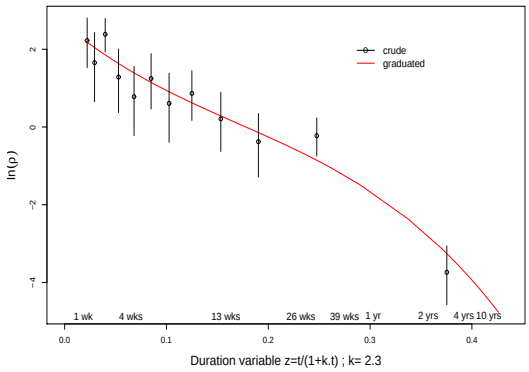
3.34.2: Male



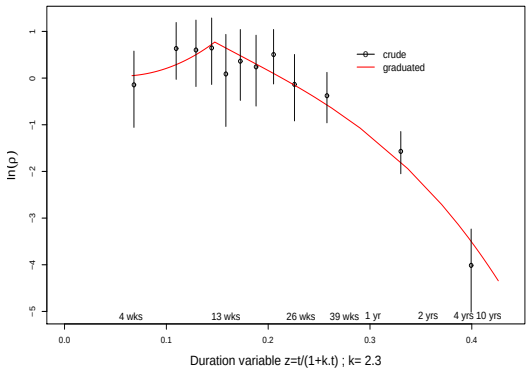
3.34.3: Non-rated



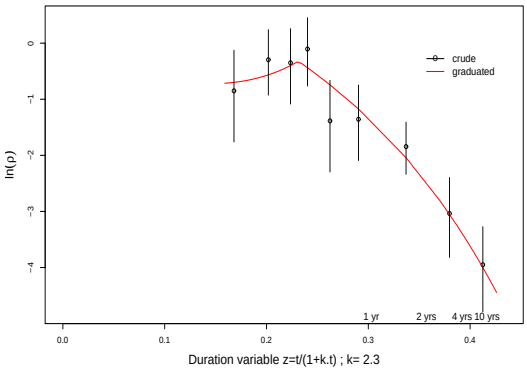
3.34.4: Rated



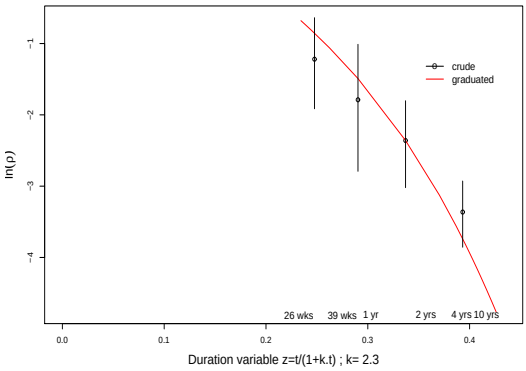
3.34.5: DP1



3.34.6: DP4



3.34.7: DP13



3.34.8: DP26

Figure 3.34: The "residual effects" and "partial residual effects" plots for the discrete covariates – cs36.

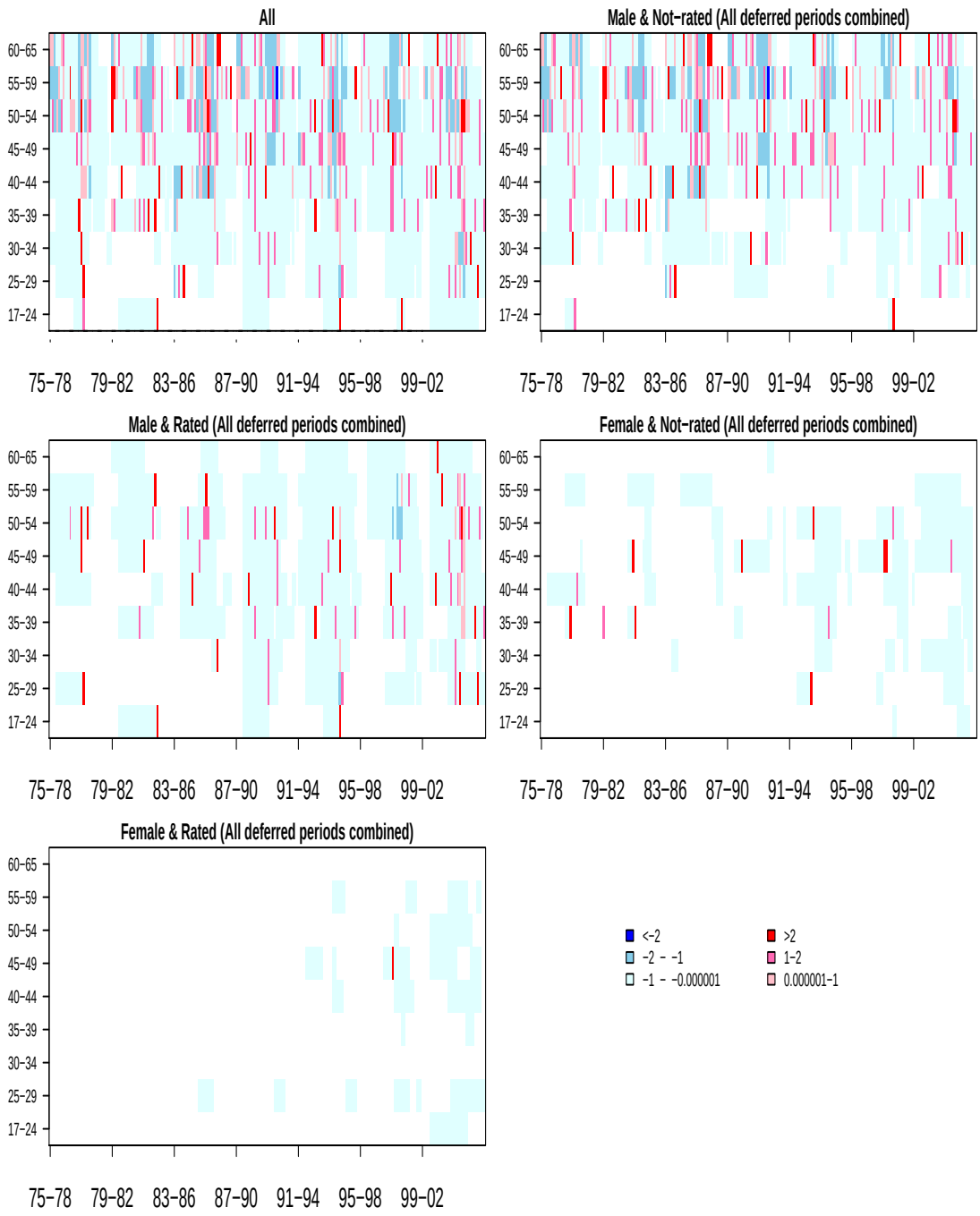


Figure 3.35: Two-dimensional plots of deviance residuals for cs36 data.

3.6.4 G10 Musculoskeletal

The causes of sickness in G10 Musculoskeletal alongside their ICD8 code, exposed to risk in units of days and the number of recoveries are presented in Table 3.12.

Table 3.12: The causes of sickness in sickness category G10 Musculoskeletal.

ICD8	Cause of sickness	Exposed to risk (days)	Recoveries
61	Arthritis and spondylitis	3,401,274	1,847
62	Other diseases of musculoskeletal system and connective tissue	6,345,553	11,284

The reference cause of sickness is cs62. The proportionality between the recovery patterns for cs61 and cs62 is assessed using the smoothed plot of the scaled Schoenfeld residuals as presented in Figure 3.36. The broken red line overlaid on this plot is the Cox's estimate. This plot suggests that the log hazard ratio between cs61 and cs62 does not stay reasonably constant at all sickness durations. Given that both causes of sickness have large amounts of data and their recovery patterns are not reasonably proportional to each other, we will estimate their recovery intensities separately.

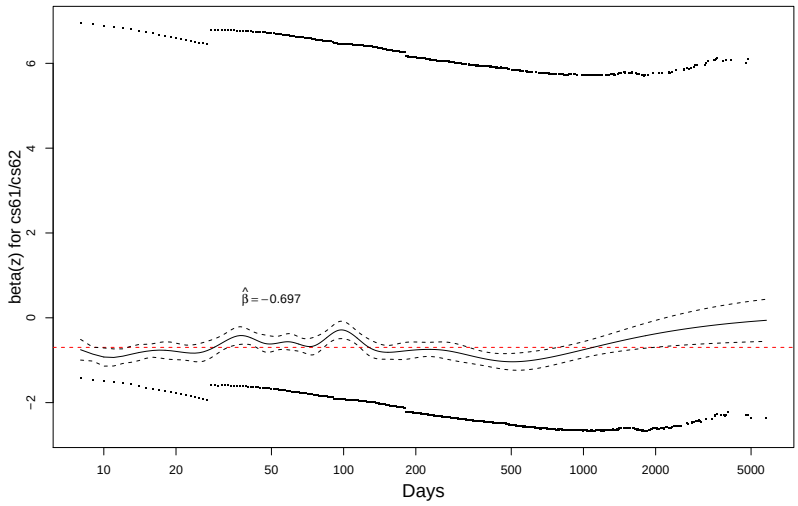


Figure 3.36: The smoothed plot of the Schoenfeld residuals for the causes of sickness in G10.

Recovery intensity model for cs61

For cs61, Table 3.13 shows, in columns two to four, the results from Models I, II and III, respectively. Model I is the Cox model which assumes that all covariate effects stay constant for all sickness durations. This assumption is not valid for $x_{\text{year}}, x_{\text{rated}}, x_{\text{dp4}}, x_{\text{dp13}}$ and x_{dp26} because the smoothed plots of their scaled Schoenfeld residuals in Figure 3.37 suggests that their respective coefficients are duration dependent and can be parameterised as

$$\begin{aligned}
\beta_{\text{year}}(z) &= \alpha_{\text{year}} + \gamma_{\text{year}}(\tau_{\text{year}} - z)_+ + \theta_{\text{year}} C(t_{1.3}(z)) \\
\beta_{\text{rated}}(z) &= \alpha_{\text{rated}} + \theta_{\text{rated}} C(t_{1.3}(z)) \\
\beta_{\text{dp4}}(z) &= \alpha_{\text{dp4}} + \gamma_{\text{dp4}_1}(\tau_{\text{dp4}_1} - z)_+ + \gamma_{\text{dp4}_2}(\tau_{\text{dp4}_2} - z)_+ \\
\beta_{\text{dp13}}(z) &= \alpha_{\text{dp13}} + \gamma_{\text{dp13}}(\tau_{\text{dp13}} - z)_+ \\
\beta_{\text{dp26}}(z) &= \alpha_{\text{dp26}} + \theta_{\text{dp26}_1} C_1(t_{1.3}(z)) + \theta_{\text{dp26}_2} C_2(t_{1.3}(z))
\end{aligned}$$

where $\tau_{\text{year}} = 32.5, \tau_{\text{dp4}_1} = 37.5, \tau_{\text{dp4}_2} = 73.5$ and $\tau_{\text{dp13}} = 135.5$

Model II is obtaind by including in Model I additional covariates created to describe the duration-varying effects of $x_{\text{year}}, x_{\text{rated}}, x_{\text{dp4}}, x_{\text{dp13}}$ and x_{dp26}. Lastly, Model III is the fully parameterised recovery intensity model with the baseline intensity described by a $\exp(\sum_{i=0}^{1} b_i C_i(t_{1.3}(z)))$ formula. This final fitted model is then assessed for its goodness-of-fit. The value of the χ^2 statistic is 125.0230. With 151 cells and 25 parameters fitted in the model, the probability value is 0.5078 on 126 degrees of freedom, indicating a reasonably good fit to the data. The "residual effects" and "partial residual effects" plots for the discrete covariates are presented in Figure 3.38. The data upon which each plot is based is indicated by the caption of the plot. Figures 3.39 and 3.40 show the "partial residual effects" plots for age and year, respectively, for each sickness duration band. The red curve overlaid on the plots in Figures 3.38–3.40 are the estimated smooth intensities according to Model III which fitted the actual experience well. The two-dimensional plots of the deviance residuals for all data as

well as its subsets are shown in Figure 3.41. The positive and negative deviance residuals for 'All' and 'Male & Not-rated' data are roughly randomly scattered. For other subsets, the plots are dominated by white and light blue cells because most of the cells have either zero exposed to risk or no recoveries.

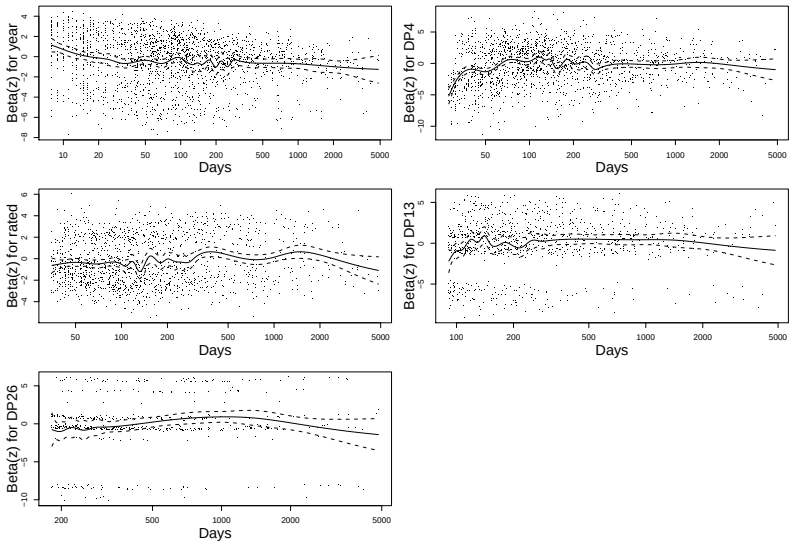
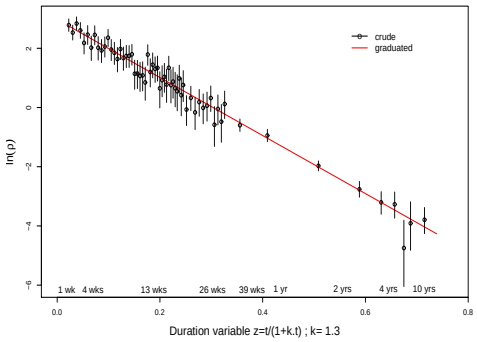


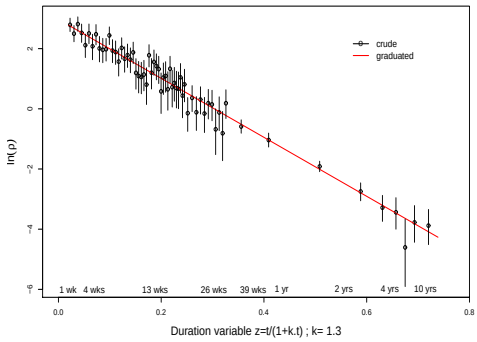
Figure 3.37: The smoothed plots of the scaled Schoenfeld residuals for $x_{\text{year}}, x_{\text{rated}}, x_{\text{dp4}}, x_{\text{dp13}}$ and x_{dp26} in cs61.

Table 3.13: Parameters in the intermediate and final recovery intensity models for cs61.

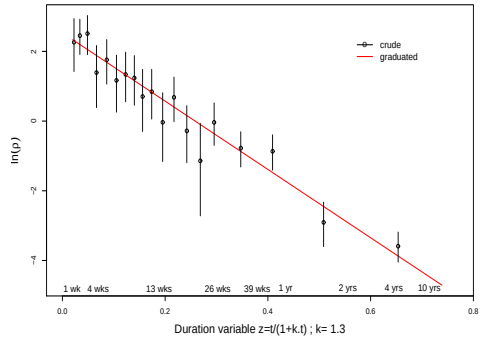
	Model I	Model II	Model III
k			1.3
s			2
b_0			2.9727
			(0.0604)
b_1			-9.8003
			(0.2229)
α_{sex}	-0.4220	-0.4390	-0.4411
	(0.1310)	(0.1307)	(0.1307)
α_{age}	-1.5370	-0.9073	-0.9037
	(0.2379)	(0.0805)	(0.0801)
α_{rated}	-0.2093	-0.6152	-0.6394
	(0.0631)	(0.1228)	(0.1173)
θ_{rated}		1.4057	1.4858
		(0.3484)	(0.3290)
α_{year}	-0.4380	-0.2200	-0.2287
	(0.0551)	(0.1073)	(0.1075)
θ_{year}		-1.6028	-1.5693
		(0.3472)	(0.3470)
τ_{year}		32.5	32.5
γ_{year}		0.0228	0.0222
		(0.0079)	(0.0077)
α_{year2}	0.4846	0.2002	0.2026
	(0.1905)	(0.0549)	(0.0549)
α_{dp4}	-0.1788	0.3338	0.2852
	(0.0705)	(0.0907)	(0.0765)
τ_{dp4_1}		37.5	37.5
γ_{dp4_1}		-0.1353	-0.1714
		(0.0615)	(0.0549)
τ_{dp4_2}		73.5	73.5
γ_{dp4_2}		-0.0358	-0.0296
		(0.0056)	(0.0041)
α_{dp13}	-0.2145	0.3857	0.2760
	(0.0952)	(0.1138)	(0.1016)
τ_{dp13}		135.5	135.5
γ_{dp13}		-0.0420	-0.0355
		(0.0073)	(0.0066)
α_{dp26}	0.7912	-10.0812	-12.4658
	(0.3780)	(6.4941)	(5.8057)
θ_{dp26_1}		15.3155	18.5551
		(8.8839)	(7.9450)
θ_{dp26_2}		-6.0138	-7.5355
		(4.3446)	(3.8761)
α_{dp52}	-0.1716	0.3590	0.3218
	(0.2856)	(0.2933)	(0.2914)
$\alpha_{\text{sex:age}}$	0.6913	0.6884	0.6884
	(0.2023)	(0.2017)	(0.2018)
$\alpha_{\text{sex:year2}}$	-0.3321	-0.3113	-0.3165
	(0.1670)	(0.1671)	(0.1671)
$\alpha_{\text{sex:dp26}}$	-0.8500	-0.7777	-0.7751
	(0.3032)	(0.3052)	(0.3053)
$\alpha_{\text{dp13:age}}$	-0.5591	-0.4783	-0.4698
	(0.1713)	(0.1719)	(0.1713)
$\alpha_{\text{dp26:age}}$	-1.4174	-1.2584	-1.2593
	(0.2942)	(0.2981)	(0.2975)
$\alpha_{\text{dp52:age}}$	-1.9150	-1.6638	-1.6507
	(0.7653)	(0.7753)	(0.7769)



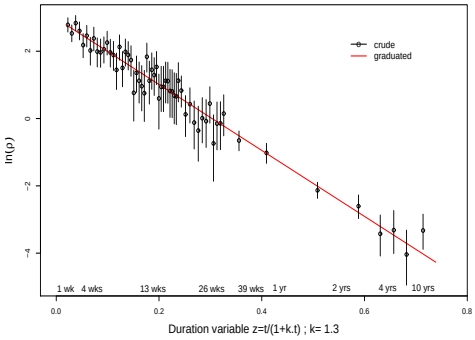
3.38.1: All



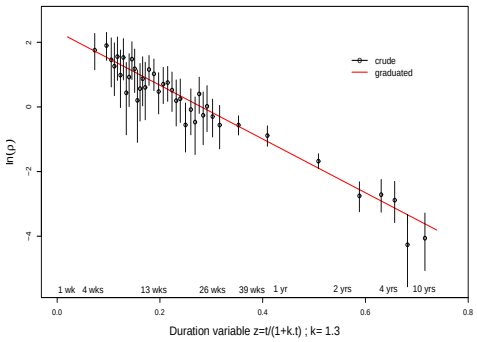
3.38.2: Male



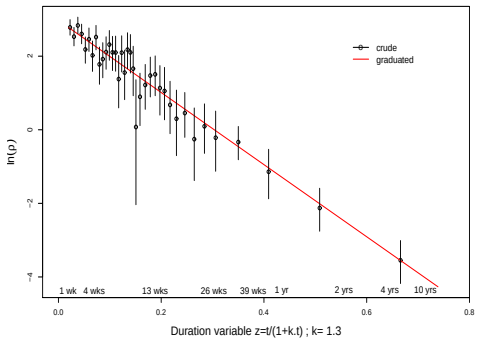
3.38.3: Female



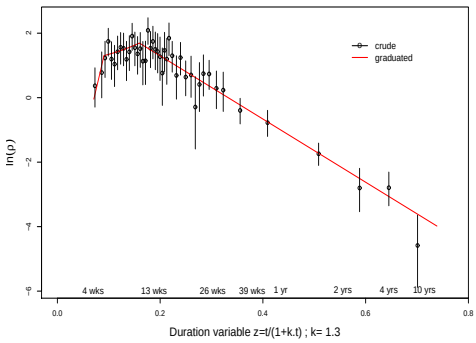
3.38.4: Non-rated



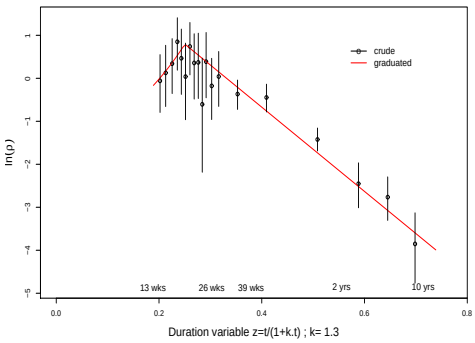
3.38.5: Rated



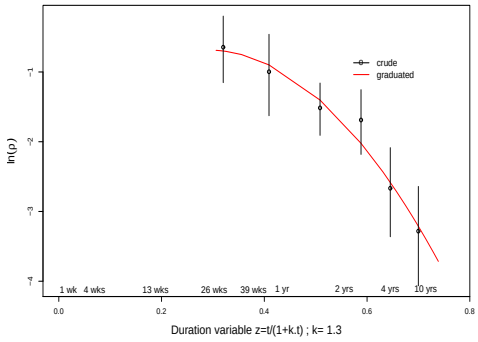
3.38.6: DP1



3.38.7: DP4



3.38.8: DP13



3.38.9: DP26

Figure 3.38: The "residual effects" and "partial residual effects" plots for the discrete covariates – cs61.

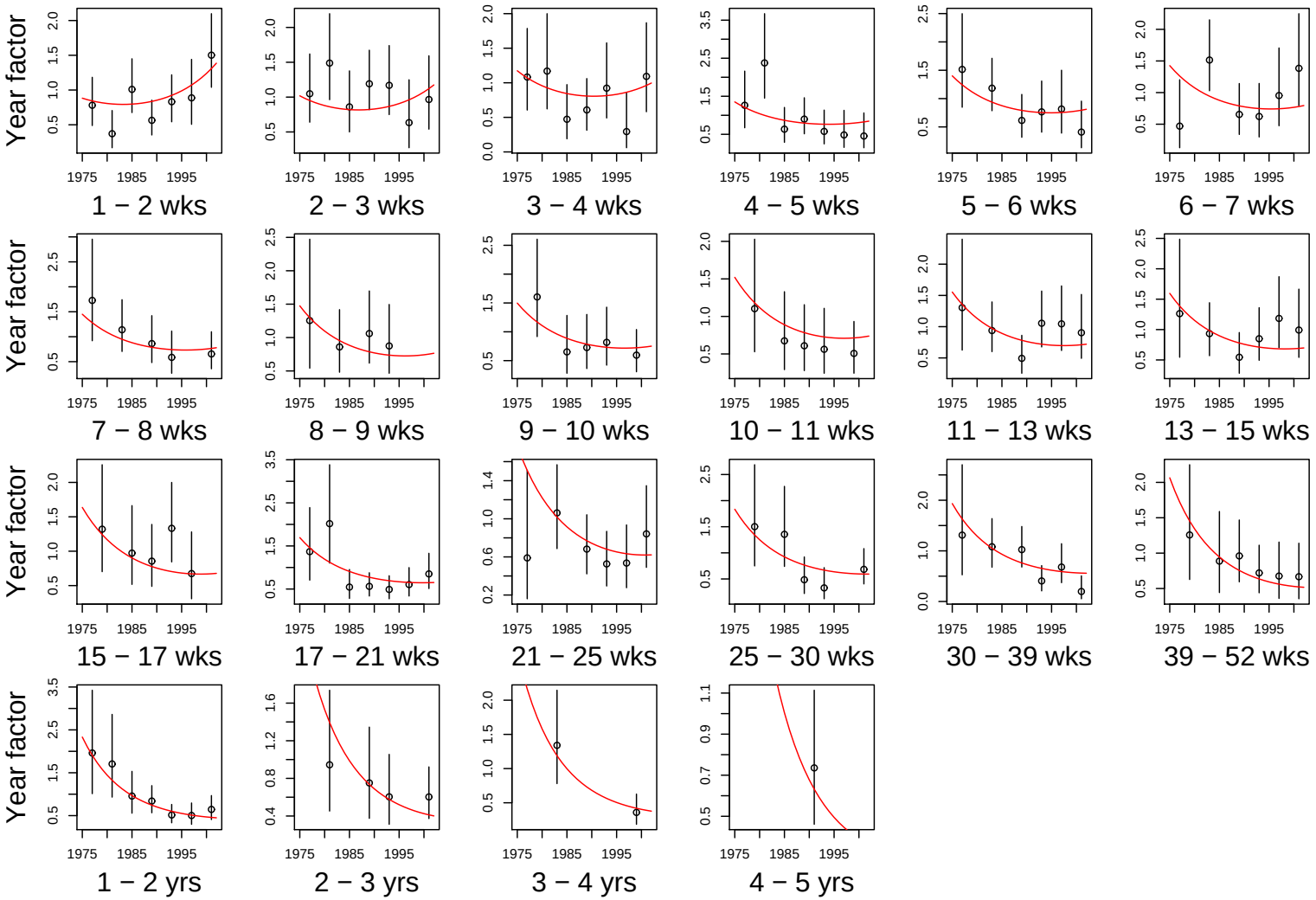


Figure 3.39: The "partial residual effects" plots for year – cs61.

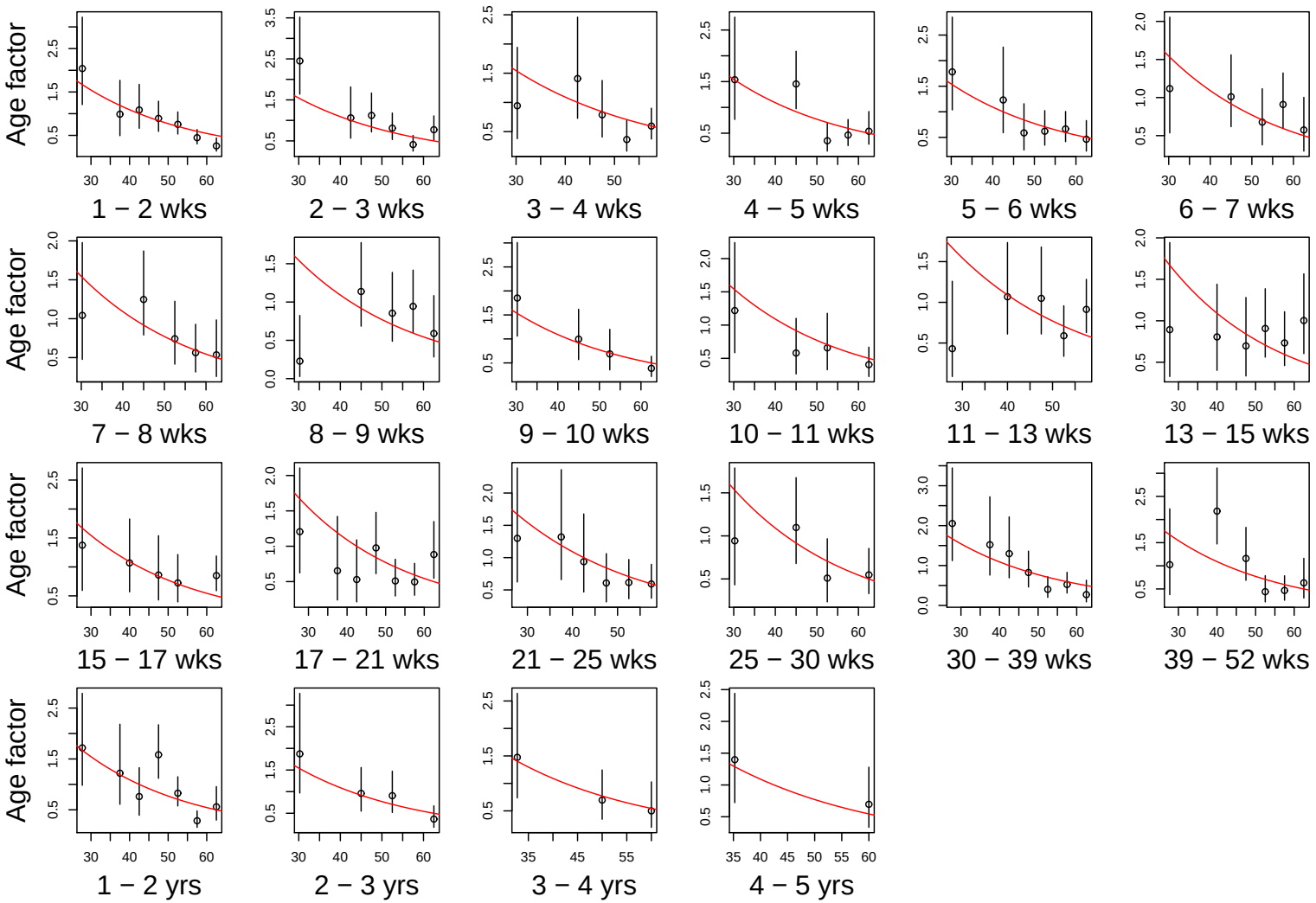


Figure 3.40: The "partial residual effects" plots for age – cs61.

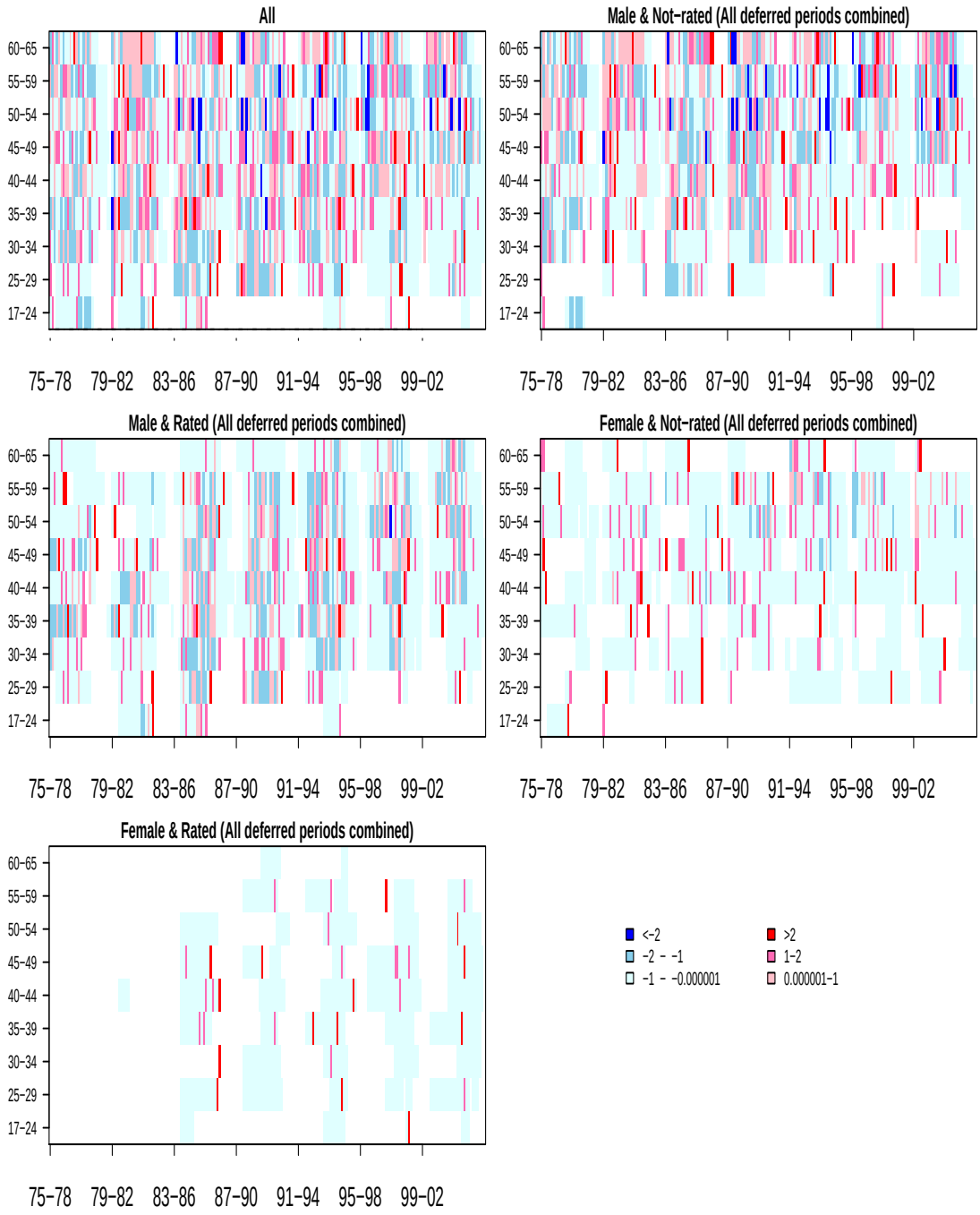


Figure 3.41: Two-dimensional plots of deviance residuals for cs61 data.

Recovery intensity model for cs62

For cs62, Table 3.14 shows, in columns two to four, the results from Models I, II and III, respectively. Model I is the Cox model which assumes that all covariate effects stay constant for all sickness durations. This assumption is not valid for $x_{\text{year}}, x_{\text{rated}}, x_{\text{dp4}}, x_{\text{dp13}}$ and x_{dp26} because the smoothed plots of their scaled Schoenfeld residuals as shown in Figure 3.42 suggests that their respective coefficients are duration dependent and can be parameterised as

$$\begin{aligned}
\beta_{\text{year}}(z) &= \alpha_{\text{year}} + \gamma_{\text{year}}(\tau_{\text{year}} - z)_+ + \theta_{\text{year}} C_1(t_{1.3}(z)) \\
\beta_{\text{rated}}(z) &= \alpha_{\text{rated}} + \phi_{\text{rated}_1}(C_1(t_{1.3}(z)) - C_1(t_{1.3}(\tau_{\text{rated}})))_+ \\
&\quad + \phi_{\text{rated}_2}(C_1(t_{1.3}(z)) - C_1(t_{1.3}(\tau_{\text{rated}})))_+^2 \\
\beta_{\text{dp4}}(z) &= \alpha_{\text{dp4}} + \gamma_{\text{dp4}_1}(\tau_{\text{dp4}_1} - z)_+ + \gamma_{\text{dp4}_2}(\tau_{\text{dp4}_2} - z)_+ \\
\beta_{\text{dp13}}(z) &= \alpha_{\text{dp13}} + \gamma_{\text{dp13}}(\tau_{\text{dp13}} - z)_+ \\
\beta_{\text{dp26}}(z) &= \alpha_{\text{dp26}} + \gamma_{\text{dp26}}(\tau_{\text{dp26}} - z)_+
\end{aligned}$$

where $\tau_{\text{year}} = 30.5, \tau_{\text{rated}} = 206.5, \tau_{\text{dp4}_1} = 40.5, \tau_{\text{dp4}_2} = 72.5, \tau_{\text{dp13}} = 131.5$ and $\tau_{\text{dp26}} = 206.5$

Model II is obtained by including in Model I additional covariates created to describe the duration-varying effects of $x_{\text{year}}, x_{\text{rated}}, x_{\text{dp4}}, x_{\text{dp13}}$ and x_{dp26}. Lastly, Model III is the fully parameterised recovery intensity model with the baseline intensity described by a $\exp(\sum_{i=0}^{3} b_i C_i(t_{1.3}(z)))$ formula. This final fitted model is then assessed for its goodness-of-fit. The value of the χ^2 statistic is 805.1444. With 866 cells and 31 parameters fitted in the model, the probability value is 0.7652 on 835 degrees of freedom, indicating a reasonably good fit to the data. The "residual effects" and "partial residual effects" plots for the discrete covariates are presented in Figure 3.43. The data upon which each plot is based on is indicated by the caption of the plot. Figures 3.44 and 3.45 show the "partial residual effects" plots for age and year, respectively, for each sickness duration band. The red curves overlaid on the plots in Figures 3.43–3.45 are the estimated smooth intensities according to Model III which

have provided a good fit to the actual experience. The two-dimensional plots of the deviance residuals for all data as well as its subset are shown in Figure 3.46. Apart from 'Female & Rated' data for which the data is sparse, the positive and negative deviance residuals for other subsets of data are roughly randomly scattered.

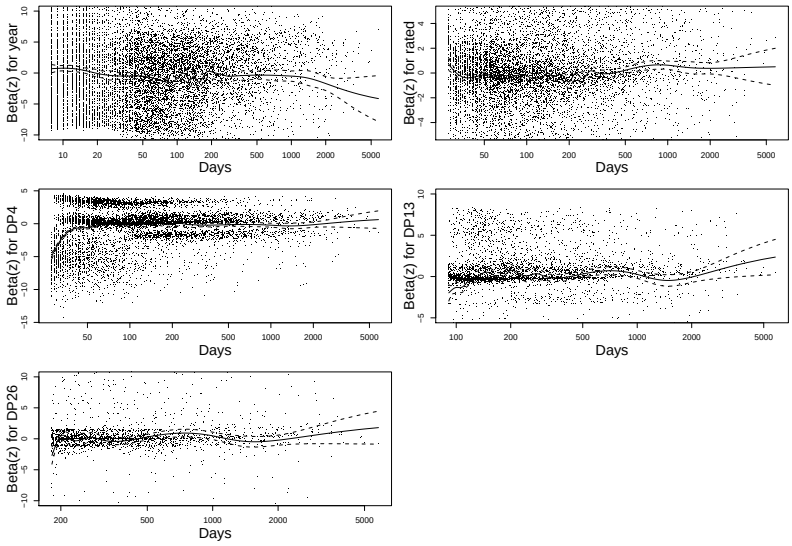
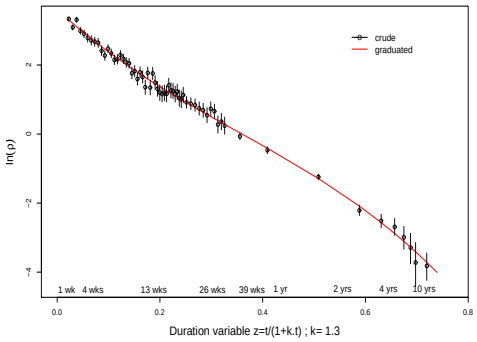


Figure 3.42: The smoothed plots of the scaled Schoenfeld residuals for $x_{\text{year}}, x_{\text{rated}}, x_{\text{dp4}}, x_{\text{dp13}}$ and x_{dp26} in cs62.

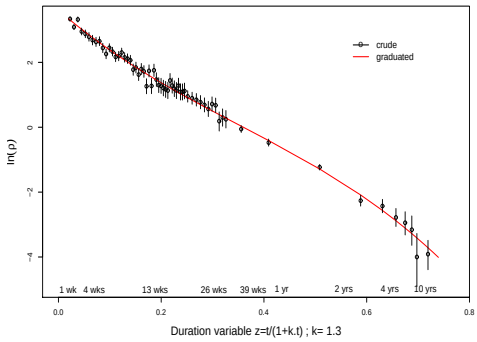
Table 3.14: Parameters in the intermediate and final recovery intensity models for cs62.

	Model I	Model II	Model III
k			1.3
b_0			11.3558
			(1.2348)
b_1			-24.7647
			(2.4823)
b_2			7.7475
			(1.2102)
b_3			-3.6943
			(0.6415)
α_{sex}	-0.2179	-0.2402	-0.2375
	(0.0352)	(0.0352)	(0.0352)
α_{age}	-0.2245	-0.2581	-0.2574
	(0.1100)	(0.1102)	(0.1100)
α_{age2}	0.1137	0.0683	0.0704
	(0.0958)	(0.0962)	(0.0962)
α_{age3}	0.1387	0.1477	0.1445
	(0.0556)	(0.0559)	(0.0558)
α_{age4}	0.1560	0.1226	0.1231
	(0.0418)	(0.0420)	(0.0420)
α_{rated}	-0.2151	-0.2969	-0.2852
	(0.0349)	(0.0382)	(0.0379)
τ_{rated}		206.5	206.5
ϕ_{rated_1}		5.9747	5.3474
		(0.9070)	(0.7839)
ϕ_{rated_2}		-13.3007	-11.2222
		(3.1531)	(2.7736)
α_{year}	-0.3237	-0.2868	-0.2713
	(0.0783)	(0.0467)	(0.0466)
θ_{year}		-0.9224	-0.9608
		(0.1758)	(0.1753)
τ_{year}		30.5	30.5
γ_{year}		0.0369	0.0345
		(0.0032)	(0.0031)
α_{year2}	-0.0447	-0.0224	-0.0220
	(0.0271)	(0.0273)	(0.0273)

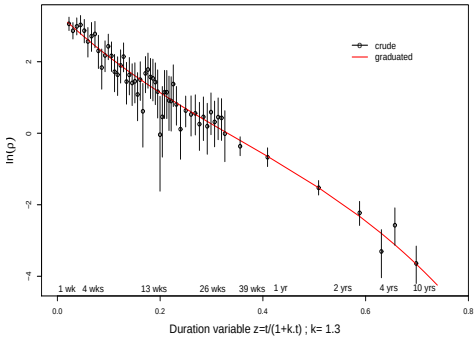
	Model I	Model II	Model III
α_{dp4}	-0.4200	0.1026	0.1096
	(0.0340)	(0.0337)	(0.0307)
τ_{dp4_1}		40.5	40.5
γ_{dp4_1}		-0.1028	-0.1021
		(0.0129)	(0.0106)
τ_{dp4_2}		72.5	72.5
γ_{dp4_2}		-0.0237	-0.0242
		(0.0023)	(0.0016)
α_{dp13}	-0.6116		
	(0.0437)		
τ_{dp13}		131.5	131.5
γ_{dp13}		-0.0358	-0.0362
		(0.0036)	(0.0033)
α_{dp26}	-0.7504	-0.1379	-0.1652
	(0.0646)	(0.0622)	(0.0612)
τ_{dp26}		206.5	206.5
γ_{dp26}		-0.0493	-0.0430
		(0.0168)	(0.0163)
α_{dp52}	-1.0478	-0.4310	-0.4208
	(0.1238)	(0.1221)	(0.1211)
$\alpha_{\text{dp4:age}}$	-0.3991	-0.4090	-0.4048
	(0.1722)	(0.1724)	(0.1723)
$\alpha_{\text{dp13:age}}$	-0.7739	-0.8002	-0.8077
	(0.2271)	(0.2279)	(0.2288)
$\alpha_{\text{dp26:age}}$	-0.8927	-0.7682	-0.7695
	(0.1497)	(0.1485)	(0.1483)
$\alpha_{\text{dp4:age3}}$	-0.2303	-0.2626	-0.2575
	(0.0867)	(0.0870)	(0.0869)
$\alpha_{\text{dp13:age3}}$	-0.2362	-0.2988	-0.2956
	(0.1175)	(0.1180)	(0.1185)
$\alpha_{\text{dp52:age3}}$	0.6734	0.5748	0.5760
	(0.1677)	(0.1675)	(0.1675)
$\alpha_{\text{sex:year}}$	0.1232	0.1985	0.1988
	(0.0685)	(0.0687)	(0.0687)
$\alpha_{\text{rated:year2}}$	0.1041	0.1942	0.1940
	(0.0429)	(0.0441)	(0.0441)



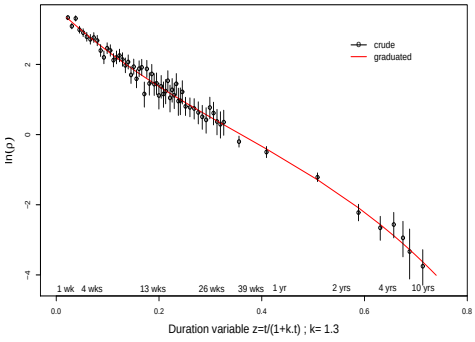
3.43.1: All



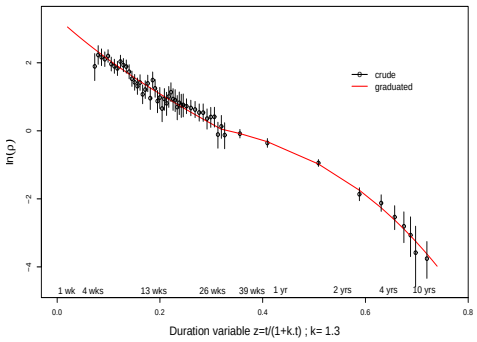
3.43.2: Male



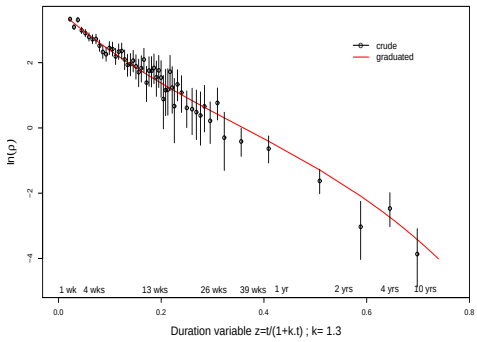
3.43.3: Female



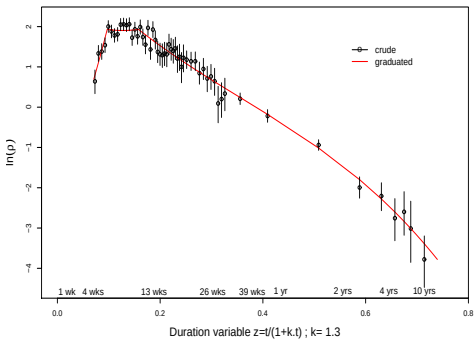
3.43.4: Non-rated



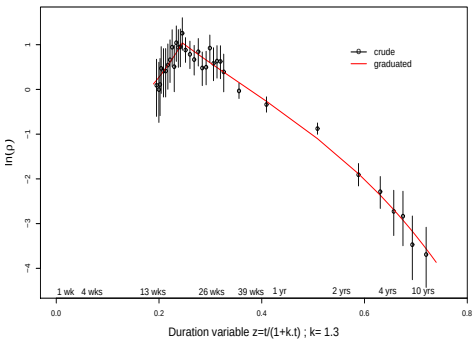
3.43.5: Rated



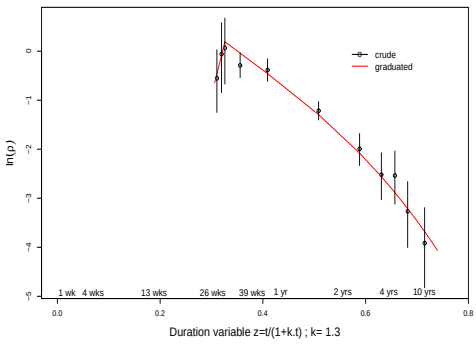
3.43.6: DP1



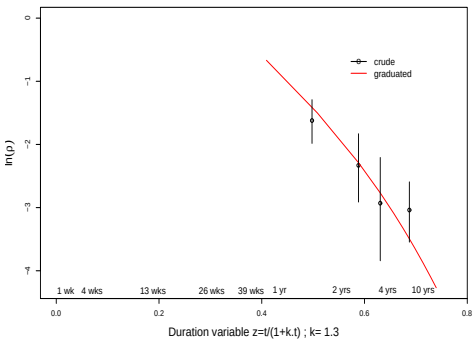
3.43.7: DP4



3.43.8: DP13



3.43.9: DP26



3.43.10: DP52

Figure 3.43: The "residual effects" and "partial residual effects" plots for the discrete covariates – cs62.

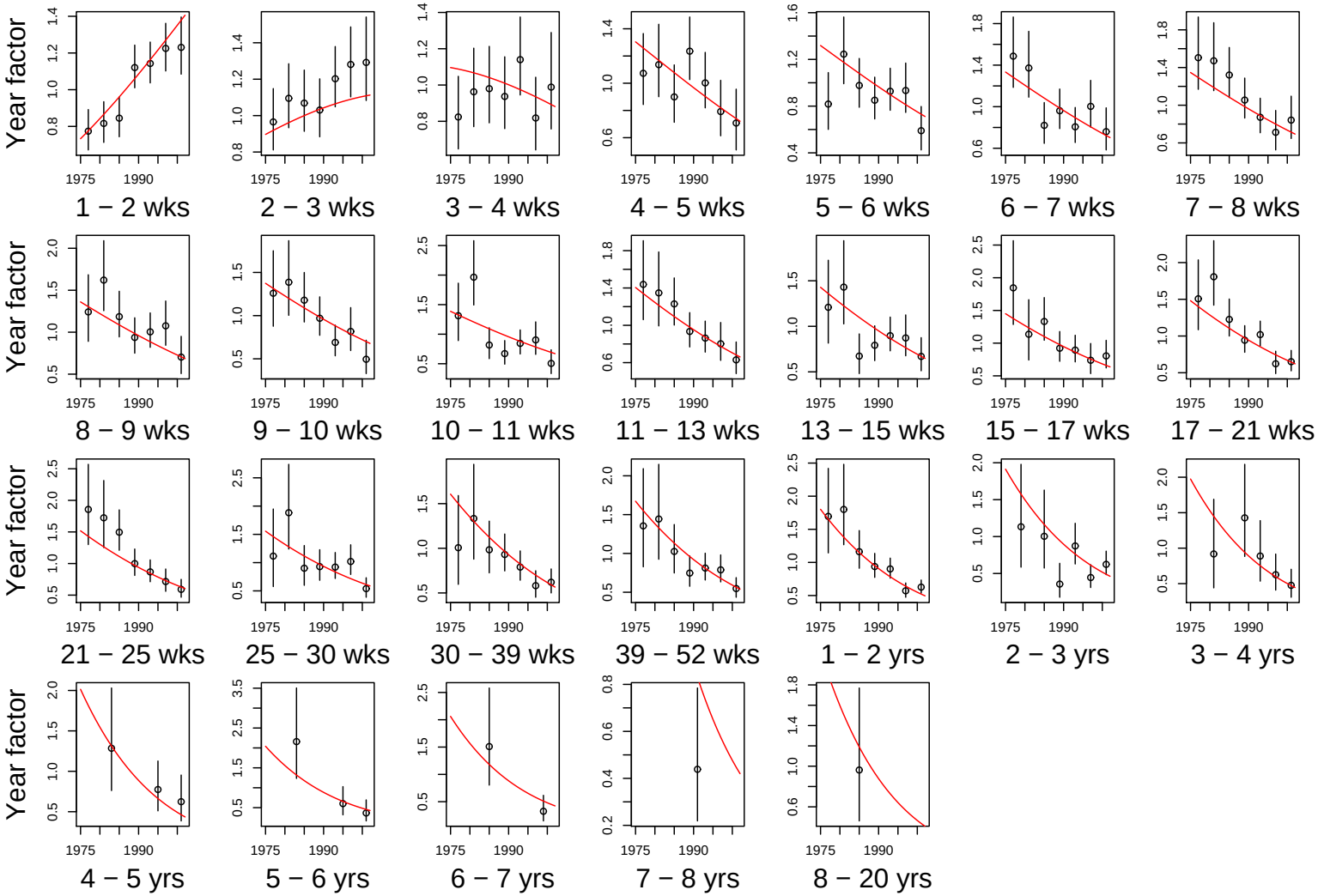


Figure 3.44: The "partial residual effects" plots for year – cs62.

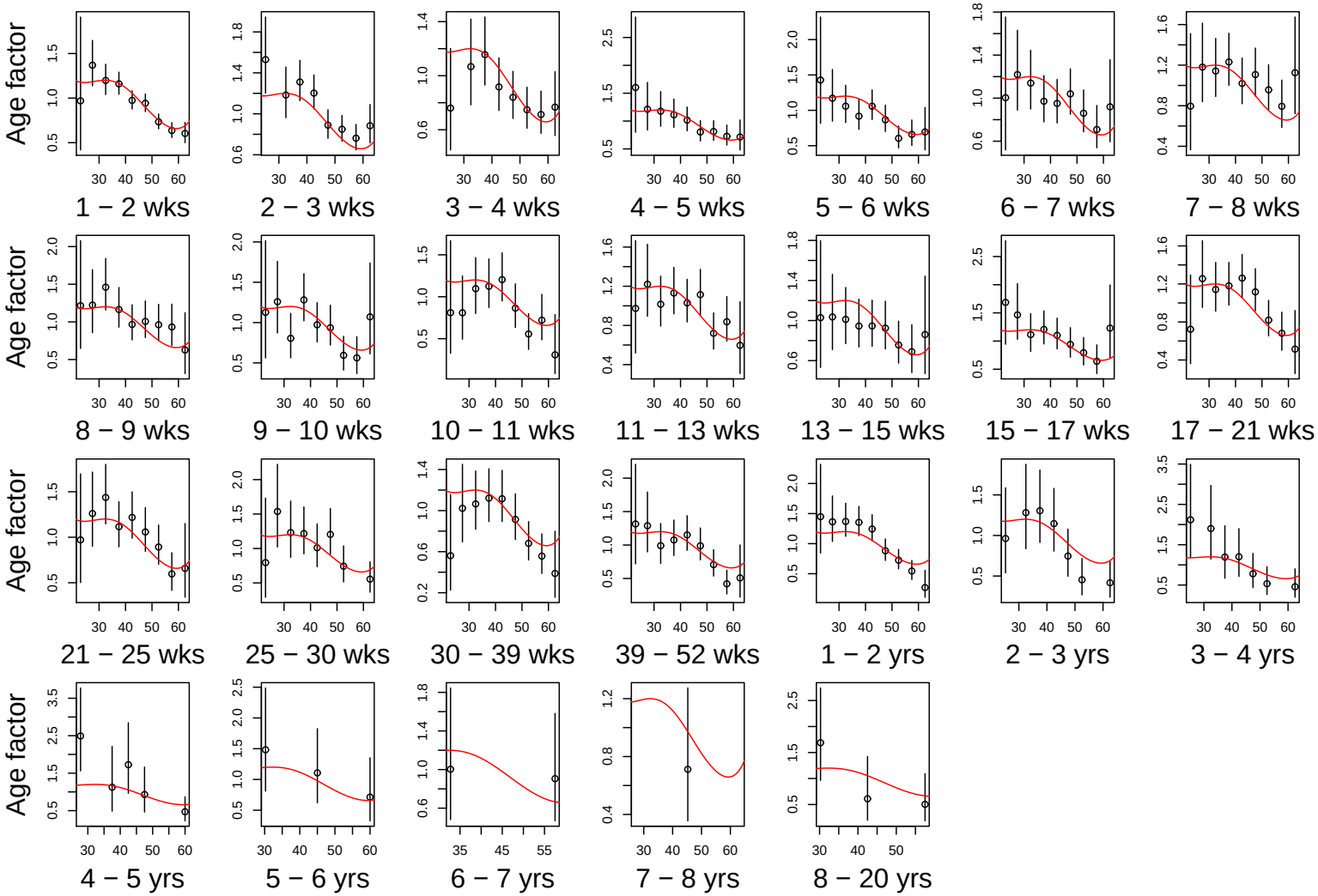


Figure 3.45: The "partial residual effects" plots for age – cs62.

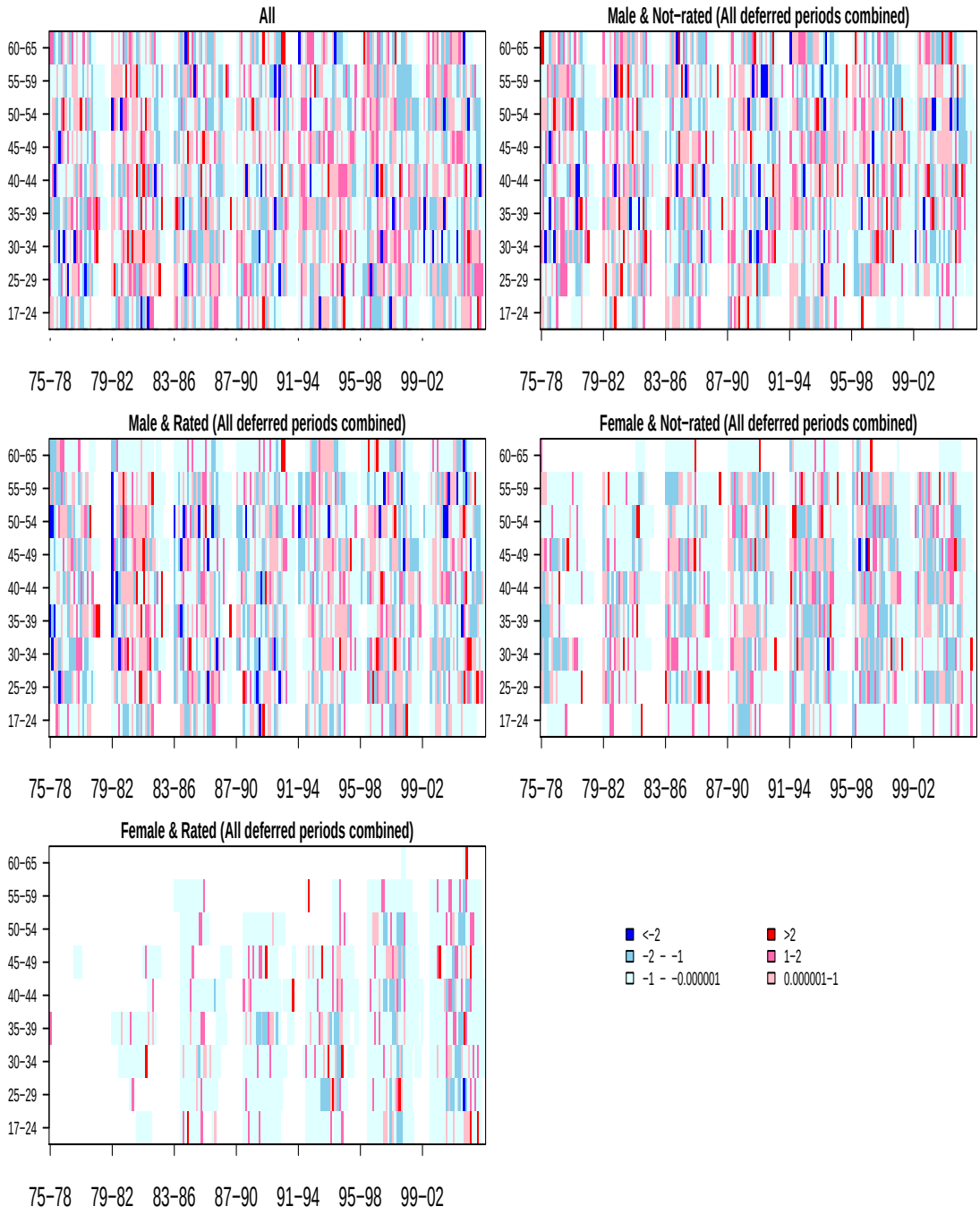


Figure 3.46: Two-dimensional plots of deviance residuals for cs62 data.

Chapter 4

Modelling of the Mortality Intensity from Sick I: Base Mortality Intensity

4.1 Introduction

This and the following chapter are dedicated to the modelling of the mortality intensity among UK IPI claimants from sick. The modelling of the mortality intensity among UK IPI claimants has been attempted before but on a smaller set of data than the one we are using. Examples include CMI Report 12 (1991), Renshaw and Haberman (1995, 2000) and CMI Working Paper 5 (2004). The latter two will be discussed in greater detail because both these studies looked at sickness data from a longer period of time and presented some interesting findings.

Renshaw and Haberman (2000) modelled the mortality intensity from sick among IPI claimants using data from 1975 to 1994 inclusive for all causes of sickness combined. They incorporated sex, sickness duration, age and calendar year as covariates in their multiplicative mortality model and fitted the model using a GLM with Poisson error structure and log link. The main results drawn from their mortality modelling exercise are as follows:

(a) Sickness duration is the dominant explanatory variable, accounting for the

biggest drop in the residual deviance. At the exploratory analysis phase, sickness duration is discretised into 11 sickness duration intervals and included in the model as categorical factors. Graphical examination of these duration-dependent factors shows that the effect of sickness duration on the mortality intensity follows a bell-shaped curve with the mortality intensity rising rapidly, peaking at around 26 weeks and falling off until it shows a slight up-turn at very long term sickness durations. However, the factors for these long term sickness duration intervals are only of marginal statistical significance. In the final model adopted, sickness duration is modelled as a continuous covariate by using piecewise linear splines with 4 knots.

(b) Age is incorporated as a linear term in the model with a positive regression coefficient, implying that the mortality intensity increases log linearly with age.

(c) Year is incorporated as a linear term in the model with a negative regression coefficient, implying that the mortality intensity declines (or improves) over the 20 years from 1975 to 1994.

(d) Females experienced a lower mortality intensity than males.

(e) There are no interaction terms between any of the covariates.

CMI Working Paper 5 (2004) presented the graduation of the mortality experience for IPI male claimants from occupational class 1 from 1991 to 1998 inclusive. Despite the data spanning eight years, the year effect is not included in the model because preliminary investigation shows that the sickness experience from 1991 to 1998 as a whole was considered rather homogeneous. A multiplicative model with attained age and sickness duration as explanatory variables was first investigated. As in Renshaw and Haberman (2000), it is of interest to note that the duration-related factors, examined when a multiplicative mortality intensity model was considered, has a bell-curved shape, with intensity rising, peaking between 18 and 22 weeks, falling off before showing an upturn at the very last interval representing sickness duration exceeding 8 years. The authors attribute this eventual up-turn to a more dominant ageing process taking effect. However, given a high χ^2 test value that indicates a poor

fit of this multiplicative model, it was abandoned in favour of an additive model. The new proposed mortality intensity model consists of two additive components. The first component is a Weibull formula, a hump-backed function of sickness duration z, designed to capture the variation of the mortality intensity with sickness duration, capped at 5 years for practical reasons. The second component is a Gompertz formula, a function of attained age alone. It was introduced in recognition of an inevitable increase in the mortality intensity with increasing attained age, which is not achievable with the first component alone. At a conceptual level, this mortality model resembles the Pollard-Heligman model (Heligman & Pollard, 1980) which has three components to deal with three phases of age-related mortality change. The final graduation formula is presented in Appendix D.

Let $\mathbf{x}$ be a covariate vector consisting of sex, age, year, rating indicator and deferred period. The notation representing these covariates and their coding are no different to those in the recovery intensity model (see Section 3.2) with the exception that the age in the mortality analysis is defined as attained age and not age at sickness inception as in the case of the recovery intensity model. The mortality intensity from sick of an individual (with covariates $\mathbf{x}$) who has been ill or disabled with a specific cause for duration z, $\lambda(z, \mathbf{x})$, is therefore modelled as

$$\lambda(z, \mathbf{x}) = \lambda^*(\mathbf{x}) + \nu(z, \mathbf{x}) \tag{4.1}$$

where $\lambda^*(\mathbf{x})$ is a base intensity that varies by sex, attained age and calendar year while $\nu(z, \mathbf{x})$ denotes the 'excess' mortality incurred from being sick with a specific cause for duration z. This 'excess' mortality can be interpreted as the mortality in excess of that experienced by a comparable population as a result of being sick for duration z. The base mortality intensity in Equation (4.1), as with the Gompertz formula in CMI Working Paper 5 (2004), provides a means to counteract the eventual fall in excess mortality at long term sickness duration. In the case of IPI claimants, we regard the UK assured lives population as a reasonably comparable group from which the base mortality, an increasing function of age, can be estimated. Unlike in CMI Working Paper 5 (2004), we do not derive this base mortality from the IPI data itself because most recoveries or deaths occur within the first few years of sickness durations, leaving

insufficient sickness data remaining at long term sickness durations at an individual cause of sickness level to make the modelling of a monotonic increasing function of age meaningful.

A mortality model that takes into account the mortality intensity of a comparable population has widespread application in cancer research and is often referred to as the 'relative survival model' in the medical statistics literature (see Dickman *et al*, 2004). There are two main classes of relative survival models. Apart from the additive mortality intensity model presented in Equation (4.1), there is the multiplicative mortality intensity model where the intensity function is the product of the 'standard' mortality and a risk factor (relative mortality) due to a particular sickness (see Andersen *et al*, 1985). Such a multiplicative model has been attempted in actuarial work. Renshaw (1988) introduced a mortality model for impaired lives with hypertension in which the impairment has a multiplicative effect on a standard mortality intensity obtained by a suitable transformation of the A1967-70 standard table. Haberman and Renshaw (1990) adopted similar approach in assessing the excess mortality among insured lives who have peptic ulcer.

The advantage of using the additive mortality intensity model in Equation (4.1) over the direct cause-specific mortality modelling in the context of cancer research is discussed in Dickman *et al* (2004). They remarked that the modelling of a cause-specific mortality intensity requires accurate information regarding the cause of death and in cancer research, such information can be unreliable when a death is due to cancer spreading to several body organs (metastasisation). With respect to IPI data, unreliability in using cause of sickness as cause of death can occur. It is usually the case that the cause of sickness recorded in the claim data remain unchanged even if the claimant eventually dies from a sickness different from it. For example, an IPI claimant who later dies from cerebrovascular disease even though the sickness that gave rise to his/her IPI claim is mental illness. Thus, Equation (4.1) provides a convenient tool to gauge the excess mortality of a sickness, regardless of whether the death is directly or indirectly related to the sickness in question.

The IPI data that we are given covers the calendar years from 1975 to 2002 during which considerable changes in the mortality of assured lives took place. In this

chapter, we focus on the modelling of assured lives mortality intensities during these years for males and females separately. In Section 4.2, we describe the assured lives data set for males and females. In Section 4.3, we present a structured approach to incorporate both age and calendar year effect in a mortality model. Such methodology is then implemented on the male and female assured lives data sets separately, the results of which are presented in Sections 4.4 and 4.5, respectively.

4.2 Data

The assured lives data set we used is provided by the CMI. The data is for all durations (0, 1 and 2+) amalgamated together and is cross-classified by

- sex : male, female
- age nearest birthday (x) : 20, 21, 22,. . . , 90
- calendar year (y): 1975, 1976, . . . , 2003 for male
  : 1983, 1984, . . . , 2003 for female

Based on this cross-classification, separately for males and females, we have values for the number of deaths (d_{xy}) at age x in calendar year y and its matching central exposed-to-risk (r_{xy}). Figures 4.1 and 4.2 present the crude mortality intensities for males and females respectively plotted on the log scale against calendar year. These graphs show a downward trend in both male and female mortality intensity at these equally spaced ages but the slope with which the mortality intensities decline is not uniform across these ages.

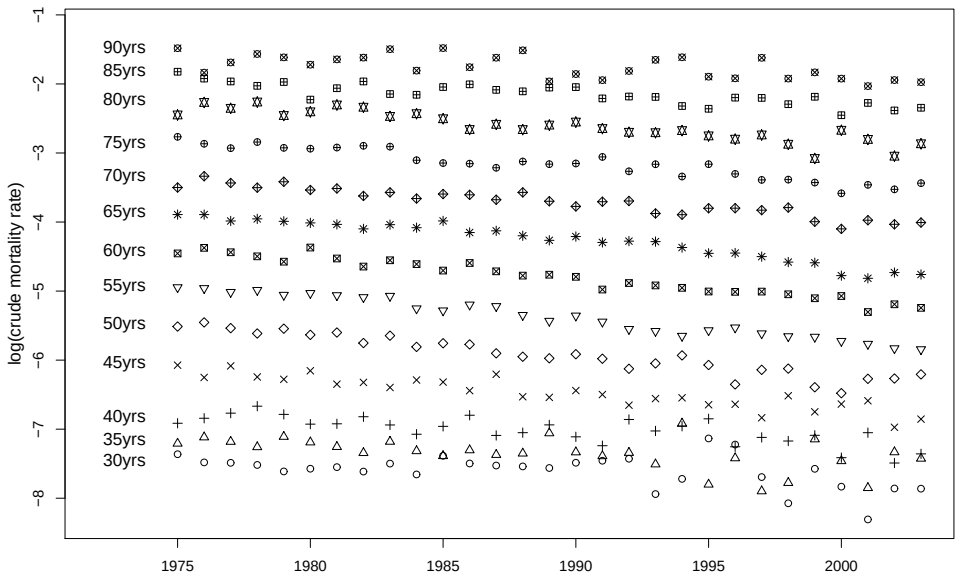


Figure 4.1: Crude mortality intensities *vs* calendar year for male assured lives plotted on a log scale for ages 30, 35, . . ., 90.

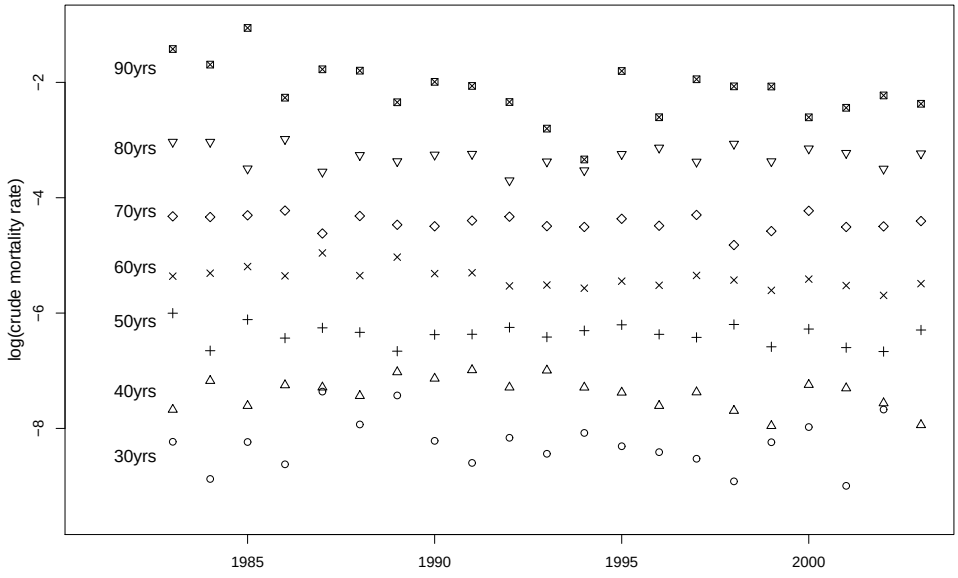


Figure 4.2: Crude mortality intensities *vs* calendar year for female assured lives plotted on a log scale for ages 30, 40, . . ., 90.

4.3 Modelling Techniques

We will model the mortality intensity from sick for males and females separately. Let $\lambda_{x,y}$ denote the mortality intensity at age x in calendar year y. For computational stability, we re-scale both age (x) and year (y) such that

$$x' = \frac{x - 55}{35}, \quad y' = \frac{y - 1989}{14} \quad .$$

We first propose a general model for $\lambda_{x,y}$ as

$$\lambda_{x,y} = \lambda_0(x') \exp(\sum_{i=1}^{n} \alpha_i(x') C_i(y')) \tag{4.2}$$

where $\lambda_0(x')$ is the baseline intensity that depends only on age, while $\exp(\sum_{i=1}^{n} \alpha_i(x') C_i(y'))$ is the relative risk for an individual in year y, when compared to the baseline intensity. In this relative risk component, $C_i(y')$ denotes the Chebycheff polymonial of degree i while $\alpha_i(x')$ is the age-dependent coefficient of the year effect. The downward trend in the UK assured lives mortality intensity over the years is not uniform across all ages and to take account of this variation in year trend, we incorporate an age-dependent year coefficient, with the following structure:

$$\alpha_i(x') = \sum_{j=1}^{p} \psi_{ij}(x') \beta_{ij} \tag{4.3}$$

where $\psi_{ij}(x')$ denotes the age-related term and β_{ij} denotes its regression coefficient.

We will first make an assumption that the baseline mortality intensity between integer age x and $x + 1$ is constant and is denoted by λ_x. The mortality intensity of an individual whose age lies in $[x, x + 1)$ is

$$\lambda_{xy} = \lambda_x \exp(\sum_{i=1}^{n} \alpha_i(x') C_i(y')) \quad . \tag{4.4}$$

Taking logs on both sides of Equation (4.4), we obtain the following additive log-linear model

$$\log(\lambda_{xy}) = \log(\lambda_x) + \sum_{i=1}^{n} \alpha_i(x') C_i(y') \quad . \tag{4.5}$$

Following this piece-wise constant baseline intensity assumption, the likelihood of the data is given by

$$L = \prod_{x,y} \exp(-r_{xy}\lambda_{xy})\lambda_{xy}^{d_{xy}} \quad . \tag{4.6}$$

Taking logs on both sides of Equation (4.6), the log likelihood of the data is given by

$$\log L = \sum_{x,y} \left(-r_{xy}\lambda_{xy} + d_{xy}\log(\lambda_{xy})\right) \quad . \tag{4.7}$$

Then Equation (4.7) can be combined with Equation (4.4) to give

$$\log L = \sum_{x,y}(-r_{xy}\lambda_x \exp(\sum_{i=1}^{n} \alpha_i(x')C_i(y')) + d_{xy}\log(\lambda_x) + d_{xy}\sum_{i}^{n} \alpha_i(x')C_i(y')) \quad . \tag{4.8}$$

Differentiating Equation (4.8) with respect to λ_x and setting it equal to zero gives

$$\hat{\lambda}_x = \frac{\sum_y d_{xy}}{\sum_y r_{xy}\exp(\sum_{i=1}^{n} \alpha_i(x')C_i(y'))} \tag{4.9}$$

which is the maximum likelihood estimate for λ_x and can be substituted into Equation (4.8) to yield a likelihood that does not depend on the baseline intensity given by

$$\log L = \sum_x \sum_y d_{xy}(\sum_{i=1}^{n} \alpha_i(x')C_i(y')) - \sum_x (\sum_y d_{xy}) \log(\sum_y r_{xy}\exp(\sum_{i=1}^{n} \alpha_i(x')C_i(y'))) + C \tag{4.10}$$

where C is a term independent of any parameter of interest. By exponentiating Equation (4.10), we obtain

$$L = \prod_x \prod_y \left(\frac{\exp(\sum_{i=1}^{n} \alpha_i(x')C_i(y'))}{\sum_y r_{xy}\exp(\sum_{i=1}^{n} \alpha_i(x')C_i(y'))} \right)^{d_{xy}} \quad . \tag{4.11}$$

The likelihood given in Equation (4.11) is the grouped data version of Cox's partial likelihood as given in Equation (3.12). To estimate all the parameters in Equation (4.4), we can obtain $\widehat{\boldsymbol{\beta}}$ from maximising Equation (4.11) and then obtain $\widehat{\boldsymbol{\lambda}}_x$ by substituting $\widehat{\boldsymbol{\beta}}$ into Equation (4.9).

Holford (1980) and Laird and Oliver (1981) show an alternative method to estimate the parameters by noting that the piece-wise constant baseline intensity model (Equation (4.4)) will produce the same estimates as the Poisson regression model because the likelihoods obtained from both approaches are proportional to each other. As such, we treat the observed number of deaths d_{xy} at age x in calendar year y as the realisation of a Poisson random variable, D_{xy}, with mean and variance

$$\mathrm{E}(D_{xy}) = \mu_{xy} = r_{xy}\lambda_{xy}, \quad \mathrm{Var}(D_{xy}) = \mu_{xy} \quad . \tag{4.12}$$

Substituting Equation (4.4) into Equation (4.12) and taking logs on both sides of the equation, we obtain

$$\log(\mu_{xy}) = \log(r_{xy}) + \log(\lambda_x) + \sum_{i=1}^{n} \alpha_i(x')C_i(y') \quad . \tag{4.13}$$

We can estimate all the parameters in Equation (4.13) using a GLM, with a Poisson error structure, log link and $\log(r_{xy})$ as an offset term. The advantage of using this approach is that we can use the convenience of the GLM framework, which is available in R, to do the estimation.

The use of GLMs in actuarial applications, including the modelling of mortality, is not unusual (see Renshaw (1991), Haberman and Renshaw (1996)). In Renshaw *et al* (1996), the mortality experience for male assured lives for duration 5 years and over, from 1958 to 1990 inclusive, is analysed using a log linear Poisson regression model in a GLM. In their approach, the multiplicative structures for both age and calendar year are modelled simultaneously, followed by an age dependent trend adjustment through the inclusion of interaction terms between age and calendar year. The approach described in this section provides an alternative way to utilise the convenience of the GLM framework to estimate the mortality intensity model in a structured manner. The general procedure is as follows:

(i) By treating the baseline intensity as piece-wise constant so that parameterisation of the baseline intensity is not necessary, we can model and estimate the parameters in the relative risk component of Equation (4.2). Any age dependent trend adjustment is included at this stage.

(ii) We choose to use a GM(r, s) structure for the baseline intensity, which is defined as follows:

$$\mathrm{GM}_{a,b}^{r,s}(x) = \sum_{i=1}^{r} a_i x^{i-1} + \exp(\sum_{i=1}^{s} b_i x^{i-1}) \quad . \tag{4.14}$$

With a GM(r, s) formula imposed on the baseline intensity, the likelihood of the data is maximised by fixing the regression coefficients in the relative risk component at the values found in (i). The optimal GM(r, s) is then chosen by examining the log likelihood values from fitting various GM(r, s) formulae.

(iii) Following (i) and (ii), we will have a complete structure for Equation (4.2). Parameter values can be determined using maximum likelihood estimation.

The results of the implementation of such an approach on both male and female assured lives data sets separately are presented in Sections 4.4 and 4.5.

4.4 Mortality Model for Male Assured lives

In this section, we focus on the modelling of the mortality experience for male assured lives aged between 20 and 90 from year 1975 to 2003 inclusive using the techniques in Section 4.3. We take as our starting point the model structure

$$\log \lambda_{xy} = h_x + \sum_{i=1}^{n} (\beta_i + \beta_{ix}) C_i(y'), \quad \text{for} \quad x = 20, 21, \ldots, 90 \tag{4.15}$$

subject to the constraints

$$\beta_{i20} = 0, \quad \text{for} \quad i = 1, 2, \ldots, n$$

where h_x is a factor indicating the constant baseline intensity and the coefficients of year-related terms are allowed to vary by every single age. In the same spirit as the Lee-Carter model (Lee and Carter, 1992), h_x is the average age-specific pattern of mortality, $C_i(y')$ $(i = 1, \ldots, n)$ captures the main year trend on the logarithmic scale in mortality intensities at all ages while β_{ix} measures the age-specific deviation from the main year trend.

Table 4.1 shows the order in which various sets of parameters in Equation (4.15) are fitted sequentially (column one), the model deviances with additional parameter(s) (column two), the number of degrees of freedom obtained after subtracting the number of parameters fitted from 2059 (=29 $\times$ 71) (column three) and the AIC (Akaike, 1974) value for each model (column four). Columns five and six of Table 4.1 are constructed by calculating the differences in deviance and degrees of freedom respectively, as a result of introducing additional parameter(s) sequentially into the model.

The model with the lowest AIC value is selected. As a result, both β_5 and β_{5x} will be excluded from our model because inclusion of these parameters increases the value of the AIC. Therefore, the model structure we have chosen is

$$\log \lambda_{xy} = h_x + \sum_{i=1}^{4} (\beta_i + \beta_{ix}) C_i(y'), \quad \text{for} \quad x = 20, 21, \ldots, 90. \tag{4.16}$$

subject to the constraints

$$\beta_{i20} = 0, \quad \text{for} \quad i = 1, 2, 3, 4.$$

This model structure involves a total of 355 parameters, entailing the need for a more parsimonious model. In the search for a suitable functional form for the age-dependent year coefficients, we plotted each of the resulting sets of parameter estimates $\{\beta_{1x}\}, \{\beta_{2x}\}, \{\beta_{3x}\}, \{\beta_{4x}\}$ (Equation (4.16)) against x. These four plots are produced in Figure 4.3. We will use a natural cubic spline basis, due to its flexibility, to model the age-dependent year coefficient. Consider k fixed knots $\tau_1, \ldots, \tau_k$ and let

$$(x' - \tau_j)_+ = \begin{cases} (x' - \tau_j) & \text{if } x' \geq \tau_j \\ 0 & \text{otherwise} \end{cases} .$$

The age-dependent coefficient of $C_i(y')$, $\alpha_i(x')$, can be represented by

$$\alpha_i(x') = \beta_i + \gamma_{i1}x' + \sum_{j=1}^{j=k-2} \gamma_{ij} w_j(x') \tag{4.17}$$

where, as given by Devlin and Weeks (1986),

$$w_j(x') = (x' - \tau_j)_+^3 - (x' - \tau_{k-1})_+^3 \frac{(\tau_k - \tau_j)}{(\tau_k - \tau_{k-1})} + (x' - \tau_k)_+^3 \frac{(\tau_{k-1} - \tau_j)}{(\tau_k - \tau_{k-1})} \quad . \tag{4.18}$$

We explore the effect of using different numbers of knots and varying the placement of knots. One method is to use a visual trial and error process. A less *ad hoc* method is to use AIC values to select the optimal number of knots. The number of knots that produced the lowest AIC is chosen. Generally, it is a standard practice to place the knots at fixed quantiles in the data. However, we shall have occasion to resort to using a visual trial and error process to guide the placement of knots.

Table 4.1: Deviance profiles associated with sequential inclusion of parameters in Equation (4.15).

				Difference	
Model terms	Deviance	D.f.	AIC	Deviance	D.f.
h_x	25760	1988	39932		
				20741.2	1
$+ \beta_1$	5018.8	1987	19193		
				1527.2	70
$+ \beta_{1x}$	3491.6	1917	17805		
				64.8	1
$+ \beta_2$	3426.8	1916	17743		
				215.4	70
$+ \beta_{2x}$	3211.4	1846	17667		
				12.5	1
$+ \beta_3$	3198.9	1845	17657		
				193.9	70
$+ \beta_{3x}$	3005.0	1775	17603		
				8.1	1
$+ \beta_4$	2996.9	1774	17597		
				142.6	70
$+ \beta_{4x}$	2854.3	1704	17594		
				0.4	1
$+ \beta_5$	2853.9	1703	17596		
				121.2	70
$+ \beta_{5x}$	2732.7	1633	17615		

First, we fitted models with three to eight knots for $\alpha_1(x')$ in turn. The quantiles of the data where knots are placed are shown in Table 4.2. The AIC values for these models are shown in the upper left corner of Table 4.3. The model with 6 knots for $\alpha_1(x')$ returns the lowest AIC value and is selected. The locations of the 6 knots are 23.5, 36.1, 48.7, 61.3, 73.9 and 86.5.

Following the inclusion of a 6-knots natural cubic spline basis for $\alpha_1(x')$ in the model, we introduce into our model sequentially three to eight knots for $\alpha_2(x')$. The resulting AIC value for each model is shown in the upper right corner of Table 4.3. We settle on a 6-knot natural spline basis since its associated model returns the lowest AIC value. The locations of the 6 knots are 23.5, 36.1, 48.7, 61.3, 73.9 and 86.5.

Focusing on $\alpha_3(x')$, a model with 8-knots for $\alpha_3(x')$ returns the lowest AIC value. However, we reckon that is too many parameters for $\alpha_3(x')$. Hence, we endeavour to achieve a similar reduction in the AIC by using a smaller number of knots placed at locations other than quantiles. After a considerable trial and error process, it transpires that we can use 6 knots located at 21.75, 50.25, 59.75, 69.25, 78.75 and 88.25 to achieve a similar fit.

As for $\alpha_4(x')$, the value of the AIC increases with each additional knot until the number of knots placed reaches 7. We felt that this sudden drop in the AIC is due to additional knots being placed at a narrow cluster of points where a considerable change of shape occurs. Again, we try to see whether we can use the same or fewer number of knots placed at locations other than quantiles to achieve a similar reduction in the AIC. We eventually settle on 6 knots located at 23.5, 45.5, 55, 65, 67.5 and 88.25.

Table 4.2: Location of knots for different numbers of knots.

k	Quantiles							
3				0.10	0.5	0.9		
4				0.05	0.35	0.65	0.95	
5			0.05	0.275	0.5	0.725	0.95	
6		0.05	0.23	0.41	0.59	0.77	0.95	
7		0.025	0.1833	0.3417	0.5	0.6583	0.8167	0.975
8	0.025	0.1607143	0.2964286	0.4321429	0.5678571	0.7035714	0.8392857	0.975

After the number of knots and the placement of knots are decided for $\alpha_i(x')$, $i = 1, \ldots, 4$, Equation (4.16) is updated to read as follows:

$$\log \lambda_{xy} = h_x + \sum_{i=1}^{4} (\beta_i + \gamma_{i1} x' + \sum_{j=2}^{j=5} \gamma_{ij} w_j(x')) C_i(y') \quad . \tag{4.19}$$

Details of the parameter estimates are set out in Table 4.4.

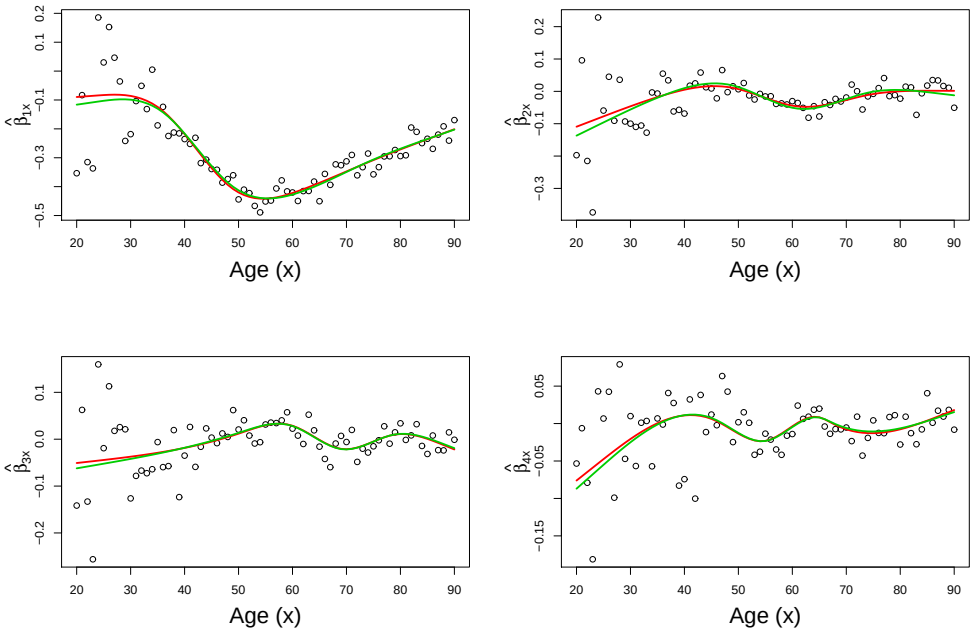


Figure 4.3: Age-varying year coefficients. The superimposed solid red line is obtained from the estimated coefficients in Table 4.4 – piece-wise constant baseline intensity; The superimposed solid green line is obtained from the estimated coefficients in Table 4.6 – GM(0,7) baseline intensity.

Table 4.3: AIC values for differing numbers of knots in each age-dependent year coefficient.

$\alpha_1(x')$		$\alpha_2(x')$	
Number of knots	AIC	Number of knots	AIC
3	17628	3	17612
4	17620	4	17587
5	17592	5	17586
6	17586	6	17566
7	17588	7	17569
8	17588	8	17569

$\alpha_3(x')$		$\alpha_4(x')$	
Number of knots	AIC	Number of knots	AIC
3	17614	3	17588
4	17612	4	17590
5	17591	5	17589
6	17593	6	17590
7	17589	7	17573
8	17578	8	17587

Table 4.4: Parameter estimates in Equation (4.19) with their standard errors.

Symbol	Estimate	Std. Error	Symbol	Estimate	Std. Error
β_1	−0.041694	0.102319	γ_{31}	0.045130	0.055406
β_2	0.114559	0.085673	γ_{32}	0.045598	0.034499
β_3	−0.005700	0.029220	γ_{33}	−1.459982	0.329608
β_4	0.124049	0.034187	γ_{34}	4.293769	0.739948
γ_{11}	0.047987	0.155690	γ_{35}	−5.552429	0.961165
γ_{12}	−1.550461	0.291322	γ_{41}	0.200165	0.057682
γ_{13}	4.816272	0.735973	γ_{42}	−0.261402	0.059528
γ_{14}	−4.822574	0.736087	γ_{43}	2.168820	0.415472
γ_{15}	1.325711	0.559288	γ_{44}	−4.808512	0.877838
γ_{21}	0.223883	0.129901	γ_{45}	11.400437	2.309315
γ_{22}	−0.124763	0.246787			
γ_{23}	−0.343810	0.631599			
γ_{24}	1.790896	0.649017			
γ_{25}	−2.352727	0.503456			

Having found a structure for the relative risk component, we focus on the graduation of the baseline intensity by using a GM(r, s) formula (see Equation (4.14)). By using a GM(r, s) formula to graduate the piece-wise constant mortality intensity, the log likelihood of the data is given by

$$\log L = \sum_x \sum_y -r_{xy}\lambda_{a,b}(x) \exp(\sum_{i=1}^{4} \alpha_i(x')C_i(y')) + d_{xy} \log(\lambda_{a,b}(x)) + d_{xy} \sum_i^4 \alpha_i(x')C_i(y') \tag{4.20}$$

where $\lambda_{a,b}(x)$ is the smooth baseline intensity taking a GM(r, s) strcuture. We first focus on finding a functional form for the baseline hazard by fixing the values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in $\alpha_i(x')$ for $i = 1, 2, 3, 4$ at the estimated regression coefficients shown in Table 4.4. Thus, the only parameters left to estimate are the **a** and **b** in the GM(r, s) formula. We fitted different GM(r, s) formulae in Equation (4.20). Table 4.5 gives twice the difference between the log likelihood from each model and the log likelihood from GM(0,2). The log likelihood is calculated by ignoring the last term in Equation (4.20). Additional parameter(s) are only worthwhile if twice the difference in log likelihood when approximated to a χ^2 distribution with the appropriate degrees of freedom, is statistically significant. With this in mind, we settled on a GM(0,7) formula for the baseline intensity.

Table 4.5: Twice the difference between the log likelihood from each each GM(r, s) formula and the log likelihood from GM(0,2).

r	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$	$s=7$	$s=8$
0	0	838	3115	3489	3588	3609	3610
1	2422	3340	3398	3557	3608	3609	
2	3507	3521	3575	3578	3608		
3	3508	3564	3575	3579			
4	3555	3565	3598				
5	3557	3565					
6	3557						

With a GM(0,7) formula chosen for the baseline hazard, we can update Equation (4.19) to give

$$\log \lambda_{xy} = \sum_{i=0}^{6} b_i C_i(x') + \sum_{i=1}^{4} (\beta_i + \gamma_{i1} x' + \sum_{j=2}^{j=5} \gamma_{ij} w_j(x')) C_i(y') \quad . \tag{4.21}$$

Note that with a GM(0,7) formula for baseline hazard, we can maintain the log-linear property of Equation (4.21) and all the parameters can be estimated in the GLM framework. The parameter estimates and their standard errors are set out in Table 4.6. A comparison between the parameter estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in Table 4.6 and their corresponding values in Table 4.4 reveals no significant discrepancy.

Table 4.6: Parameter estimates in Equation (4.21) with their standard errors.

Symbol	Estimate	Std. Error	Symbol	Estimate	Std. Error
b_0	−5.008173	0.007158	γ_{21}	0.167464	0.123156
b_1	3.112883	0.013179	γ_{22}	0.028993	0.227137
b_2	0.449331	0.010706	γ_{23}	−0.800877	0.573081
b_3	−0.338143	0.008173	γ_{24}	2.362281	0.580483
b_4	0.092917	0.006055	γ_{25}	−2.815544	0.462222
b_5	−0.037030	0.004121	γ_{31}	0.043105	0.055076
b_6	−0.019695	0.003306	γ_{32}	0.044698	0.034306
β_1	−0.088583	0.097105	γ_{33}	−1.420710	0.327058
β_2	0.071923	0.080383	γ_{34}	4.172928	0.732956
β_3	−0.005599	0.029076	γ_{35}	−5.386457	0.953359
β_4	0.120374	0.034071	γ_{41}	0.194873	0.057476
γ_{11}	−0.016607	0.148566	γ_{42}	−0.254007	0.059278
γ_{12}	−1.390934	0.274117	γ_{43}	2.108121	0.413333
γ_{13}	4.361826	0.689810	γ_{44}	−4.670528	0.873335
γ_{14}	−4.299791	0.693569	γ_{45}	11.037816	2.300782
γ_{15}	0.963943	0.544455			

Figures 4.4 – 4.8 depict the crude and graduated mortality intensities (on the log scale) plotted against age for calendar year 1975 to 2003. Table 4.7 gives a summary of some of the formal statistical tests of a graduation, applied separately to all 29 calendar years. These tests are carried out on the relative deviation after fitting the model in Equation (4.21) to the data. The p-values for the sign test, run test, Kolmogorov-Smirnov test, serial correlation test (with lag 1, 2, 3) are recorded in column one to column six of Table 4.7. These statistical test are discussed in detail in Forfar *et al.* (1988). Any p-value that is less than 5%, indicating failure of the test concerned, is highlighted by an asterisk.

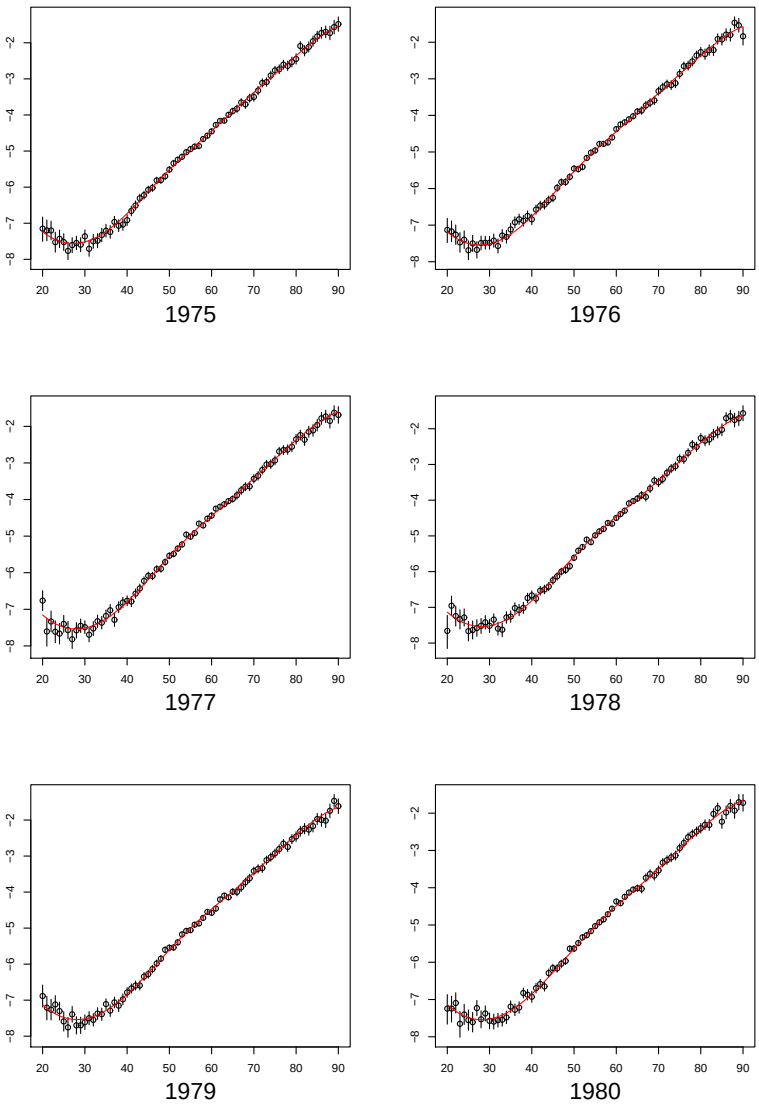


Figure 4.4: Log graduated and crude mortality intensities *vs* age for male assured lives from 1975–1980. The vertical lines are the 95% confidence interval based on the crude estimates.

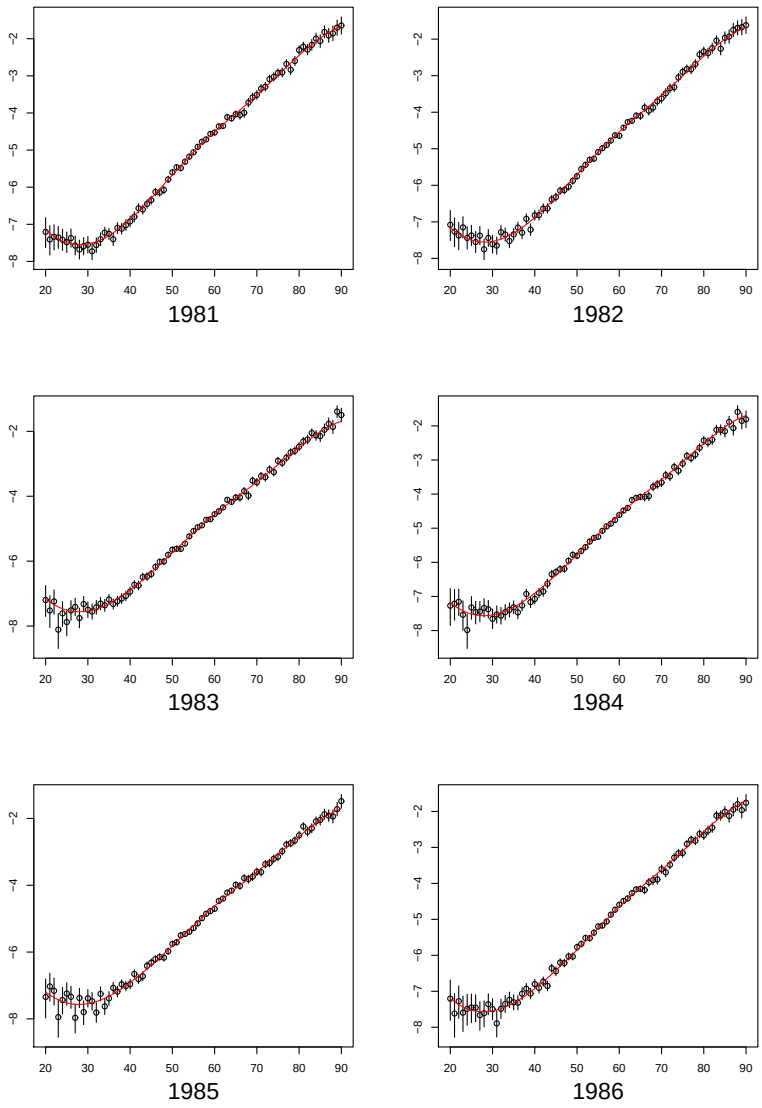


Figure 4.5: Log graduated and crude mortality intensities *vs* age for male assured lives from 1981–1986. The vertical lines are the 95% confidence interval based on the crude estimates.

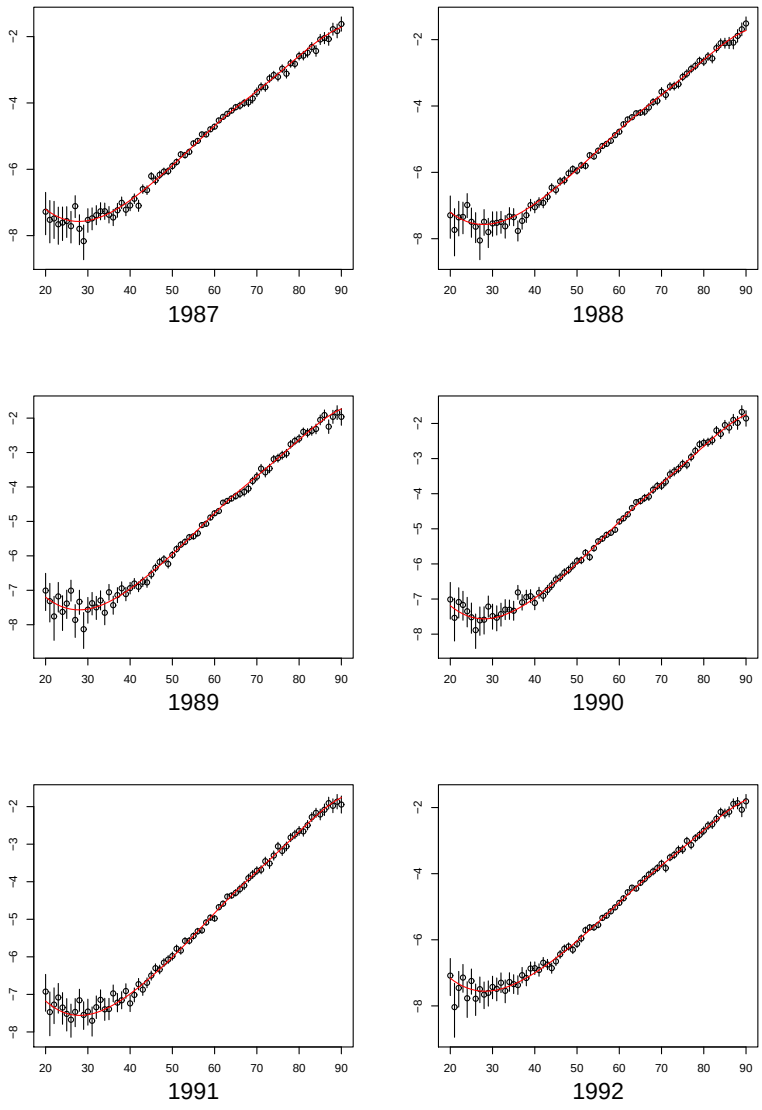


Figure 4.6: Log graduated and crude mortality intensities *vs* age for male assured lives from 1987–1992. The vertical lines are the 95% confidence interval based on the crude estimates.

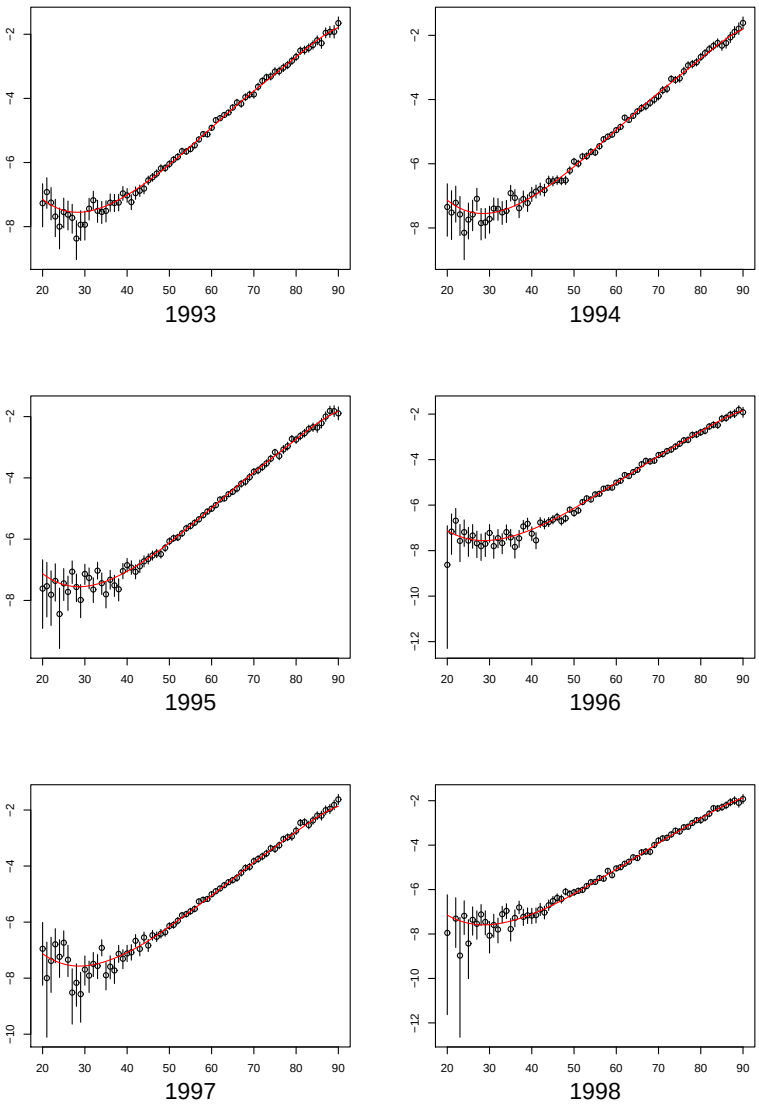


Figure 4.7: Log graduated and crude mortality intensities *vs* age for male assured lives from 1993-1998. The vertical lines are the 95% confidence interval based on the crude estimates.

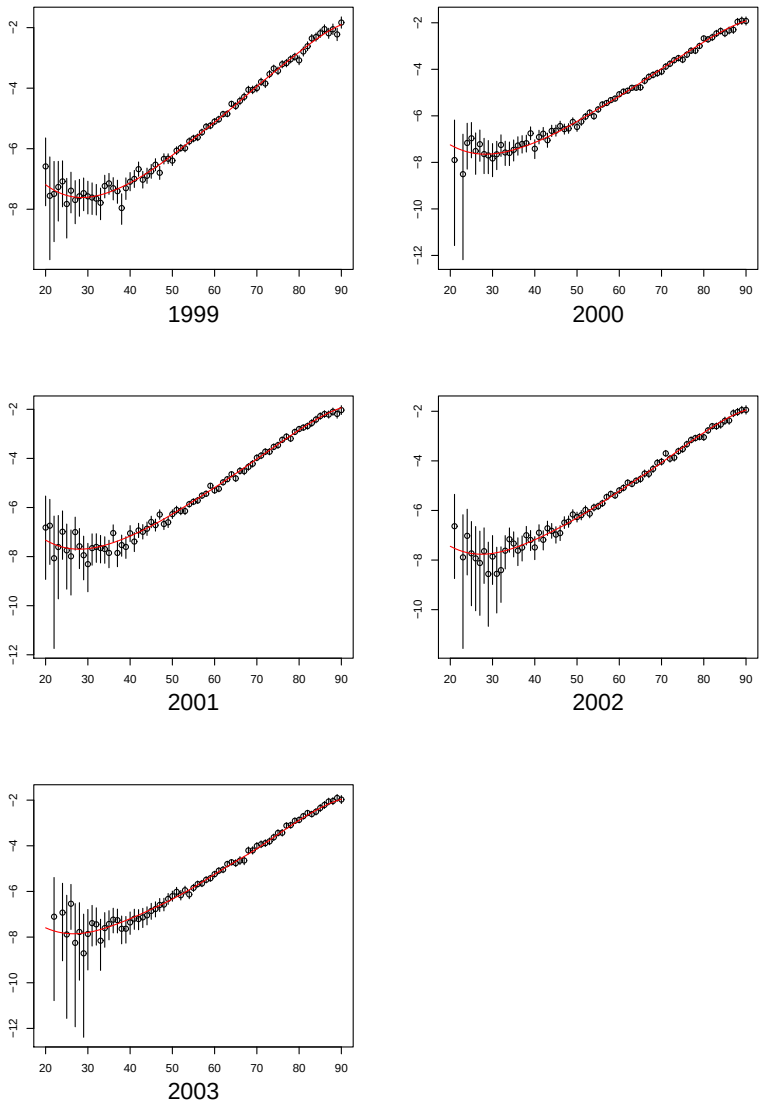


Figure 4.8: Log graduated and crude mortality intensities *vs* age for male assured lives from 1999-2003. The vertical lines are the 95% confidence interval based on the crude estimates.

Table 4.7: Percentage p-value for graduation tests for each calendar year separately on male assured lives.

Year	Sign Test	Run Test	KS Test	Serial Cor. Test 1	2	3
1975	50.0	23.7	78.5	4.9*	23.6	69.8
1976	95.2	17.3	86.9	3.6*	88.4	85.3
1977	11.7	38.8	66.0	81.8	85.1	21.4
1978	23.8	18.9	81.5	21.0	45.8	40.4
1979	11.7	38.8	29.0	27.9	95.1	54.1
1980	50.0	11.6	96.2	61.3	95.1	13.5
1981	31.7	42.1	48.9	24.6	59.6	90.7
1982	50.0	83.1	39.8	77.0	14.1	39.7
1983	82.9	85.3	66.0	73.7	12.7	92.7
1984	4.8*	55.8	64.3	24.4	80.3	28.5
1985	59.4	83.1	26.4	71.2	63.0	34.4
1986	76.2	61.0	7.2	33.0	10.4	74.9
1987	7.7	41.9	6.5	50.8	2.9*	57.3
1988	40.6	32.2	48.8	12.4	12.9	97.2
1989	31.8	33.1	95.2	37.8	39.6	40.4
1990	99.2	55.0	75.3	28.6	19.5	77.3
1991	82.9	71.4	65.0	70.8	79.5	14.1
1992	50.0	76.3	98.0	36.8	88.9	91.4
1993	23.8	8.6	35.9	1.0*	60.4	41.4
1994	11.8	22.0	76.9	8.1	76.5	86.6
1995	0.8*	51.2	94.1	27.6	89.9	23.9
1996	7.7	99.9	30.9	48.2	96.4	63.6
1997	59.4	59.4	81.3	22.1	21.0	85.9
1998	97.2	87.6	44.4	69.8	27.3	62.1
1999	68.2	60.0	99.7	56.0	48.1	71.3
2000	31.8	1.7*	9.5	6.0	48.8	13.1
2001	7.7	61.4	98.4	83.5	68.7	83.0
2002	40.6	24.1	30.8	36.1	80.9	87.7
2003	76.2	12.3	87.1	17.0	9.7	94.1

Figures 4.4- 4.8 show that the graduated values fitted the crude estimates adequately. From these figures, the graduated values for 1986 seemingly fit the crude estimates better than those for 1997 because the graduated values for 1997 fit poorly the crude estimates at younger age range. However, the results from the Kolmogorov-Smirnov test seem to indicate the opposite. The Kolmogorov-Smirnov test we use considers the distribution of the maximum deviation between the distributions of actual and expected deaths and therefore more weight is given to older ages with a lot of deaths than to younger ages does with few deaths. In other words, this test tends to be more sensitive to ages with a lot of deaths (i.e. near the median of the distribution) than to ages with few deaths (i.e. at the tail of the distribution). In Table 4.7, of the 174 tests conducted, only 7 tests produced p-value which is less than the 5% significance level.

For each distinct combination of age and calendar year, we will calculate its deviance residual (see Equation (3.26)). Figure 4.9 shows the two-dimensional plot of the deviance residuals, with calendar years as rows and ages as columns. The red and blue rectangles represent positive and negative deviance residuals, respectively. Each colour is expressed in three different intensities, representing the different ranges of values for the deviance residual (see legend of graph). There is a total of 2059 (29 $\times$ 71) deviance residuals, of which 983 are positive and 1076 are negative. Assuming that the number of positive signs is binomially distributed as $B(2059, 0.5)$, the p-value is 2(1-0.97871)=0.043, suggesting the observed number of positive signs is slightly less than expected. However, the positive and negative signs seems fairly randomly scattered.

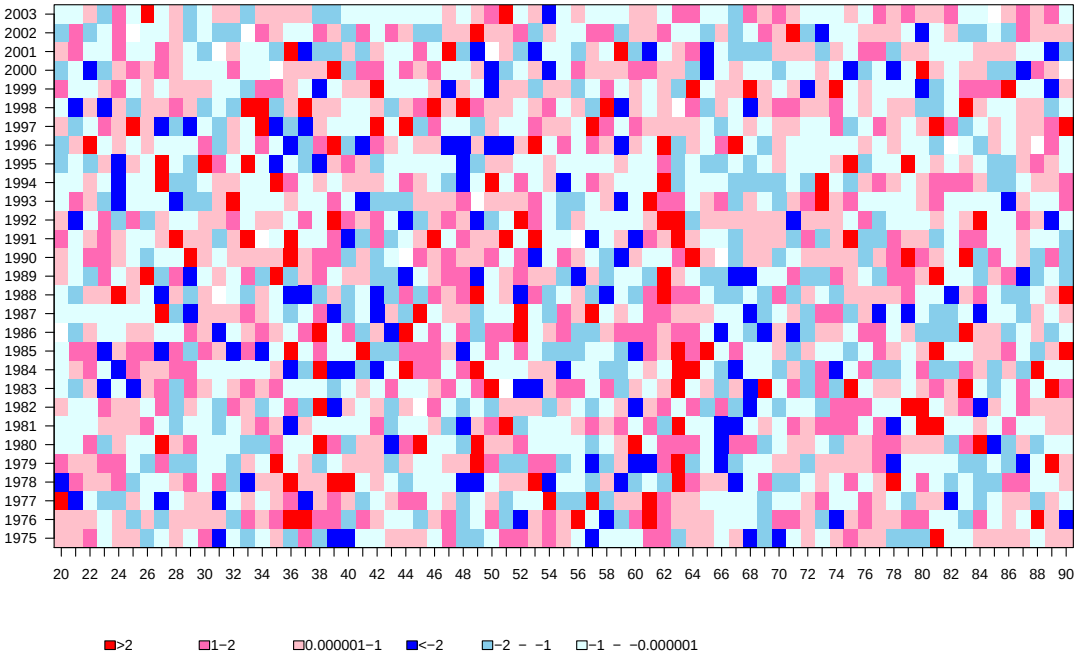


Figure 4.9: A two-dimensional plot of deviance residuals for male assured lives aged 20 to 90 from calendar year 1975 to 2003.

4.5 Mortality Model for Female Assured lives

In this section, we focus on the modelling of the female assured lives mortality intensity. The UK female assured lives data is available only from 1983 to 2003. Thus, we need to make sure that the mortality model fitted to the 1983-2003 data will produce sensible results when the mortality intensities are extrapolated back to 1975. In this respect, we are guided by CMI Report 6 (1983) that produced a graduation formula for the ultimate mortality probability, q_x, based on female assured lives data from 1975-1978. For all ages $x \geq 0$, the ultimate mortality probability q_x is given by the formula

$$q_x = \frac{\exp(f(x))}{1 + \exp(f(x))}$$

where

$$f(x) = a_1 + a_2 t + a_3(2t^2 - 1) + a_4(4t^3 - 3t) + a_5(8t^4 - 8t^3 + 1)$$

$$t = \frac{x - 70}{50}$$
$$a_1 = -2.96537254$$
$$a_2 = 6.23522259$$
$$a_3 = 1.18884477$$
$$a_4 = 0.39070030$$
$$a_5 = 0.19540908 \quad .$$

We can then obtain an estimate for $\mu_x (x = 20, 21, 22, \ldots, 90)$ by using

$$\mu_x = -\ln(1 - q_{x-1/2}) \quad . \tag{4.22}$$

The mortality intensities from Equation (4.22) will be taken as the intensities at 1977, the mid point of the quadrennia 1975-1978. We denote this set of mortality intensities as FA77 and aim to fit a female mortality model such that when the mortality intensities are extrapolated back to 1977, the intensities produced will be very close to FA77.

First, we take as our starting point the model structure

$$\log \lambda_{xy} = h_x + \phi_y, \quad \text{for} x = 20, 21, \ldots, 90 \tag{4.23}$$

where h_x and ϕ_y are the factors representing age x and year y, respectively. We estimated all the parameters by using a Poisson regression model and plotted $\hat{h}_{45} + \hat{\phi}_y$ against year y for $y = 1983, 1984, \ldots, 2003$ in Figure 4.10. The triangle shape symbol is the FA77 intensity at age 45. To make sure the mortality intensity varies linearly by year outside the data range, we will use natural cubic splines to model the year effect. Table 4.8 shows the AIC values obtained by replacing the $\{\phi_y\}$ with a natural cubic spline with differing numbers of knots.

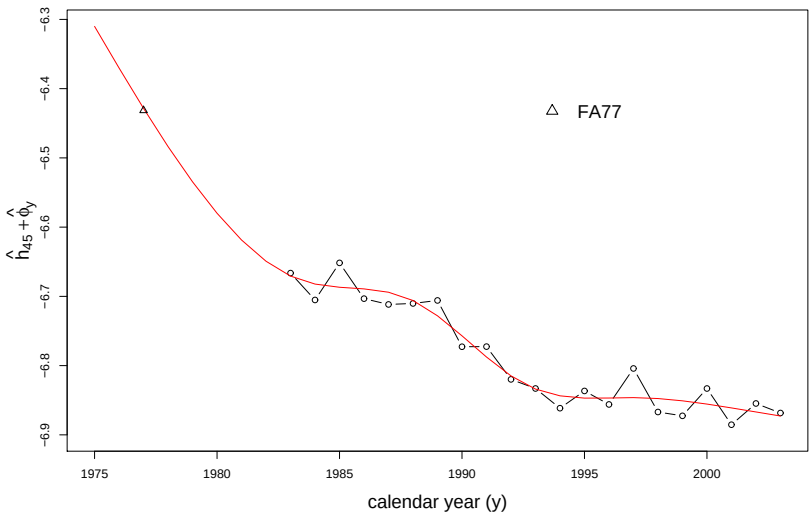


Figure 4.10: Plotting of $\hat{h}_{45} + \hat{\phi}_y$ against year y. The superimposed solid red line is obtained from the estimated coefficients of year related terms in Equation (4.24). The triangle symbol is the FA77 rate at age 45.

Table 4.8: AIC values for different numbers of knots to model the year effect in Equation (4.23).

Number of knots	Placement of knots	AIC
4	1975, 1985, 1993, 2001	9620.0
5	1975, 1984, 1990, 1996, 2002	9623.9
6	1975, 1983.7, 1988.5, 1993.0, 1997.5, 2002	9619.5
7	1975, 1984, 1987, 1991, 1994, 1998, 2002	9622.8
8	1975, 1983.5, 1986.666, 1989.834, 1993, 1996.166, 1999.334, 2002.5	9622.0

We settled on 6-knot natural spline for the calendar year effect since it gives the lowest AIC value. The estimated regression coefficients of the 6-knots spline basis are given in the following equation.

$$\log \lambda_{xy} = h_x - 0.8509y' + 0.5613w_1(y') - 2.7045w_2(y') + 4.4548w_3(y') - 3.6786w_4(y') \tag{4.24}$$

where $w(y')$ is a natural spline basis as defined in Equation (4.18). The fitted mortality intensities at age 45 from 1975 to 2003, given in Equation (4.24), are overlaid in Figure 4.10. In this same graph, note that the extrapolated value at year 1977 is close to the corresponding FA77 rate (triangle symbol).

We will further refine our model by allowing for age-dependent year coefficients. In the absence of data from 1975-1983, we felt that it is reasonable not to allow the coefficients of y' and $w_1(y')$, the two year terms that dictate the year trend between 1975 and 1983, to vary by age. With coefficients of $w_2(y'), w_3(y')$ and $w_4(y')$ varying by every single age, the structure of our model then becomes

$$\log \lambda_{xy} = h_x + \beta_1 w_1(y') + \beta_2 w_2(y') + \sum_{i=3}^{5} \sum_{x=21}^{90} (\beta_i + \beta_{ix}) w_i(y'). \tag{4.25}$$

We plotted the resulting sets of parameter estimates $\{\beta_{3x}\}, \{\beta_{4x}\}, \{\beta_{5x}\}$ (Equation 4.25) against x. These three plots are produced in Figure 4.11.

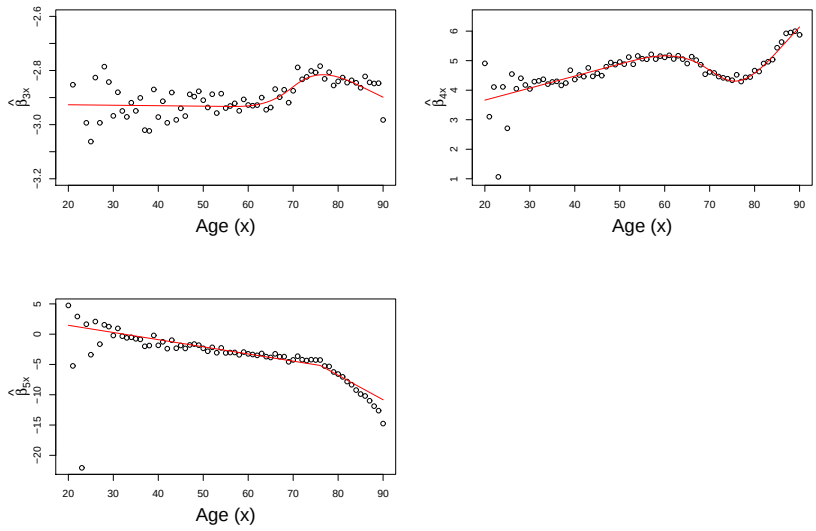


Figure 4.11: Age-dependent year coefficients. The superimposed solid red line is obtained from the estimated coefficients of year-related terms of Equation (4.26) in Table 4.11.

Next, we denote by $\alpha_i(x')$ the age-dependent coefficient of $w_i(y')$ for $i = 3, 4, 5$. We use a natural cubic spline to model both $\alpha_3(x')$ and $\alpha_4(x')$ and a linear spline for $\alpha_5(x')$. After a lengthy trial-and-error process guided by the plots in Figure 4.11, the number of knots chosen together with their locations are shown in Table 4.9.

Table 4.9: Location of knots for the age-dependent year coefficients in Equation (4.26).

	Number of knots	Placement of knots
$\alpha_3(x')$	4	55, 68, 72, 86
$\alpha_4(x')$	4	50, 68, 75, 86
$\alpha_5(x')$	1	76

With a complete structure for $\alpha_i(x')$ $(i = 3, 4, 5)$, we can re-write Equation (4.25) to give

$$\begin{aligned} \log \lambda_{xy} &= h_x + \sum_{i=3}^{4}(\beta_i + \gamma_{i1}x' + \sum_{j=2}^{3}\gamma_{ij}w_j(x'))w_i(y') \\ &\quad +(\beta_5 + \gamma_{51}x' + \gamma_{52}(x' - (76-55)/35)_+)w_5(y') \\ &\quad +\beta_1 w_1(y') + \beta_2 w_2(y'). \end{aligned} \tag{4.26}$$

Details of the parameter estimates and their standard errors are given in columns two and three of Table 4.11. As for the male mortality intensity, we find a suitable GM(r, s) formula to smooth the $\{h_x\}$ by fixing the values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in Equation (4.26) at the estimated regression coefficients in column two of Table 4.11. We fitted different GM(r, s) formulae and the log likelihood from each model is shown in Table 4.10.

Table 4.10: Twice the difference between the log likelihood from each each GM(r, s) formula and the log likelihood from GM(0,2).

r	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$	$s=7$	$s=8$	$s=9$
0	0	190.5	238.3	284.6	304.4	310.9	319.8	320.8
1	102.7	206.5	297.2	299.5	305.4	311.4	321.3	
2	130.5	282.9	297.9	300.3	310.5	318.0		
3	271.5	289.7	297.9	311.3	317.3			
4	302.6	303.1	307.8					
5	303.3	319.7						
6	306.7							

We settled on a GM(0,8) formula for the baseline intensity and update Equation (4.26) to give

$$\begin{aligned}\log \lambda_{xy} &= \sum_{i=0}^{7} b_i C_i(x') + \beta_1 w_1(y') + \beta_2 w_2(y') + \sum_{i=3}^{4}(\beta_i + \gamma_{i1}x' + \sum_{j=2}^{3}\gamma_{ij} w_j(x'))w_i(y') \\ &\quad +(\beta_5 + \gamma_{51}x' + \gamma_{52}(x' - (76-55)/35)_+)w_5(y').\end{aligned}$$

The parameter estimates in Equation (4.27) and their standard errors are shown in the fourth and fifth columns of Table 4.11.

Table 4.11: Parameter estimates in Equation (4.26) and (4.27) alongside their standard errors.

	Piece-wise constant baseline		GM(0,8) baseline	
Symbol	Estimate	Std. Error	Estimate	Std. Error
b_0	–	–	−5.939330	0.614450
b_1	–	–	3.367757	0.032079
b_2	–	–	0.309219	0.025177
b_3	–	–	−0.061689	0.021643
b_4	–	–	0.111592	0.018500
b_5	–	–	−0.078545	0.015911
b_6	–	–	−0.005641	0.011097
b_7	–	–	−0.028507	0.010097
β_1	−0.7590	1.1666	−0.748567	1.166394
β_2	0.5062	0.7566	0.496324	0.756464
β_3	−2.4934	2.5506	−2.416903	2.550120
β_4	3.9214	2.5417	3.758088	2.540982
β_5	−2.7202	1.3849	−2.651587	1.384525
γ_{31}	−0.7413	0.2627	−0.693227	0.259576
γ_{32}	5.6975	1.1565	4.177638	1.009590
γ_{33}	−67.3409	11.2553	−50.898874	9.424659
γ_{41}	2.7578	1.0809	2.755805	1.069499
γ_{42}	−5.3424	1.3588	−3.670702	1.209876
γ_{43}	92.1564	17.6679	68.604318	15.027415
γ_{51}	−2.8993	1.9056	−3.534043	1.879462
γ_{52}	−19.7049	6.4042	−15.510799	5.744604

Figures 4.12 – 4.15 show the crude and graduated mortality intensities (on the log scale) plotted against age from year 1983 to 2003 inclusive.

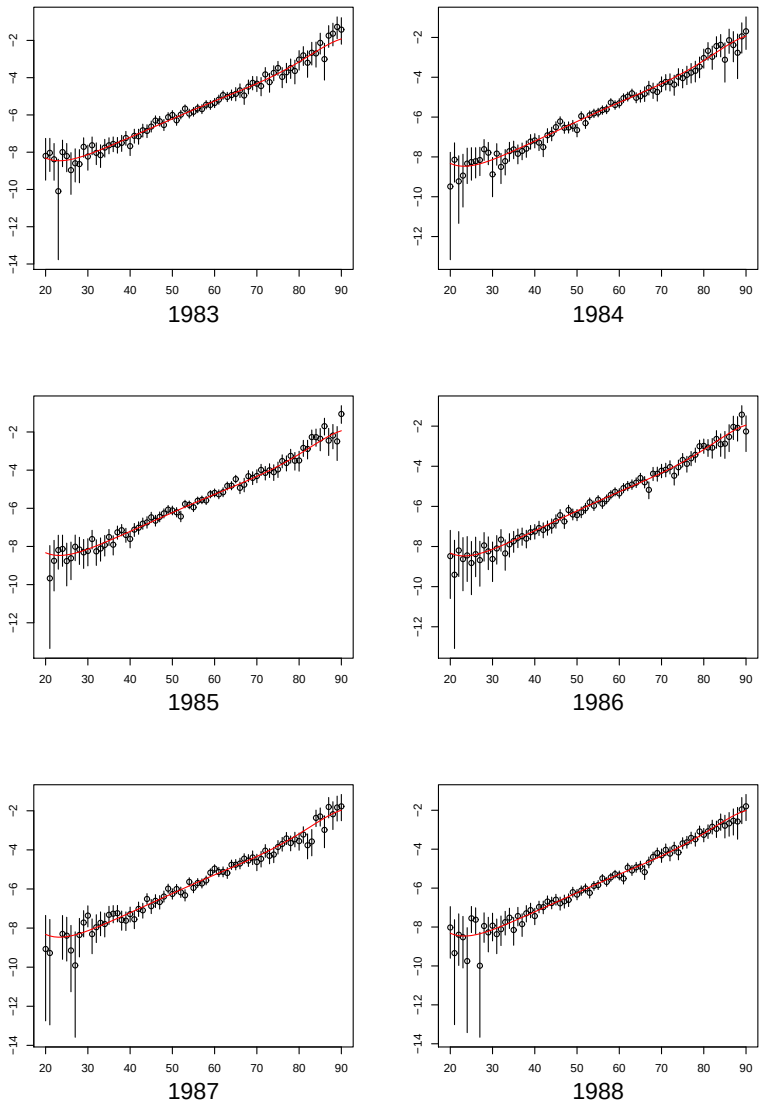


Figure 4.12: Log graduated and crude mortality intensities *vs* age for female assured lives from 1983-1988. The vertical lines are the 95% confidence interval based on the crude estimates.

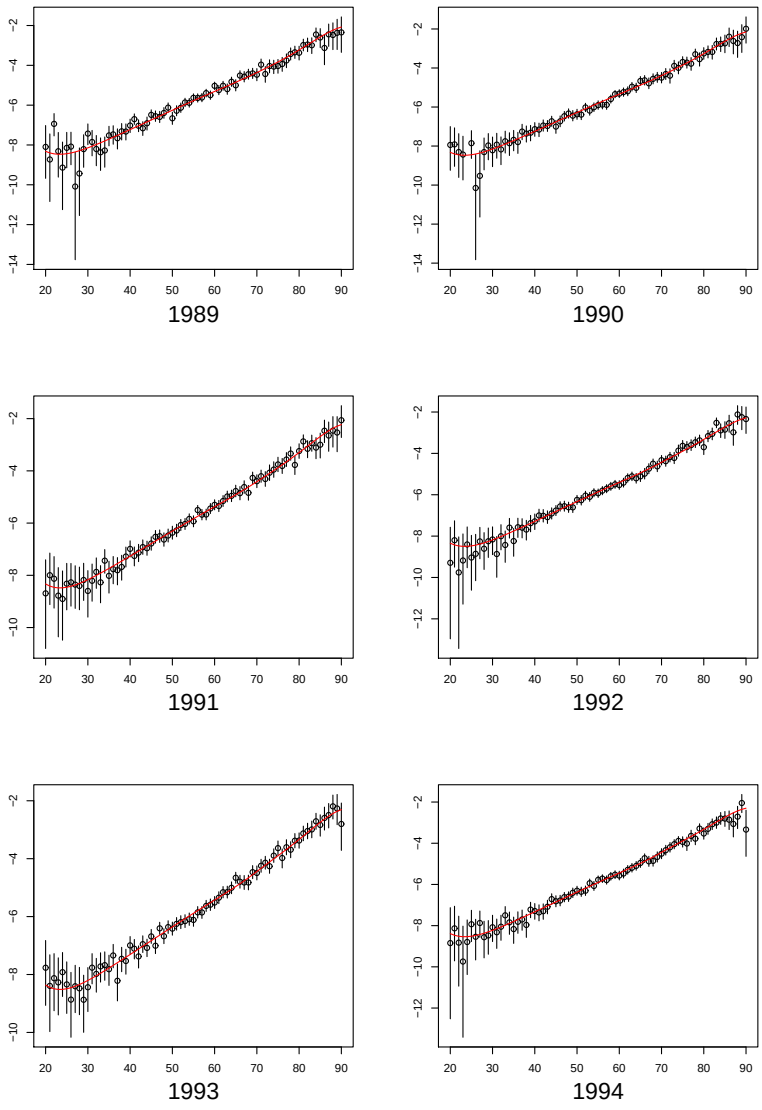


Figure 4.13: Log graduated and crude mortality intensities *vs* age for female assured lives from 1989-1994. The vertical lines are the 95% confidence interval based on the crude estimates.

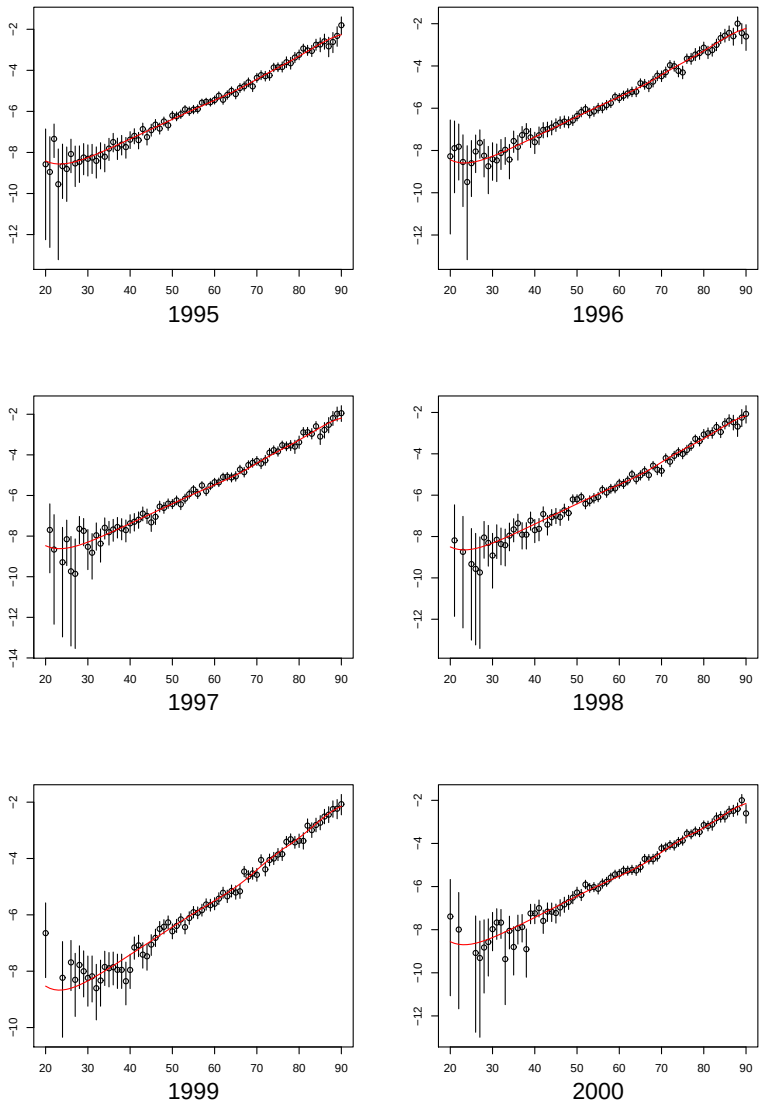


Figure 4.14: Log graduated and crude mortality intensities *vs* age for female assured lives from 1995-2000. The vertical lines are the 95% confidence interval based on the crude estimates.

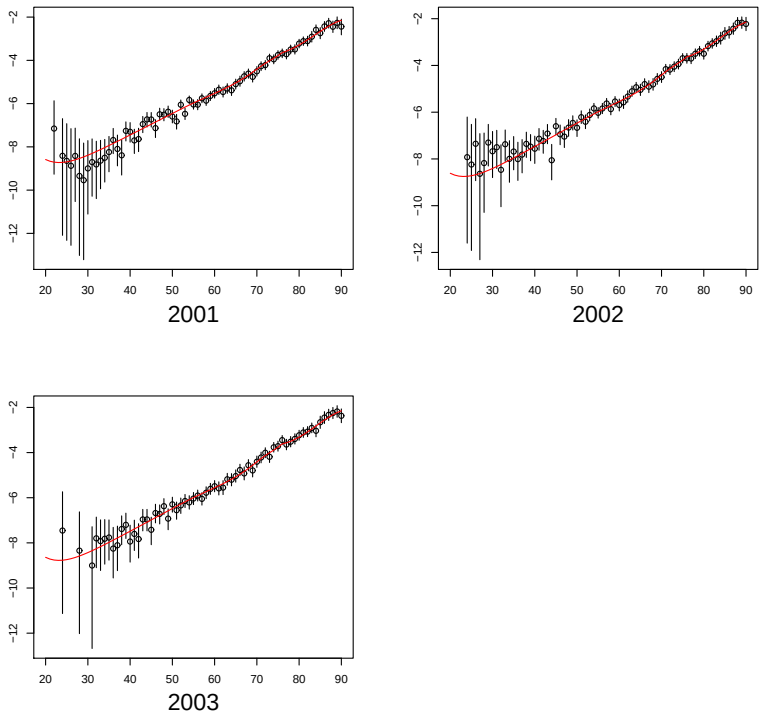


Figure 4.15: Log graduated and crude mortality intensities *vs* age for female assured lives from 2001-2003. The vertical lines are the 95% confidence interval based on the crude estimates.

We also show in Figures 4.16 and 4.17 the crude and graduated female mortality intensities (on a log scale) for 5 yearly spaced ages, $x = 30, 35, 40, \ldots, 90$ over the years 1975-2003. Included in these figures are the FA77 intensities at these ages. Such plots provide a visual check on how close the extrapolated intensities from Equation (4.27) are to their corresponding FA77 intensities. Table 4.12 gives a summary of some of the formal statistical tests of a graduation, applied separately to all 21 calendar years. The tests are carried out on the relative deviation after fitting the model in Equation (4.27) to the data. The p-values expressed as percentages for the sign test, run test, Kolmogorov-Smirnov, serial correlation test (with lag 1 to 3) are recorded in column one to column six of Table 4.12. Any p-value that is less than 5% is highlighted by an asterisk.

Table 4.12: Percentage p-value of graduation test for each calendar year separately on female assured lives.

Year	Sign Test	Run Test	KS Test	Serial Cor. Test 1	2	3
1983	40.6	60.0	93.0	57.5	63.8	4.8*
1984	7.7	11.7	78.9	16.2	33.9	41.1
1985	68.2	17.3	90.9	58.8	80.4	51.6
1986	11.8	48.5	78.5	68.9	74.7	73.7
1987	17.1	20.2	34.6	9.4	35.8	67.6
1988	59.4	88.4	99.4	92.1	47.9	71.8
1989	50.0	68.5	51.0	56.3	69.1	79.8
1990	40.6	95.5	77.2	50.3	20.3	29.1
1991	50.0	92.5	99.2	96.0	67.5	57.7
1992	76.2	95.8	82.7	66.8	32.7	95.0
1993	88.2	55.4	59.4	89.0	6.5	78.4
1994	7.7	85.1	87.3	70.0	75.6	35.9
1995	50.0	83.1	98.3	95.8	58.0	37.7
1996	68.2	7.8	65.6	6.4	95.8	90.2
1997	59.4	83.1	89.0	49.6	86.9	90.7
1998	4.8*	65.1	63.9	55.3	31.4	15.5
1999	11.8	29.9	93.2	50.4	66.4	90.6
2000	31.8	8.1	97.2	37.1	22.3	86.6
2001	7.7	70.6	99.9	84.6	46.8	65.4
2002	50.0	40.6	64.9	85.0	11.0	7.4
2003	59.4	59.4	96.9	96.0	44.3	98.8

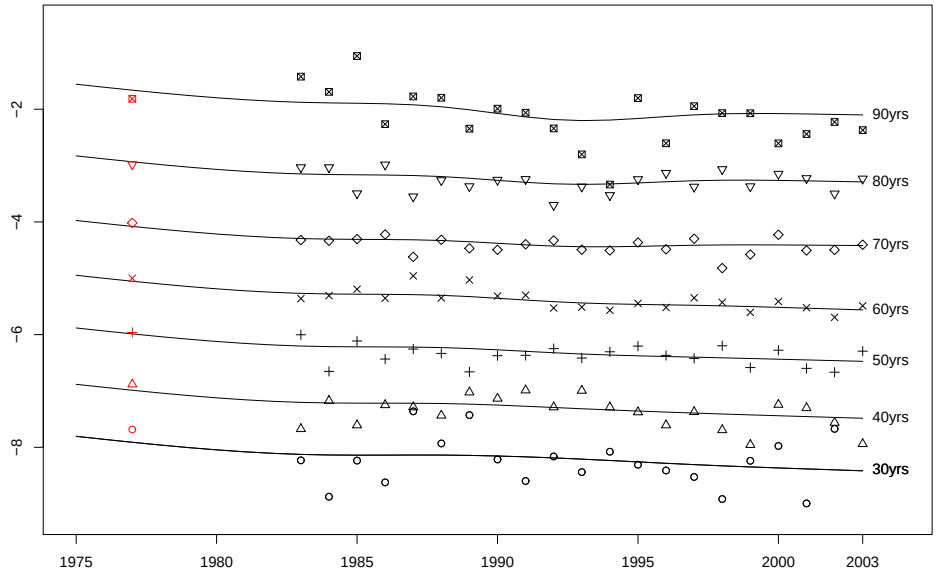


Figure 4.16: Log graduated and crude female mortality intensities *vs* calendar year, ages 30, 40, 50, 60, 70, 80, 90. The red symbols are FA77 rates.

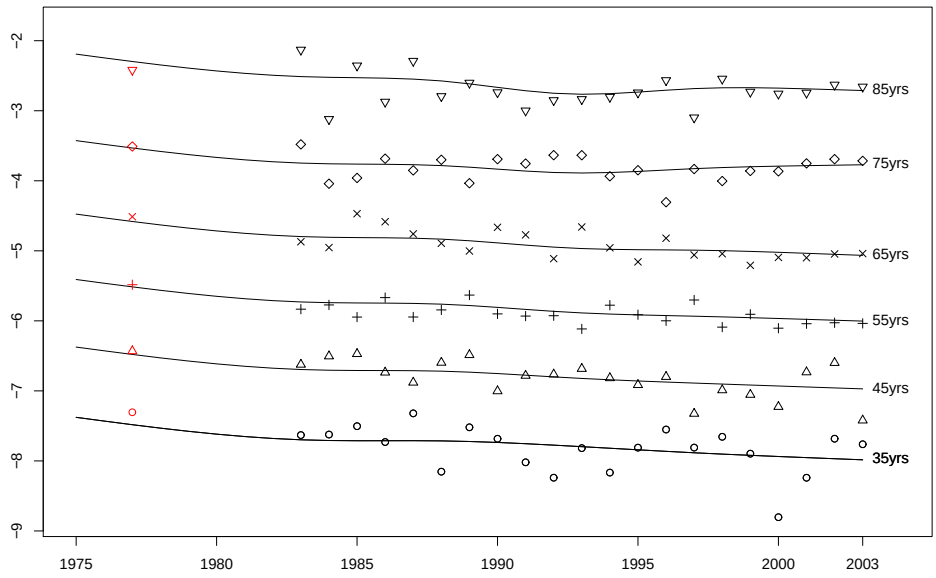


Figure 4.17: Log graduated and crude mortality intensities vs calendar year, ages 35, 45, 55, 65, 75, 85. The red symbols are FA77 rates.

Figures 4.12- 4.15 show that the graduated values fitted the crude estimates adequately. Table 4.12 shows that of the 126 tests conducted, only 2 tests produced $p-$value which are less than the 5% significance level. For each distinct combination of age and calendar year, we calculate its deviance residual (see Equation (3.26)). Figure 4.18 shows the two-dimensional plot of the deviance residuals, with calendar years as rows and ages as columns. There is a total of 1491 (21×71) deviance residuals, of which 711 are positive and 780 are negative. Assuming that the number of positive signs is binomially distributed as $B(1491, 0.5)$, the $p-$value is 2(1-0.9609)=0.078, suggesting the observed number of positive signs is not less than expected. The positive and negative signs are fairly randomly scattered.

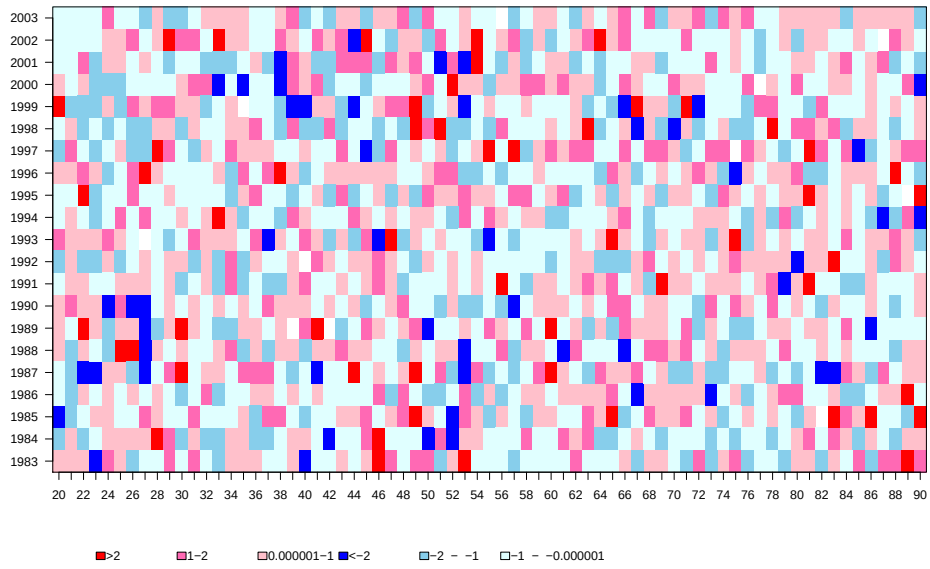


Figure 4.18: A two-dimensional plot of deviance residuals for female assured lives aged 20 to 90 from calendar year 1983 to 2003.

Chapter 5

Modelling of the Mortality Intensity from Sick II: Excess Mortality Intensity

5.1 Introduction

As mentioned in Chapter 4, the mortality intensity from sick is split into two additive components. The first component is a base mortality that varies by sex, age and calendar year while the second component is the excess mortality intensity entailed with being sick with a specific cause. The mortality intensity at time z since sickness inception for an individual with covariates $\mathbf{x}$ is $\lambda(z, \mathbf{x})$, is

$$\lambda(z, \mathbf{x}) = \lambda^*(\mathbf{x}) + \nu(z, \mathbf{x}) \tag{5.1}$$

in which $\lambda^*(\mathbf{x})$ is the base mortality as discussed in Chapter 4 while $\nu(z, \mathbf{x})$ is the excess mortality intensity that will be the focus of this chapter. In Section 5.2, we will present the methodology to estimate the excess mortality intensity and describe the transformation the data undergoes to make it compatible with the estimation method. The results of the implementation of the methods described in Section 5.2 on IPI data by cause of sickness are presented in Section 5.3.

5.2 Estimation Method

We consider the covariates $\mathbf{x}$ in the excess mortality component to act multiplicatively on the baseline intensity, $\nu_0(z)$, that depends only on sickness duration z. We re-write Equation (5.1) to give

$$\lambda(z, \mathbf{x}) = \lambda^*(\mathbf{x}) + \nu_0(z) \exp(\mathbf{x}\boldsymbol{\beta}) \tag{5.2}$$

where $\boldsymbol{\beta}$ denotes the regression coefficients. Further, we partition the sickness duration into J discrete intervals with τ_j $(j = 0, 1, \ldots, J)$ as the interval end-points. We then assume that the baseline intensity is constant in each discrete sickness duration interval such that

$$\nu_0(z) = \nu_j \quad \text{for } z \in [\tau_{j-1}, \tau_j) \quad . \tag{5.3}$$

Following this assumption, the mortality intensity in sickness duration interval j, z_j, is written as

$$\lambda(z_j, \mathbf{x}) = \lambda^*(\mathbf{x}) + \nu_j \exp(\mathbf{x}\boldsymbol{\beta}) \quad . \tag{5.4}$$

Estève *et al.* (1990) use a full likelihood approach based on individual level data to estimate the parameters in Equation (5.4). The same likelihood can be obtained using grouped data. In grouped data, we have a value for the central exposed-to-risk (R_{jl}) and number of deaths (d_{jl}) for each distinct covariate pattern, indexed by l, in sickness duration interval z_j. The base mortality intensity, $\lambda^*(x_l)$, is assumed constant in each distinct covariate pattern l. The likelihood of the data is therefore given by

$$L = \prod_{j,l} \{\exp[-R_{jl}(\lambda^*(\mathbf{x}_l) + \nu_j \exp(\mathbf{x}_l\boldsymbol{\beta}))](\lambda^*(\mathbf{x}_l) + \nu_j \exp(\mathbf{x}_l\boldsymbol{\beta}))^{d_{jl}}\} \quad . \tag{5.5}$$

Taking logs on both sides of Equation (5.5) and removing constant terms, we have

$$\log L = \sum_{j,l} \{-R_{jl}\nu_j \exp(\mathbf{x}_l\boldsymbol{\beta}) + d_{jl} \log[\lambda^*(\mathbf{x}_l) + \nu_j \exp(\mathbf{x}_l\boldsymbol{\beta})]\} \quad . \tag{5.6}$$

Taking the derivative of Equation (5.6) with respect to ν_j and equating it to zero, the MLE for ν_j, $\hat{\nu_j}$, can be found by solving the following equation iteratively

$$\sum_l \frac{d_{jl} \exp(\mathbf{x}_l \boldsymbol{\beta})}{\lambda^*(\mathbf{x}_l) + \hat{\nu_j} \exp(\mathbf{x}_l \boldsymbol{\beta})} = \sum_l R_{jl} \exp(\mathbf{x}_l \boldsymbol{\beta}) \quad . \tag{5.7}$$

Dickman *et al.* (2004) show that we can use the convenience of GLMs to maximise the likelihood in Equation (5.5). We do that by first assuming that the number of deaths d_{jl} follows a Poisson distribution with mean

$$\mu_{jl} = \lambda(z_j, x_l) R_{jl} \quad . \tag{5.8}$$

The kernel of the log-likelihood by treating d_{jl} as a Poisson observation with the above mean is identical to the log-likelihood in Equation (5.6). Denoting d^*_{jl} as the expected number of deaths under model for base mortality intensity and substituting Equation (5.4) into Equation (5.8), we obtain

$$\begin{aligned} \mu_{jl}/R_{jl} &= \lambda^*(x_l) + \nu_j \exp(\mathbf{x}_l \boldsymbol{\beta}) \\ &= d^*_{jl}/R_{jl} + \nu_j \exp(\mathbf{x}_l \boldsymbol{\beta}) \end{aligned}$$

which, after some simple algebraic manipulation and let $\gamma_j = \log(\nu_j)$, can be written as

$$\begin{aligned} \log(\mu_{jl} - d^*_{jl}) &= \log(R_{jl}) + \log(\nu_j) + \mathbf{x}_l \boldsymbol{\beta} \\ &= \log(R_{jl}) + \gamma_j + \mathbf{x}_l \boldsymbol{\beta} \quad . \end{aligned} \tag{5.9}$$

The parameters in Equation (5.9) can be estimated by using a GLM with Poisson error structure, outcome d_{jl}, link $\log(\mu_{jl} - d^*_{jl})$ and $\log R_{jl}$ as an offset term.

The parameterisation of the excess mortality intensity can be split into two parts. The first part is to graduate the piece-wise constant baseline intensities by incorporating a rich smooth parametric formula such as a GM$(0, s)$ formula in the model and let the data provide a smooth estimator. With a GM$(0, s)$ formula for the baseline intensity, the additive main effects model structure therefore takes the following form:

$$\lambda(z, \mathbf{x}) = \lambda^*(\mathbf{x}) + \exp\left(\sum_{i=1}^{s} b_i (t_k(z))^{i-1}\right) \exp(\mathbf{x}\boldsymbol{\beta})) \tag{5.10}$$

where the duration variable $t_k(z)$ is defined as $t_k(z) = z/(1 + kz)$. This duration variable is used in CMI Working Paper 15 (2004) and in the modelling of the recovery intensity (see Section 3.3). The covariate vector $\mathbf{x}$ consist of $\{x_{\text{sex}}, x_{\text{rated}}, x_{\text{age}}, x_{\text{year}}, x_{\text{dp4}}, x_{\text{dp13}}, x_{\text{dp26}}, x_{\text{dp52}}\}$. In the case that excess mortality intensities for several causes of sickness are being modelled together, a set of binary indicators representing these causes of sickness will be added to the covariate vector $\mathbf{x}$.

Substituting Equation (5.10) into Equation (5.8) and after some algebraic manipulation, we obtain the following additive log linear model which, as in Equation (5.9), can be estimated in the framework of a GLM

$$\log(\mu_{jl} - d^*_{jl}) = \log(R_{jl}) + b_0 + b_1 t_k(z_j) + \ldots + b_s (t_k(z_j))^s + \mathbf{x}_l\boldsymbol{\beta} \quad . \tag{5.11}$$

The value of k used in the duration transformation is not pre-specified but will be determined by the data. For each GM$(0, s)$ formula ($s = 1, 2, \ldots, 6$), we will try out a range of k values ($k = 0, 0.1, 0.2, \ldots, 2.9, 3.0$). The differences in the residual deviance produced for two different GM$(0, s)$ formulae is approximated by a χ^2 distribution with degrees-of-freedom equal to the difference in the number of parameters used.

The second part of the parameterisation process is to select which covariates from $\mathbf{x}$ to include in the model after keeping the chosen GM$(0, s)$ formula and its optimal k value in the model. The variable selection on the basis of AIC statistics is carried out by using the 'stepAIC' function in the 'R' statistical package, which enables the main effects and all possible two-way interaction terms to be explored and their statistical significance assessed in an automated manner. Note that in this process, the detection and fixing of non-proportional hazards is carried out at the same time through the inclusion of time by covariate interaction term(s). We are aware that a form of residual akin to the Schoenfeld residuals (Section 3.2), designed to check the proportional hazard assumption in the Cox model, has been developed in the context of additive relative survival models (see Stare *et al.*, 2005). However, due to the absence of evidence regarding a time varying effect for any covariate in the mortality intensity among UK IPI claimants in the actuarial literature, we decided not to carry out such tests in a separate exercise. Selection of variables is data-driven and not guided by medical opinion or aimed to uncover hitherto unknown relationships. The goodness-of-fit for the chosen model is then assessed by using the χ^2 statistic and a two-dimensional plot of the deviance residuals (see Section 3.5).

5.2.1 Data Preparation

The data we have from the CMI does not lend itself readily to GLM analysis as described in Section 5.2. Thus, the format of the data needs to be suitably transformed.

First, we partitioned sickness duration into the following 41 intervals: 1–2 weeks, 2–3 weeks, ..., 29–30 weeks, 30-39 weeks, 39 weeks – 1year, 1–2 years, 2–3 years, ..., 7–8 years, 8–12 years, 12–16 years and 16–20 years. These sickness duration intervals are indexed by j. Then, each claim record is split into multiple observations, one observation for each interval visited by the claimant. For each pseudo-observation generated, we calculate the exposed-to-risk (R) and record the number of deaths (d). Consider, for example, the claim record for an individual who died after being sick for 38 days. This claim record will generate five pseudo-observations for which $R = 7$ days and $d = 0$ for the first four observations whereas $R = 3$ days and $d = 1$ for the fifth observation. These pseudo-observations will be given the same covariate value for sex, DP and rating indicator as the original observation. The covariates age and

year are defined as the exact age and exact year at the start of the sickness duration interval respectively. Note that this definition of age in the mortality intensity from sick model is different to that in the recovery intensity model. Their values in these pseudo-observations are calculated accordingly.

Next, the expected number of deaths, d^*, according to the assured lives mortality intensity (see Chapter 4), is calculated for each pseudo-observation. Let x_s and y_s be the age and year at the start of sickness duration interval j respectively. The expected number of deaths, d^*, for each pseudo-observation is given by

$$d^* = \int_0^{R_j} \lambda^*(x_s + t, y_s + t)dt \tag{5.12}$$

where R_j is the exposed-to-risk in sickness duration interval j and $\lambda^*(x, y)$ is the assured lives mortality intensity for a person aged x in year y as presented in Chapter 4.

Integration in the right-hand side of Equation (5.12) is calculated numerically. First, age (17 to 65) and year (1975 to 2002) are discretised into intervals of 0.1 year. Then, the assured lives mortality intensity is assumed constant in each age–year interval and is taken as the intensity at the midpoint of the age–year interval. These rates are then arranged in a two dimensional array with age as the row dimension and year as the column dimension. Once the starting age–year interval for a person is located, the assured lives mortality intensities applicable to the person move diagonally downwards as sickness duration advances. The expected number of deaths is the sum of these assured lives mortality intensities multiplied by the corresponding central exposed-to-risk in each age–year interval visited.

Generation of pseudo-observations for each original observation can substantially increase the size of the data set and lead to increased computational time. To avoid this, we sum the central exposed-to risk (R), number of deaths (d) and expected number of deaths (d^*) of the pseudo-observations within each distinct covariate pattern to give only one observation for each distinct covariate pattern in each sickness duration interval. The covariates are sex (2 levels), deferred period (5 levels), occupational rating (2 levels), integer age last birthday (49 levels) and integer calendar year (28 levels) at the start of the sickness duration interval. The data is processed in a way

such that the same age last birthday is applicable to the entire sickness duration interval. These five covariates yield a maximum of $2 \times 5 \times 2 \times 49 \times 28 = 27{,}440$ distinct covariate patterns, indexed by l. For each covariate pattern l, we have R_{jl}, d_{jl} and d^*_{jl} denoting the exposed-to-risk, observed number of deaths and expected number of deaths in sickness duration interval j, respectively.

5.3 Results

5.3.1 Grouping of Causes of sickness

There are 70 causes of sickness. The total number of deaths for all causes of sickness is 3,498, of which 1,665 deaths are due to malignant neoplasm. Excluding malignant neoplasm, only three causes of sickness have 200-300 deaths each while 41 causes of sickness have less than ten deaths each. Given the low number of deaths in the large majority of causes of sickness, separate modelling for each cause of sickness is impractical. We therefore decided to classify the IPI data into fewer groups and use amalgamated IPI data from each group in our analysis.

With the exception of cs61 Arthritis, any individual cause of sickness which has more than 50 deaths is allowed to form its own group and there are six of them. Despite having 59 deaths, data from cs61 Arthritis is combined with cs62 Musculoskeletal because it has as many as 50.079 expected deaths on the basis of the assured lives mortality, leaving less than 10 deaths to be accounted for.

The other causes of sickness are bundled together in accordance to the grouping of causes of sickness presented in Tables 2.3 and 3.2. In respect of the analysis of mortality from sick, Table 5.1 shows the classification of IPI data into 15 different sickness groups (column 1), the ICD8 codes of causes of sickness which constitute each group (column 2), their respective numbers of deaths (column 3) and their respective expected numbers of deaths according to the assured lives mortality intensity (column 4). The cause of sickness represented by each ICD8 code in column 2 is listed in Appendix A.

Table 5.1: Grouping of causes of sickness in the modelling of mortality intensity from sick.

Name	Sickness Group	ICD8 code	Deaths (d)	Expected deaths (d^*)
H1	Infective	1–19	40	5.2021
H2	Malignant neoplasm	20	1665	21.9678
H3	Benign neoplasm	21	101	3.1365
H4	Endocrine & Metabolic	22–26	48	8.0040
H5	Mental illness	27	219	90.0761
H6	Nervous system & sense organs	28–31	220	49.3129
H7	Other circulatory diseases	32-34, 37-38	127	30.4939
H8	Ischaemic heart disease	35	289	85.8681
H9	Cerebrovascular disease	36	136	24.7038
H10	Respiratory	39–45	102	15.6277
H11	Digestive	47–51	114	12.2847
H12	Genito-urinary	52–55	58	5.9798
H13	Musculoskeletal	61, 62	144	107.7378
H14	Injuries	66–70	88	33.4236
H15	All Others	46, 56–60, 63–65	147	28.2141
	All causes of sickness		3498	522.0329

We estimate the piece-wise constant baseline intensity, ν_j, for each sickness duration interval z_j $(j = 1, 2, \ldots, 41)$ in the absence of covariates. The MLE for ν_j, $\hat{\nu}_j$, can be obtained by solving the following equation iteratively

$$\sum_l \frac{d_{jl}}{\lambda^*(\mathbf{x}_l) + \hat{\nu_j}} = \sum_l R_{jl} \quad . \tag{5.13}$$

Note that the above equation is the same as Equation (5.7) but with $\exp(\mathbf{x}\boldsymbol{\beta}) = 1$. We can also estimate ν_j using a Poisson regression model by creating a set of indicator variables, f_j, one for each sickness duration interval j (excluding the reference interval) where $f_j = 1$ for interval j and 0 otherwise. Both approaches will produce identical estimates. There are no estimates for sickness duration intervals with no deaths.

Figures 5.1–5.4 show the values of $\hat{\nu}_j$ $(j = 1, 2, \ldots, 41)$ plotted against the mid-point of the sickness duration intervals for each sickness group. Sickness groups which exhibit the same mortality pattern are plotted together on the same graph. As far as possible, we will attempt to relate some features of the excess mortality intensity to any known medical opinion expressed in the medical literature. Some features present in each figure are as follows:

(a) Figure 5.1: 'H1 Infective', 'H3 Benign neoplasm', 'H4 Endocrine & Metabolic', 'H10 Respiratory', 'H11 Digestive', 'H12 Genito-urinary' and 'H14 Injuries'

  (i) The mortality curves exhibit a hump-backed feature with a peak at around 200 days. A similar feature is reported in the mortality intensity for all sicknesses combined (Renshaw and Haberman (2000), CMI Working Paper 5 (2005)).

  (ii) The levels of the excess mortality intensity for 'H3 Benign neoplasm' and 'H14 Injuries' are noticeably higher and lower, respectively, than the rest.

  (iii) Apart from 'H1 Infective', 'H10 Respiratory' and 'H11 Digestive', there are no estimates for very short sickness durations.

(b) Figure 5.2: 'H5 Mental illness', 'H6 Nervous system & sense organs' and 'H13 Musculoskeletal'

  (i) The mortality curves exhibit a U-shaped relationship with sickness duration. The decline in the excess mortality intensity is followed by an up-turn at around 3 years of sickness duration. In relation to 'H5 Mental illness', the rise in excess mortality intensity after a few years of sickness may not be surprising given that Stark *et al.* (2003), a study based on psychiatric patients discharged from psychiatric hospitals in Scotland after a stay exceeding one year, reported that *"...Deaths from respiratory disease were four times higher than expected, and deaths from other causes, including cardiovascular disease, were also elevated..."*. They attributed this marked increase in mortality rates to alcohol abuse, poor diet, smoking and possibly antipsychotic drugs.

(ii) The level of excess mortality intensity is highest in 'H6 Nervous system & sense organs', followed by 'H5 Mental illness' and 'H13 Musculoskeletal'.

(c) Figure 5.3: 'H8 Ischaemic heart disease', 'H9 Cerebrovascular disease' and 'H7 Other circulatory diseases'

(i) The excess mortality intensity declines with increasing sickness duration until it remains relatively constant at very long term sickness durations.

(ii) The level of excess mortality intensity for 'H8 Ischaemic heart disease' is noticeably lower than the rest.

(d) Figure 5.4: 'H2 Malignant neoplasm'

(i) The excess mortality intensity follows a bell-shaped curve with the intensity rising, peaking at around 150 days and falling off before showing an up-turn at the very last point. The striking similarity of this feature to that found in CMI Working Paper 5 is expected given that deaths from malignant neoplasm makes up the bulk of the total deaths reported.

(e) Figure 5.5: 'H15 All others'

(i) Sicknesses with few deaths that are not readily classifiable into well-recognised sickness groups are bundled into this sickness group.

(ii) The excess mortality intensity remains relatively constant until it starts to decline at around 200 days.

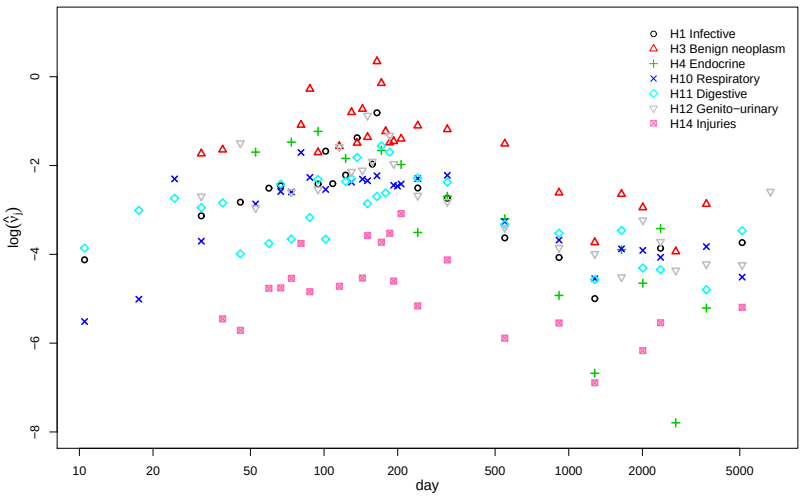


Figure 5.1: MLE, $\hat{\nu}_j$, for sickness groups: 'H1 Infective', 'H3 Benign neoplasm', 'H4 Endocrine & Metabolic', 'H10 Respiratory', 'H11 Digestive', 'H12 Genito-urinary' and 'H14 Injuries'.

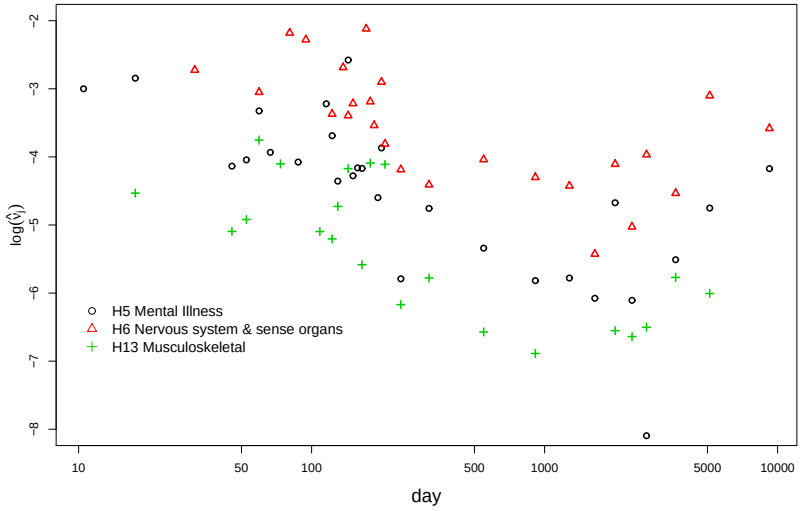


Figure 5.2: MLE, $\hat{\nu}_j$, for sickness groups: 'H5 Mental illness', 'H6 Nervous system & sense organs' and 'H13 Musculoskeletal'.

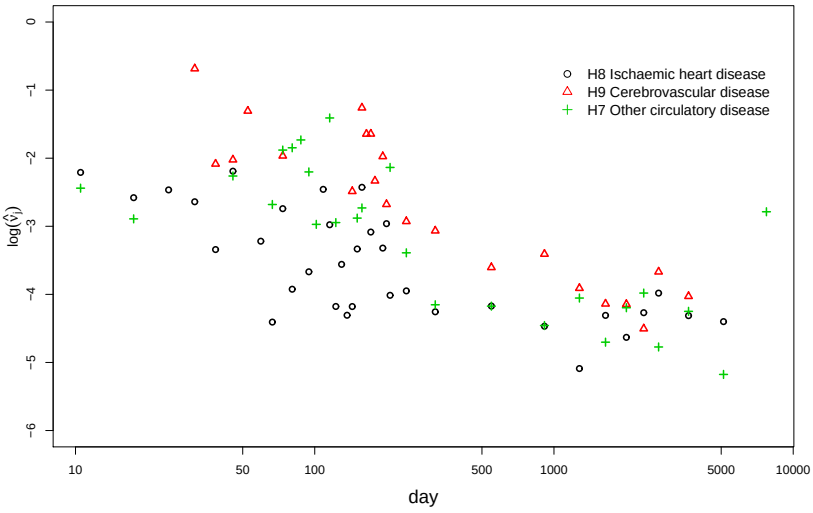


Figure 5.3: MLE, $\hat{\nu}_j$, for sickness groups: 'H8 Ischaemic heart disease', 'H9 Cerebrovascular disease' and 'H7 Other circulatory diseases'.

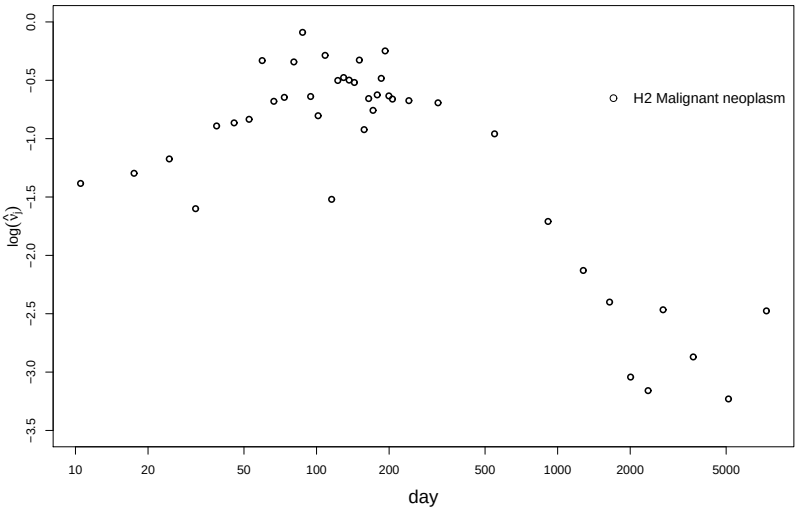


Figure 5.4: MLE, $\hat{\nu}_j$, for sickness group: 'H2 Malignant neoplasm'.

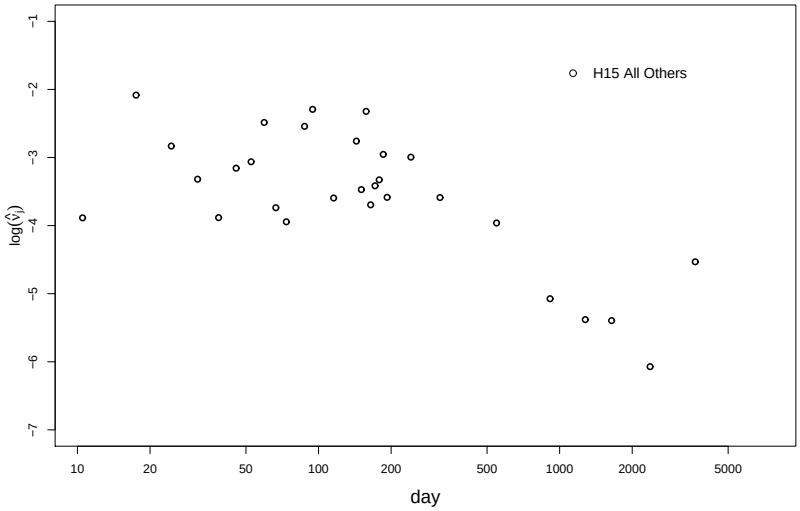


Figure 5.5: MLE, $\hat{\nu}_j$, for sickness group: 'H15 All others'.

As shown in Figures 5.1 – 5.5, the mortality curves (on a log-log scale) for different sickness groups in each graph have the same shape and are parallel to one another. This observation points the way to a proportional hazard model incorporating cause of sickness as a factor. The 15 sickness groups listed in Table 5.1 are further classified into five separate sickness categories (MI to MV) as set out in Table 5.2. Sickness groups belonging to the same category will be modelled together.

Table 5.2: The five sickness categories in the modelling of mortality intensity from sick.

Sickness Category	Sickness Group
MI	H1 Infective, H3 Benign neoplasm, H4 Endocrine & metabolic, H10 Respiratory, H11 Digestive, H12 Genito-urinary, H14 Injuries
MII	H5 Mental illness, H6 Nervous system & sense organs, H13 Musculoskeletal
MIII	H7 Other circulatory disease, H8 Ischaemic heart disease, H9 Cerebrovascular disease
MIV	H2 Malignant Neoplasm
MV	H15 All Others

5.3.2 Sickness Category MI

The seven sickness groups in this sickness category MI are 'H1 Infective', 'H3 Benign neoplasm', 'H4 Endocrine & metabolic', 'H10 Respiratory', 'H11 Digestive', 'H12 Genito-urinary' and 'H14 Injuries'. There are a total of 551 deaths and 83.66 expected deaths in this category. Apart from 'H11 Digestive' which is the reference sickness group, the other sickness groups are coded using indicator variables and are added as covariates in Equation (5.11). For example, 'H12 Genito-urinary' is represented by I_{H12}. The parameterisation and estimation of the excess mortality intensity are described in Section 5.2. In respect of finding a suitable GM$(0, s)$ formula to the baseline intensity, we fitted Equation (5.11) to different GM$(0, s)$ formula at a range of k values ($k = 0, 0.1, 0.2, 0.3, \ldots, 2.9, 3.0$). Table 5.3 shows the deviance profile of fitting different GM$(0, s)$ formula at selected k values.

Table 5.3: Deviance profile for k with different GM(0,s) as baseline intensity for sickness category MI.

k	0.0	0.5	1.0	1.5	1.8	2.0	2.5	3.0
GM(0,3)	6249.4	6215.5	6187.5	6178.2	6177.2	6177.6	6180.6	6185.1
GM(0,4)	6247.1	6172.2	6173.5	6176.8	6177.2	6177.1	6176.4	6176.0
GM(0,5)	6239.8	6169.9	6169.6	6171.0	6172.6	6173.6	6175.3	6176.0

From Table 5.3, we note that the smallest residual deviance for GM(0,3) is produced at $k = 1.8$. The additional term in a GM(0,4) formula fitted at $k = 0.5$ is marginally significant and is positive, meaning that the mortality curves show an upturn at very long sickness durations, a feature present in Figure 5.1. The difference between the smallest residual deviance for GM(0,5) yielded at $k = 1.0$ and the smallest residual deviance for GM(0,4) produced at $k = 0.5$ is not statistically significant. Although GM(0,4) is the best formula statistically, we decided against using it and opted for a GM(0,3) formula instead because we felt that there is no reason why the mortality curves for these sickness groups should show an eventual mild upturn at very long sickness durations.

Thus, with a GM(0,3) formula and $k = 1.8$ as the baseline intensity, the excess mortality model is

$$\log(\mu_{jl}-d^*_{jl}) = \log(R_{jl})+b_0+b_1 t_{1.8}(z_j)+b_2(t_{1.8}(z_j))^2+I_{H1}+I_{H3}+I_{H4}+I_{H10}+I_{H12}+I_{H14}+\mathbf{x}_l\boldsymbol{\beta} \tag{5.14}$$

where the covariate vector $\mathbf{x}$ consist of $\{x_{\text{sex}}, x_{\text{rated}}, x_{\text{age}}, x_{\text{year}}, x_{\text{dp4}}, x_{\text{dp13}}, x_{\text{dp26}}, x_{\text{dp52}}\}$. To select which covariates to include from the above model, we assess the statistical significance of the main effects and all two-way interaction terms (including interaction term between duration variable and covariates) using the AIC statistics. The model after the AIC selection was applied is as follow:

$$\begin{aligned} \log(\mu_{jl} - d^*_{jl}) \;=\; & \log(R_{jl}) + b_0 + b_1 t_{1.8}(z_j) + b_2(t_{1.8}(z_j))^2 + I_{H3} + I_{H4} + I_{H10} \\ & + I_{H14} + \beta_{\text{sex}} x_{\text{sex}} + \beta_{\text{age}} x_{\text{age}} + \beta_{\text{year}} x_{\text{year}} + \beta_{\text{sex:H4}} x_{\text{sex}} I_{H4} \\ & + \beta_{\text{sex:H10}} x_{\text{sex}} I_{H10}. \end{aligned} \tag{5.15}$$

All the estimated parameters in Equation (5.15) are set out in Table 5.4. We note that

(i) Excess mortality intensity for 'H1 Infective' and 'H12 Genito-urinary' are not statistically different from 'H11 Digestive', the reference sickness group.

(ii) Excess mortality intensity for 'H3 Benign neoplasm' is elevated by a factor of 4.313 ($= \exp(1.4617)$) while the intensity for 'H14 Injuries' is reduced by a factor of 0.144 ($= \exp(-1.9396)$).

(iii) Female excess mortality intensity is 0.202 ($= \exp(-1.6018)$) times the intensity for males, but this factor is increased to 0.808 ($= \exp(-1.6018 + 1.3888)$) and 0.878 ($= \exp(-1.6018 + 1.4720)$) in 'H4 Endocrine & Metabolic' and 'H10 Respiratory', respectively.

(iv) A deterioration of the excess mortality intensity with an increase in age, given a positive age coefficient. The excess mortality intensity increases by a factor of 1.025 ($= \exp(0.6482/26)$) with every increment in age. This means that the age effect from the assured lives mortality alone does not account fully for the variation of mortality intensity by age.

(v) An improvement in the excess mortality intensity over the years, given a negative year coefficient. The excess mortality intensity decreases by a factor of 0.948 ($= \exp(-0.7008/13)$) with every passing year, in addition to the improvement in assured lives mortality intensity.

Table 5.4: Parameters in the excess mortality intensity model for sickness category MI.

Symbol	Estimate	Std.Error
b_0	-4.8649	0.2884
b_1	17.3217	1.9526
b_2	-31.9593	3.0942
I_{H3}	1.4617	0.1295
I_{H14}	-1.9396	0.1878
I_{H4}	-0.3329	0.2041
I_{H10}	-0.1696	0.1462
β_{sex}	-1.6018	0.3287
β_{age}	0.6482	0.1452
β_{year}	-0.7008	0.1090
$\beta_{\text{sex:H4}}$	1.3888	0.5698
$\beta_{\text{sex:H10}}$	1.4720	0.5332

The mortality curve for the baseline profile (male, non-rated, DP1, aged 43, year 1988) for each sickness group is shown in Figure 5.6. The value of the χ^2 statistic is 37.29565. With 47 cells, obtained in accordance to the grouping algorithm described in Appendix E, and 12 parameters fitted in the model, the probability value is 0.364 on 35 degrees of freedom, indicating a good fit to the data. Figure 5.7 shows the two-dimensional plot of the deviance residuals (as described in Section 3.6) for each sickness group, using combined data from both sexes, rating indicators and all deferred periods. These two-dimensional residuals plots are dominated by blue color because most of the cells have zero deaths. The white space are due to cells with no exposed-to-risk and for which deviance residuals cannot be calculated. The blocks of blue cells seemingly give the impression of inadequate fit. However, we reckon that in order for more meaningful conclusion to be drawn from these two-dimensional residuals plots, the number of deaths have to be sufficiently large or the age, year and sickness duration bands are partitioned in such a way that there are sufficient deaths in each cell.

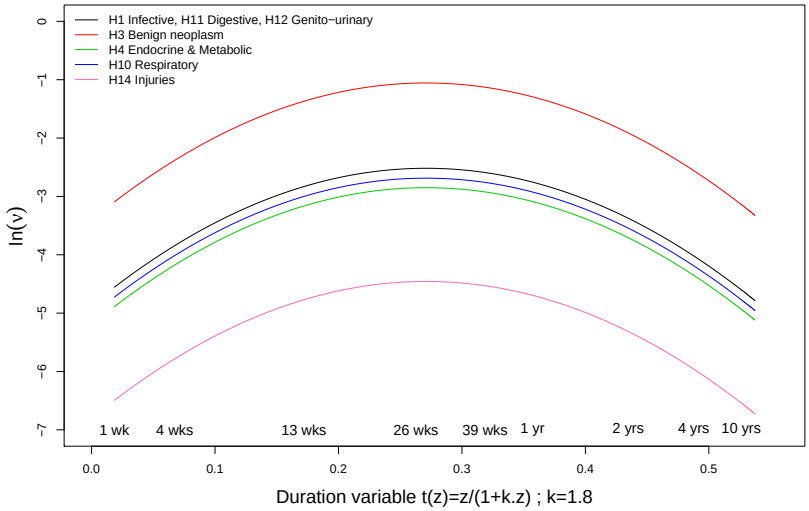


Figure 5.6: Fitted excess mortality intensity (on log scale) for sickness groups in sickness category MI using parameters in Table 5.4: male, non-rated, DP1, aged 43, year 1988.

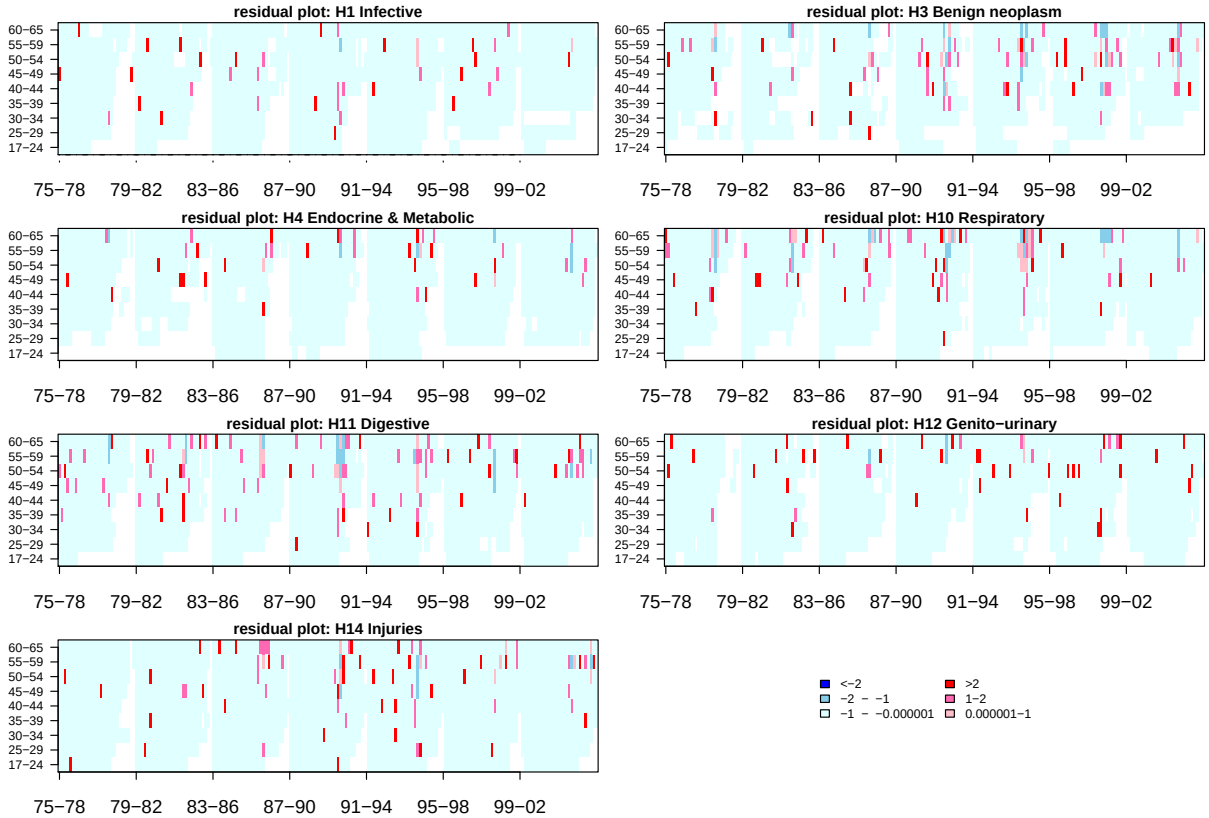


Figure 5.7: A two-dimensional plot of deviance residuals for each sickness group in sickness category MI.

5.3.3 Sickness Category MII

There are three sickness groups in this category with a total of 583 deaths and 247.13 expected deaths. These sickness groups are 'H5 Mental illness', 'H6 Nervous system & sense organs' and 'H13 Musculoskeletal'. Apart from 'H5 Mental illness' which is the reference sickness group, 'H6 Nervous system & sense organs' and 'H13 Musculoskeletal' are coded using indicator variables I_{H6} and I_{H13}, respectively, and are added as covariates in Equation (5.11). For the parameterisation of the baseline intensity, we fitted Equation (5.11) to different GM$(0, s)$ formula at a range of k values ($k = 0, 0.1, 0.2, 0.3, \ldots, 2.9, 3.0$). Table 5.5 shows the deviance profile of fitting different GM$(0, s)$ formula at selected k values.

Table 5.5: Deviance profile for k with different GM(0,s) as baseline intensity for sickness category MII.

k	0.0	0.2	0.5	1.0	1.5
GM(0,3)	6539.0	6524.7	6530.4	6537.0	6539.8
GM(0,4)	6536.2	6523.7	6528.4	6536.0	6538.8
GM(0,5)	6535.3	6523.6	6527.8	6535.5	6537.3

From Table 5.5, we note that the smallest residual deviance for GM(0,3) is produced at $k = 0.2$ and other GM$(0, s)$ functions of higher degree of varying k values do not give a significantly better fit. With a GM(0,3) formula at $k = 0.2$ for the baseline intensity, the excess mortality model is

$$\log(\mu_{jl} - d^*_{jl}) = \log(R_{jl}) + b_0 + b_1 t_{0.2}(z_j) + b_2 (t_{0.2}(z_j))^2 + I_{H6} + I_{H13} + \mathbf{x}_l \boldsymbol{\beta} \quad (5.16)$$

where the covariate vector $\mathbf{x}$ consist of $\{x_{\text{sex}}, x_{\text{rated}}, x_{\text{age}}, x_{\text{year}}, x_{\text{dp4}}, x_{\text{dp13}}, x_{\text{dp26}}, x_{\text{dp52}}\}$. To select which covariates to include from the above model, we assess the statistical significance of the main effects and all two-way interaction terms (including interaction term between duration variable and covariates) using the AIC statistics. The model after the AIC selection was applied is as follow:

$$\log(\mu_{jl} - d^*_{jl}) = \log(R_{jl}) + b_0 + b_1 t_{0.2}(z_j) + b_2 (t_{0.2}(z_j))^2 + I_{H6} + I_{H13} + \beta_{\text{sex}} x_{\text{sex}} + \beta_{\text{year}} x_{\text{year}} \quad (5.17)$$

All the estimated parameters in Equation (5.17) are set out in Table 5.6. We note that

(i) The excess mortality intensity for 'H6 Nervous system & sense organs' and 'H13 Musculoskeletal', compared to 'H5 Mental illness', are multiplied by factors of 2.453 (= exp(0.89720)) and 0.278 (= exp(−1.27924)), respectively.

(ii) Female excess mortality intensity is 0.5086 (= exp(−0.67619)) times the male intensity.

(iii) Year effect is significant and negative. The excess mortality intensity is reduced by a factor of 0.9213 (= exp(−1.06559/13)) with every single increase in year.

(iv) Age is not included as an explanatory variable, indicating that much of the age variation in the mortality intensity is already explained by the underlying assured lives mortality. This is not surprising given that over 42% (247.13/583 = 0.424) of the observed deaths were expected on the basis of the assured lives mortality.

Table 5.6: Parameters in the excess mortality intensity model for sickness category MII.

Symbol	Estimate	Std.Error
b_0	-4.09059	0.19514
b_1	-0.94310	0.20655
b_2	0.24153	0.04945
I_{H6}	0.89720	0.13815
I_{H13}	-1.27924	0.26424
β_{sex}	-0.67619	0.24625
β_{year}	-1.06559	0.13509

The mortality curves for the baseline profile (male, non-rated, DP1, aged 43, year 1988) for all sickness groups are shown in Figure 5.8. The value of the χ^2 statistic is 48.33466. With 50 cells and 7 parameters fitted, the probability value is 0.266 on 43 degrees of freedom, indicating a good fit to the data. Figure 5.9 shows the two-dimensional plot of the deviance residuals for each sickness group, using combined data from both sexes, rating indicators and all deferred periods. These two-dimensional residuals plots are dominated by blue because the majority of the cells have zero deaths.

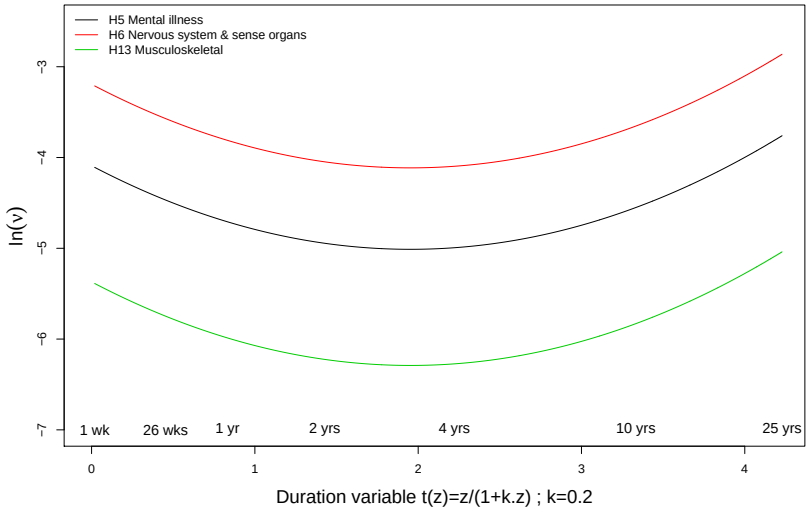


Figure 5.8: Fitted excess mortality intensity (on log scale) for each sickness group in sickness category MII using parameters in Table 5.6: male, non-rated, DP1, aged 43, year 1988.

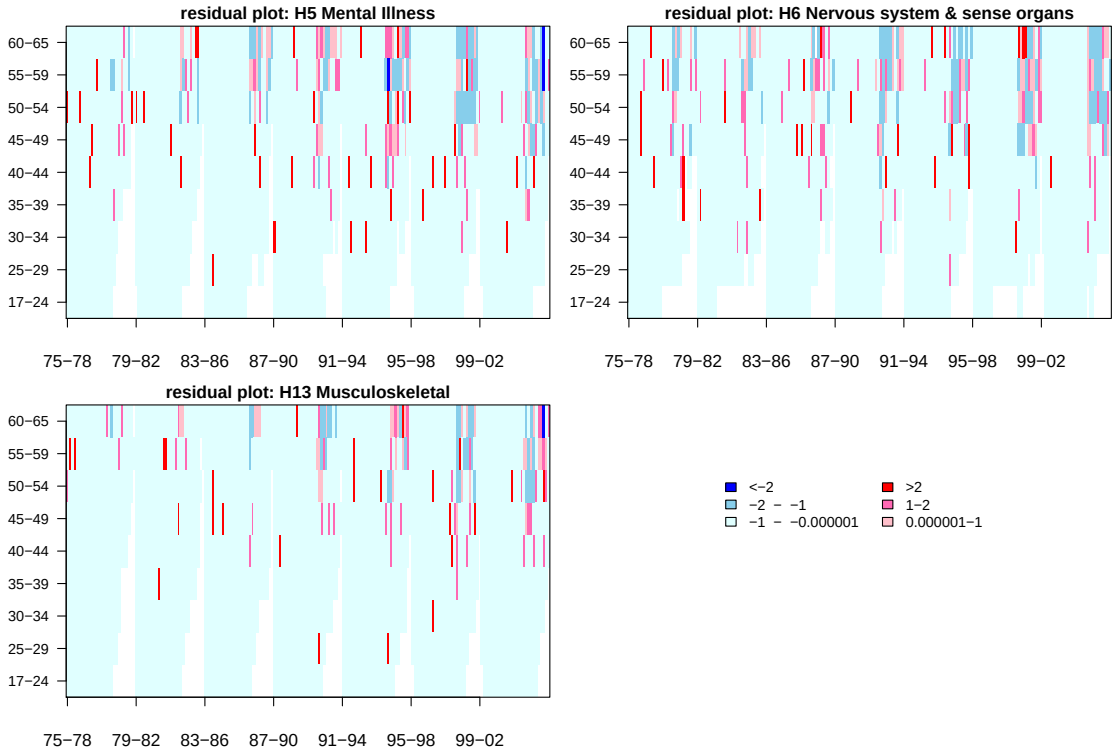


Figure 5.9: A two-dimensional plot of deviance residuals for each sickness group in sickness category MII.

5.3.4 Sickness Category MIII

There are three sickness groups in this category with a total of 552 deaths and 141.07 expected deaths. These sickness groups are 'H7 Other circulatory disease', 'H8 Ischaemic heart disease' and 'H9 Cerebrovascular disease'. Apart from 'H7 Other circulatory disease' which is the reference sickness group, 'H8 Ischaemic heart disease' and 'H9 Cerebrovascular disease' are coded using indicator variables I_{H8} and I_{H9}, respectively, and are added as covariates in Equation (5.11). For the parameterisation of the baseline intensity, we fitted Equation (5.11) to different GM$(0, s)$ formula at a range of k values ($k = 0, 0.1, 0.2, 0.3, \ldots, 2.9, 3.0$). Table 5.7 shows the deviance profile of fitting different GM$(0, s)$ formula at selected k values.

Table 5.7: Deviance profile for k with different GM(0,s) as baseline intensity for sickness category MIII.

k	0.0	0.5	1.0	1.5	1.8	2.0	2.5	3.0
GM(0,2)	5917.2	5897.7	5892.2	5890.2	5889.3	5889.5	5889.4	5889.6
GM(0,3)	5914.3	5888.3	5888.7	5889.2	5889.3	5889.3	5889.4	5889.3
GM(0,4)	5889.9	5888.2	5888.1	5888.2	5888.4	5888.5	5888.7	5888.9

From Table 5.7, we note that the smallest residual deviance for GM(0,2) is produced at $k = 1.8$ and no other GM$(0, s)$ functions of higher degree of different k values give a significantly better fit. Therefore, with a GM(0,2) formula at $k = 1.8$ as the baseline intensity, the excess mortality model is

$$\log(\mu_{jl} - d^*_{jl}) = \log(R_{jl}) + b_0 + b_1 t_{1.8}(z_j) + I_{H8} + I_{H9} + \mathbf{x}_l\boldsymbol{\beta} \tag{5.18}$$

where the covariate vector $\mathbf{x}$ consist of $\{x_{\text{sex}}, x_{\text{rated}}, x_{\text{age}}, x_{\text{year}}, x_{\text{dp4}}, x_{\text{dp13}}, x_{\text{dp26}}, x_{\text{dp52}}\}$. To select which covariates to include from the above model, we assess the statistical significance of the main effects and all two-way interaction terms (including interaction term between duration variable and covariates) using the AIC statistics. The model after the AIC selection was applied is as follow:

$$\begin{aligned}\log(\mu_{jl} - d^*_{jl}) \;=\; & \log(R_{jl}) + b_0 + b_1 t_{1.8}(z_j) + I_{H8} + I_{H9} + \beta_{\text{sex}} x_{\text{sex}} + \beta_{\text{age}} x_{\text{age}} \\ & + \beta_{\text{year}} x_{\text{year}} + \beta_{\text{year:H8}} x_{\text{year}} I_{H8} + \beta_{\text{year:t(z)}} x_{\text{year}} t_{1.8}(z_j) \\ & + \beta_{\text{year:age}} x_{\text{year}} x_{\text{age}}. \end{aligned} \tag{5.19}$$

All the estimated parameters in Equation (5.19) are set out in Table 5.8. We note that

(i) Compared to the excess mortality intensity of 'H7 Other circulatory diseases', the intensity for 'H8 Ischaemic heart disease' is reduced by a factor of 0.7756 ($= \exp(-0.2541)$) while the intensity for 'H9 Cerebrovascular disease' is reduced by a factor of 1.540 ($= \exp(0.4316)$).

(ii) Female excess mortality intensity is 0.4398 ($= \exp(-0.8215)$) times that of the male intensity.

(iii) Age coefficient is dependent on year. The coefficient of x_{age} is $0.5776+1.3444x_{\text{year}}$, resulting in an age factor ranging from $0.9709 = \exp((0.5776 - 1.3444)/26)$ in year 1975 to $1.0810 = \exp((0.5776 + 1.3444 \times 1.077)/26)$ in year 2002.

(iv) Year coefficient for 'H8 Ischaemic heart disease' is -1.0110 ($= -0.3745-0.6374$).

(v) Year coefficient varies linearly with $t_{1.8}(z)$. In the case of 'H7 Other circulatory diseases' and 'H9 Cerebrovascular disease', the linear dependency is given by $-0.3745-2.5180t_{1.8}(z)$. As sickness duration goes from 0 to the limit 1/1.8, the year factor reduces from 0.9716 ($= \exp(-0.3745/13)$) to 0.87248 ($= \exp((-0.3745 - 2.5180 \times (1/1.8)))/13$).

(vi) Year coefficient varies linearly with age according to the relationship $-0.3745\ -1.3444x_{\text{age}}$. As age increases from 20 to 70, the year factor increases from 0.8867 ($= \exp((-0.3745 + 1.3444 \times 0.8846)/13)$) to 1.0818 ($= \exp((-0.3745 + 1.3444 \times 1.0385))/13$).

The mortality curves for the baseline profile (male, non-rated, DP1, aged 43, year 1988) for all sickness groups are shown in Figure 5.10. The value of the χ^2 statistic is 27.17197. With 49 cells and 10 parameters fitted, the probability value is 0.923 on 39 degrees of freedom, indicating a good fit to the data. Figure 5.11 shows the two-dimensional plot of the deviance residuals for each sickness group, using combined data from both sexes, rating indicators and all deferred periods. These two-dimensional residuals plots are dominated by blue because most of the cells have zero deaths.

Table 5.8: Parameters in the excess mortality intensity model for sickness category MIII.

Symbol	Estimate	Std.Error
b_0	-2.6850	0.2032
b_1	-3.3266	0.4181
I_{H8}	-0.2541	0.1426
I_{H9}	0.4316	0.1566
β_{sex}	-0.8215	0.4149
β_{age}	0.5776	0.2255
β_{year}	-0.3745	0.3968
$\beta_{\text{year:H8}}$	-0.6374	0.2448
$\beta_{\text{year:t(z)}}$	-2.5180	0.8640
$\beta_{\text{year:age}}$	1.3444	0.4705

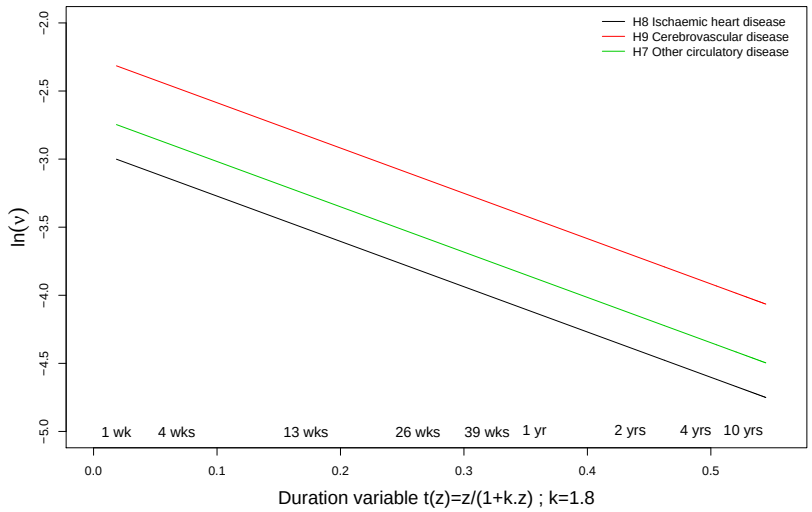


Figure 5.10: Fitted excess mortality intensity (on log scale) for each sickness group in sickness category MIII using parameters in Table 5.8: male, non-rated, DP1, aged 43, year 1988.

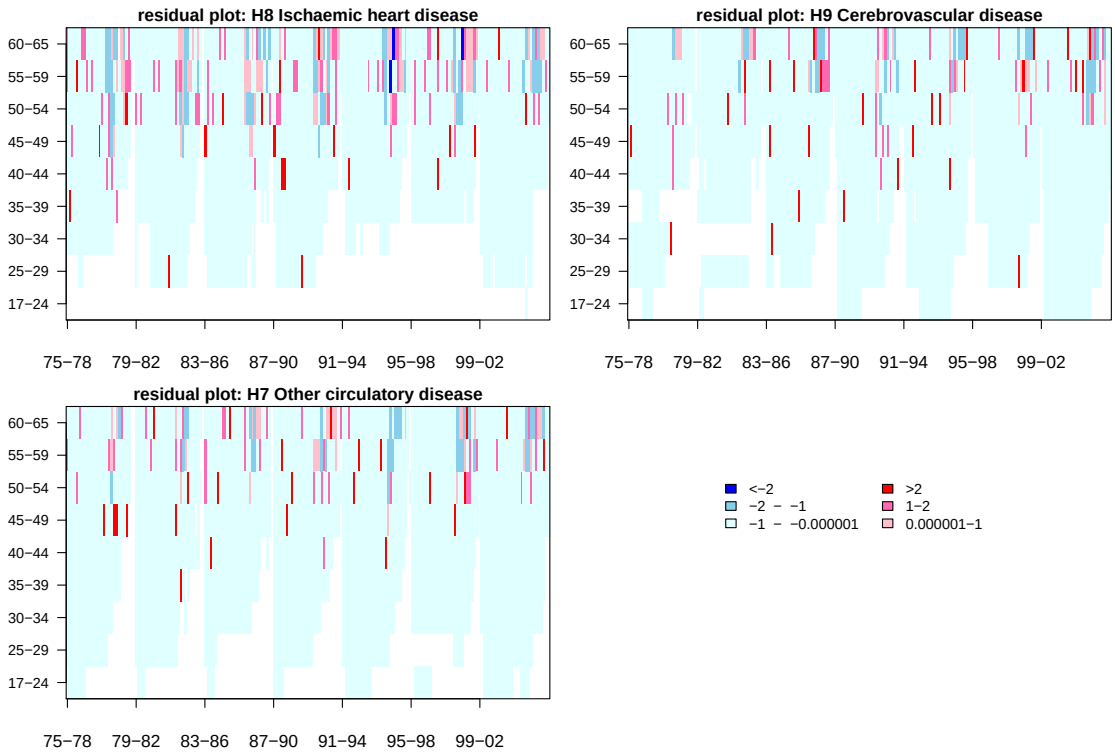


Figure 5.11: A two-dimensional plot of deviance residuals for each sickness group in sickness category MIII.

5.3.5 Sickness Category MIV

The sickness group in this category is 'H2 Malignant neoplasms' which has 1665 deaths and 21.97 expected deaths. With regards to the parameterisation of the baseline intensity, we fitted Equation (5.11) to different GM$(0, s)$ formula at a range of k values ($k = 0, 0.1, 0.2, 0.3, \ldots, 2.9, 3.0$). Table 5.9 shows the deviance profile of fitting different GM$(0, s)$ formula at selected k values.

Table 5.9: Deviance profile for k with different GM(0,s) as the baseline intensity for sickness category MIV.

k	0.0	0.5	1.0	1.5	1.8	2.0	2.5
GM(0,3)	12077	12044	12040	12053	12063	12069	12085
GM(0,4)	12075	12037	12038	12040	12043	12046	12052
GM(0,5)	12052	12036	12038	12035	12033	12033	12033
GM(0,6)	12041	12032	12027	12030	12032	12032	12032

From Table 5.9, we note that GM(0,6) at $k = 1.0$ is the optimal fit for the baseline intensity. The excess mortality model is therefore given by

$$\log(\mu_{jl} - d^*_{jl}) = \log(R_{jl}) + \sum_{i=0}^{5} b_i (t_{1.8}(z_j))^i + \mathbf{x}_l \boldsymbol{\beta} \tag{5.20}$$

where the covariate vector $\mathbf{x}$ consist of $\{x_{\text{sex}}, x_{\text{rated}}, x_{\text{age}}, x_{\text{year}}, x_{\text{dp4}}, x_{\text{dp13}}, x_{\text{dp26}}, x_{\text{dp52}}\}$. To select which covariates to include from the above model, we assess the statistical significance of the main effects and all two-way interaction terms (including interaction term between duration variable and covariates) using the AIC statistics. The model after the AIC selection was applied is as follow:

$$\begin{aligned} \log(\mu_{jl} - d^*_{jl}) \;=\; & \log(R_{jl}) + \sum_{i=0}^{5} b_i (t_{1.8}(z_j))^i + \beta_{\text{sex}} x_{\text{sex}} + \beta_{\text{age}} x_{\text{age}} + \beta_{\text{rated}} x_{\text{rated}} \\ & + \beta_{\text{year}} x_{\text{year}} + \beta_{\text{age:t(z)}} x_{\text{age}} t_{1.0}(z_j) + \beta_{\text{year:t(z)}} x_{\text{year}} t_{1.0}(z_j) \\ & + \beta_{\text{age:rated}} x_{\text{age}} x_{\text{rated}}. \end{aligned} \tag{5.21}$$

All the estimated parameters in Equation (5.21) are set out in Table 5.10. We note that

(i) Female excess mortality intensity is reduced by a factor 0.6008 ($= \exp(-0.50942)$) when compared to the male intensity.

(ii) Age coefficient varies linearly with $t_{1.0}(z)$, with the linear dependency described by $0.63060 - 1.54019 t_{1.0}(z)$. The age coefficient will become negative once sickness duration exceeds 253 days. A negative correlation between age coefficient and follow-up time is found in Dickman *et al.* (2004) where the excess mortality intensity for patients diagnosed with localised colon carcinoma and skin melanoma are investigated. The reason given by the author is that cancer patients with the worst prognosis or those who are frail and elderly are likely to die very soon after diagnosis while those who survive after a year will have a higher chance of survival.

(iii) Rated coefficient varies linearly with age and the relationship is given by $-0.23475 + 0.51680 x_{\text{age}}$. The coefficient ranges from -0.6919 to 0.3019 as age increases from 20 to 70.

(iv) Year coefficient varies linearly with $t_{1.0}(z)$ according to $-1.20784 + 1.08863 t_{1.0}(z)$, giving rise to the year coefficient ranging from -1.20784 ($t_{1.0}(z) = 0$) to -0.11922 ($t_{1.0}(z) = 1$).

(v) The up-turn in the excess mortality intensity (given a positive $(t_{1.0}(z))^5$ term) is probably due to heterogeneity in the type of cancer and the frailty of claimants.

Table 5.10: Parameters in the excess mortality intensity model for sickness category MIV.

Symbol	H2 Malignant neoplasms	Std. Error
b_0	−2.78945	0.53791
b_1	30.66597	8.03569
b_2	−149.28489	42.94951
b_3	351.54265	103.26848
b_4	−392.05986	114.04745
b_5	160.38971	46.99728
β_{sex}	−0.50942	0.07385
β_{age}	0.63060	0.20648
β_{rated}	−0.23475	0.07832
β_{year}	−1.20784	0.14423
$\beta_{\text{year}:t_{1.0}(z)}$	1.08862	0.29414
$\beta_{\text{age}:t_{1.0}(z)}$	−1.54019	0.40082
$\beta_{\text{age:rated}}$	0.51680	0.17477

The mortality curves for the baseline profile (male, non-rated, DP1, aged 43, year 1988) are shown in Figure 5.12. The value of the χ^2 statistic is 123.8620. With 130 cells and 12 parameters fitted, the probability value is 0.338 on 118 degrees of freedom, indicating a good fit to the data. Figure 5.13 shows the two-dimensional plot of the deviance residuals using amalgamated data from both sexes, rating indicators and all deferred periods. The two-dimensional residuals plot for female is dominated by blue because most of the cells have zero deaths. The two-dimensional plot of the deviance residuals for males and females separately are produced in Figure 5.13.

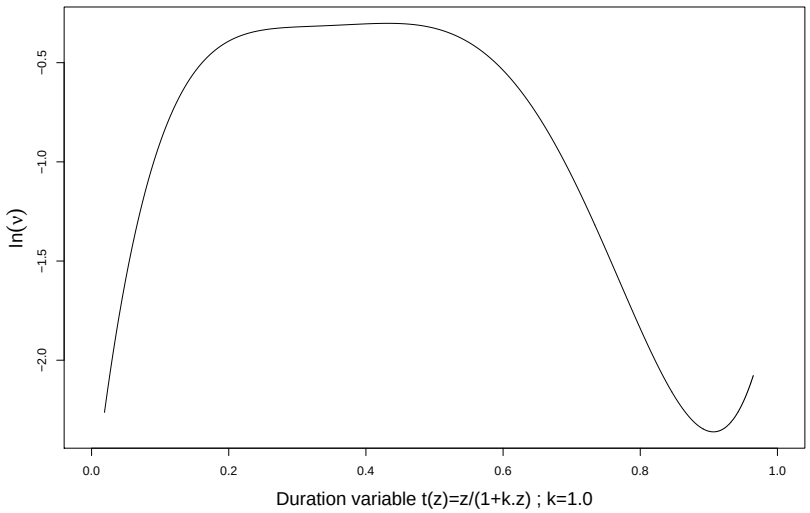


Figure 5.12: Fitted excess mortality intensity (on log scale) for 'H2 Malignant neoplasm' using parameters in Table 5.10: male, non-rated, DP1, aged 43, year 1988.

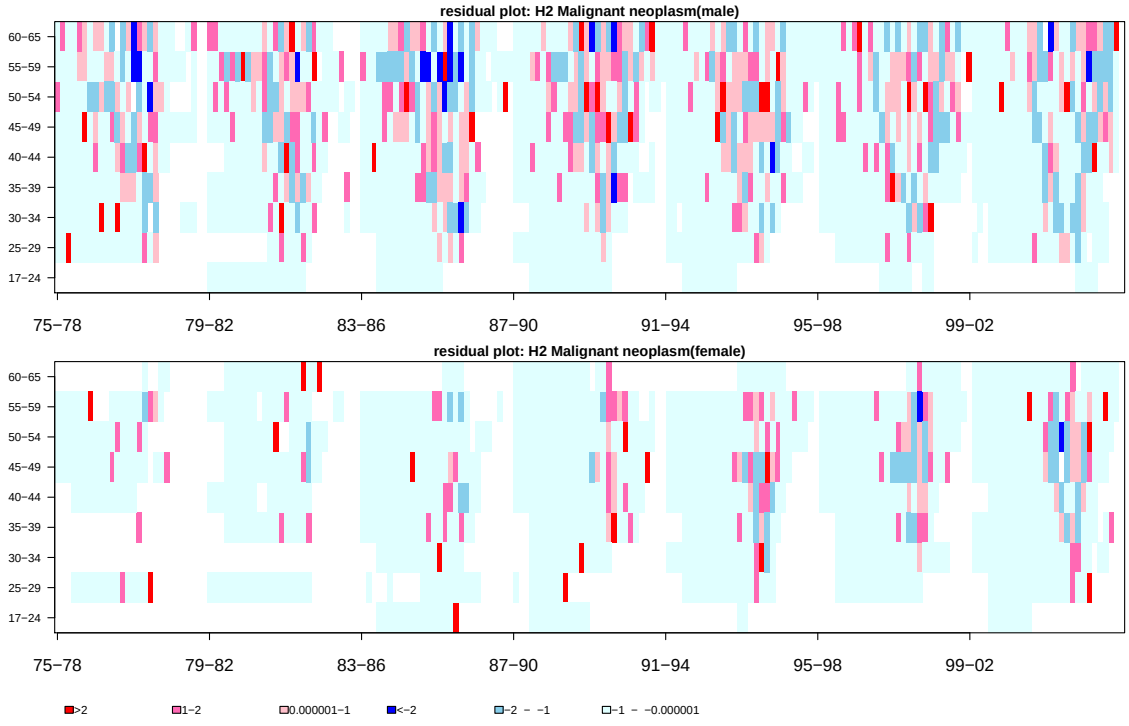


Figure 5.13: A two-dimensional plot of deviance residuals for 'H2 Malignant neoplasm', separately for males and females.

5.3.6 Sickness Category MV

The sickness group in this category is 'H15 All others' which has 147 deaths and 28.21 expected deaths. In respect to the baseline intensity, Table 5.11 shows the deviance profile of fitting different GM$(0, s)$ formulae to the baseline intensity at varying k values in Equation (5.10).

Table 5.11: Deviance profile for k with different GM(0,s) as the baseline intensity for sickness category MV.

k	0.0	0.2	0.5	1.0	1.5	2.0
GM(0,2)	1742.1	1740.4	1743.3	1748.7	1752.8	1755.9
GM(0,3)	1740.5	1740.1	1738.4	1737.9	1739.1	1740.9
GM(0,4)	1738.4	1736.9	1737.0	1737.6	1736.9	1736.6
GM(0,5)	1733.6	1735.8	1735.7	1735.0	1736.1	1736.6

With a GM(0,2) formula and $k = 0.2$ for the baseline intensity, the excess mortality model is therefore given by

$$\log(\mu_{jl} - d^*_{jl}) = \log(R_{jl}) + b_0 + b_1(t_{0.2}(z_j)) + \mathbf{x}_l\boldsymbol{\beta} \tag{5.22}$$

where the covariate vector $\mathbf{x}$ consist of $\{x_{\text{sex}}, x_{\text{rated}}, x_{\text{age}}, x_{\text{year}}, x_{\text{dp4}}, x_{\text{dp13}}, x_{\text{dp26}}, x_{\text{dp52}}\}$. To select which covariates to include from the above model, we assess the statistical significance of the main effects and all two-way interaction terms (including interaction term between duration variable and covariates) using the AIC statistics. The model after the AIC selection was applied is as follow:

$$\log(\mu_{jl} - d^*_{jl}) \;=\; \log(R_{jl}) + b_0 + b_1(t_{0.2}(z_j)) + \beta_{\text{sex}}x_{\text{sex}} + \beta_{\text{age}}x_{\text{age}} + \beta_{\text{year}}x_{\text{year}} \tag{5.23}$$

All the estimated parameters in Equation (5.23) are set out in Table 5.12. We note that

(i) Female excess mortality intensity is 0.3311 (= exp(−1.1052)) times that of the male intensity.

(ii) Age effect is significant and positive. Excess mortality intensity is elevated by a factor of 1.0538 (= exp(1.3633/26)) with every increment in age.

(iii) Year effect is significant and negative. Excess mortality intensity is reduced by a factor of 0.9333 (= exp(−0.8971/13)) with every single increase in year.

Table 5.12: Parameters in the excess mortality intensity model for sickness category MV.

Symbol	H15 All others	Std. Error
b_0	−3.0139	0.1834
b_1	−0.8778	0.1393
β_{sex}	−1.1052	0.4181
β_{age}	1.3633	0.2987
β_{year}	−0.8971	0.2249

The mortality curves for the baseline profile (male, non-rated, DP1, aged 43, year 1988) are shown in Figure 5.14. The value of the χ^2 statistic is 4.407133. With 13 cells and 5 parameters fitted, the probability value is 0.819 on 8 degrees of freedom, indicating a good fit to the data. Figure 5.15 shows the two-dimensional plot of the deviance residuals using combined data from both sexes, rating indicators and all deferred periods. This two-dimensional residuals plot is dominated by blue because most of the cells have zero deaths.

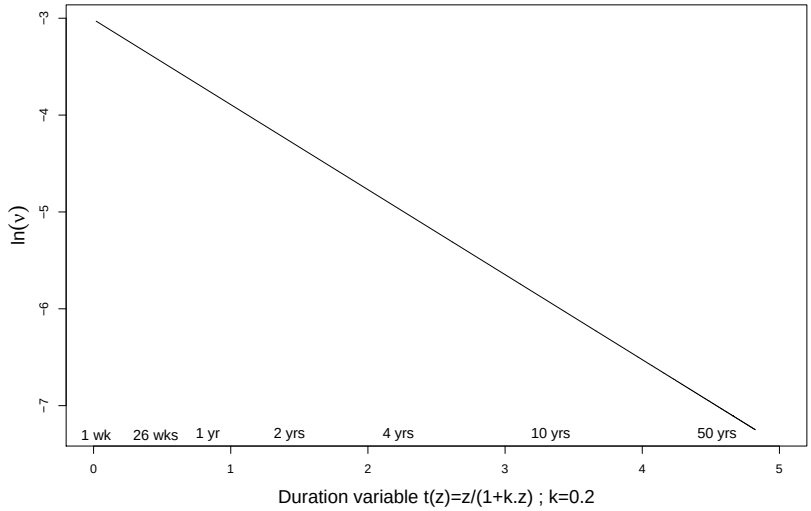


Figure 5.14: Fitted excess mortality intensity (on log scale) for sickness group 'H15 All Others' using parameters in Table 5.12: male, non-rated, DP1, aged 43, year 1988.

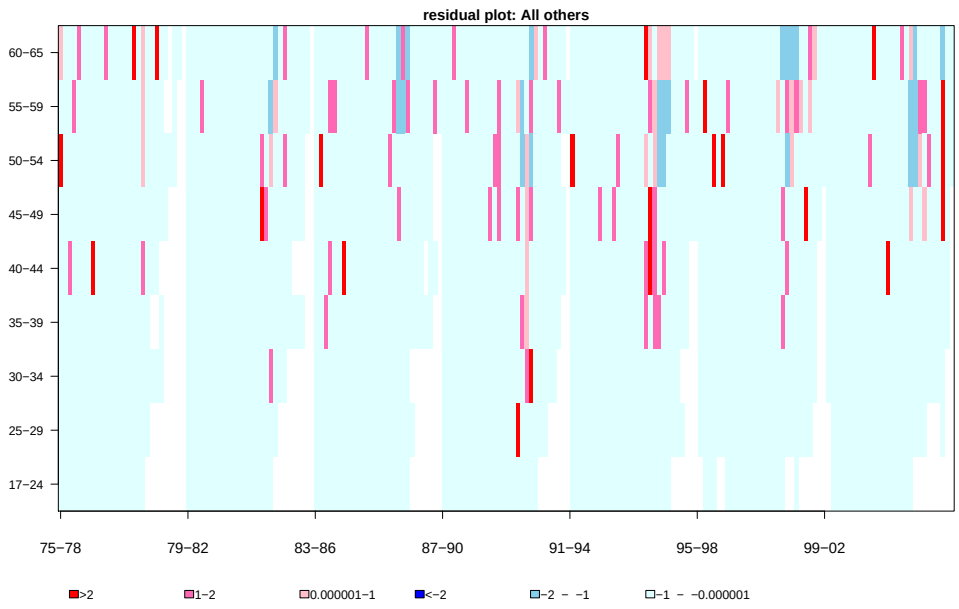


Figure 5.15: A two-dimensional plot of deviance residuals for sickness group 'H15 All Others'.

Chapter 6

Some Applications

6.1 Introduction

In this chapter we present two applications of the recovery intensity and mortality intensity from sick by cause of sickness, estimated in earlier chapters, in two aspects. In Section 6.2 we present the expected present values of annuities by cause of sickness. In Section 6.3 we derive the aggregate recovery and mortality intensities from sick and compare them against their corresponding graduated intensities from CMI Working Paper 5 (2004).

6.2 Expected Present Values of Annuities by Cause of Sickness

Under an IPI policy the life office is obliged to pay the IPI claimant a regular income (annuity) until he/she recovers, dies or reaches the retirement age of 60 (for females) or 65 (for males), whichever happens earliest. The claim termination intensity is therefore the sum of the recovery intensity and the mortality intensity from sick. For most causes of sickness, the recovery intensity forms the bulk of the claim termination intensity.

For reserving purposes, we wish to calculate the expected present value of liability for current claims. We consider the expected present value (EPV) of an annuity of £1 per annum payable continuously to a person who is currently aged x (in years) in year y and sick from cause of sickness i with duration of sickness z (in years). For a male IPI claimant, the maximum period for the annuity payment is $65-x$ years. We denote the EPV of this annuity by $\bar{a}^{\overline{S_iS_i}}_{x,y,z:\overline{65-x|}}$. The formula for $\bar{a}^{\overline{S_iS_i}}_{x,y,z:\overline{65-x|}}$ is given by

$$\bar{a}^{\overline{S_iS_i}}_{x,y,z:\overline{65-x|}} = \int_0^{65-x} v^t {}_tp^{\overline{S_iS_i}}_{x,y,z} dt \tag{6.1}$$

where v is the discounting factor and ${}_tp^{\overline{S_iS_i}}_{x,y,z}$ represents the probability that an individual currently aged x in year y with exact sickness duration z will remain sick continuously to exact sickness duration $z+t$.

To evaluate the above expression, we first need to calculate ${}_tp^{\overline{S_iS_i}}_{x,y,z}$, the formula for which is given by

$${}_tp^{\overline{S_iS_i}}_{x,y,z} = \exp\left\{-\int_0^t (\rho(i)_{x+s,y+s,z+s} + \lambda(i)_{x+s,y+s,z+s})ds\right\} \tag{6.2}$$

where $\rho(i)$ and $\lambda(i)$ are the recovery intensity and mortality intensity from being sick with cause of sickness i, respectively.

The above probability can be approximated by using a numerical method. For this purpose, CMI Report 12 (1991), in Part D, presented the following recursive approximate formula:

$${}_{t+h}p^{\overline{S_iS_i}}_{x,y,z} \approx \frac{{}_tp^{\overline{S_iS_i}}_{x,z}\left\{1-\frac{h}{2}(\rho(i)_{x+t,y+t,z+t}+\lambda(i)_{x+t,y+t,z+t})\right\}}{\left\{1+\frac{h}{2}(\rho(i)_{x+t+h,y+t+h,z+t+h}+\lambda(i)_{x+t+h,y+t+h,z+t+h})\right\}} \tag{6.3}$$

with ${}_0p^{\overline{S_iS_i}}_{x,y,z} = 1$. The derivation of this recursive formula is given in Appendix C. We then approximate the integral in Equation (6.1) by using Simpson's rule for numerical integration. Let $t_0 < t_1 < t_2 < \ldots < t_n$ and $t_{i+1} - t_i = h$ for $i = 0, 1, 2, \ldots, n-1$, where t_i and h are in unit of year and $n = (65-x)/h$. By using the repeated Simpson's rule, the approximation for the integral formula for the EPV of annuity in Equation (6.1) is given by

$$\bar{a}^{\overline{S_iS_i}}_{x,y,z:\overline{65-x}|} \approx \frac{h}{6}\sum_{i=0}^{n-1}\left(v^{t_i}{}_{t_i}p^{\overline{S_iS_i}}_{x,y,z} + 4v^{(t_i+\frac{h}{2})}{}_{t_i+\frac{h}{2}}p^{\overline{S_iS_i}}_{x,y,z} + v^{t_{i+1}}{}_{t_{i+1}}p^{\overline{S_iS_i}}_{x,y,z}\right) \tag{6.4}$$

We use the same step size h for both the approximations in Equations (6.3) and (6.4). The initial step size for h is 1/5840 of a year. The desired accuracy for $\bar{a}^{\overline{S_iS_i}}_{x,y,z:\overline{65-x}|}$ is attained by halving the step size until successive approximations are within £10^{-6} (absolute error) of each other.

6.2.1 Results

In this section we present the results of using Equations (6.3) and (6.4) to approximate the EPV of continuous annuity for each cause of sickness. One of the inputs in the claim termination intensity model is "year", which can be treated in two ways. We can either allow the year to advance along with duration z, which is the case in Section 6.2, or let time stand still at the starting date of the annuity. The former case may be more realistic for past periods, but for periods outside the range of the data, it involves projecting the intensities outside the range of the data, which, depending on the functional form of the year effect incorporated in the intensities models, may produce implausible results. For example, if the recovery intensity model that includes a non-linear year effect which is not a monotonic decreasing function (e.g. quadratic) is projected to the future, we may eventually obtain increasing intensities with year, a contrary to the falling year trend observed in the IPI data. Thus, the joint modelling of year and duration that resulted in a non-linear year effect may not be reliable for forecasting future intensities. All the results presented in this section are obtained by letting time stand still at the starting date of the annuity which is fixed at 2002. The current age increases along with duration z, but this is allowed for in the intensities anyway.

Tables 6.1 – 6.5 show the EPV of continuous annuities of £1000 per annum by cause of sickness, payable to a person with a non-rated occupation immediately after the end of deferred period for DP1, DP4, DP13, DP26 and DP52, respectively, with the payment ceasing at recovery, death or age 65. The values are calculated for age 20, 40 and 60 at sickness inception, separately for males and females. We use the same retirement age for both males and females so that a fair comparison can be made between their EPV of annuities.

Table 6.1: Expected present values of continuous annuities of £1000 per year payable to a DP1 IPI claimant with a non-rated occupation who falls sick at exact age 20, 40 and 60, calculated separately for males and females using rate of interest 4%.

	Males Age			Females Age				Males Age			Females Age		
cs	20	40	60	20	40	60	cs	20	40	60	20	40	60
1	197.82	219.20	313.97	432.16	460.35	509.44	37	629.77	562.42	971.76	668.26	596.83	1007.32
2	313.81	334.91	410.00	646.17	664.95	640.05	38	230.79	217.27	615.52	242.88	228.82	637.09
3	7.75	18.73	132.66	11.71	41.16	240.69	39	8.36	18.01	60.62	9.41	23.42	79.48
4	309.93	1238.02	1246.69	312.08	1328.00	1338.96	40	6.74	10.05	20.88	9.06	16.32	40.67
5	353.14	1491.87	1375.71	355.99	1605.80	1479.03	41	25.14	41.81	95.34	25.14	41.84	95.91
6	45.67	55.54	126.64	111.53	131.91	230.98	42	8.58	36.51	196.11	9.69	51.09	246.72
7	45.67	55.54	126.64	111.53	131.91	230.98	43	10.04	26.80	89.21	11.66	36.68	116.12
8	45.67	55.54	126.64	111.53	131.91	230.98	44	21.61	94.22	221.89	29.16	132.82	277.24
9	1439.52	1343.62	969.72	2368.68	2168.92	1326.76	45	18.27	76.93	758.83	18.32	77.72	770.78
10	45.67	55.54	126.64	111.53	131.91	230.98	46	8.84	22.49	63.10	7.24	15.23	42.37
11	45.67	55.54	126.64	111.53	131.91	230.98	47	92.13	482.45	388.23	93.62	524.25	416.83
12	406.07	1447.14	897.68	422.69	1585.17	965.41	48	16.48	24.09	36.11	26.13	40.31	74.27
13	183.07	511.02	752.47	183.72	537.56	803.40	49	18.75	27.54	42.41	29.96	49.33	99.63
14	45.67	55.54	126.64	111.53	131.91	230.98	50	35.38	54.93	86.69	35.42	55.60	88.40
15	191.89	213.14	308.43	420.74	449.20	501.73	51	81.89	457.67	637.89	1199.55	884.77	339.89
16	45.67	55.54	126.64	111.53	131.91	230.98	52	4.41	51.83	398.06	4.41	54.17	423.39
17	45.67	55.54	126.64	111.53	131.91	230.98	53	2.98	27.50	121.80	2.99	29.25	129.21
18	45.67	55.54	126.64	111.53	131.91	230.98	54	239.01	222.01	152.73	247.67	232.91	157.49
19	45.67	55.54	126.64	111.53	131.91	230.98	55M	13.97	93.77	187.90	-	-	-
20	165.26	765.11	1216.99	214.57	695.06	966.29	55F	-	-	-	39.78	170.56	740.39
21	24.40	145.80	330.23	294.45	243.74	134.57	56	-	-	-	392.55	1092.44	901.51
22	6718.59	2510.31	1098.92	6832.39	2559.33	1120.16	57	-	-	-	392.55	1092.44	901.51
23	1341.75	1307.04	1531.07	1361.56	1331.37	1561.37	58	-	-	-	392.55	1092.44	901.51
24	191.10	222.44	710.42	193.26	225.82	723.65	59	34.88	136.95	190.90	35.11	140.20	198.21
25	191.10	222.44	710.42	193.26	225.82	723.65	60	69.79	257.10	290.68	70.47	263.86	302.57
26	844.60	860.68	1282.01	856.49	876.14	1307.07	61	176.44	1403.80	1439.87	4841.50	4640.03	1711.56
27	953.35	1603.61	1813.03	1159.44	1869.38	1936.90	62	328.96	371.94	587.39	423.91	466.82	668.10
28	10.72	233.60	657.24	77.78	405.92	488.53	63	1658.15	2862.07	1562.92	1207.32	2358.79	1431.15
29	340.85	306.17	148.14	361.61	317.75	149.88	64	-	-	-	392.55	1092.44	901.51
30	5.48	63.21	344.50	18.17	123.00	231.86	65	606.49	1448.48	1039.81	392.55	1092.44	901.51
31	70.06	928.72	1251.62	469.24	1375.94	1033.27	66	122.14	237.79	255.78	196.42	379.53	354.30
32	29.35	30.36	181.37	29.81	30.89	186.57	67	52.79	203.38	319.12	44.54	170.34	286.14
33	2081.57	1749.50	1642.09	2223.88	1867.26	1704.54	68	1072.22	1176.83	1003.69	1747.50	1810.89	1296.10
34	499.94	451.85	876.61	529.64	478.81	908.41	69	258.41	334.48	480.29	505.76	610.80	690.62
35	11136.01	1110.18	1311.40	11490.31	1136.95	1334.99	70	64.52	90.79	208.44	131.52	181.03	331.61
36	2909.52	5309.38	2689.80	7358.61	8937.35	3453.91							

Table 6.2: Expected present values of continuous annuities of £1000 per year payable to a DP4 IPI claimant with a non-rated occupation who falls sick at exact age 20, 40 and 60, calculated separately for males and females using rate of interest 4%.

	Males			Females				Males			Females		
	Age			Age				Age			Age		
cs	20	40	60	20	40	60	cs	20	40	60	20	40	60
1	968.79	1003.76	949.53	1614.14	1632.76	1282.43	37	2730.00	2353.82	1970.69	2897.98	2500.12	2041.41
2	3435.85	3112.19	1786.67	4852.77	4344.42	2202.22	38	1637.06	1464.59	1611.27	1728.63	1549.05	1667.54
3	1165.17	2080.10	1826.47	1900.52	3076.90	2243.76	39	2970.68	1609.57	1800.83	3444.74	1910.28	1946.10
4	394.06	1732.14	1691.69	397.20	1862.52	1820.71	40	2160.69	897.72	1298.79	3303.91	1516.36	1646.12
5	442.28	2034.40	1818.62	446.31	2194.28	1958.91	41	1196.43	1322.55	734.58	1207.26	1338.92	744.78
6	1435.11	1430.27	1160.42	2282.05	2225.85	1524.09	42	3066.61	2340.46	2322.77	3548.42	2709.91	2464.49
7	1435.11	1430.27	1160.42	2282.05	2225.85	1524.09	43	3628.18	2028.24	1969.86	4150.70	2371.07	2115.14
8	1435.11	1430.27	1160.42	2282.05	2225.85	1524.09	44	5555.49	3326.66	2380.26	6170.36	3759.59	2520.98
9	6485.61	5444.91	2426.21	8335.40	6984.85	2852.83	45	1004.89	2079.35	2301.30	1017.04	2111.58	2340.15
10	1435.11	1430.27	1160.42	2282.05	2225.85	1524.09	46	164.41	450.47	607.12	121.45	320.16	499.73
11	1435.11	1430.27	1160.42	2282.05	2225.85	1524.09	47	712.17	2785.17	1463.18	727.20	3040.22	1579.87
12	747.68	2718.06	1566.24	781.61	2984.11	1689.20	48	73.25	71.92	65.59	69.99	69.15	63.61
13	250.72	803.32	1165.70	251.82	848.80	1248.82	49	212.66	199.50	140.13	202.23	192.92	136.72
14	1435.11	1430.27	1160.42	2282.05	2225.85	1524.09	50	41.55	86.29	138.19	41.82	88.97	142.94
15	2784.97	2583.90	1612.69	4052.99	3706.11	2018.68	51	489.99	1445.08	1295.89	2986.15	2288.25	912.73
16	1435.11	1430.27	1160.42	2282.05	2225.85	1524.09	52	132.55	833.95	1120.69	136.11	900.48	1202.62
17	1435.11	1430.27	1160.42	2282.05	2225.85	1524.09	53	2468.80	2690.21	1249.02	2647.48	2946.24	1343.04
18	1435.11	1430.27	1160.42	2282.05	2225.85	1524.09	54	418.83	380.90	226.15	438.07	405.09	236.57
19	1435.11	1430.27	1160.42	2282.05	2225.85	1524.09	55M	929.57	1198.63	716.02	-	-	-
20	1060.02	1826.33	1544.93	1602.48	2079.58	1424.56	55F	-	-	-	335.02	499.00	413.88
21	875.84	1994.76	1401.78	2636.66	2070.97	821.53	56	-	-	-	2627.55	4070.53	2146.58
22	8246.09	3734.21	1638.36	8380.18	3803.79	1669.53	57	-	-	-	2627.55	4070.53	2146.58
23	6871.61	5851.55	2952.33	6980.41	5966.67	3012.41	58	-	-	-	2627.55	4070.53	2146.58
24	3161.23	2969.67	2279.40	3205.47	3023.34	2324.54	59	1032.79	1950.93	1202.33	1045.56	2004.88	1253.82
25	3161.23	2969.67	2279.40	3205.47	3023.34	2324.54	60	1561.65	2608.24	1434.74	1582.16	2682.34	1497.41
26	5733.95	5000.04	2781.67	5822.30	5096.63	2837.94	61	1158.85	3735.76	2349.32	8626.47	7598.63	2596.16
27	1808.68	3276.77	2742.62	2949.09	4595.66	3068.26	62	916.27	2450.31	1755.76	1103.11	2758.36	1870.64
28	2140.63	3942.66	2178.27	4702.96	4853.40	1976.39	63	5518.57	6509.07	2678.96	4713.70	5991.74	2638.79
29	1575.86	1395.78	575.32	1682.80	1455.22	583.99	64	-	-	-	2627.55	4070.53	2146.58
30	1074.21	2546.05	1720.00	2929.14	3299.90	1501.27	65	3276.23	4625.36	2234.05	2627.55	4070.53	2146.58
31	4314.02	6088.50	2751.36	7542.58	7112.79	2595.52	66	525.33	881.37	650.24	815.51	1271.11	830.96
32	377.30	370.37	859.31	391.52	385.70	886.56	67	1126.08	1459.68	859.69	2005.66	2362.56	1197.41
33	5572.59	4583.13	2625.74	5948.83	4888.61	2722.60	68	2569.84	2602.10	1714.95	3622.65	3510.81	2024.94
34	3820.79	3226.53	2262.55	4065.88	3433.55	2344.75	69	786.41	925.00	965.24	1297.55	1440.90	1240.38
35	12894.84	2206.20	1863.54	13302.39	2260.49	1896.74	70	351.37	444.29	643.76	597.52	727.55	865.49
36	3686.21	6086.25	2951.84	8242.16	9573.47	3607.54							

Table 6.3: Expected present values of continuous annuities of £1000 per year payable to a DP13 IPI claimant with a non-rated occupation who falls sick at exact age 20, 40 and 60, calculated separately for males and females using rate of interest 4%.

	Males Age			Females Age				Males Age			Females Age		
cs	20	40	60	20	40	60	cs	20	40	60	20	40	60
1	6868.09	5904.34	2620.46	8672.70	7425.71	3020.13	37	5701.14	4854.79	2756.13	6040.23	5147.78	2847.59
2	7628.88	6456.20	2734.27	9491.15	8014.30	3130.91	38	7186.94	5983.30	2982.54	7637.94	6358.09	3082.69
3	2025.52	3124.56	2273.95	2977.74	4290.94	2677.34	39	14166.80	7103.69	3409.97	14694.06	7590.26	3507.62
4	497.59	2698.48	2365.53	502.61	2911.17	2549.49	40	13422.66	5858.43	3208.54	14591.76	7043.08	3400.73
5	554.23	3067.53	2467.56	560.46	3317.28	2660.69	41	1171.94	2662.07	1887.68	1180.12	2694.85	1917.32
6	4774.41	4322.80	2255.99	6335.79	5683.08	2659.34	42	14242.63	8013.14	3573.02	14767.04	8489.19	3662.65
7	4774.41	4322.80	2255.99	6335.79	5683.08	2659.34	43	14648.76	7654.04	3466.80	15157.39	8135.24	3561.71
8	4774.41	4322.80	2255.99	6335.79	5683.08	2659.34	44	15716.49	8959.66	3588.99	16179.58	9416.37	3677.80
9	10696.35	8588.07	3128.60	12667.70	10215.05	3508.64	45	4168.37	5407.00	3085.48	4218.44	5490.04	3137.16
10	4774.41	4322.80	2255.99	6335.79	5683.08	2659.34	46	1246.48	2378.14	1703.75	942.84	1947.19	1580.25
11	4774.41	4322.80	2255.99	6335.79	5683.08	2659.34	47	1582.85	6172.92	2802.67	1618.93	6742.05	3029.71
12	1580.70	5441.50	2608.93	1663.48	5983.25	2818.07	48	332.93	865.66	1012.59	1156.98	2445.28	1798.54
13	327.01	1443.31	1899.75	329.07	1536.30	2042.19	49	473.19	1173.78	1194.25	1576.40	3049.39	2004.32
14	4774.41	4322.80	2255.99	6335.79	5683.08	2659.34	50	1014.28	1666.45	1061.13	1075.39	1810.33	1134.73
15	6820.06	5869.14	2613.00	8620.55	7387.86	3012.83	51	1793.26	3439.23	2015.09	6233.63	4773.74	1677.87
16	4774.41	4322.80	2255.99	6335.79	5683.08	2659.34	52	914.99	2772.66	2070.86	953.14	3003.44	2226.87
17	4774.41	4322.80	2255.99	6335.79	5683.08	2659.34	53	6194.99	5791.25	2287.18	6637.21	6338.04	2462.62
18	4774.41	4322.80	2255.99	6335.79	5683.08	2659.34	54	1978.81	1757.10	843.29	2088.61	1894.59	899.91
19	4774.41	4322.80	2255.99	6335.79	5683.08	2659.34	55M	3200.80	3417.69	1651.22	-	-	-
20	874.79	2094.16	1852.00	1266.39	2484.23	1916.44	55F	-	-	-	4954.52	2595.88	634.02
21	3640.66	4594.75	2145.68	6763.44	5299.31	1909.04	56	-	-	-	6953.15	7848.56	3101.43
22	14549.12	9390.11	3192.00	14794.27	9575.05	3255.41	57	-	-	-	6953.15	7848.56	3101.43
23	9502.62	8041.91	3376.98	9646.12	8195.37	3444.70	58	7774.34	8265.62	3089.25	6953.15	7848.56	3101.43
24	5706.85	5261.54	2962.80	5781.48	5352.90	3020.91	59	2391.05	3648.99	1954.52	2416.33	3739.94	2031.34
25	5706.85	5261.54	2962.80	5781.48	5352.90	3020.91	60	6333.95	7056.22	2757.95	6417.36	7250.78	2871.94
26	8422.86	7282.80	3276.74	8545.96	7418.89	3342.12	61	2354.05	7403.89	3465.79	10664.04	10730.35	3604.16
27	3895.69	6177.86	3516.99	4381.74	6644.45	3598.16	62	2103.98	5902.19	3042.09	2421.00	6310.64	3132.32
28	4207.60	7366.80	3366.37	7178.86	8361.68	3299.34	63	10254.16	9827.19	3345.87	9495.80	9551.32	3396.81
29	6065.66	5332.09	1990.51	6493.06	5567.27	2022.33	64	-	-	-	6953.15	7848.56	3101.43
30	2707.00	5912.83	3120.12	5234.73	6877.66	3019.17	65	7774.34	8265.62	3089.25	6953.15	7848.56	3101.43
31	6664.72	9175.78	3627.94	9904.28	10157.57	3601.09	66	3817.84	4310.61	1920.57	4801.20	5164.44	2161.09
32	3870.91	3416.67	2409.70	4079.84	3609.13	2488.01	67	2427.45	3807.36	2026.30	2215.62	3607.24	1985.31
33	11632.12	9215.15	3504.69	12441.05	9833.93	3625.04	68	6387.58	5755.26	2627.74	7822.81	6858.42	2887.51
34	8552.48	6996.50	3162.85	9110.18	7446.42	3269.94	69	3562.71	3515.19	2065.56	4756.46	4511.82	2355.07
35	14787.13	4135.16	2486.41	15247.67	4239.38	2529.29	70	1819.13	1977.41	1551.86	2651.97	2746.75	1844.77
36	6878.93	8513.55	3387.31	11574.39	11439.10	3823.48							

Table 6.4: Expected present values of continuous annuities of £1000 per year payable to a DP26 IPI claimant with a non-rated occupation who falls sick at exact age 20, 40 and 60, calculated separately for males and females using rate of interest 4%.

	Males Age			Females Age				Males Age			Females Age		
cs	20	40	60	20	40	60	cs	20	40	60	20	40	60
1	11222.85	9062.40	3159.69	13124.00	10617.06	3488.00	37	5039.62	8377.17	3638.39	5294.28	8899.31	3752.12
2	11858.04	9477.87	3218.81	13757.38	11030.02	3543.01	38	3466.36	7153.77	3555.82	3618.84	7584.13	3666.55
3	5794.87	6552.53	2969.22	7406.81	8045.01	3309.46	39	677.43	5102.56	3296.03	799.25	5552.02	3384.40
4	1345.58	6929.96	3205.55	1374.66	7544.84	3439.50	40	481.64	3914.07	3110.27	759.95	4997.57	3285.86
5	1515.52	7340.69	3249.12	1550.91	7998.69	3486.66	41	1061.03	2295.13	1943.54	1065.55	2315.75	1970.80
6	9259.37	7740.68	2958.87	11125.25	9280.46	3299.70	42	702.78	6051.94	3446.44	828.86	6515.73	3527.34
7	9259.37	7740.68	2958.87	11125.25	9280.46	3299.70	43	861.03	5668.91	3348.45	1012.82	6128.45	3434.28
8	9259.37	7740.68	2958.87	11125.25	9280.46	3299.70	44	1548.98	7112.19	3461.17	1796.94	7578.09	3541.32
9	14137.42	10926.94	3412.60	15984.65	12446.80	3722.08	45	12466.57	10785.78	3557.64	12632.42	10956.38	3614.66
10	9259.37	7740.68	2958.87	11125.25	9280.46	3299.70	46	885.76	3373.82	2621.94	697.20	2886.26	2575.44
11	9259.37	7740.68	2958.87	11125.25	9280.46	3299.70	47	1886.45	7823.71	3288.05	1926.46	8506.10	3529.35
12	6677.01	10335.63	3405.74	7136.89	11339.38	3657.14	48	1456.49	2687.36	1948.61	3436.43	5190.48	2675.40
13	820.72	5149.23	2986.29	833.22	5580.74	3202.19	49	1872.79	3241.95	2116.86	4176.92	5913.72	2820.17
14	9259.37	7740.68	2958.87	11125.25	9280.46	3299.70	50	1165.42	1890.27	1217.51	1223.68	2034.11	1293.14
15	11181.57	9035.21	3155.76	13082.63	10589.91	3484.33	51	6076.72	7225.08	2810.52	11672.03	8847.79	2683.31
16	9259.37	7740.68	2958.87	11125.25	9280.46	3299.70	52	1889.00	4266.20	2463.63	1975.67	4613.25	2635.43
17	9259.37	7740.68	2958.87	11125.25	9280.46	3299.70	53	8197.47	7331.32	2633.52	8762.85	7990.15	2819.26
18	9259.37	7740.68	2958.87	11125.25	9280.46	3299.70	54	4851.18	4281.70	1872.81	5117.05	4610.73	1996.30
19	9259.37	7740.68	2958.87	11125.25	9280.46	3299.70	55M	4988.25	4979.25	2113.52	-	-	-
20	1311.53	2823.13	2045.70	2082.57	3604.13	2211.88	55F	-	-	-	253.01	2297.56	2233.89
21	5442.62	5856.30	2366.57	9124.11	7101.91	2355.15	56	-	-	-	5174.04	8824.26	3512.99
22	16606.98	11453.82	3469.04	16882.74	11675.01	3533.92	57	-	-	-	5174.04	8824.26	3512.99
23	13127.78	10556.12	3557.97	13330.52	10755.97	3624.82	58	-	-	-	5174.04	8824.26	3512.99
24	9856.51	8422.60	3354.44	9995.11	8573.15	3416.78	59	11352.11	10456.39	3341.64	11503.90	10733.69	3465.78
25	9856.51	8422.60	3354.44	9995.11	8573.15	3416.78	60	12333.50	11005.31	3416.28	12502.07	11299.91	3543.63
26	12274.04	10016.84	3510.16	12459.56	10204.03	3575.94	61	1564.22	9217.87	3819.89	13658.64	13697.69	3997.50
27	4187.59	8211.60	3728.16	4677.07	8689.71	3788.33	62	6033.27	8564.87	3550.86	6561.44	8953.71	3611.28
28	10467.77	11319.71	3765.75	13416.71	12203.25	3787.47	63	3114.89	6856.94	3192.33	2560.46	6351.84	3221.82
29	9620.92	8421.39	2958.42	10303.65	8792.50	3003.81	64	-	-	-	5174.04	8824.26	3512.99
30	8870.99	10509.71	3687.60	11951.39	11424.58	3695.57	65	5939.48	9175.52	3443.84	5174.04	8824.26	3512.99
31	12375.45	12182.20	3844.06	15062.92	13024.42	3879.92	66	1681.68	5685.97	3006.35	2291.99	6564.21	3175.63
32	1343.77	4621.97	3331.21	1382.23	4871.48	3433.84	67	5236.84	6414.37	2641.64	4953.87	6236.42	2627.26
33	7878.42	10033.75	3733.81	8340.26	10683.62	3851.01	68	8475.44	7363.46	2907.75	9913.10	8422.02	3116.62
34	4626.22	8084.81	3619.74	4852.97	8584.81	3732.79	69	5433.68	5092.40	2461.17	6794.30	6162.78	2707.58
35	16052.00	4688.90	3311.01	16541.57	4795.71	3365.90	70	3230.90	3291.58	2014.28	4342.90	4242.74	2284.29
36	10183.39	10401.71	3528.12	14352.81	12707.36	3817.72							

Table 6.5: Expected present values of continuous annuities of £1000 per year payable to a DP52 IPI claimant with a non-rated occupation who falls sick at exact age 20, 40 and 60, calculated separately for males and females using rate of interest 4%.

	Males Age			Females Age				Males Age			Females Age		
cs	20	40	60	20	40	60	cs	20	40	60	20	40	60
1	8992.64	7756.10	2860.41	10706.97	9154.54	3107.59	37	15107.64	11736.67	3401.38	16128.83	12479.85	3491.92
2	9693.54	8234.71	2918.02	11433.61	9640.67	3160.47	38	14185.53	11155.91	3359.41	15120.42	11851.55	3448.59
3	3918.60	5100.33	2675.97	5088.12	6346.78	2936.91	39	10348.73	9986.73	3183.45	10917.19	10376.42	3247.12
4	705.87	5368.58	3018.02	712.71	5739.34	3181.14	40	9363.71	9086.28	3071.06	10772.69	9999.90	3188.18
5	787.05	5793.88	3052.43	795.38	6202.40	3217.72	41	2312.02	3895.50	2455.74	2321.62	3927.79	2485.75
6	6968.44	6306.47	2666.00	8550.37	7647.57	2927.62	42	10453.43	10599.16	3272.71	11021.20	10969.10	3331.49
7	6968.44	6306.47	2666.00	8550.37	7647.57	2927.62	43	11028.20	10361.33	3214.73	11590.64	10739.31	3276.70
8	6968.44	6306.47	2666.00	8550.37	7647.57	2927.62	44	12657.26	11203.80	3281.37	13191.22	11551.31	3339.66
9	12394.51	9987.90	3108.04	14154.02	11380.44	3333.48	45	10340.20	9786.06	3238.03	10456.30	9920.28	3281.05
10	6968.44	6306.47	2666.00	8550.37	7647.57	2927.62	46	4513.23	5954.42	2622.11	10504.90	10328.98	3222.47
11	6968.44	6306.47	2666.00	8550.37	7647.57	2927.62	47	1982.08	9023.28	3294.04	2014.67	9704.44	3474.76
12	4834.29	10698.26	3272.51	5086.27	11590.28	3452.18	48	954.30	1697.08	1685.74	2002.21	3342.27	2305.95
13	453.27	3659.37	2843.82	456.28	3884.13	2995.93	49	1175.10	2052.55	1838.64	2430.61	3914.48	2441.73
14	6968.44	6306.47	2666.00	8550.37	7647.57	2927.62	50	4369.62	5147.58	2204.97	4591.08	5512.26	2316.70
15	8947.87	7725.15	2856.58	10660.24	9122.93	3104.07	51	6820.61	7868.64	2793.20	12202.57	9368.20	2705.64
16	6968.44	6306.47	2666.00	8550.37	7647.57	2927.62	52	4277.09	6733.16	2831.08	4463.37	7216.61	2981.25
17	6968.44	6306.47	2666.00	8550.37	7647.57	2927.62	53	10987.42	9378.28	2921.03	11672.14	10112.26	3077.02
18	6968.44	6306.47	2666.00	8550.37	7647.57	2927.62	54	7212.54	6344.55	2471.47	7566.93	6770.38	2597.71
19	6968.44	6306.47	2666.00	8550.37	7647.57	2927.62	55M	7969.65	7400.49	2633.95	-	-	-
20	1834.32	3651.12	2207.87	2906.11	4631.86	2377.40	55F	-	-	-	2144.84	6277.71	2958.34
21	6067.96	6525.77	2472.47	9511.91	7619.92	2484.41	56	-	-	-	16665.81	13470.60	3506.53
22	16292.69	11287.05	3237.81	16535.73	11477.63	3287.18	57	-	-	-	16665.81	13470.60	3506.53
23	12570.18	10366.09	3302.01	12738.12	10535.98	3352.61	58	12097.92	11043.40	3222.77	16665.81	13470.60	3506.53
24	9282.70	8226.04	3154.61	9390.52	8349.73	3202.39	59	10825.47	10313.35	3161.30	10947.76	10548.26	3253.41
25	9282.70	8226.04	3154.61	9390.52	8349.73	3202.39	60	11795.87	10854.90	3212.70	11933.82	11105.81	3306.66
26	11691.99	9818.59	3267.53	11843.23	9976.40	3317.47	61	2679.00	11385.03	3566.54	10383.81	13211.62	3607.74
27	6648.30	10678.26	3527.12	7262.60	11126.40	3569.65	62	15211.82	11207.56	3464.74	15703.49	11521.35	3504.23
28	8246.78	10159.08	3377.64	11179.86	11041.87	3388.22	63	6188.76	7336.36	2817.47	12292.07	11327.53	3319.52
29	11377.18	9891.98	3129.31	12190.90	10325.34	3172.44	64	-	-	-	16665.81	13470.60	3506.53
30	6620.36	9151.42	3302.88	9488.25	10052.32	3300.58	65	12097.92	11043.40	3222.77	16665.81	13470.60	3506.53
31	10416.07	11283.66	3452.80	13250.81	12130.57	3476.73	66	10553.59	9467.06	2918.04	11638.65	10208.56	3043.34
32	11906.99	9674.75	3242.93	12636.29	10251.62	3328.30	67	11801.61	10854.60	3192.28	11621.75	10835.99	3215.70
33	16233.42	12432.67	3449.31	17362.14	13233.45	3541.43	68	15895.31	12341.06	3364.05	16795.36	12951.51	3453.43
34	14895.20	11603.76	3391.94	15896.36	12336.01	3482.17	69	13943.04	11127.77	3236.21	15042.95	11873.65	3343.08
35	18442.43	11146.21	3350.00	19010.04	11443.77	3399.61	70	11892.51	9808.52	3086.13	13154.16	10677.40	3212.48
36	16261.47	13007.51	3479.74	18494.88	14244.77	3599.96							

We first examine the variation of the EPV of annuities by cause of sickness. We present in Figure 6.1 the claim termination intensity for cs35 (Ischaemic heart disease), cs61 (Arthritis & spondylitis) and cs62 (Musculoskeletal), separately for males and females, with respect to a DP1 IPI policyholder aged 40 at sickness inception and has a non-rated occupation. Given an inverse relationship between the claim termination intensity and the probability of remaining sick, it is expected that the higher the claim termination intensity, the lower the probability of remaining sick, which in turn leads to a lower EPV of annuity. It is shown in Figure 6.1 that cs39 has the highest level of claim termination intensity, followed by cs62 and cs61. This is also the order in which their corresponding EPV of annuities in Table 6.1 are arranged from lowest to highest.

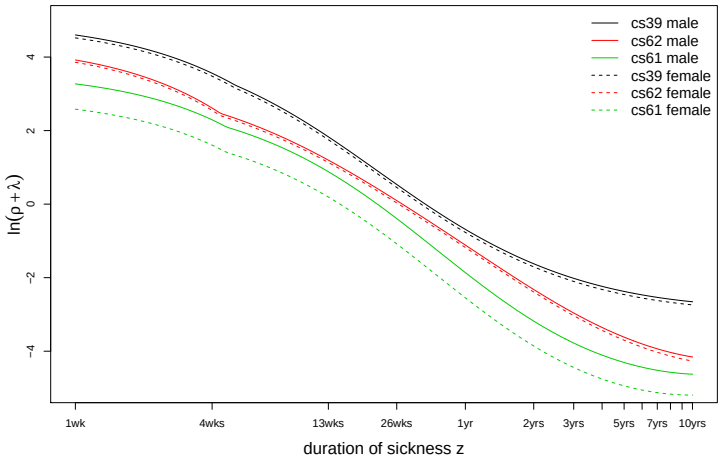


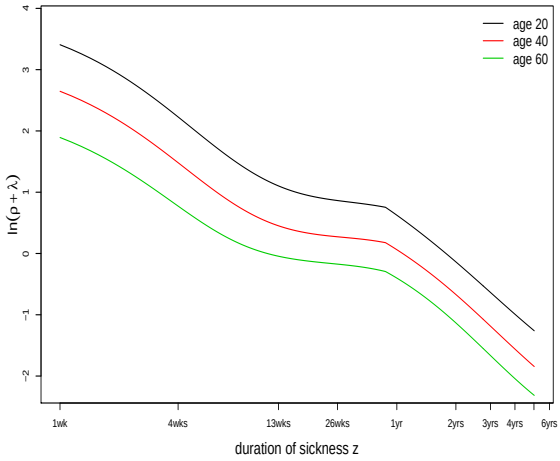
Figure 6.1: The claim termination intensity for cs39, cs61 and cs62, separately for males and females, with respect to a DP1 IPI policyholder aged 40 at sickness inception holding a non-rated occupation.

We also examine the variation of EPV of annuities by age at sickness inception. We present in Figure 6.2 the claim termination intensity for up to 5 years of sickness durations for a male DP1 IPI policyholder with a non-rated occupation who falls sick at ages 20, 40 and 60, separately for cs20 (Malignant neoplasms), cs35 (Ischaemic heart disease) and cs36 (Cerebrovascular disease). In general, the recovery intensity decreases with age while the mortality intensity from sick increases with age. Since the recovery intensity is the major component in the claim termination intensity, the variation of the claim termination intensity by age, at least for the initial period of sickness durations, is usually dictated by the age effect included in the recovery

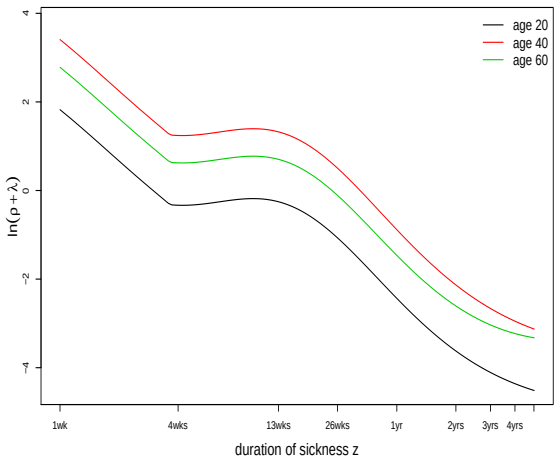
intensity model. For cs20 and cs36, due to a negative linear age effect included in their recovery intensity models, their claim termination intensities decrease as age at sickness inception increases. For cs35, due to a non-linear age effect that follows a humpbacked shape, the claim termination intensity for age 20 is lowest, followed by that for ages 60 and then 40. We also observe that the claim termination intensities for cs35 and cs36 at age 60 show an upturn at longer sickness durations, which is most likely due to the ageing process (which is modelled by the 'base' mortality component of the mortality intensity from sick model) taking effect. The maximum durations of payment for a person falling sick at exact age 20, 40 and 60 are 45, 20 and 5 years less 7 days, respectively. To make a fair comparison between the EPV of annuities for these different ages, we decided to restrict the payment period to at most 5 years less 7 days. Table 6.6 shows the EPV of such annuities for different ages at sickness inception, separately for cs20, cs35 and cs36. As expected, for each cause of sickness, the order in which the values are arranged from lowest to highest is opposite to the order in which the level of their corresponding claim termination intensity in Figure 6.2 are arranged from lowest to highest.

Table 6.6: Expected present values of continuous annuities of £1000 per year lasting for at most 5 years less 7 days for male DP1 policyholder with a non-rated occupation who falls sick at exact age 20, 40 and 60, separately for cs20, cs35 and cs36. rate of interest 4%.

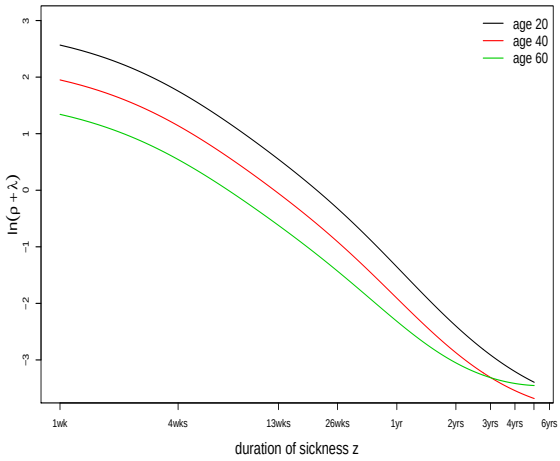
Age	20	40	60
cs20	157.05	562.17	1216.99
cs35	2790.01	510.90	1311.40
cs36	955.13	1888.93	2689.80



6.2.1: cs20 (Malignant neoplasms)



6.2.2: cs35 (Ischaemic heart disease)



6.2.3: cs36 (Cerebrovascular disease)

Figure 6.2: The claim termination intensity for male DP1 policyholder with non-rated occupation falling sick at ages 20, 40 and 60, separately for cs20, cs35 and cs36.

Finally, we examine the variation of EPV of annuities by deferred period. In Table 6.7, we present the EPV of continuous annuities payable to a male IPI claimant aged 40 at sickness inception holding a non-rated occupation and with the payment starting at 1, 4, 13, 26 and 52 weeks of sickness duration and ceasing at age 65 for different deferred periods, separately for cs35 (Ischaemic heart disease) and cs39 (Acute respiratory infections). The claim termination intensities by deferred period for both cs35 and cs39, upon which the values in Table 6.7 are calculated, are presented in Figure 6.3. From Table 6.7, we see that for both causes of sickness, the longer the sickness duration at which the payment starts for each deferred period, the higher the annuity value because the claim termination intensity generally decreases with sickness duration.

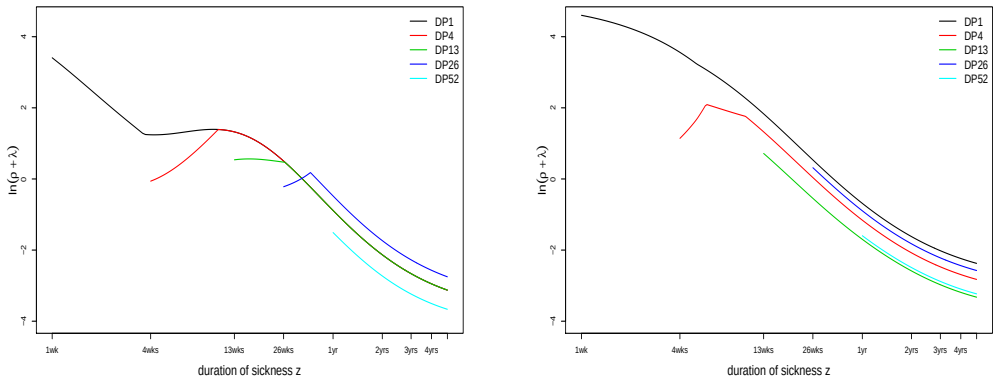
For cs35, apart from the 'run-in' period for DP4 and DP13, the termination intensities for DP4 and DP13 thereafter are the same as DP1. As a result, the EPV of annuities for DP4 with payment starting after the 'run-in period' are the same as their corresponding DP1 values. For annuity payment starting at 26 weeks of sickness duration, DP13 has a slightly different EPV of annuity than DP1 because the 'run-in' period for DP13 lasted until 185 days while DP26 has the lowest EPV of annuity because it has higher claim termination intensity than other deferred periods.

Focusing on cs39, for all sickness durations at which annuity payments start, DP1 has a lower EPV of annuity than those from higher deferred periods because the DP1 claim termination intensity lies above all other deferred periods. We also note that the EPV of annuities for DP1 payable immediately after the end of 1 week and 4 weeks are much lower than the values for a different starting duration from DP1 or the same starting duration from higher deferred periods. This is most likely because the vast majority of the recoveries for cs39 are from DP1 and they are concentrated in the first few weeks of sickness duration, leaving relatively few recoveries at longer sickness durations. As a result, after the initial few weeks of very high recovery intensity, the shape of the curve thereafter is very much guided by the presence of a few recoveries at very long sickness durations, resulting in an intensity curve that is extrapolated downwards quickly.

Table 6.7: Expected present values of continuous annuities of £1000 per year payable to a male aged 40 at sickness inception holding a non-rated occupation and with the payment starting at selected sickness duration and ceasing at age 65 for different deferred periods, separately for cs35 and cs39. rate of interest 4%.

			cs35		
Sickness duration (weeks)	DP1	DP4	DP13	DP26	DP52
1	1110.18				
4	1814.24	2206.20			
13	3291.79	3291.79	4135.16		
26	6072.87	6072.87	6074.24	4688.90	
52	8830.26	8830.26	8830.26	6921.29	11146.21

			cs39		
Sickness duration (weeks)	DP1	DP4	DP13	DP26	DP52
1	18.01				
4	191.22	1609.57			
13	1936.54	4099.73	7103.69		
26	4062.94	6469.04	9124.22	5102.56	
52	5904.83	8148.81	10363.37	6911.17	9986.73



6.3.1: cs35 6.3.2: cs39

Figure 6.3: The claim termination intensities by deferred period, separately for cs35 and cs39, for a male IPI claimant aged 40 at sickness inception holding a non-rated occupation.

6.3 Aggregate Recovery and Mortality Intensities from Sick

The graduated recovery and mortality intensities for IPI claimant presented in CMI Working Paper 5 (2004) are obtained by using all the sickness data regardless of the cause of sickness. The full graduation formulae for the recovery and mortality intensities in CMI Working Paper 5 (2004) are presented in Appendix D. In this section, we wish to compare these graduated intensities and their corresponding aggregate intensities obtained from using the cause-specific recovery and mortality intensities. We let the aggregate recovery and mortality intensities from sick for a portfolio of sicknesses consisting of different causes of sickness be represented by ρ^A and λ^A respectively. Denote cause of sickness by i, $i = 1, 2, \ldots I$. We assume that at sickness duration z_0, which is the starting point of annuity payment, the cause of sickness i is included in the portfolio with proportion $r(z_0, i)$, with $\sum_i r(z_0, i) = 1$. The aggregate recovery and mortality intensities from sick at the starting point z_0 are

$$\rho^A_{x,y,z_0} = \sum_{i=1}^{I} r(z_0, i)\rho(i)_{x,y,z_0}$$

and

$$\lambda^A_{x,y,z_0} = \sum_{i=1}^{I} r(z_0, i)\lambda(i)_{x,y,z_0}.$$

The probability of sickness surviving to duration z $(=z_0 + t)$, conditional on sickness existing at the starting duration z_0, is given by ${}_tp_{x,y,z_0}$ where

$${}_tp_{x,y,z_0} = \sum_{i=1}^{I} r(z_0, i)\,{}_tp^{\overline{S_i S_i}}_{x,y,z_0}$$

and the proportionate distribution of causes at sickness duration z is given by

$$r(z, i) = \frac{r(z_0, i)\,{}_tp^{\overline{S_i S_i}}_{x,y,z_0}}{{}_tp_{x,y,z_0}}.$$

Thus, the aggregate recovery intensity and mortality intensity form sick at sickness duration z are given by

$$\rho^A_{x,y,z} = \sum_{i=1}^{I} r(z,i)\rho(i)_{x,y,z}$$

and

$$\lambda^A_{x,y,z} = \sum_{i=1}^{I} r(z,i)\lambda(i)_{x,y,z}.$$

For the purposes of example, we will calculate the aggregate recovery and mortality intensities for a male aged 40 at sickness inception with a non-rated occupation, separately for each deferred period. In order to compare them against that in CMI Working Paper 5 (2004), the year at which these aggregate intensities are calculated is 1995, i.e the midpoint between 1991 and 1998. As in Section 6.2.1, we let the year stand still at the start of the annuity payment.

Apart from having cause-specific termination assumptions, we also need to know the proportion with which each cause of sickness is included in the portfolio of sickness belonging to male aged 40 at sickness inception with a non-rated occupation at the start of the annuity payment, i.e. at the end of deferred period. These proportions are taken as the proportion of claim inceptions by cause of sickness for each deferred period in the IPI male occupational class 1 data belonging to ages 37 to 43 from years 1991 to 1998, results of which are shown in Table 6.8.

Table 6.8: The proportion in percentage with which each cause of sickness is included in the portfolio belonging to male aged 40 at sickness inception with a non-rated occupation at the start of annuity payment for DP1, DP4, DP13, DP26 and DP52.

cs	DP1	DP4	DP13	DP26	DP52	cs	DP1	DP4	DP13	DP26	DP52
1	0.639	0.463	0.000	0.000	0.000	34	0.246	1.389	0.000	1.250	1.389
2	0.000	0.000	0.000	0.000	0.000	35	0.344	3.935	3.376	3.750	1.389
3	3.291	0.463	0.000	0.000	1.389	36	0.049	1.157	0.844	1.250	2.778
4	0.000	0.231	0.000	0.000	0.000	37	0.246	0.231	0.422	0.000	0.000
5	0.000	0.000	0.000	0.000	0.000	38	1.179	1.389	0.844	0.000	1.389
6	0.000	0.000	0.000	0.000	0.000	39	7.613	1.620	0.000	1.250	0.000
7	0.000	0.000	0.000	0.000	0.000	40	17.976	0.000	0.000	0.000	0.000
8	0.000	0.000	0.000	0.000	0.000	41	0.835	0.926	0.000	0.000	0.000
9	0.049	0.231	0.000	0.000	0.000	42	0.982	0.000	0.422	0.625	0.000
10	0.049	0.000	0.000	0.000	0.000	43	2.652	0.000	0.422	0.000	0.000
11	0.049	0.000	0.000	0.000	0.000	44	0.098	0.000	0.000	0.000	0.000
12	0.098	0.000	0.422	0.625	0.000	45	0.688	0.000	0.422	0.000	0.000
13	0.147	0.463	0.000	0.000	0.000	46	0.344	0.000	0.000	0.000	0.000
14	0.000	0.000	0.000	0.000	0.000	47	0.147	0.694	0.000	0.000	0.000
15	0.049	0.000	0.000	0.000	0.000	48	0.344	0.231	0.000	0.000	0.000
16	0.000	0.000	0.000	0.000	0.000	49	1.572	2.778	0.844	1.250	0.000
17	0.000	0.000	0.000	0.000	0.000	50	0.098	0.694	0.000	0.000	0.000
18	0.000	0.000	0.422	0.000	0.000	51	0.933	2.083	2.532	0.625	0.000
19	7.466	1.157	1.266	0.000	0.000	52	0.393	0.231	0.000	0.000	0.000
20	0.786	6.019	10.127	12.500	15.278	53	0.049	0.000	0.000	0.000	0.000
21	0.295	0.694	0.844	0.625	1.389	54	0.000	0.000	0.000	0.000	0.000
22	0.000	0.231	0.422	0.000	0.000	55M	1.916	1.157	0.000	0.000	0.000
23	0.000	0.231	0.000	0.000	0.000	59	1.523	0.463	0.000	0.000	0.000
24	0.000	0.000	0.000	0.000	0.000	60	1.277	0.694	0.422	0.000	1.389
25	0.295	0.694	0.422	1.250	0.000	61	0.835	1.157	0.844	1.875	1.389
26	0.000	0.000	0.000	0.000	0.000	62	21.906	20.833	16.878	10.625	11.111
27	5.501	20.370	29.958	35.625	40.278	65	1.130	4.398	3.797	9.375	8.333
28	1.081	0.231	0.844	1.250	0.000	66	5.157	9.259	8.439	6.250	2.778
29	0.098	0.000	0.000	0.000	0.000	67	2.603	7.176	8.017	2.500	2.778
30	0.000	0.000	0.000	0.000	1.389	68	0.196	0.000	0.000	0.625	0.000
31	4.224	3.472	5.063	6.250	4.167	69	0.000	0.463	0.000	0.000	0.000
32	0.000	0.000	0.000	0.000	0.000	70	2.554	2.083	1.688	0.625	0.000
33	0.000	0.000	0.000	0.000	1.389						

Recovery intensity

With respect to the recovery intensity, Figures 6.4–6.8 show the aggregate recovery intensities for a male DP1 IPI policyholder aged 40 at sickness inception with a non-rated occupation for DP1, DP4, DP13, DP26 and DP52, respectively. We also overlaid on these graphs the graduated recovery intensities from CMI Working Paper 5 (2004). From these figures, we note that the aggregate recovery intensities for DP1, DP4 and DP13 are rather close to their corresponding intensities from CMI Working Paper 5 (2004). For DP26 and DP52, the difference between both set of intensities

are greater, most likely due to fewer number of recoveries underlying both set of intensities. The number of recoveries belonging to male aged 40 at sickness inception for DP1, DP4, DP13, DP26 and DP52 are 1365, 807, 275, 93 and 20, respectively. For DP4, each cause of sickness has a different break point and for the majority of them, the breakpoint occurs between 40 to 50 days, resulting in a 'smooth' curve in the aggregate recovery intensity that occurs before the distinct break point at 56 days (8 weeks) from CMI Working Paper 5 (2004). For DP13, we also observe a 'smooth' curve in the aggregate recovery intensity as opposed to a distinct break point and the location of the 'smooth' curve is close to the distinct break point at 119 days from CMI Working Paper 5 (2004) because the break point for each individual cause of sickness is close to 119 days.

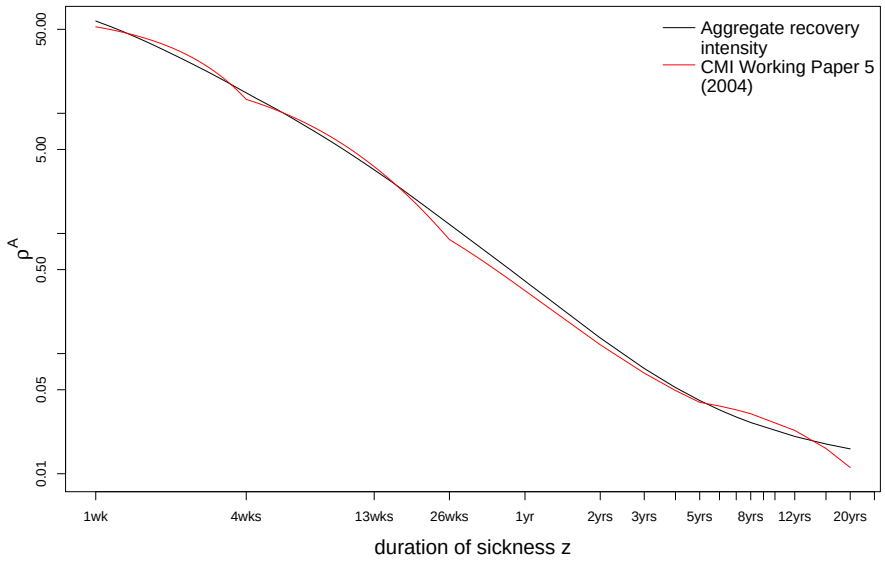


Figure 6.4: A comparison between the aggregate recovery intensities obtained given cause-specific recovery intensities and those from CMI Working Paper 5 (2004) with respect to a male DP1 IPI policyholder aged 40 at sickness inception with a non-rated occupation.

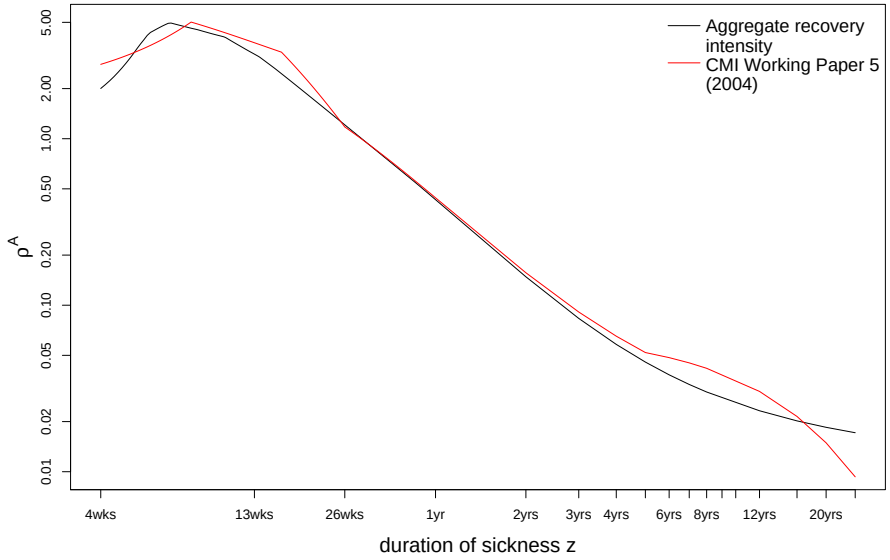


Figure 6.5: A comparison between the aggregate recovery intensities obtained given cause-specific recovery intensities and those from CMI Working Paper 5 (2004) with respect to a male DP4 IPI policyholder aged 40 at sickness inception with a non-rated occupation.

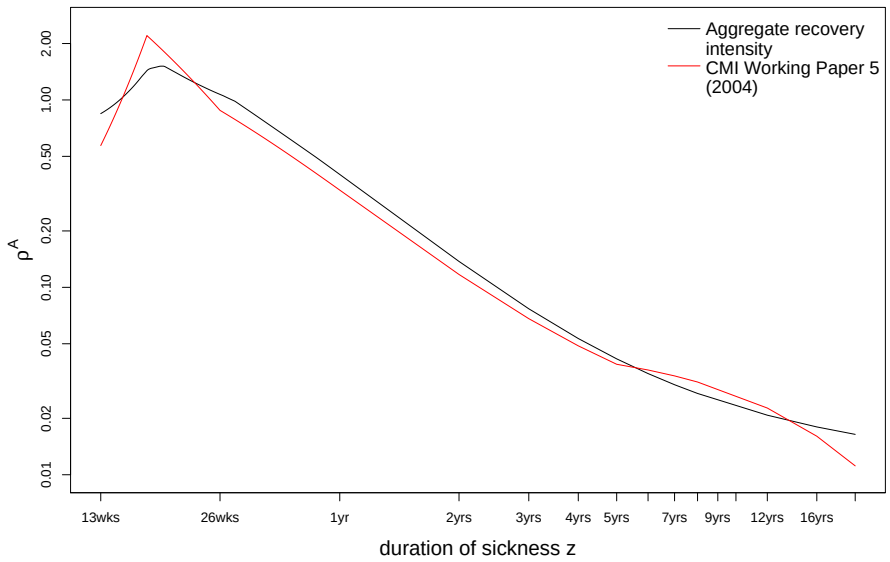


Figure 6.6: A comparison between the aggregate recovery intensities obtained given cause-specific recovery intensities and those from CMI Working Paper 5 (2004) with respect to a male DP13 IPI policyholder aged 40 at sickness inception with a non-rated occupation.

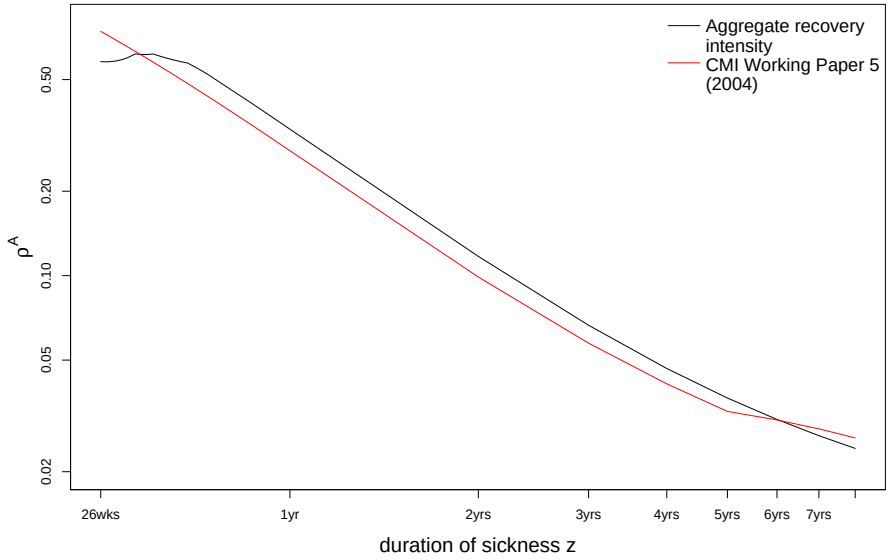


Figure 6.7: A comparison between the aggregate recovery intensities obtained given cause-specific recovery intensities and those from CMI Working Paper 5 (2004) with respect to a male DP26 IPI policyholder aged 40 at sickness inception with a non-rated occupation.

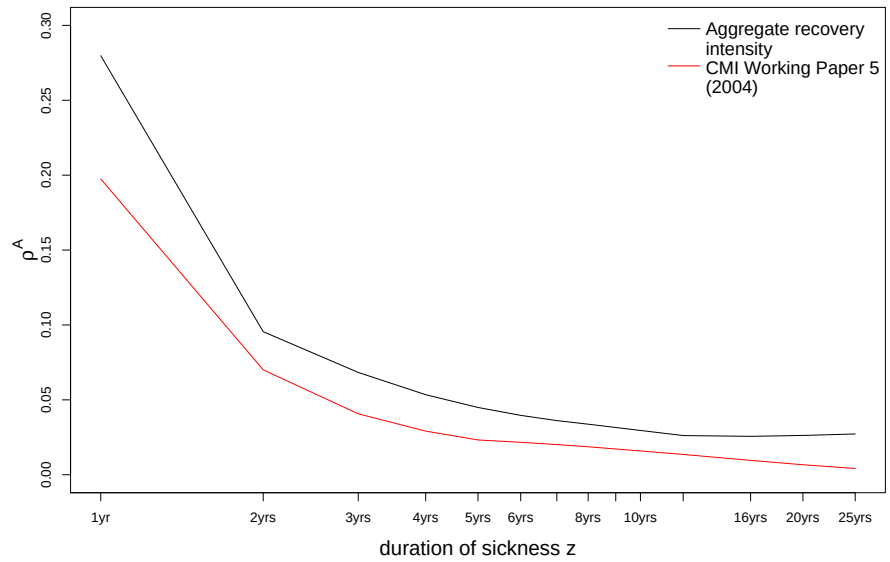


Figure 6.8: A comparison between the aggregate recovery intensities obtained given cause-specific recovery intensities and those from CMI Working Paper 5 (2004) with respect to a male DP52 IPI policyholder aged 40 at sickness inception with a non-rated occupation.

Mortality intensity from sick

With respect to the mortality intensity from sick, Figures 6.9–6.13 show the aggregate mortality intensities for up to 25 years of sickness duration in relation to a male IPI claimant aged 40 at sickness inception with a non-rated occupation for DP1, DP4, DP13, DP26 and DP52 respectively. We also overlaid on these graphs the graduated mortality intensities from CMI Working Paper 5 (2004). For DP1, although both set of intensities follow a hump-backed shape, the intensities from CMI Working Paper 5 (2004) are higher than the aggregate intensities for most sickness durations, with the difference being most marked in the region near the peak value at around 22 weeks of sickness duration. For both DP4 and DP13, the intensities from CMI Working Paper 5 (2004) lie above the aggregate intensities until both intensities become very close to each other after around 1 year of sickness duration. For both DP26 and DP52, the aggregate intensities are generally higher than the intensities from CMI Working Paper 5 (2004). The eventual upturn in both sets of intensities for all deferred periods is due to the more dominant ageing process taking effect.

In general, for all deferred periods, both sets of mortality intensities are more similar at the later sickness durations, during which most of the deaths are concentrated, than at the initial sickness durations when relatively few deaths occur. Nevertheless, for all deferred periods, the difference between both sets of mortality intensities is more significant than the difference between their counterparts for recovery intensities. This is likely due to a small number of deaths, particularly at younger ages, underlying both sets of mortality intensities. Therefore, these estimated mortality intensities may not be as 'reliable' as their corresponding recovery intensities which are estimated based on a large number of recoveries. In the IPI data used by CMI Working Paper 5 (2004), the number of deaths (recoveries) belonging to those under age 40 at sickness inception for DP1, DP4, DP13, DP26 and DP52 are 7(2366), 14(568), 13(130), 9(55) and 4(11), respectively.

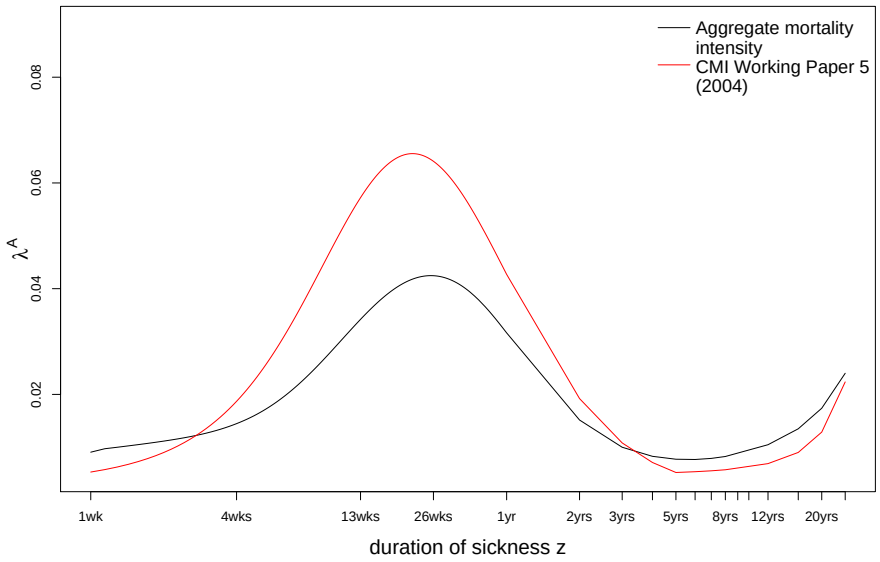


Figure 6.9: A comparison between the aggregate mortality intensities obtained given cause-specific recovery intensities and those from CMI Working Paper 5 (2004) with respect to a male DP1 IPI policyholder aged 40 at sickness inception with a non-rated occupation.

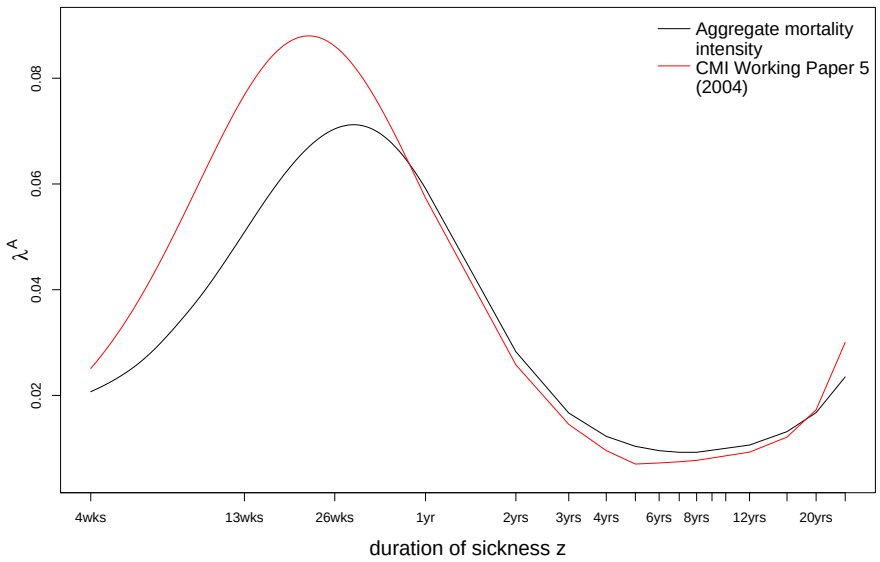


Figure 6.10: A comparison between the aggregate mortality intensities obtained given cause-specific recovery intensities and those from CMI Working Paper 5 (2004) with respect to a male DP4 IPI policyholder aged 40 at sickness inception with a non-rated occupation.

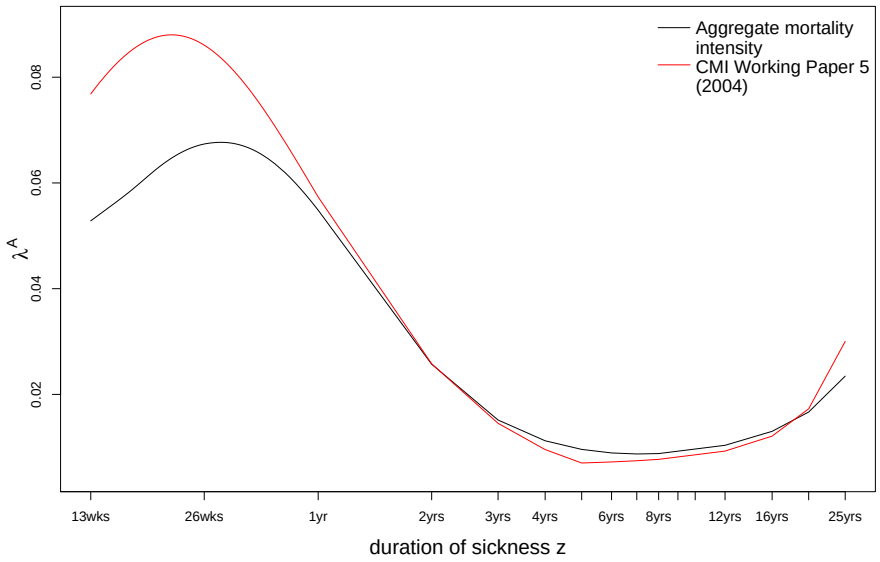


Figure 6.11: A comparison between the aggregate mortality intensities obtained given cause-specific recovery intensities and those from CMI Working Paper 5 (2004) with respect to a male DP13 IPI policyholder aged 40 at sickness inception with a non-rated occupation.

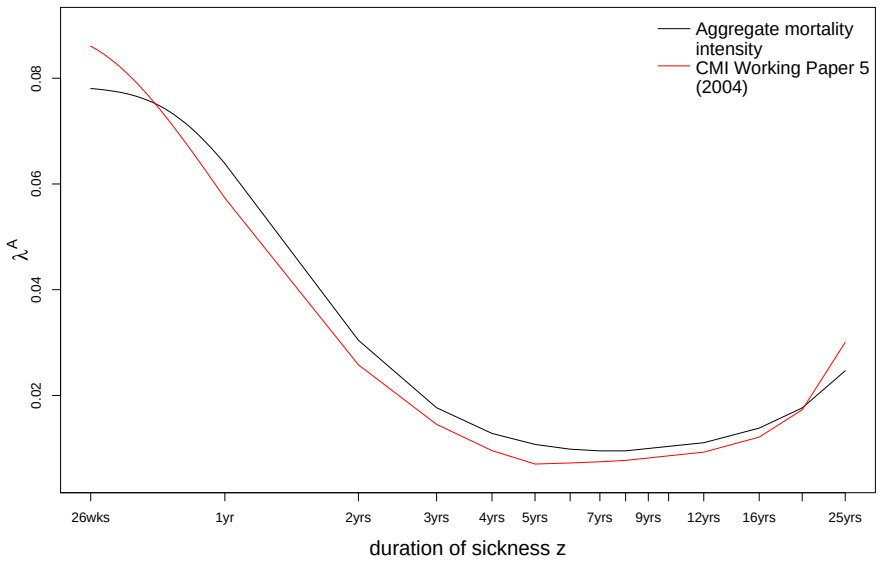


Figure 6.12: A comparison between the aggregate mortality intensities obtained given cause-specific recovery intensities and those from CMI Working Paper 5 (2004) with respect to a male DP26 IPI policyholder aged 40 at sickness inception with a non-rated occupation.

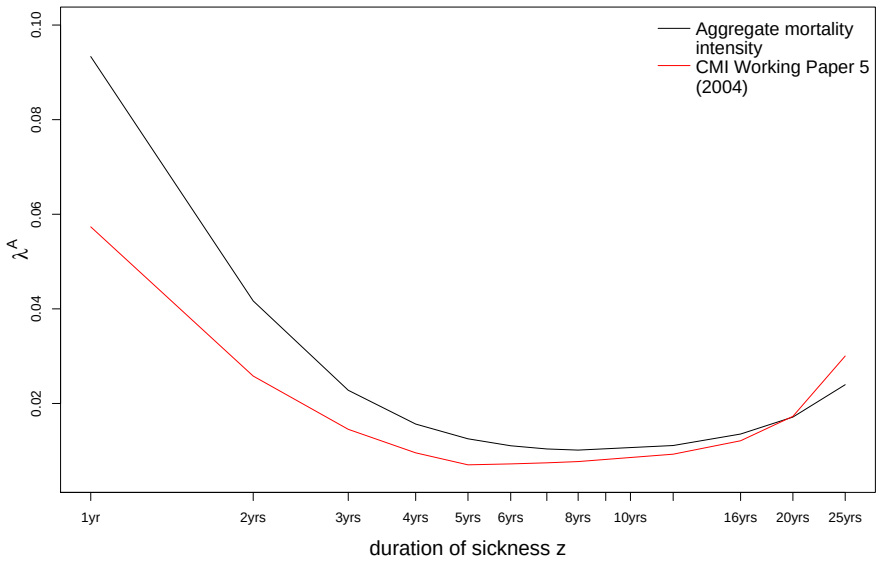


Figure 6.13: A comparison between the aggregate mortality intensities obtained given cause-specific recovery intensities and those from CMI Working Paper 5 (2004) with respect to a male DP52 IPI policyholder aged 40 at sickness inception with a non-rated occupation.

Comparison at age 60

In the IPI data used by CMI Working Paper 5 (2004), the number of deaths belonging to those over age 60 at sickness inception for DP1, DP4, DP13, DP26 and DP52 are 19, 20, 18, 20 and 10, respectively. Since there are relatively more deaths at higher ages, we carry out the same comparison in relation to a male IPI claimant aged 60 at sickness inception. The distribution of sicknesses at the start of annuity payment for each deferred period is taken as the proportion of claim inceptions by cause of sickness, separately for each deferred period, in the IPI male occupational class 1 data from years 1991 to 1998 belonging to ages 57 to 63, results of which are shown in Table 6.9.

Table 6.9: The proportion in percentage with which each cause of sickness is included in the portfolio belonging to male aged 60 at sickness inception with a non-rated occupation at the start of annuity payment for DP1, DP4, DP13, DP26 and DP52.

cs	DP1	DP4	DP13	DP26	DP52	cs	DP1	DP4	DP13	DP26	DP52
1	0.138	0.175	0.000	0.000	0.000	34	1.866	1.573	1.509	3.738	5.455
2	0.069	0.000	0.000	0.000	0.000	35	9.053	19.755	21.983	23.598	16.970
3	1.244	0.000	0.000	0.000	0.000	36	1.106	3.497	5.819	5.841	4.848
4	0.069	0.000	0.000	0.467	0.000	37	0.415	1.224	0.647	0.701	0.606
5	0.000	0.000	0.216	0.000	0.000	38	2.695	1.573	2.802	0.701	0.606
6	0.000	0.000	0.000	0.000	0.000	39	5.045	0.350	0.216	0.701	0.606
7	0.000	0.000	0.000	0.000	0.000	40	7.187	0.000	0.000	0.000	0.000
8	0.000	0.000	0.000	0.000	0.000	41	0.691	0.175	0.216	0.234	1.212
9	0.000	0.000	0.000	0.000	0.000	42	2.695	0.524	1.293	2.804	3.636
10	0.000	0.000	0.000	0.000	0.000	43	0.484	0.000	0.000	0.000	0.000
11	0.000	0.000	0.000	0.000	0.000	44	0.000	0.000	0.000	0.000	0.000
12	0.000	0.000	0.431	0.000	0.000	45	0.622	0.350	0.647	0.467	0.000
13	0.069	0.350	0.216	0.467	0.606	46	0.207	0.175	0.000	0.000	0.000
14	0.000	0.000	0.000	0.000	0.000	47	0.346	0.175	0.000	0.000	0.000
15	0.207	0.000	0.216	0.000	0.000	48	0.138	0.000	0.000	0.000	0.000
16	0.000	0.000	0.000	0.000	0.000	49	3.317	3.147	1.078	0.234	0.000
17	0.000	0.000	0.000	0.000	0.000	50	0.691	0.350	0.216	0.000	0.000
18	0.000	0.000	0.000	0.000	0.000	51	1.797	1.399	0.862	1.168	3.030
19	3.870	0.874	0.431	0.000	0.606	52	0.553	0.000	0.216	0.000	0.000
20	3.179	10.315	14.224	11.682	9.697	53	0.000	0.175	0.000	0.000	0.000
21	0.415	1.399	1.724	0.467	0.606	54	0.346	0.000	0.000	0.000	0.000
22	0.138	0.175	0.216	0.467	0.000	55M	5.598	2.622	1.078	1.168	1.212
23	0.276	0.524	0.647	0.000	1.212	59	1.106	0.699	0.431	0.000	0.606
24	0.069	0.000	0.000	0.000	0.000	60	1.589	0.350	0.431	0.234	1.818
25	0.691	0.000	0.000	0.467	0.000	61	5.252	10.839	7.543	7.477	7.273
26	0.138	0.175	0.431	0.234	0.606	62	16.517	10.490	8.190	7.009	10.909
27	5.045	11.538	13.362	13.084	15.152	65	1.451	2.273	2.586	2.570	3.030
28	0.968	0.699	0.647	0.234	0.606	66	3.525	3.322	1.724	0.467	0.000
29	2.972	1.573	0.216	0.467	0.000	67	1.175	2.448	1.293	0.701	0.606
30	0.000	0.000	0.000	0.000	0.000	68	0.069	0.000	0.000	0.000	0.000
31	4.216	4.371	5.172	11.916	7.879	69	0.000	0.000	0.000	0.000	0.000
32	0.000	0.000	0.000	0.000	0.000	70	0.691	0.350	1.078	0.000	0.606
33	0.000	0.000	0.000	0.234	0.000						

There are notable differences between the distribution of sicknesses at the start of annuity payment for male IPI claimants aged 60 and those for aged 40 as tabulated in Table 6.8. For example, the proportions for cs34–cs38 (heart-related diseases), cs55M (genito-urinary related diseases) and cs61 (arthritis) are higher for age 60 than for age 40 for all deferred periods. On the other hand, the proportions for cs27 (mental illness), cs40 (influenza) and cs66 (road transport accident) are higher for age 40 than for age 60 for all deferred periods.

Figures 6.14 – 6.18 give a visual comparison between both sets of intensities for up to 12 years of sickness durations for DP1, DP4, DP13, DP26 and DP52, respectively.

For DP1, the aggregate intensities lie above the corresponding intensities from CMI Working Paper 5 (2004), with the gap between them narrowing with increasing sickness duration until about 5 year of sickness duration, from which the intensities from CMI Working Paper 5 (2004) become increasingly higher than the aggregate intensities.

For DP4, DP13 and DP52, the same pattern as in DP1 is observed but the sickness duration at which their respective intensities from CMI Working Paper 5 (2004) cross over and become increasingly higher than their respective aggregate intensities are 0.90 year, 2.28 years and 1.73 years. For DP26, the intensities from CMI Working Paper 5 (2004) are higher than their corresponding aggregate intensities and the gap between them remains relatively constant until it starts to widen from around 2 years of sickness duration.

We note that, for each deferred period, the difference between both sets of mortality intensities in the region near the peak value is smaller for age 60 than for age 40 as the estimates for age 60 are based on greater numbers of deaths. On the other hand, for each deferred period, the difference between both sets of intensities at the longer period of sickness durations (i.e. during the upturn of the intensity) is greater for age 60 than for age 40.

For a male age 60 at sickness inception, the mortality intensities during the upturn apply to a male with attained age over 65. For the aggregate intensities, this upturn is described by the 'base' mortality intensity derived from assured lives data set and therefore gives a more realistic set of intensities outside the age range of the data than the extrapolated values from the Gompertz formula of CMI Working Paper 5 (2004).

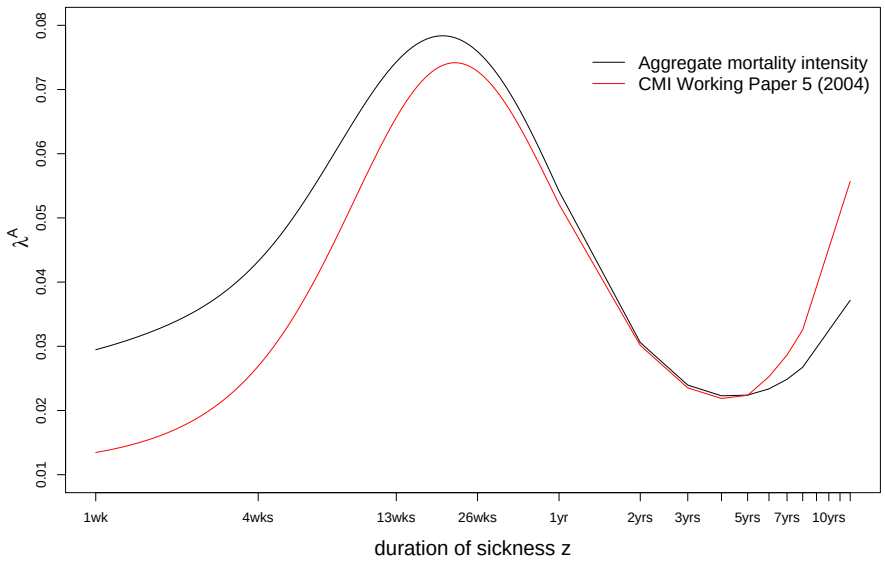


Figure 6.14: A comparison between the aggregate mortality intensities obtained given cause-specific recovery intensities and those from CMI Working Paper 5 (2004) with respect to a male DP1 IPI policyholder aged 60 at sickness inception with a non-rated occupation.

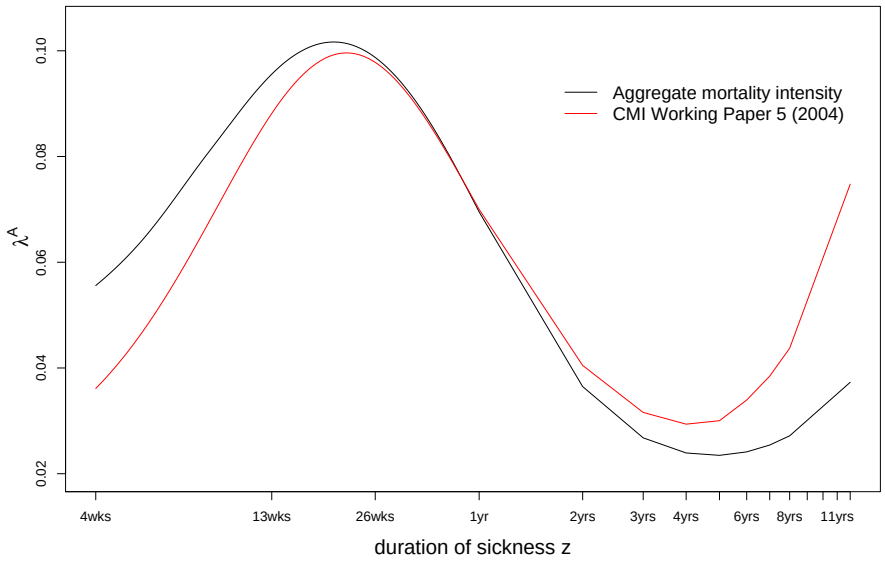


Figure 6.15: A comparison between the aggregate mortality intensities obtained given cause-specific recovery intensities and those from CMI Working Paper 5 (2004) with respect to a male DP4 IPI policyholder aged 60 at sickness inception with a non-rated occupation.

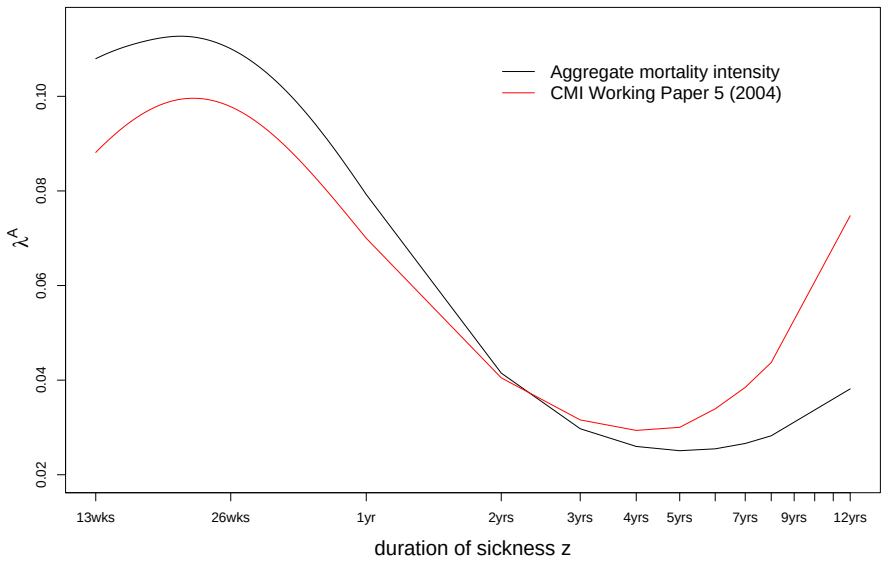


Figure 6.16: A comparison between the aggregate mortality intensities obtained given cause-specific recovery intensities and those from CMI Working Paper 5 (2004) with respect to a male DP13 IPI policyholder aged 60 at sickness inception with a non-rated occupation.

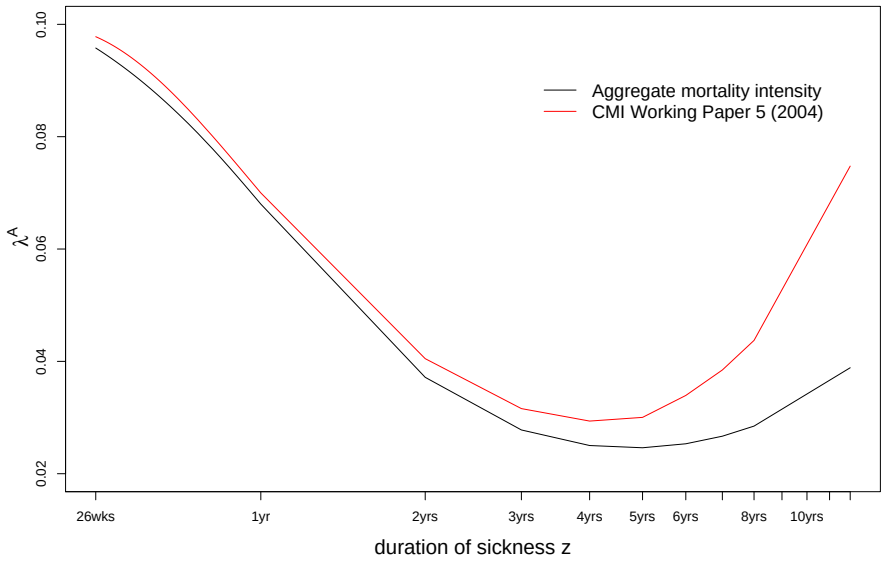


Figure 6.17: A comparison between the aggregate mortality intensities obtained given cause-specific recovery intensities and those from CMI Working Paper 5 (2004) with respect to a male DP26 IPI policyholder aged 60 at sickness inception with a non-rated occupation.

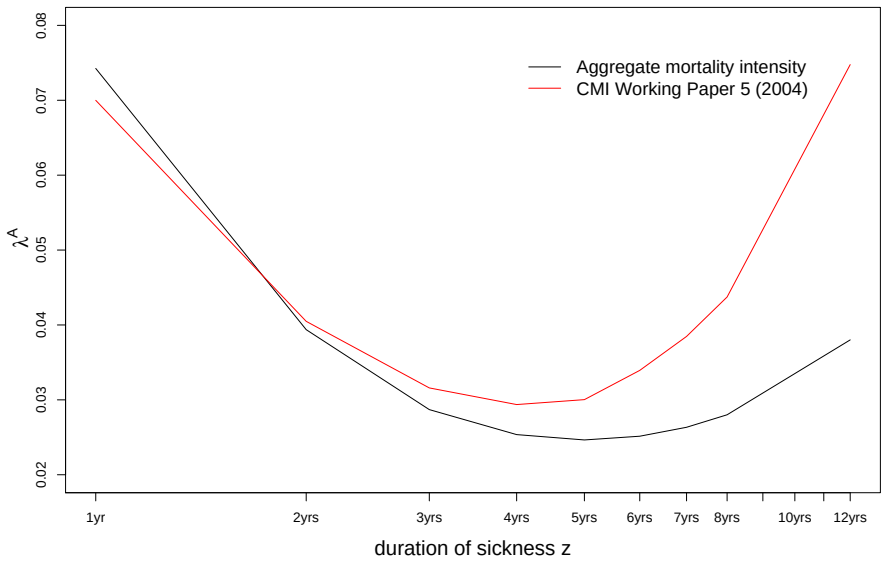


Figure 6.18: A comparison between the aggregate mortality intensities obtained given cause-specific recovery intensities and those from CMI Working Paper 5 (2004) with respect to a male DP52 IPI policyholder aged 60 at sickness inception with a non-rated occupation.

Chapter 7

Contributions and Ideas for Further Research

The purpose of this chapter is to provide an overview of the main research findings of this thesis as well as some avenues for further research.

Contributions

The main contribution of this thesis is the estimation of the recovery and mortality intensities from sick by cause of sickness as well as highlighting the role a number of well known statistical techniques such as the Cox model, generalised linear models and relative survival models have to play in the analysis of IPI data.

In terms of methodology, the CMI's usual approach to analysing IP data with covariates is to split the data into homogeneous subgroups and compare the sickness experience for each subgroup to the standard experience presented in CMI Report 12 (1991) or CMI Working Paper 5 (2004) by using the Actual/Expected ratios. Such comparisons are usually one-way analyses which do not give a complete picture of how the covariates jointly relate to the recovery and mortality intensities from sick. In this thesis, we incorporate covariates in the regression model for the recovery and mortality intensities from sick and estimate the covariate effects in a structured manner. The covariates we consider are sex, rating indicator, deferred period, calendar

year, age at sickness inception (for recovery intensity), attained age (for mortality intensity from sick) and sickness duration. Of all these covariates, sickness duration has the most impact on both recovery and mortality intensities from sick and the variation of these intensities with sickness duration is very often rapid (for the initial period of sickness duration) and irregular.

Apart from introducing a regression-type model, we use the Cox model heavily in the parameterisation of the recovery intensity model. In particular, we show in Chapter 3 that by modelling the variation of sickness duration in the baseline intensity, we are able to avoid specifying its precise functional form and estimate the remaining covariate effects by using the partial likelihood of the Cox model. In addition, the phenomenon of a 'run-in' period for DP4, DP13 and DP26 and the duration-dependent year effect, found in the IP related literature, are similarly revealed by using a standard diagnostic tool associated with the Cox model. These duration-dependent covariate effects are estimated using a generalised linear model with Poisson error structure which is shown to be somewhat equivalent to the Cox model. For the baseline intensity, the irregularity in the variation for sickness duration is reduced by using the transformed sickness duration variable (see Section 3.3).

We use an additive relative survival model in the modelling of the mortality intensity from sick. As far as we are aware, this form of modelling has not been attempted before in insurance modelling and is an attractive alternative to the usual direct cause-specific modelling or the multiplicative relative survival model for reasons stated in Section 4.1.

In terms of estimation results, the recovery intensity model by cause of sickness is presented in Section 3.6 and Appendix B. From these results, we note that causes of sickness which belong to the same sickness category do not necessarily have the same recovery patterns with sickness duration. In sickness category G5 Nervous system, cs29 (Cataract) has a quadratic-shaped recovery pattern while cs28 (Inflammatory diseases of eye), cs30 (Otitis media and mastoiditis) and cs31 (Other diseases of nervous system and sense organs) have recovery patterns which vary linearly with the transformed duration variable. The difference in recovery pattern can probably be

explained by the dissimilar treatment for these causes of sickness. For cs29 (Cataract), surgery is normally involved and therefore hospitalisation and waiting time for surgery are required. This may explain the recovery intensity for cs29, in the case of DP1, that increases with sickness duration until it starts to decrease from around 4 weeks of sickness duration. Similar quadratic-shaped recovery patterns are also observed for both cs48 (Appendicitis) and cs49 (Hernia) from sickness category G8 Digestive as well as cs54 (Hyperplasia of prostate) from sickness category G9 Genito-urinary, all of which involve surgery as part of their treatment.

We also note that most of the causes of sickness have recovery pattern which vary linearly with the transformed duration variable. Examples are all causes of sickness in G3 Endocrine & Metabolic, 5 out of the 7 causes of sickness in G7 Respiratory and 3 out of the 4 causes of sickness in G5 Nervous system & sensory organs. We note that these causes of sickness are normally treated with medication.

For cs35 (Ischaemic heart disease), the recovery pattern for DP1 decreases before becoming relatively constant between 4 weeks and 13 weeks of sickness duration and falling of thereafter (see Figure 3.29). This peculiar recovery pattern may be due to treatment varying with the seriousness of the disease. For example, for people who receive medication to treat angina, their recovery pattern may be a decreasing function of sickness duration. For people with a more serious heart condition who are in need of surgery, their recovery pattern may be quadratic-shaped. Thus, a combination of both types of people receiving different treatments may result in the peculiar recovery pattern we see in Figure 3.29. A similar recovery pattern as in cs35 is also observed in cs55F (Other diseases of genito-urinary system (female)). This is also likely to be due to the different types of female genito-urinary disease in cs55F that involve either surgery, such as hysterectomy, or medication.

It is therefore reasonable to say that the recovery pattern for a cause of sickness, to a certain extent, is influenced by the type of medical treatment a person receives. In view of this, it is essential for IPI underwriters or actuaries to be aware of any changes in medical treatment that could affect the recovery pattern.

For the mortality intensity from sick by cause of sickness, although it consists of the sum of the 'base' mortality and 'excess mortality' intensities, its overall shape is very likely to follow that of the 'excess mortality' at least for the initial period of sickness duration during which sickness duration is the more dominant explanatory variable. As described in Section 5.3, in modelling the 'excess mortality' intensity, the causes of sickness are classified into five sickness categories, i.e. MI to MV, depending on the shape of their mortality curves. For causes of sickness in sickness categories MI and MIV, their mortality curves have a hump-backed feature. For causes of sickness in sickness categories MIII and MV, their mortality curves decrease linearly with transformed sickness duration. For cause of sickness in sickness category MII (cs20 Malignant neoplasm), the mortality curves follow a U-shape. No easy explanation can be given as to the shape of the mortality curve in each sickness category.

In general, sickness duration has a more dominant effect on mortality intensity to begin with, but its effect wears off over time until attained age becomes more important . Given that deaths from cs20 Malignant neoplasm constitute almost half of the total deaths, the aggregate mortality intensity for DP1, given cause-specific mortality intensity, (see Figure 6.9 and 6.14) follows the same hump-backed shape as cs20.

Ideas for Further Research

Actuaries are required to assess the magnitude, timing and duration of future claim payments associated with IPI so that sufficient reserves are set aside to cover these future cash outflows. In respect of reserving more reliably for claims in payment, we have produced in this thesis the estimated recovery and mortality intensities by cause of sickness which will enable an actuary to calculate a more reliable reserve for a portfolio of claims resulting from different causes of sickness. In respect of reserving more reliably for future claims, apart from having cause-specific termination assumptions, it is useful to know the sickness inception intensity by cause of sickness, although in computing reserves a breakdown of sickness inception by cause will produce the same result as one calculated without the breakdown. However, the modelling of sickness inception by cause of sickness with year trend incorporated will shed light on the relative importance of different cause of sickness over the years. These sickness inception intensities are defined in the multiple state model by cause of sickness in Chapter 1 (i.e. transitions from Healthy to each of the Sick states). Thus, a natural extension of this thesis is to estimate these sickness inception intensities using IPI in-force data so that the multiple state model by cause of sickness can become fully operational for IPI business.

We calculated the expected present values of annuities by cause of sickness for claims in payment and presented the results in Chapter 6. Another important quantity worth investigating is the average duration of a claim by cause of sickness which is important for the underwriting and claim control stages of IPI business. It is also possible to construct confidence interval for the aggregate recovery and mortality intensities which are derived using cause-specific intensities by using bootstrapping.

The graduation formulae for recovery and mortality intensities from sick presented in CMI Report 12 (1991) and CMI Working Paper 5 (2004) assume homogeneity in that all IP claimants, regardless of their cause of sickness, are subject to the same recovery and mortality intensities. In this thesis, we have relaxed this assumption by introducing an observed source of heterogeneity, cause of sickness, into our model. It would be interesting to incorporate into the model unobserved sources of heterogeneity (frailty) that are not readily captured by covariates. The idea is that an individual frailty will influence the occurrence of recovery, death or sickness.

Finally, the impact of economic variables, such as the level of interest rates and unemployment in particular, on the recovery intensity could also be investigated.

Appendix A

ICD8 Code for Causes of Sickness

The list below shows the causes of sickness, grouping within each of the 12 sickness categories as presented in Table 2.3 as well as their codes according to Abbreviated List C in the Eight Revision of the *Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death* (ICD8).

G1 Infections & acute respiratory

ICD8 code	Cause of sickness
1	Typhoid, paratyphoid fever, other salmonella infections
2	Bacillary dysentery and amoebiasis
3	Enteritis and Other diarrhoeal diseases
4	Tuberculosis of respiratory system
5	Other tuberculosis, including late effect
6	Brucellosis
7	Diphteria
8	Whooping cough
9	Streptococcal sore throat and scarlet fever
10	Small pox
11	Measles
12	Viral encephalitis
13	Infectious hepatitis
14	Typhus and other rickettsioses

ICD8 code	Cause of sickness
15	Malaria
16	Syphilis and its sequelae
17	Gonnococcal infections
18	Helminthiases
19	All other infective and parasitic diseases

G2 Neoplasms

ICD8 code	Cause of sickness
20	Maglinant neoplasm, including neoplasms of lymphatic and haematopoietic tissue
21	Benign neoplasms and neoplasms of unspecified nature

G3 Endocrine & Metabolic

ICD8 code	Cause of sickness
22	Thyrotoxicosis with or without goitre
23	Diabetes Mellitus
24	Avitaminoses and other nutritional deficiency
25	Other endocrine and metabolic diseases
26	Anaemias

G4 Mental Illness

ICD8 code	Cause of sickness
27	Psychoses and non psychotic mental disorders

G5 Nervous system & sensory organs

ICD8 code	Cause of sickness
28	Inflammatory diseases of eye
29	Cataract
30	Otitis media and mastoiditis
31	Other diseases of nervous system and sense organs

G6 Circulatory

ICD8 code	Cause of sickness
32	Active rheumatic fever
33	Chronic rheumatic heart disease
34	Hypertensive disease
35	Ischaemic heart disease
36	Cerebrovascular disease
37	Venous thrombosis and embolism
38	Other diseases of respiratory system

G7 Respiratory

ICD8 code	Cause of sickness
39	Acute respiratory infections
40	Influenza
41	Pneumonia
42	Bronchitis, emphysema and astham
43	Hypertrophy of tonsils and adenoids
44	Pneumoconioses and related diseases
45	Other diseases of respiratory system

G8 Digestive

ICD8 code	Cause of sickness
47	Peptic Ulcer
48	Appendicitis
49	Intestinal obstruction and hernia
50	Cholelithiasis and cholecystitis
51	Other diseases of digestive system

G9 Genito-Urinary

ICD8 code	Cause of sickness
52	Nephritis and nephrosis
53	Calculus of urinary system
54	Hyperplasia of prostate
55	Other diseases of genito-urinary system

G10 Musculoskeletal

ICD8 code	Cause of sickness
61	Arthritis and spondylitis
62	Other diseases of musculoskeletal system and connective tissue

G11 Injuries

ICD8 code	Cause of sickness
66	Road transport accident
67	All other accidents
68	Attempted suicide and self-inflicted injuries
69	Attempted homicide and injury purposely inflicted by other persons; legal intervention
70	All other external causes

G12 All other known causes

ICD8 code	Cause of sickness
46	Diseases of teeth and supporting structures
56	Abortion
57	Other complications of pregnancy, childbirth and the puerperium
58	Delivery without mention of complication
59	Infections of skin and subcutaneous tissue
60	Other diseases of skin and subcutaneous tissue
63	Congenital anomalies
64	Certain causes of perinatal morbidity
65	Other specified and ill-defined diseases

Appendix B

The Recovery Intensity Models for Other Sickness Categories

In this Appendix we present the estimated recovery intensity models for causes of sickness in Sickness Categories 'G1 Infections & acute respiratory', 'G3 Endocrine & Metabolic', 'G5 Nervous system & sensory organs', 'G7 Respiratory', 'G8 Digestive', 'G9 Genito-urinary', 'G11 Injuries' and 'G12 All other known causes'. For each recovery intensity model, the parameters for the baseline intensity are first given, followed by those for the duration-fixed covariates, the interaction terms between duration-fixed covariates, and the duration-dependent covariates. In terms of the goodness-of-fit for each recovery intensity model, the χ^2 statistic is given. The relevant "partial residual effects" plot,'residual effects" plot and two-dimensional plot of deviance residuals (constructed wheneve the data is sufficiently large), can be found in Ling (2008).

G1 Infections & acute respiratory

There are 19 causes of sickness in sickness category G1 Infections & acute respiratory and they are cs1 – cs19. The data belonging to cs7, cs14, cs16, cs17 and cs18 are amalgamated with cs19 data because their exposed-to-risk in days (number of recoveries) are 195(1), 329(4), 2,558(3), 57(3) and 769(3), respectively. Of the remaining causes of sickness, those with recovery intensities which are proportional to each other are modelled together using a proportional hazards model with

cause of sickness as a factor. As such, the 14 causes of sickness can be split into two groups, G1sub1 and G1sub2, and a separate recovery intensity model is fitted to each group. The causes of sickness in each group are as follows:

G1sub1: cs4, cs5, cs12 and cs13
G1sub2: cs1, cs2, cs3, cs6, cs8, cs9, cs10, cs11, cs15 and cs19

For G1sub1, the χ^2 statistic for the fitted recovery intensity model is 80.4461. With 39 cells and 14 parameters fitted in the model, the probability value is 0.8992 on 25 degrees of freedom. For G1sub2, the χ^2 statistic for the fitted recovery intensity model is 80.4461. With 233 cells and 23 parameters fitted in the model, the probability value is 0.2329 on 210 degrees of freedom. Table B.1 and B.2 show the exposed-to-risk in days, the number of recoveries and the parameters in the recovery intensity model for each cause of sickness in G1sub1 and G1sub2, respectively.

Table B.1: Parameters in the recovery intensity models for causes of sickness in G1sub1.

	cs4	cs5	cs12	cs13
Exposed to risk (days)	36,361	17,198	73,977	78,112
Number of recoveries	66	30	59	330
k	0.5	0.5	0.5	0.5
b_0	1.8335	1.7434	2.4943	2.2079
b_1	-3.0257	-3.0257	-4.9218	-3.0257
α_{rated}	-0.3231	-0.3231	-0.3231	-0.3231
α_{dp26}	-0.7099	-0.7099	-0.7099	-0.7099
α_{year}	-0.4472	-0.4472	-0.4472	-0.4472
α_{age}	-0.1061	-0.1061	-0.1061	-0.1061
τ_{dp4}	50.5	50.5	50.5	50.5
γ_{dp4}	-0.0329	-0.0329	-0.0329	-0.0329
τ_{dp13}	122.5	122.5	122.5	122.5
γ_{dp13}	-0.0549	-0.0549	-0.0549	-0.0549
τ_{year}	27.5	27.5	27.5	27.5
γ_{year}	0.0403	0.0403	0.0403	0.0403
τ_{age}	27.5	27.5	27.5	27.5
ϕ_{age}	-2.0455	-2.0455	-2.0455	-2.0455

Table B.2: Parameters in recovery intensity models for causes of sickness in G1sub2.

	cs1	cs2	cs3	cs6	cs8	cs9	cs10	cs11	cs15	cs19
Exposed to risk (days)	14,183	11,683	30,668	4,155	385	1,905	753	3,175	2,240	262,666
Number of recoveries	106	48	947	21	12	58	12	30	29	3,135
k	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3
b_0	29.9825	29.8673	30.5319	30.3101	30.3101	30.3101	30.3101	30.3101	29.9898	30.3101
b_1	-89.6762	-89.6762	-89.6762	-89.6762	-89.6762	-89.6762	-89.6762	-89.6762	-89.6762	-89.6762
b_2	25.9040	25.9040	25.9040	25.9040	25.9040	25.9040	25.9040	25.9040	25.9040	25.9040
b_3	-22.0979	-22.0979	-22.0979	-22.0979	-22.0979	-22.0979	-22.0979	-22.0979	-22.0979	-22.0979
α_{sex}	-0.1842	-0.1842	-0.1842	-0.1842	-0.1842	-0.1842	-0.1842	-0.1842	-0.1842	-0.1842
α_{rated}	-0.2666	-0.2666	-0.2666	-0.2666	-0.2666	-0.2666	-0.2666	-0.2666	-0.2666	-0.2666
α_{dp4}	0.3488	-0.1372	-0.5933	-0.1372	-0.1372	-0.1372	-0.1372	-0.1372	-0.1372	-0.1372
α_{dp13}	-0.3030	-0.3030	-0.3030	-0.3030	-0.3030	-0.3030	-0.3030	-0.3030	-0.3030	-0.3030
α_{dp26}	-0.7125	-0.7125	-0.7125	-0.7125	-0.7125	-0.7125	-0.7125	-0.7125	-0.7125	-0.7125
α_{year}	-0.5801	-0.5801	-0.5801	-0.5801	-0.5801	-1.6387	-0.5801	-0.5801	-0.5801	-0.5801
α_{age}	-0.3495	-0.3495	-0.7117	-0.3495	-0.3495	-0.3495	-0.3495	-0.3495	-0.3495	-0.3495
α_{age2}	-0.1331	-0.1331	-0.1331	-0.1331	-0.1331	-0.1331	-0.1331	-0.1331	-0.1331	-0.1331
τ_{dp4}	47.5	47.5	47.5	47.5	47.5	47.5	47.5	47.5	47.5	47.5
γ_{dp4}	-0.0800	-0.0800	-0.0800	-0.0800	-0.0800	-0.0800	-0.0800	-0.0800	-0.0800	-0.0800
τ_{dp13}	119.5	119.5	119.5	119.5	119.5	119.5	119.5	119.5	119.5	119.5
γ_{dp13}	-0.0421	-0.0421	-0.0421	-0.0421	-0.0421	-0.0421	-0.0421	-0.0421	-0.0421	-0.0421
τ_{year}	47.5	47.5	47.5	47.5	47.5	47.5	47.5	47.5	47.5	47.5
γ_{year}	0.0247	0.0247	0.0247	0.0247	0.0247	0.0247	0.0247	0.0247	0.0247	0.0247

G3 Endocrine & Metabolic

There are 5 causes of sickness in sickness category G3 Endocrine & Metabolic and they are cs22 – cs26. The recovery intensities for these 5 causes of sickness are modelled together using a proportional hazards model with cause of sickness as a factor because they are found to be proportional to each other. Table B.3 shows for each cause of sickness in G3 Endocrine & Metabolic the exposed-to-risk in days, the number of recoveries and the parameters in the recovery intensity model. The χ^2 statistic for the fitted model is 43.9113. With 50 cells and 16 parameters fitted in the model, the probability value is 0.1188 on 34 degrees of freedom.

Table B.3: Parameters in the recovery intensity models for causes of sickness in G3 Endocrine & Metabolic.

	cs22	cs23	cs24	cs25	cs26
Exposed-to-risk (days)	209,164	246,938	6,044	137,380	51,597
Number of recoveries	91	137	9	292	67
k	2.3	2.3	2.3	2.3	2.3
b_0	2.7287	2.9605	3.5392	3.5392	3.1256
b_1	-14.5645	-14.5645	-14.5645	-14.5645	-14.5645
α_{dp4}	0.4594	-0.2807	-0.2807	-0.2807	-0.2807
α_{dp26}	-0.3933	-0.3933	-0.3933	-0.3933	-0.3933
α_{year}	0.0339	0.0339	0.0339	0.0339	0.0339
α_{age}	0.1178	-0.6799	-0.6799	-0.6799	-0.6799
α_{age2}	-0.2940	-0.2940	-0.2940	-0.2940	-0.2940
τ_{dp4}	52.5	52.5	52.5	52.5	52.5
γ_{dp4}	-0.0569	-0.0569	-0.0569	-0.0569	-0.0569
τ_{dp13}	140.5	140.5	140.5	140.5	140.5
γ_{dp13}	-0.0372	-0.0372	-0.0372	-0.0372	-0.0372
τ_{year}	20.5	20.5	20.5	20.5	20.5
γ_{year}	0.0453	0.0453	0.0453	0.0453	0.0453
θ_{year}	-3.0624	-3.0624	-3.0624	-3.0624	-3.0624

G5 Nervous system & sensory organs

There are 4 causes of sickness in sickness category G5 Nervous system & sensory organs and they are cs28 – cs31. The recovery intensities for cs28, cs30 and cs31 are modelled together because they are found to be proportional to each other. The χ^2 statistic for the fitted recovery intensity model for these 3 causes of sickness is 203.3780. With 204 cells and 17 parameters fitted in the model, the probability value is 0.1956 on 187 degrees of freedom. The recovery intensity for cs29 is modelled separately and the fitted model has a χ^2 statistic of 11.4250. With 23 cells and 6 parameters fitted in the model, the probability value is 0.8336 on 17 degrees of freedom. Table B.4 shows for each cause of sickness in G5 Nervous system & sensory organs the exposed-to-risk in days, the number of recoveries and the parameters in the recovery intensity model.

Table B.4: Parameters in the recovery intensity models for causes of sickness in G5 Nervous system & sensory organs.

	cs28	cs29	cs30	cs31
Exposed-to-risk(days)	235,093	92,897	44,719	3,587,559
Number of recoveries	424	285	148	2,008
k	2.3	6.7	2.3	2.3
b_0	4.1016	-308.3070	4.4025	3.6802
b_1	-18.0868	63.1057	-18.0868	-18.0868
b_2		-309.1824		
α_{sex}	-0.0937		-0.0937	-0.0937
α_{rated}	0.1686		0.1686	0.1686
α_{dp4}	-0.1120		-0.1120	-0.1120
α_{dp26}	-0.3538		-0.3538	-0.3538
α_{year}	-0.6030	1.2748	-0.6030	-0.6030
α_{age}	-0.9358		-0.9358	-0.9358
$\alpha_{\text{sex:age}}$	0.3825		0.3825	0.3825
$\alpha_{\text{dp4:age}}$	0.3385		0.3385	0.3385
$\alpha_{\text{dp26:year}}$	-0.4569		-0.4569	-0.4569
τ_{rated}	68.5		68.5	68.5
γ_{rated}	-0.0245		-0.0245	-0.0245
τ_{dp4}	46.5	56.5	46.5	46.5
γ_{dp4}	-0.0698	-0.0342	-0.0698	-0.0698
τ_{dp13}	121.5		121.5	121.5
γ_{dp13}	-0.0541		-0.0541	-0.0541
τ_{year}	45.5		45.5	45.5
γ_{year}	0.0302		0.0302	0.0302
θ_{year}		-12.7439		

G7 Respiratory

There are 7 causes of sickness in sickness category G7 Respiratory and they are cs39 – cs45. The recovery intensities for cs39, cs40, cs42, cs43 and cs44 are modelled together because they are found to be proportional to each other. The χ^2 statistic for the fitted recovery intensity model for these 5 causes of sickness is 283.0298. With 291 cells and 25 parameters fitted in the model, the probability value is 0.2261 on 266 degrees of freedom. The recovery intensities for cs41 and cs45 are fitted separately. For cs41, the fitted recovery intensity model has a χ^2 statistic of 30.8606. With 53 cells and 11 parameters fitted in the model, the probability value is 0.8977 on 42 degrees of freedom. For cs45, the fitted recovery intensity model has a χ^2 statistic of 66.8385. With 65 cells and 13 parameters fitted in the model, the probability value is 0.0808 on 52 degrees of freedom. Table B.5 shows for each cause of sickness in G7 Respiratory the exposed-to-risk in days, the number of recoveries and the parameters in the recovery intensity model.

Table B.5: Parameters in the recovery intensity models for causes of sickness in G7 Respiratory.

	cs39	cs40	cs41	cs42	cs43	cs44	cs45
Exposed-to-risk(day)	127,930	41,123	67,836	505,352	14,379	7,373	210,108
Number of recoveries	2,639	4,504	630	1,419	329	38	881
k	2.3	2.3	6.7	2.3	2.3	2.3	2.3
b_0	4.5644	4.8340	-302.4215	4.3385	4.4458	4.1399	29.1540
b_1	-17.7844	-17.7844	61.1308	-17.7844	-17.7844	-17.7844	-103.2332
b_2			-304.1462				25.1611
b_3							-27.5685
α_{sex}	-0.0796	-0.2212		-0.0796	-0.0796	-0.0796	
α_{rated}			-0.4901				
α_{dp4}	-0.5160	-0.5160	-0.1596	-0.5160	-0.5160	-0.5160	-0.4317
α_{dp13}	-1.1482	-1.1482	-0.6582	-1.1482	-1.1482	-1.1482	
α_{dp26}	-0.3686	-0.3686		-0.3686	-0.3686	-0.3686	-1.0302
α_{dp52}	-1.0029	-1.0029		-1.0029	-1.0029	-1.0029	
α_{year}	-0.0206	-0.0206	0.1820	-0.0206	-0.0206	-0.0206	-0.0037
α_{age}	-0.5603	-0.4438	-0.6164	-0.7953	-0.5603	-0.5603	-0.7969
α_{age2}	-0.0347	-0.0347		-0.0347	-0.0347	-0.0347	-0.2947
$\alpha_{\text{dp4:age}}$			0.4928				
$\alpha_{\text{dp26:age}}$	-1.2614	-1.2614		-1.2614	-1.2614	-1.2614	
$\alpha_{\text{dp4:year}}$	-0.4343	-0.4343	-0.7182	-0.4343	-0.4343	-0.4343	
$\alpha_{\text{dp13:year}}$	-0.9366	-0.9366		-0.9366	-0.9366	-0.9366	
$\alpha_{\text{dp4:age2}}$	-0.3969	-0.3969		-0.3969	-0.3969	-0.3969	
$\alpha_{\text{dp13:age2}}$	-0.8687	-0.8687		-0.8687	-0.8687	-0.8687	
$\alpha_{\text{year:age}}$							-0.3893
τ_{dp4_1}	40.5	40.5	59.5	40.5	40.5	40.5	42.5
γ_{dp4_1}	-0.1001	-0.1001	-0.0496	-0.1001	-0.1001	-0.1001	-0.1187
τ_{dp4_2}	70.5	70.5		70.5	70.5	70.5	
γ_{dp4_2}	-0.0156	-0.0156		-0.0156	-0.0156	-0.0156	
τ_{dp13}							220.5
γ_{dp13}							-0.0111
$\tau_{\text{year}_\gamma}$	35.5	35.5		35.5	35.5	35.5	
γ_{year}	0.0114	0.0114		0.0114	0.0114	0.0114	
τ_{year_ϕ}							20.5
ϕ_{year}							-3.2097

G8 Digestive

There are 5 causes of sickness in sickness category G8 Digestive and they are cs47 – cs51. The recovery intensities for cs48 and cs49 are modelled together because they are found to be proportional to each other. The χ^2 statistic for the fitted recovery intensity model for both causes of sickness is 255.0009. With 282 cells and 22 parameters fitted in the model, the probability value is 0.5759 on 260 degrees of freedom. The recovery intensities for cs47, cs50 and cs51 are fitted separately. For cs47, the fitted recovery intensity model has a χ^2 statistic of 23.7274. With 30 cells and 9 parameters fitted in the model, the probability value is 0.3064533 on 21 degrees of freedom. For cs50, the fitted recovery intensity model has a χ^2 statistic of 32.1172. With 49 cells and 11 parameters fitted in the model, the probability value is 0.7375 on 38 degrees of freedom. For cs51, the fitted recovery intensity model has a χ^2 statistic of 107.4741. With 128 cells and 15 parameters fitted in the model, the probability value is 0.6289974 on 113 degrees of freedom. Table B.6 shows for each cause of sickness in G8 Digestive the exposed-to-risk in days, the number of recoveries and the parameters in the recovery intensity model.

Table B.6: Parameters in the recovery intensity models for causes of sickness in G8 Digestive.

	cs47	cs48	cs49	cs50	cs51
Exposed-to-risk (days)	91,916	24,189	268,100	70,661	583,935
Number of recoveries	350	608	2,618	612	1,565
k	1.3	6.7	6.7	2.3	2.3
b_0	3.0712	-375.3728	-375.5133	-28.1944	55.5717
b_1	-8.0560	93.3918	93.3918	13.0201	-177.4550
b_2		-375.9880	-375.9880	-30.1205	51.8543
b_3					-50.4236
α_{sex}		-0.4958	-0.4958		-0.1155
α_{rated}	-0.3572	-0.0171	-0.0171	-0.5489	-0.2075
α_{dp4}		-0.2369	-0.4718		0.1674
α_{dp13}		-0.0213	-0.0213	-0.5161	
α_{dp26}		-0.8313	-0.8313		-0.4412
α_{year}	-1.4777	0.2038	0.2038	-0.5034	-0.4041
α_{year2}		0.1512	0.1512	0.2684	
α_{age}	-6.8031	-0.5330	-0.5330	-0.3858	-0.6502
$\alpha_{\text{sex:age}}$					0.6943
$\alpha_{\text{dp4:age}}$		0.5364	0.5364		
$\alpha_{\text{dp4:year}}$		-0.3284	-0.3284		
$\alpha_{\text{dp13:year}}$		-0.7464	-0.7464		
$\alpha_{\text{rated:age}}$		-0.2655	-0.2655		
$\alpha_{\text{sex:dp4}}$		0.5211	0.5211		
τ_{rated}		75.5	75.5		
γ_{rated}		-0.0269	-0.0269		
τ_{dp4_1}	49.5	40.5	40.5		42.5
γ_{dp4_1}	-0.0620	-0.0912	-0.0912	-0.1707	-0.1064
τ_{dp4_2}					90.5
γ_{dp4_2}					-0.0091
τ_{dp13}		128.5	128.5		119.5
γ_{dp13}		-0.0329	-0.0329		-0.0469
$\tau_{\text{year}_\gamma}$		26.5	26.5		32.5
γ_{year}		0.0667	0.0667		0.0351
τ_{year_ζ}	135.5			52.5	
ζ_{year_1}	9.7968			49.9025	
ζ_{year_2}				-460.9677	
θ_{age_1}	4.1772				
θ_{age_2}	-6.3111				

G9 Genito-urinary

There are 4 causes of sickness in sickness category G8 Digestive and they are cs52 – cs55. For cs55, the recovery intensity for males and females are fitted separately because their recovery pattern are very dissimilar to each other. We refer to cs55 for males and females as cs55M and cs55F, respectively.

The recovery intensities for cs52, cs53 and cs55M are modelled together because they are found to be proportional to each other. The χ^2 statistic for the fitted recovery intensity model for these causes of sickness is 80.4461. With 125 cells and 16 parameters fitted in the model, the probability value is 0.9816 on 109 degrees of freedom.

The recovery intensities for cs54 and cs55F are estimated separately. For cs47, the fitted recovery intensity model has a χ^2 statistic of 10.8174. With 21 cells and 4 parameters fitted in the model, the probability value is 0.8659 on 17 degrees of freedom. For cs55F, the fitted recovery intensity model has a χ^2 statistic of 46.8785. With 60 cells and 13 parameters fitted in the model, the probability value is 0.4775 on 47 degrees of freedom. Table B.7 shows for each cause of sickness in G9 Genito-urinary, the exposed-to-risk in days, the number of recoveries and the parameters in the recovery intensity model.

Table B.7: Parameters in the recovery intensity models for causes of sickness in G9 Genito-urinary.

	cs52	cs53	cs54	cs55M	cs55F
Exposed-to-risk(days)	102,633	30,929	23,737	211,813	157,067
Number of recoveries	155	163	259	1,211	818
k	1.3	1.3	6.7	1.3	6.7
b_0	3.1425	3.5696	-404.5940	3.6364	1690.9520
b_1	-9.5505	-9.5505	103.1949	-9.5505	-12041.2100
b_2			-404.0155		1684.5420
b_3					-3940.3580
α_{rated}	-0.2545	-0.2545		-0.2545	-0.8554
α_{dp4}	-0.1718	-0.1718		-0.1718	-0.1288
α_{dp13}					-0.5375
α_{year}	0.3411	-0.8032		-0.3135	-0.3416
α_{age}	-0.8187	-0.2099		-0.2099	-1.1964
$\alpha_{\text{dp4:age}}$					0.9581
$\alpha_{\text{dp13:age}}$					1.5520
$\alpha_{\text{rated:year}}$					0.5599
τ_{dp4}	43.5	43.5	47.5	43.5	49.5
γ_{dp4}	-0.1051	-0.1051	-0.0659	-0.1051	-0.0787
τ_{dp13}	155.5	155.5		155.5	
γ_{dp13}	-0.0218	-0.0218		-0.0218	
τ_{year}	50.5	50.5		50.5	
γ_{year}	0.0168	0.0571		0.0168	
τ_{age}	38.5	38.5		38.5	
γ_{age}	-0.0330	-0.0330		-0.0330	

G11 Injuries

There are 5 causes of sickness in sickness category G11 Injuries and they are cs66 – cs70.

The recovery intensities for cs68, cs69 and cs70 are modelled together because they are found to be proportional to each other. The χ^2 statistic for the fitted recovery intensity model for these causes of sickness is 118.6597. With 165 cells and 21 parameters fitted in the model, the probability value is 0.9396 on 144 degrees of freedom. Table B.8 shows for each cause of sickness in G9 Genito-urinary, the exposed-to-risk in days, the number of recoveries and the parameters in the recovery intensity model.

The recovery intensities for cs66 and cs67 are estimated separately. For cs66, the fitted recovery intensity model has a χ^2 statistic of 293.528. With 332 cells and 20 parameters fitted in the model, the probability value is 0.7666 on 312 degrees of freedom. For cs66, the fitted recovery intensity model has a χ^2 statistic of 571.7478. With 645 cells and 19 parameters fitted in the model, the probability value is 0.9407 on 626 degrees of freedom.

Table B.8: Parameters in the recovery intensity models for causes of sickness in G11 Injuries.

	cs66	cs67	cs68	cs69	cs70
Exposed-to-risk (days)	1,458,667	2,383,238	18,158	39,193	476,814
Number of recoveries	3,893	7,695	60	117	1,993
k	2.3	2.3	2.3	2.3	2.3
b_0	-220.4817	26.8452	-449.0517	-448.6254	-448.2689
b_1	250.2069	-102.6865	550.8088	550.8088	550.8088
b_2	-280.6401	23.4335	-570.4026	-570.4026	-570.4026
b_3	81.8599	-29.3062	181.6597	181.6597	181.6597
b_4	-57.7592		-118.9095	-118.9095	-118.9095
α_{sex}	-0.1429	0.0511	-0.1819	-0.1819	-0.1819
α_{rated}	-0.2727	-0.1240	-0.2593	-0.2593	-0.2593
α_{dp4}	-0.1762	-0.1896	0.2256	0.2256	0.2256
α_{dp13}	-0.2107	-0.2924			
α_{dp26}	-0.4365	-0.4144			
α_{dp52}	-0.5208	-0.8429	-0.8803	-0.8803	-0.8803
α_{year}	0.2242	-0.1959	-0.3274	-0.3274	-0.3274
α_{year2}		0.1159	0.1697	0.1697	0.1697
α_{age}	-0.3068	-0.5003	-0.4127	-0.4127	-0.4127
α_{age2}			-0.1332	-0.1332	-0.1332
$\alpha_{\text{dp4:age}}$		0.2699			
$\alpha_{\text{dp26:age}}$	-0.8826				
$\alpha_{\text{dp13:year}}$	-0.3122				
$\alpha_{\text{rated:year}}$	-0.3346				
$\alpha_{\text{sex:dp4}}$		-0.2848			
τ_{dp4_1}	47.5	47.5	44.5	44.5	44.5
γ_{dp4_1}	-0.0803	-0.0869	-0.0775	0.0248	-0.0775
τ_{dp4_2}			92.5	92.5	92.5
γ_{dp4_2}			-0.0157	-0.0157	-0.0157
τ_{dp13}	119.5	128.5	115.5	115.5	115.5
γ_{dp13}	-0.0579	-0.0325	-0.0878	-0.0878	-0.0878
$\tau_{\text{year}_\gamma}$	28.5	28.5	24.5	24.5	24.5
γ_{year}	-0.0163	0.0316	0.0280	0.0280	0.0280
τ_{year_ϕ}	67.5				
ϕ_{year}	-3.0723				
θ_{year}		-1.2328			

G12 All other known causes

There are 9 causes of sickness in sickness category G11 Injuries and they are cs46, cs56 – cs60 and cs63 – cs65.

The recovery intensities for cs46, cs56, cs57, cs58, cs63, cs64 and cs65 are modelled together because they are found to be proportional to each other. The χ^2 statistic for the fitted recovery intensity model for these causes of sickness is 242.1508. With 279 cells and 27 parameters fitted in the model, the probability value is 0.6608 on 252 degrees of freedom.

The recovery intensity for cs59 and cs60 are estimated together since they are found to be proportional to each other. The χ^2 statistic for the fitted recovery intensity model for these causes of sickness is 77.6994. With 97 cells and 11 parameters fitted in the model, the probability value is 0.7268 on 86 degrees of freedom.

The recovery intensities for cs66 and cs67 are estimated separately. For cs66, the fitted recovery intensity model has a χ^2 statistic of 293.528. With 332 cells and 20 parameters fitted in the model, the probability value is 0.7666 on 312 degrees of freedom. For cs66, the fitted recovery intensity model has a χ^2 statistic of 571.7478. With 645 cells and 19 parameters fitted in the model, the probability value is 0.9407 on 626 degrees of freedom.

Table B.9 shows for each cause of sickness in G9 Genito-urinary, the exposed-to-risk in days, the number of recoveries and the parameters in the recovery intensity model.

Table B.9: Parameters in the recovery intensity models for causes of sickness in G12 All other known causes.

	cs46	cs56	cs57	cs58	cs59	cs60	cs63	cs64	cs65
Exposed-to-risk (days)	7,763	1,965	14,971	1,233	126,352	194,363	89,725	36,664	2,452,853
Number of recoveries	201	18	108	2	701	642	102	140	3177
k	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3
b_0	29.8043	28.8907	28.8907	28.8907	4.1147	3.9567	28.5392	28.8907	28.8907
b_1	-89.4239	-89.4239	-89.4239	-89.4239	-16.4267	-16.4267	-89.4239	-89.4239	-89.4239
b_2	25.1623	25.1623	25.1623	25.1623			25.1623	25.1623	25.1623
b_3	-22.5471	-22.5471	-22.5471	-22.5471			-22.5471	-22.5471	-22.5471
α_{sex}	-0.0563	-0.0563	-0.0563	-0.0563			-0.0563	-0.0563	-0.0563
α_{rated}	-1.2448	-0.0209	-0.0209	-0.0209			-0.0209	-0.0209	-0.0209
α_{dp4}	-0.0552	-0.0552	-0.0552	-0.0552			-0.0552	-0.0552	-0.0552
α_{dp13}	0.0118	0.0118	0.0118	0.0118	0.5062	0.0120	0.0118	0.0118	0.0118
α_{dp26}	-0.2239	-0.2239	-0.2239	-0.2239	-0.4631	-0.4631	0.6170	-0.2239	-0.2239
α_{dp52}	-0.3710	-0.3710	-0.3710	-0.3710			0.8936	-0.3710	-0.3710
α_{year}	-0.4669	-0.7892	-0.7892	-0.7892	0.1197	0.1197	-0.7892	-0.7892	-0.7892
α_{age}	-0.4937	-0.4937	-0.4937	-0.4937	-0.4512	-0.4512	-0.4937	-0.4937	-0.4937
$\alpha_{\text{dp26:age}}$	-0.5812	-0.5812	-0.5812	-0.5812			-0.5812	-0.5812	-0.5812
$\alpha_{\text{dp13:year}}$	-0.3462	-0.3462	-0.3462	-0.3462			-0.3462	-0.3462	-0.3462
$\alpha_{\text{rated:year}}$	0.2946	0.2946	0.2946	0.2946			0.2946	0.2946	0.2946
$\alpha_{\text{sex:year}}$	0.2106	0.2106	0.2106	0.2106			0.2106	0.2106	0.2106
$\alpha_{\text{sex:dp52}}$	-1.0490	-1.0490	-1.0490	-1.0490			-1.0490	-1.0490	-1.0490
τ_{dp4}	47.5	47.5	47.5	47.5	63.5	63.5	47.5	47.5	47.5
γ_{dp4}	-0.1016	-0.1016	-0.1016	-0.1016	-0.0396	-0.0396	-0.1016	-0.1016	-0.1016
τ_{dp13}	121.5	121.5	121.5	121.5	119.5	119.5	121.5	121.5	121.5
γ_{dp13}	-0.0572	-0.0572	-0.0572	-0.0572	-0.0773	-0.0773	-0.0572	-0.0572	-0.0572
τ_{dp26}	220.5	220.5	220.5	220.5			220.5	220.5	220.5
γ_{dp26}	-0.0512	-0.0512	-0.0512	-0.0512			-0.0512	-0.0512	-0.0512
$\tau_{\text{year}_\gamma}$	34.5	34.5	34.5	34.5			34.5	34.5	34.5
γ_{year}	0.0313	0.0313	0.0313	0.0313			0.0313	0.0313	0.0313
τ_{year_ϕ}					32.5	32.5			
ϕ_{year}					-4.1303	-4.1303			

Appendix C

The Derivation of the Recursive Formula for ${}_tp^{\overline{S_iS_i}}_{x,y,z}$

In this Appendix we present the deriviation of the recurrence relations for the evaluation of ${}_tp^{\overline{S_iS_i}}_{x,y,z}$ in Equation (6.3).

The Kolmogorov forward equation for ${}_tp^{\overline{S_iS_i}}_{x,y,z}$ is given by

$$\frac{\partial}{\partial t}{}_tp^{\overline{S_iS_i}}_{x,y,z} = -{}_tp^{\overline{S_iS_i}}_{x,y,z}(\rho(i)_{x+t,y+t,z+t} + \nu(i)_{x+t,y+t,z+t})$$

and similarly,

$$\frac{\partial}{\partial t}{}_{t+h}p^{\overline{S_iS_i}}_{x,y,z} = -{}_{t+h}p^{\overline{S_iS_i}}_{x,y,z}(\rho(i)_{x+t+h,y+t+h,z+t+h} + \nu(i)_{x+t+h,y+t+h,z+t+h})$$

We then take the "average" value of the derivative over $(t, t+h)$ as

$$\left\{\frac{\partial}{\partial t}{}_tp^{\overline{S_iS_i}}_{x,y,z} + \frac{\partial}{\partial t}{}_{t+h}p^{\overline{S_iS_i}}_{x,y,z}\right\}/2$$

and put, approximately:

$$
\begin{aligned}
&\left\{ {}_{t+h}p^{\overline{S_iS_i}}_{x,y,z} - {}_{t}p^{\overline{S_iS_i}}_{x,y,z} \right\} /h \\
&= \left\{ \frac{\partial}{\partial t} {}_{t}p^{\overline{S_iS_i}}_{x,y,z} + \frac{\partial}{\partial t} {}_{t+h}p^{\overline{S_iS_i}}_{x,y,z} \right\} /2 \\
&= \left\{ -{}_{t}p^{\overline{S_iS_i}}_{x,y,z} (\rho(i)_{x+t,y+t,z+t} + \nu(i)_{x+t,y+t,z+t}) \right\} /2 \\
&\quad + \left\{ -{}_{t+h}p^{\overline{S_iS_i}}_{x,y,z} (\rho(i)_{x+t+h,y+t+h,z+t+h} + \nu(i)_{x+t+h,y+t+h,z+t+h}) \right\} /2
\end{aligned}
$$

After some algebraic manipulation, we get

$$
\begin{aligned}
&{}_{t+h}p^{\overline{S_iS_i}}_{x,y,z} \left\{ 1 + \frac{h}{2} (\rho(i)_{x+t+h,y+t+h,z+t+h} + \nu(i)_{x+t+h,y+t+h,z+t+h}) \right\} \\
&= {}_{t}p^{\overline{S_iS_i}}_{x,y,z} \left\{ 1 - \frac{h}{2} (\rho(i)_{x+t,y+t,z+t} + \nu(i)_{x+t,y+t,z+t}) \right\}
\end{aligned}
$$

and obtain the following recursive formula:

$$
{}_{t+h}p^{\overline{S_iS_i}}_{x,y,z} = \frac{{}_{t}p^{\overline{S_iS_i}}_{x,y,z} \left\{ 1 - \frac{h}{2} (\rho(i)_{x+t,y+t,z+t} + \nu(i)_{x+t,y+t,z+t}) \right\}}{\left\{ 1 + \frac{h}{2} (\rho(i)_{x+t+h,y+t+h,z+t+h} + \nu(i)_{x+t+h,y+t+h,z+t+h}) \right\}}
$$

Appendix D

The Graduation Formulae for the Recovery and Mortality Intensities from Sick in CMI Working Paper 5 (2005)

In this Appendix we present the graduation formula for recovery and mortality intensitiies from sick as reported in CMI Working Paper 5 (2005). The general functional form for the recovery intensity is discussed in Sections 1.5 while that for the mortality intensity from sick is discussed in Section 4.1. Both sets of graduated intensities are used to compare against the aggregate intensities in Section 6.3.

Graduation formula for the recovery intensity

The graduation formula for the recovery intensity is represented by the following symbolic form:

$$\log(\rho(d, y, z)) = s_d + g_z + q_z + f_{yz} + h_{yz}$$

where z is the exact sickness duration in years, such that

$$Z = \begin{cases} z & \text{for} \quad z \leq 5 \\ 5 & \text{for} \quad z > 5 \end{cases}$$

where y is the exact age (in years) at the date of falling sick, such that

$$Y = \begin{cases} y - 50 & \text{for} \quad z \leq 5 \\ y - 55 + z & \text{for} \quad z > 5 \end{cases}$$

$w = (365/7)Z$ i.e. Z is translated into units of weeks

$t(Z) = w/(1 + kw)$

The full details of the component terms in the above graduation formula are as follows:

$$s_d = s(d) \quad \text{and} \quad d \in \{\text{DP1, DP4, DP13, DP26, DP52}\}$$

$$g_z = \begin{cases} -b_1 t(Z) & \text{for} \quad w \leq 26 \\ -b_1 t(26) - b_2\{t(Z) - t(26)\} & \text{for} \quad w > 26 \end{cases}$$

$$q_z = \begin{cases} -r_1(16 - w)/8 & \text{for DP4 if} \quad 8 \leq w < 16 \\ 0 & \text{otherwise} \end{cases}$$

$$r_z = \begin{cases} -r_2(8 - w)/4 - r_1 & \text{for DP4 if} \quad 4 \leq w < 8 \\ -r_3(17 - w)/4 & \text{for DP13 if} \quad 13 \leq w < 17 \\ 0 & \text{otherwise} \end{cases}$$

$$f_{yz} = a_1(Y/100) + a_2(Y/100)^2 + a_3(Y/100)^3 + a_4(Y/100)t(Z)$$

$$h_{yz} = \begin{cases} \{t(4) - t(Z)\}\{h_0 + h_1 Y/100 + h_2 t(Z)\} & \text{for} \quad w < 4 \\ 0 & \text{for} \quad w > 4 \end{cases}$$

The parameter values in the above graduation formula are as follows:

$$s(1) = 3.036467 \qquad s(2) = 3.316474 \qquad s(3) = 3.025743 \qquad s(4) = 2.856549$$
$$s(5) = 2.511347 \qquad k = 0.016000 \qquad a_1 = -3.080944 \qquad a_2 = -6.419924$$
$$a_3 = 20.048953 \qquad a_4 = -0.113352 \qquad b_1 = 0.195291 \qquad b_2 = 0.108662$$
$$h_0 = 0.198289 \qquad h_1 = -0.724805 \qquad h_2 = 0.047682 \qquad r_1 0.622543$$
$$r_2 = 1.197880 \qquad r_3 = 1.830356$$

Graduation formula for the mortality intensity from sick

The graduation formula for the mortality intensity from sick is given by

$$\nu(y+z,z) = \left(\frac{a \exp\{-b/(Z+c)\}}{(Z+c)^2} + (r/100) \exp\{s(Y+Z)\} \right) q(d)$$

where z is the exact sickness duration in years, such that

$$Z = \begin{cases} z & \text{for} \quad z \leq 5 \\ 5 & \text{for} \quad z > 5 \end{cases}$$

where y is the exact age (in years) at the date of falling sick, such that

$$Y = \begin{cases} y - 50 & \text{for} \quad z \leq 5 \\ y - 55 + z & \text{for} \quad z > 5 \end{cases}$$

$$\text{where } q(d) = \begin{cases} q & \text{for} \quad DP1 \\ 1 & \text{for other deferred periods} \end{cases}$$

The parameter values in the above graduation formula are as follows:

$$a = 0.188906 \qquad b = 1.081708 \qquad c = 0.132474$$
$$r = 0.257331 \qquad s = 0.149466 \qquad q = 0.744739$$

Appendix E

The Grouping Algorithm

It was explained in Section 3.5 that in order to carry out the χ^2 test when the data is sparse, it was desirable to group cells in a systematic and reasonable way so that the number of events in each cell is sufficiently large. The purpose of this appendix is to present the grouping algorithm used in the merger of cells.

In a tableau, data is arranged with columns representing sickness duration bands and rows representing age bands. Let k_{column}, k_{row} and k_{cell} be integers representing the minimum numbers of expected events for any column, row and cell in the final compressed tableau, respectively. As in CMI Report 15 (1996), we choose $k_{\text{column}} = k_{\text{row}} = 15$ and $k_{\text{cell}} = 8$.

The tableau is first traversed from left to right, i.e. from the lowest sickness duration band to the highest. Any column with fewer than k_{column} expected events is merged with the column to the right, i.e. the next higher sickness duration band. If the total number of expected number of events in this newly merged column is still fewer than k_{column}, then it is merged with subsequent columns to the right until at least k_{column} expected events is obtained in the new column. Such a grouping procedure from left to right may result in the last column having fewer than k_{column} expected events but can no longer be moved further to the right. In this case, this last column will be merged together with the last preceding non-zero column with enough events. Once this is done, no column in the tableau has fewer than k_{column} expected events.

The same procedure is then used to group the rows by traversing the tableau from top to bottom, i.e. from the lowest age band to the highest age band. If the total number of expected events in any row is fewer than k_{row}, it is added to the subsequent rows until at least k_{row} expected events is obtained. This procedure may leave the last row with fewer than k_{row} expected events, in which case, it is merged with the preceding non-zero row with enough events. At the end of this procedure, each row has no fewer than k_{row} expected events.

Lastly, individual cells within each row are compressed in a way such that when the row is traversed from left to right, any cell with fewer than k_{cell} expected events is added to the subsequent cells to the right until k_{cell} expected events is obtained. If the last cell were left with fewer than k_{cell} expected events, it is added to the preceding cell on its left. At the end of this procedure, each cell has at least k_{cell} expected events.

In order to group cells in a tableau, there should be at least k_{cell} expected events in a tableau. For each IP dataset by cause of sickness, we have a total of $2 \times 2 \times 5 \times 9 = 140$ possible tableaux, one for each distinct combination of sex, rating indicator, deferred period and year band.

We first checked that the total expected events in the tableaux belonging to either sex is at least k_{cell}. If the total expected events in the tableaux belonging to any sex is fewer than k_{cell}, these tableaux are merged with their corresponding tableaux from the opposite sex. Once this check is done, the total expected events in the resulting tableaux for male, female, or both combined is greater than k_{cell}.

The same procedure is then applied to non-rated and rated tableaux. If the total expected events in the tableaux belonging to either rated or not-rated is fewer than k_{cell}, these tableaux are combined with their corresponding tableaux from another rating indicator. At the end of this procedure, the total expected events in the resulting tableaux for not-rated, rated, or both combined is greater than k_{cell}.

We then apply the same procedure on tableaux belonging to each deferred period, from DP1 to DP52, so that if the total expected events for a tableaux belonging to any of the deferred periods is fewer than k_{cell}, they are merged together with the corresponding tableaux from the next higher deferred period until k_{cell} expected events is obtained. At the end of this procedure, the total expected events for tableaux belonging to each deferred period or merged deferred periods is at least k_{cell}.

Lastly, the same procedure is then used to group tableaux for year bands. There are 9 tableaux, one for each year band, and they are arranged from the lowest year band to the highest. If any of these 9 tableaux has fewer than k_{cell} expected events, it is combined with the tableau on its right (i.e. the tableau for the higher year band) until the total combined expected events is at least k_{cell}. Such a grouping procedure may leave the final tableau (i.e. the tableau for the highest year band) with fewer than k_{cell} expected events, in which case, it is merged with the preceding non-zero tableau with enough events.

At the end of these procedures, each of the resulting tableaux has at least k_{cell} expected events. If the total expected number of events in a tableau is fewer than twice k_{cell}, then the entire tableau is compressed to a single cell.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study. *Annals of statistics*, **10(4)**, 1100–1120.

Andersen, P. K.; Borch-Johnsen, K.; Deckert, T.; Green, A.; Hougaard, P.; Keiding, N. and Kreiner, S. (1985). A Cox regression model for relative mortality and its application to diabetes mellitus survival data. *Biometrics*, **41(4)**, 921–932.

Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics*, **30**, 89–100.

Breslow, N. E. (1985). Cohort analysis in epidemiology. *A Celebration of Statistics: The ISI Centenary Volume. A. C. Atkinson and S. E. Fienberg, eds. Springer-Verlag*, , 109–143.

Continuous Mortality Investigation Committee (1976). *Continuous Mortality Investigation Reports: Number 2.* The Institute of Actuaries and the Faculty of Actuaries.

Continuous Mortality Investigation Committee (1983). *Continuous Mortality Investigation Reports: Number 6.* The Institute of Actuaries and the Faculty of Actuaries.

Continuous Mortality Investigation Committee (1986). *Continuous Mortality Investigation Reports: Number 8.* The Institute of Actuaries and the Faculty of Actuaries.

Continuous Mortality Investigation Committee (1991). *Continuous Mortality Investigation Reports: Number 12.* The Institute of Actuaries and the Faculty of Actuaries.

Continuous Mortality Investigation Committee (1996). *Continuous Mortality Investigation Reports: Number 15.* The Institute of Actuaries and the Faculty of Actuaries.

Continuous Mortality Investigation Committee (2000). *Continuous Mortality Investigation Reports: Number 18.* The Institute of Actuaries and the Faculty of Actuaries.

Continuous Mortality Investigation Committee (2005). *Continuous Mortality Investigation Reports: Number 22.* The Institute of Actuaries and the Faculty of Actuaries.

Continuous Mortality Investigation Committee (2004). *Continuous Mortality Investigation Working Paper: Number 5.* The Institute of Actuaries and the Faculty of Actuaries.

Continuous Mortality Investigation Committee (2004). *Continuous Mortality Investigation Working Paper: Number 7.* The Institute of Actuaries and the Faculty of Actuaries.

Continuous Mortality Investigation Committee (2006). *Continuous Mortality Investigation Working Paper: Number 23.* The Institute of Actuaries and the Faculty of Actuaries.

Cordeiro, I.M.F. (1998). A stochastic model for the analysis of permanent health insurance claims by cause of disability. *Ph.D. Thesis, Heriot-Watt University, Edinburgh.*

Cordeiro, I.M.F. (2002). A multiple state model for the analysis of permanent health insurance claims by cause of disability. *Insurance: Mathematics and Economics*, **30**, 167-186.

Cox, D. R. (1972). Regression Models and Life Tables (with Discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, **34(2)**, 187–220.

Czabo, C. and Rudolph, F. (2002). Application of survival analysis to long-term care insurance. *Insurance: Mathematics and Economics*, **31(3)**, 395–413.

Devlin, T.F. and Weeks, B.J (1986). Spline functions for logistic regression modeling. *Proc 11th Annual SAS Users Group Intnl Conf. Cary NC: SAS Institute, Inc.*, , 646-651.

Dickman, P.W.; Sloggett, A.; Hills, M; and Hakulinen, T. (2004). Regression models for relative survival. *Statistics in Medicine*, **23**, 51–64.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, **72(359)**, 557–565.

Estève, J., Benhamou, E., Croasdale, M. and Raymond, M. (1990). Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine*, **9(5)**, 529–538.

Forfar, D.O, McCutcheon, J.J and Wilkie, A.D. (1988). On graduation by mathematical formula. *Journal Of The Institute of Actuaries*, **115(I)**, 1–149.

Gray, R. J. (1990). Some diagnostic methods for Cox regression models through hazard smoothing. *Biometrics*, **46(1)**, 93–102.

Haberman, S. and Renshaw, A.E. (1990). Generalised linear models and excess mortality from peptic ulcers. *Insurance: Mathematics and Economics*, **9(1)**, 21–32.

Haberman, S. and Renshaw, A.E. (1996). Generalized linear models and actuarial science. *The Statistician*, **20(2)**, 407–436.

Heligman, L. and Pollard, J.H. (1980). The age pattern of mortality. *Journal of the Institute of Actuaries*, **107**, 49–80.

Hess, K.R. (1995). Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Statistics in Medicine*, **14(15)**, 1707–1723.

Holford, T. R. (1976). Life tables with concomitant information. *Biometrics*, **32(3)**, 587–597.

Holford, T. R. (1980). The analysis of rates and survivorship using loglinear models. *Biometrics*, **36(2)**, 299–305.

Jones, B.L. (1992). An analysis of long-term care data from Hamilton-Wentworth, Ontario. *Actuarial Research Clearing House*, **1**, 337–352.

Jones, B.L. (1995). A stochastic population model for high demand CCRCs. *Insurance: Mathematics and Economics*, **16**, 69–77.

Jones, B.L. (1997). Methods for the Analysis of CCRC data. *North American Actuarial Journal*, **1**, 40–54.

Kluwer (2001). *Income Protection Insurance 2001.* Croner Publications and Kluwer Publishing, United Kingdom.

Laird, N. and Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, **76**, 231–240.

Lee, R. D. and Carter, L. (1992). Modeling and Forecasting the Time Series of U.S. Mortality. *Journal of the American Statistical Association*, **87**, 659-671.

Ling, S.Y. (2008). Supporting document for Ph.D thesis at http://www.ma.hw.ac.uk/∼singyee/.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

Pitt, D. G. W. (2007). Modelling the claim duration of income protection insurance policyholders using parametric mixture models. *Annals of Actuarial Science*, **2(I)**, 1–24.

Renshaw, A. E. (1988). Modelling excess mortality using GLIM. *Journal of the Institute of Actuaries*, **115(2)**, 299–315.

Renshaw, A.E. (1991). Actuarial graduation practice and generalized linear & nonlinear models. *Journal of the Institute of Actuaries*, **118(11)**, 295–312.

Renshaw, A. E. and Haberman, S. (1995). On the graduation associated with a multiple state model in permanent health insurance. *Insurance: Mathematics and Economics*, **17(1)**, 1–17.

Renshaw, A.E., Haberman, S., Hatzopoulos, P. (1996). The modelling of recent mortality trends in U.K. male assured lives. *British Actuarial Journal*, **2(2)**, 449–477.

Renshaw, A.E. and Haberman, S. (2000). Modelling the recent time trends in UK permanent health insurance recovery, mortality and claim inception transition intensities. *Insurance: Mathematics and Economics*, **27**, 365-396.

Sanders, A.J. and Silby, N.F. (1988). Actuarial aspects of PHI in the UK. *Journal of the Institute of Actuaries Students' Society*, **31**, 1–57.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, **69(1)**, 239–241.

Stare, J., Pohar, M., Henderson, R. (2005). Goodness of fit of relative survival models. *Statistics in Medicine*, **24(24)**, 3911–3925.

Stark, C., MacLeod, M., Hall, D., O'Brien, F. and Pelosi, A. (2003). Mortality after discharge from long-term psychiatric care in Scotland, 1977 94: a retrospective cohort study. *BMC Public Health*, **3:30**, .

Therneau, T. and Grambsch, P. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81(3)**, 515–526.