Journal of Demographic Economics (2023), **89**, 419–463
doi:10.1017/dem.2023.3

CAMBRIDGE UNIVERSITY PRESS

RESEARCH PAPER

Survival analysis of longitudinal data: the case of English population aged 50 and over

Marjan Qazvini

Department of Actuarial Mathematics and Statistics, School of Mathematical and Computer Sciences, Heriot-Watt University, Dubai, UAE
Corresponding author: Email: M.Qazvini@hw.ac.uk, Marjan.Qazvini@gmail.com

(Received 3 April 2023; revised 3 April 2023; accepted 3 April 2023)

Abstract

This study considers data from 5 waves of the English Longitudinal Study of Ageing (ELSA). We aim to study the impact of demographic and self-rated health variables including disability and diseases on the survival of the population aged 50+. The disability variables that we consider are mobility impairment, difficulties in performing Activities of Daily Living (ADL) and Instrumental Activities of Daily Living (IADL). One of the problems with the survey study is missing observations. This may happen due to different reasons, such as errors, nonresponse and temporary withdrawals. We address this problem by applying single and multiple imputation methods. We then fit a Generalized Linear model (GLM) and Generalized Linear Mixed model (GLMM) to our data and show that a GLMM performs better than a GLM in terms of information criteria. We also look at the predictability of our models in terms of the time-dependent receiver operating characteristic (ROC) and the area of ROC, i.e. AUC. We conclude that among the disability factors, IADL and among the diseases, cancer significantly affect the survival of the English population aged 50 and older.

Keywords: Longitudinal data; survival analysis; random effects model; multiple imputation

1. Introduction

Survival analysis is the study of the survival of a member and the variables that affect survival until the event of interest occurs. In our study the event of interest is death and we aim to investigate the impact of demographic factors such as age, marital status, employment status and self-reported health factors such as mobility, Activities of Daily Living (ADL) and Instrumental Activities of Daily Living (IADL) impairment and Noncommunicable Diseases (NCD) such as Cardiovascular Diseases (CVD) and lung diseases on the lives of the English population aged 50+. We consider ELSA which is a panel study of a representative cohort of individuals aged 50 or older living in private households in England and is conducted every two years since 2002. ELSA has been extensively studied by researchers in medical and social sciences. For example, Demakakos *et al.* (2016) study the relationship between wealth and all-cause and cause-specific mortality using Cox proportional regression analysis and

find that wealth is strongly associated with CVD and other non-cancer mortality among people aged 50–64, whereas there is a weak relationship between wealth and cancer mortality. In another study, Demakakos *et al.* (2018) investigate the relationship between self-perceptions of socio-economic status and all-cause and cause-specific mortality using Cox proportional regression hazard model and chained equations to impute missing values. Some studies use ELSA to consider the problem of disability and factors that affect disability among the elderly. For example, d'Orsi *et al.* (2014) apply Generalized Estimating Equation (GEE) models with 2-year lagged Poisson regression to look at the impacts of socio-economic, demographic and lifestyle on IADL, its duration and speed of recovery. Potente and Monden (2016) use a multinomial logistic model to study the relationship between socio-economic status and disability before death in ELSA dataset. They look at the association between income, education, wealth and disability factors such as ADL, IADL and mobility impairment among deceased individuals. Steptoe and Zaninotto (2020) study the relationship between socio-economic status, represented by wealth and physical capability, sight, hearing impairment, physiological function, cognitive performance, emotional well-being and social function using covariance and logistic regression analysis while controlling for age, gender, ethnicity, education and long-term health conditions. They find that lower wealth is associated with all these problems and hence will increase the rate of ageing. In both studies, among socio-economic indicators, the impact of wealth is more significant than education and income. Torres *et al.* (2016) consider the problem of wealth inequality and disability by taking into account depression and social support using multinomial logistic regressions. Guzman-Castillo *et al.* (2017) predict the life expectancy with and without disability using Sullivan's index (1971). They use ELSA dataset and apply a discrete-time probabilistic Markov model to examine the individual and collective impacts of CVD, dementia, disability and other diseases on life expectancy. Some of the studies in social sciences consider the impact of social interaction and well-being on survival among people in old age. Davies *et al.* (2021) apply linear mixed models and Cox proportional hazard model to study the impact of social isolation and loneliness on frailty and use multiple imputations to impute missing data. [See, also Steptoe *et al.* (2015); Khondoker *et al.* (2017); Rafnsson *et al.* (2020) and the references therein]. There are comparative studies that combine ELSA with data from other countries. Aida *et al.* (2018) consider the impact of social and behavioral factors on survival among old-aged population in Japan and England. They fill in missing values using multiple imputation methods with chained equations and apply Laplace regression models to estimate the percentile of survival time and Cox proportional hazards for sensitivity analysis. Donati *et al.* (2019) use ELSA and The Irish Longitudinal Study of Ageing to predict the development of functional disability through deep and shallow artificial neural networks. They consider declining ADLs and IADLs between two consecutive waves as a measure of frailty among participants and apply Minimum Redundancy Maximum Relevance (MRMR) algorithm to select a subset of the most significant features. Kessler *et al.* (2020) focus on the comparison of risk factors such as physical inactivity, smoking, hypertension, diabetes and high BMI which are the causes of NCD among the population aged 60+ from ELSA and Bagé Cohort study of Ageing (SIGa-Bagé). They conclude that the level of these risks among the Brazilian population is higher than the English population. They ascribe their results to the quality of healthcare and economic situations in England. Stamate *et al.* (2022) also apply machine

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

learning algorithms to predict the development of dementia among participants. They consider patients' general health, mental health, life satisfaction, mobility, socio-economic status, etc and apply Cox proportional hazard model with Elastic Net regularization and a Random Forest algorithm where the best split is obtained by maximizing the survival difference between subsequent nodes. They removed rows with missing values at 51% or greater and apply K-nearest neighbors with $K = 5$ to impute the rest and use concordance measure c-statistics which is a generalization of the Receiver Operating Characteristic curve (ROC) and the area under curve (AUC) [Heagerty and Zheng (2005)] to evaluate the performance of their models.

After reviewing the literature related to our dataset (ELSA), we look at the literature that considers time to event analysis. Cox (1972) introduces Cox regression models for the analysis of the effect of categorical and quantitative variables on survival in continuous time. In our case, the exact time of death and withdrawal, i.e. censored time is unknown and we only know the year of death or whether an interviewee has participated in the follow-up interview or not. Therefore, we need to carry out a discrete-time survival analysis. Thompson (1977) considers the problem of ties and introduces a model based on the grouping of failure times. He applies a logistic model based on Cox's binary hazard model and uses a modified Newton–Raphson's method to estimate the likelihood equations. Friedman (1982) considers a piecewise exponential model for the analysis of censored data with covariates and uses iterative numerical methods to find Maximum Likelihood (ML) estimates. Discrete-time survival analysis has been considered by Allison (1982) where he compares discrete-time and continuous-time survival models and shows that the likelihood function is similar to the function for binary responses in GLMs and therefore similar estimation methods can be applied. Discrete-time survival model in the context of job duration and unemployment has been considered for example, by Petersen (1986), Ham and Rea (1987), and Singer and Willet (1993). Discrete-time frailty models also known as survival models with random effects can account for unobserved heterogeneity which cannot be described by existing covariates. Scheike and Jensen (1997) study the time to pregnancy and point out that random effects models can capture the heterogeneity due to biological variation. The likelihood function of random effects models involves numerical integrations and can be solved by numerical methods such as Laplace methods [Breslow and Clayton (1993)], Gauss–Hermite quadrature (GHQ), adaptive GHQ [Davidian and Giltinan (2003); Bolker *et al.* (2009)], and Markov Chain Monte Carlo methods [Fahrmeir and Knorr-Held (1997); Bolker *et al.* (2009)]. Davidian and Giltinan (2003) provide a review of mixed effects models and discuss different estimation methods and the supported software. Bolker *et al.* (2009) consider GLMMs in the context of ecology and evolution and look at different inference methods for hypothesis testing and model selection. [See, also, Tutz and Schmid (2016) and the references therein for a detailed discussion of discrete-time survival models and their applications in R].

In our dataset, the maximum number of observations for each participant is 5 waves, i.e. 10 person-years. However, some participants may contribute to less than 5 waves due to death, permanent or temporary withdrawals. For example, some individuals may participate only in wave 1 and then re-join in wave 5 and that means we do not have any information in waves 2, 3 and 4. In these cases, we consider the withdrawal periods as missing records and apply Multivariate Imputation by Chained Equations (MICE) to impute these records. [See, for example, Lee and Carlin (2010); Azur *et al.* (2011); van Buuren and Groothuis-Oudshoorn (2011); van Buuren (2018)].

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

Longitudinal study or panel study is a useful tool in order to study the developmental pattern of the same variables over a short or long period. It is often used in clinical psychology to study changes in behavior, thoughts and emotions and in economics to study consumer trends. In actuarial science, longitudinal data has been considered by Frees (2004), Antonio and Valdez (2012), Antonio and Zhang (2014) in the context of experience rating and credibility theory and by Renshaw and Haberman (2000), Richayzen and Walsh (2002), Li *et al.* (2017) and Hanewald *et al.* (2019) in the context of mortality and disability trend.

In this study, we use ELSA dataset and perform a discrete-time survival analysis to examine the impact of demographic and self-reported health factors on the survival of the population aged 50+. We fit a random effects model to our imputed datasets. The predictability of our model will be examined in terms of time-dependent ROC and AUC. The rest of the paper is organized as follows: in Section 2 we explain ELSA dataset and the variables of our study. Section 3 discusses the models and algorithms we use to analyze our dataset. In Section 4 we discuss our results and Section 5 concludes.

2. Data and data preparation

The English Longitudinal Study of Ageing (ELSA) is a collection of economic, social, psychological, cognitive, health, biological and genetic data. The study commenced in 2002 and the sample has been followed up every 2 years. The first cohort was selected from respondents to the Health Survey for England (HSE) in 1998, 1999 and 2001 and included people born on or before February 29, 1952, i.e. aged 50 and older. The first ELSA wave was in 2002–2003. Wave 2 took place in 2004–2005, wave 3 in 2006–2007, wave 4 in 2008–2009 and wave 5 in 2010–2011. To make sure ELSA is designed to be representative of people aged 50 and over in England, in waves 3 and 4, a refreshment cohort of people just entering their 50 s was introduced. These new cohorts are called "Cohort 3" and "Cohort 4". The cohort number was chosen to reflect the wave in which the new sample was introduced. There is no "Cohort 2" or "Cohort 5" as no new sample was issued at waves 2 and 5. In wave 2, an End of Life interview was conducted with the purpose of finding out about the health and socio-economic situations of people just before their death. End of Life interviews have been carried out at waves 2, 3, 4 and 6. [For more information on ELSA, sampling and interview process see, for example, Steptoe *et al.* (2013); Blake *et al.* (2015)]. Table 1 shows the number of participants in each cohort and the number of deaths among "core members" reported in waves 2, 3, 4 and 6. Core members are age-eligible sample members who participated in the HSE and are interviewed in the first wave of ELSA when invited to join [Bridges *et al.* (2015)]. Here, we only focus on core members and do not consider their partners.

In this study, we collected information about age, gender, marital status, employment status, self-rated physical and health conditions of core members from waves 1, 2, 3, 4 and 5 and retrieve information regarding their status from waves 2, 3, 4 and 6. The variables that we consider in our study are presented in Table 2. These variables are extracted from "core data" files that can be obtained from ELSA website after registration. The participants have been asked to fill in a questionnaire. The information provided is based on participants' self-assessment of their health conditions. Responses such as "refusal", "don't know" and "schedule not applicable" are considered missing values. Our dependent variable is the *status* of the individuals

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

Table 1. Number of core members and end of life interviews in each wave

Waves	Cohort 1	Cohort 3[2]	Cohort 4[2]	End of life interviews[3]
(1) 2002–2003	11,391	–	–	–
(2) 2004–2005	8, 780[1]	–	–	133
(3) 2006–2007	7,535	1,275	–	369
(4) 2008–2009	6,623	972	2,291	234
(5) 2010–2011	6,242	936	1,912	–
(6) 2012–2013	–	–	–	240

NatCen: Technical Report (Wave 6).
[1] New cohorts were only introduced in waves 3 and 4.
[2] The number of end of life interviews of the core members.
[3] In wave 2 one member aged below 50 has been removed.

Table 2. Dependent and independent variables of the study

Age	Continuous	Age above 90 is recorded as 90 to avoid disclosure (50 – 90)
Gender	Categorical	Male (1), female (2)
Marital status	Categorical	Single (0), couple (1)
Employment	Categorical 1, 4, 3, 5, 2.	Employed, retired, permanently sick or disabled, self-employed other (looking after home or family, semi-retired, unemployed and other).
Mobility	Discrete 0, …, 10	Walking 100 yards, sitting for about two hours, getting up from chair climbing several flights of stairs without resting, stooping, kneeling, reaching or extending your arms, pulling or pushing large objects, lifting or carrying weights, picking up a 5p coin
ADL	Discrete 0, …, 6	Dressing including putting on shoes and socks, walking across room, bathing or showering, eating such as cutting up your food, getting in or out of bed, using toilet
IADL	Discrete 0, …, 7	Using a map, preparing a hot meal, shopping for groceries, making telephone calls, taking medications, doing work around the house or gardens, managing money such as paying bills
CVDs	Yes (1) No (0)	High blood pressure, angina, heart attack, heart failure, heart murmur, abnormal heart rhythm, diabetes, stroke, other heart diseases
Other diseases	Yes (1) No (0)	Chronic lung diseases, asthma, arthritis, osteoporosis, cancer, Parkinson's, psychiatric diseases, Alzheimer and dementia
Status	Death (1) Survival (0)	Dependent variable

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

that can be obtained from "eol" files for different waves and is coded 1 if death occurs and 0 otherwise. The *age* is a continuous variable which is limited to 90 to avoid disclosure. The age average ranges from 65.2 in wave 1 to 67.8 in wave 5. The *gender* is coded 1 for males and 2 for females. From wave 3, "civil partnership" has been added to the legal marital status. In our study, the *marital status* has two categories: "married" (being in a partnership) and "single" (no partnership including divorced and widowed). There are 8 categories of *employment status*. We combine "semi-retired", "other", "looking after home or family" and "unemployed" as "other".

The questions relating to health factors are different in different waves. We selected questions which are common among all waves. In wave 1 participants were asked to select the physical difficulties and diseases that they suffered, from a list and from wave 2 they were asked about the health conditions carried forward from the previous wave and the newly-diagnosed conditions. In other words, if a participant confirms a particular disease, this can be either the disease from the previous year or the newly-diagnosed disease. Therefore, for these variables, we consider "not applicable" as no experience of physical difficulty or disease. The variables *mobility*, *ADL* and *IADL* score are discrete. They represent the number of activities with which participants have difficulties or need help. For example, a score of "0" means that the participants have not reported difficulties in performing any activities related to mobility and a score of "10" means they have reported difficulties in performing all 10 activities. The scores for *ADL* and *IADL* are similarly defined. The remaining variables are dichotomous.

Figure 1 shows the number of times that a problem related to mobility, ADL and IADL has been reported. We observe that difficulties in performing physical activities, which are represented by green circles, are more frequent than difficulties in performing IADL, which are represented by red circles. Further, difficulty with "stooping" is the most reported problem and difficulty with "picking up a 5p coin" is the least reported problem among mobility impairment. Out of activities related to ADL, difficulty with "dressing" is the most reported problem and difficulty with "eating" is the least reported problem. The most reported problem under IADL is difficulty with "pushing or pulling large objects" and the least reported problem is difficulty with "taking medications".

Table 3 shows the "number of death", "at risk", i.e. the number of members exposed to risk and "hazard probability", i.e. the probability of event occurrence in each interval. Figure 2 shows the baseline hazard function in our model which is the hazard function for the reference individual in the model (see Section 3).

2.1. Missing values

Table 4 shows an example of the type of data and missing values that we are dealing with in our study in long format. In this table, we have 5 periods with repeated observations of variables such as V_1, V_2, … for individuals with ID numbers 1, 2, 3 and 4. Individual 1 participates in interviews within the periods $[t_1, t_2)$, …, $[t_4, t_5)$ and dies in the period $[t_4, t_5)$. Individual 2 participates within the period $[t_1, t_2)$ and $[t_2, t_3)$ and withdraws in the period $[t_2, t_3)$, i.e. right-censored. Individual 3 participates within the period $[t_2, t_3)$ and dies in the same period. Therefore, we do not have any information about this individual for the period $[t_1, t_2)$ except for age, gender and status assuming that in the period $[t_1, t_2)$ the individual was 2 years younger and the gender is fixed. This individual is left-censored. Individual 4

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

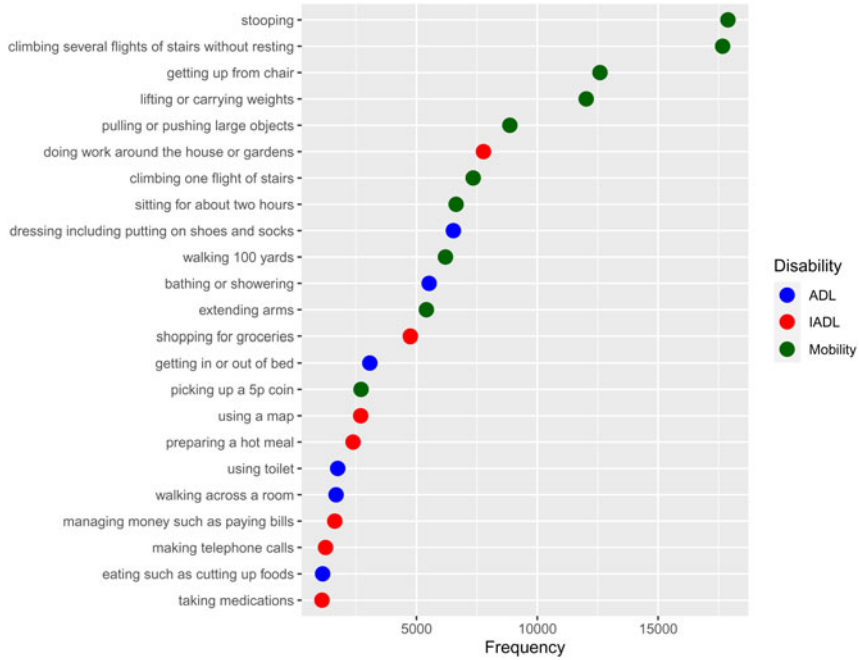

Figure 1. The number of times a problem has been reported.

participates within the periods $[t_1, t_2)$ and $[t_2, t_3)$, withdraws temporary in the period $[t_3, t_4)$, rejoins the study in the period $[t_4, t_5)$ and dies in that period. Therefore, the only information that we have from this individual in the period $[t_3, t_4)$ is age, gender and status. In this study, all *NA*s are considered missing values.

We assume that missing values are independent of the occurrence of death, that is participants have not withdrawn from the study because they were more or less likely to die. Given that the interviews are conducted every two years, we expect each participant to be two years older in the follow-up interview, so the missing values of the variable *age* can be easily filled in. However, this means that the minimum age is now below 50 and the age is not capped at 90. The variable *gender* is "time-invariant" and complete. The next variable to consider is the *employment status*. We set all missing values as "retired" for participants aged 60 and above. To fill in the missing values of *employment status* and *marital status* we use Last Observation Carried Forward (LOCF) method. LOCF is a single imputation method in which the last observed record of an individual is used to fill in subsequent missing values. However, it is not recommended by the National Academy of Sciences for the imputation of missing values in clinical trials and/or without justifications[1]. Here we do not expect much changes in *employment status* and *marital status* among the elderly. After the implementation of LOCF, we still have two participants with missing values that we remove from our dataset. For other variables such as physical disability and diseases, we use multiple imputation methods. MICE, known as "fully conditional

[1]https://www.nap.edu/catalog/12955/the-prevention-and-treatment-of-missing-data-in-clinical-trials

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

Table 3. Number of deaths and at risk in each time interval

Wave	Number of death	Number at risk	Hazard probability
[1, 2)	322	14,966	0.0215
[2, 3)	231	12,978	0.0178
[3, 4)	190	11,759	0.0162
[4, 5)	158	10,477	0.0151
[5, 6)	75	9,090	0.0083
Total	976	14,966	

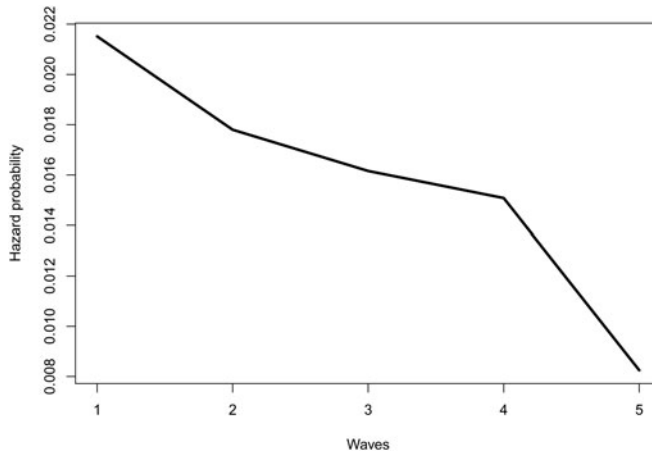

Figure 2. Baseline hazard function: log $(-\log(1-\lambda(t)) = \gamma_1 T_1 + \gamma_2 T_2 + \gamma_3 T_3 + \gamma_4 T_4 + \gamma_5 T_5$.

specification" or "sequential regression multiple imputation", can handle different types of variables such as continuous, binary and categorical variables. In R package *mice*, we use the "predictive mean matching" (pmm) method for the imputation of *mobility*, *ADL* and *IADL* and "logistic regression" for the imputation of other variables. One of the advantages of pmm is that it always finds values that have been actually observed in the data, i.e. all imputed values are based on the possible observed values and this reduces the possibility of model misspecification and allows the algorithm to be used for any data types. The assumption underlying this method is that the distribution of the missing data is the same as the distribution of the observed data. Figure A3 in the Appendix compares the distribution of the imputed values of these 3 covariates with the distribution of the observed values. We explain pmm and logistic algorithms in Section 3.4. We create 5 complete datasets with 10 iterations. The number of iterations can be determined by inspecting the trace lines generated by the algorithm. In Figures A1 and A2 we cannot detect any trends as expected and therefore 10 iterations is reasonable [van Buuren and Groothuis-Oudshoorn (2011)]. MICE is computationally expensive for a large number of variables. Therefore, we consider a model with a very simple predictor matrix by setting all variables except *status* equal to zero [van Buuren (2018), page 300].

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

Table 4. An example of longitudinal data with missing values (long format). V_1, V_2, … are variables such as age, ADL score, etc.

ID	Time	V_1	V_2	V_3	…	Status
1	$[t_1 - t_2)$	1	0	0	…	0
1	$[t_2 - t_3)$	*NA*	0	0	…	0
1	$[t_3 - t_4)$	1	*NA*	0	…	0
1	$[t_4 - t_5)$	0	1	0	…	1
2	$[t_1 - t_2)$	0	*NA*	0	…	0
2	$[t_2 - t_3)$	1	0	0	…	0
3	$[t_1 - t_2)$	*NA*	*NA*	*NA*	…	0
3	$[t_2 - t_3)$	1	0	0	…	1
4	$[t_1 - t_2)$	0	1	0	…	0
4	$[t_2 - t_3)$	1	0	0	…	0
4	$[t_3 - t_4)$	*NA*	*NA*	*NA*	…	0
4	$[t_4 - t_5)$	0	1	0	…	1

The 5 datasets are identical for the observed data but differ in the imputed values. We store these 5 complete datasets and perform our analysis on each dataset instead of combining the results using Rubin's rules. Table 5 shows the mean for *mobility*, *ADL* and *IADL* scores and the number of participants with a particular disease for 5 datasets after imputation, i.e. datasets I, II, III, IV and V.

3. Models

In this section, we look at discrete-time survival models, also known as time-to-event models. When we are working with survey data the information is not available at the exact point in time and we only know the period in which the event of interest occurs, which is usually every one year or in our case every two years. We can divide the underlying continuous-time process into intervals $[0, a_1)$, $[a_1, a_2)$, …, $[a_{t-1}, a_t)$, $[a_t, \infty)$. Let T be a discrete-time random variable, where $T = t$ means the event has happened in the interval $[a_{t-1}, a_t)$. Censoring is common in survival analysis and right-censoring occurs when the participants are lost to follow up or when the study ends. For individual i, $i = 1, \ldots, n$, let T_i denote duration times and U_i be right-censoring times. Let $t = \min(T_i, U_i)$ be the observed discrete time and δ_i be the censoring indicator:

$$\delta_i = \begin{cases} 1, & T_i < U_i, \text{ i.e. observation is uncensored} \\ 0, & T_i \geq U_i, \text{ i.e. observation is censored.} \end{cases} \tag{1}$$

Let $y_{it} \in \{0, 1\}$ be the event indicator. We then have

$$y_{it} = \begin{cases} 1, & \text{event occurs in } [a_{t-1}, a_t), \\ 0, & \text{event does not occur in } [a_{t-1}, a_t). \end{cases} \tag{2}$$

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

Table 5. Mean and frequency of the variables for 5 new datasets. 1: Mean.

	Status	Mobility[1]	ADL[1]	IADL[1]	High pressure	Angina	Heart attack	Heart failure
Dataset I	0	2.028	0.4091	0.4450	38,327	55,489	57,890	58,979
	1				20,938	3,776	1,375	286
Dataset II	0	2.029	0.4062	0.4428	38,437	55,525	57,934	58,968
	1				20,828	3,740	1,331	297
Dataset III	0	2.011	0.4047	0.4457	38,445	55,499	57,903	58,969
	1				20,820	3,766	1,362	296
Dataset IV	0	2.013	0.4070	0.4417	38,381	55,481	57,874	58,971
	1				20,884	3,784	1,391	294
Dataset V	0	2.024	0.4054	0.4451	38,326	55,486	57,909	58,966
	1				20,939	3,779	1,356	299
		Heart murmur	Heart rhythm	Diabetes	Stroke	Other	Lung	Asthma
Dataset I	0	57,178	55,443	53,999	57,994	57,610	56,220	52,854
	1	2,087	3,822	5,266	1,271	1,655	3,045	6,401
Dataset II	0	57,184	55,468	53,995	57,987	57,624	56,184	52,833
	1	2,081	3,797	5,270	1,278	1,641	3,081	6,432
Dataset III	0	57,191	55,441	53,972	57,968	57,610	56,221	52,919
	1	2,074	3,824	5,293	1,297	1,655	3,044	6,346
Dataset IV	0	57,169	55,460	53,975	57,945	57,608	56,186	52,822
	1	2,096	3,805	5,290	1,320	1,657	3,079	6,443
Dataset V	0	57,195	55,467	53,945	57,958	57,610	56,203	52,882
	1	2,070	3,798	5,320	1,307	1,655	3,062	6,383
		Arthritis	Osteoporosis	Cancer	Parkinson's	Psychiatric	Alzheimer	Dementia
Dataset I	0	38,560	55,390	56,763	58,903	54,590	59,081	58,702
	1	20,705	3,875	2,502	362	4,675	184	563

Dataset II	0	38,541	55,448	56,726	58,885	54,624	59,094	58,684
	1	20,724	3,817	2,539	380	4,641	171	581
Dataset III	0	38,483	55,451	56,764	58,880	54,622	59,094	58,681
	1	20,782	3,814	2,501	385	4,643	171	584
Dataset IV	0	38,456	55,479	56,721	58,867	54,629	59,099	58,695
	1	20,809	3,786	2,544	398	4,636	166	570
Dataset V	0	38,611	55,453	56,741	58,895	54,575	59,100	58,684
	1	20,654	3,812	2,524	370	4,690	165	581

In the following, we explain survival models and the estimation methods in GLM [McCullagh and Nelder (1989)] and GLMM framework.

3.1. GLMs with time-varying covariates

Let $\mathbf{x}_{it} = (x_{i1}, \ldots, x_{it})^T$ be a vector of covariates and $i = 1, 2, \ldots, n$ be the sample size. For individual i, we then define the hazard function in discrete time by

$$\lambda(t|\mathbf{x}_{it}) = \Pr(T_i = t|T_i \geq t, \mathbf{x}_{it}) = h(\eta_{it}), \tag{3}$$

which is the conditional probability that an event occurs at time t given that it has not already occurred and η_{it} is the linear predictor, given by

$$\eta_{it} = \gamma_{0t} + \mathbf{x}_{it}^T\boldsymbol{\gamma}. \tag{4}$$

In Equation (3), $h(.)$ is a function, which is assumed to be strictly monotonically increasing with the inverse function $g = h^{-1}$ which is known as the link function. Therefore, Equation (3) can also be written as

$$g(\lambda(t|\mathbf{x}_{it})) = \gamma_{0t} + \mathbf{x}_{it}^T\gamma.$$

The function $h(.)$ is selected such that the value of the discrete-time hazard is restricted to the interval [0, 1]. Common candidates are logistic, probit, Gompertz (clog-log), and Gumbel (log-log) link functions. Further, γ_{0t} is the intercept which may vary over time and is interpreted as a baseline hazard and $\boldsymbol{\gamma}$ is the vector of parameters. Thompson (1977) shows that logistic and clog-log link functions give rise to similar results. In this study, we choose clog-log link function, where $h(\eta) = 1 - \exp(-\exp(\eta))$. Hence, the hazard function is given by

$$\lambda(t|\mathbf{x}_{it}) = 1 - \exp(-\exp(\gamma_{0t} + \mathbf{x}_{it}^T\boldsymbol{\gamma})), \tag{5}$$

which can also be written as

$$\log(-\log(1 - \lambda(t|\mathbf{x}_{it}))) = \gamma_{0t} + \mathbf{x}_{it}^T\boldsymbol{\gamma}. \tag{6}$$

We can also define the discrete-time survival function by

$$S(t|\mathbf{x}_{it}) = \Pr(T_i > t|\mathbf{x}_{it}) = \prod_{k=1}^{t}[1 - \lambda(k|\mathbf{x}_{it})], \tag{7}$$

which is the probability that death occurs after time t given the covariates. Hence the unconditional probability of death at time t, i.e. the probability that death occurs within the interval $[a_{t-1}, a_t)$ is given by

$$\Pr(T_i = t|\mathbf{x}_{it}) = \prod_{k=1}^{t-1}[1 - \lambda(k|\mathbf{x}_{it})]\lambda(t|\mathbf{x}_{it}). \tag{8}$$

Assuming that censoring is non-informative and does not depend on the parameters, the likelihood function for observation i is given by

$$L = \prod_{i=1}^{n}[\Pr(T_i = t_i)]^{y_{it}}[\Pr(T_i > t_i)]^{1-y_{it}}. \tag{9}$$

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

Substituting (7) and (8) and taking the logarithm yields the log-likelihood function [Allison (1982); Singer and Willet (1993)]

$$l = \sum_{i=1}^{n}\left[y_{it} \log \lambda(t|\mathbf{x}_{it}) + y_{it} \sum_{k=1}^{t_i-1} \log\left[1 - \lambda(k|\mathbf{x}_{it})\right] + (1 - y_{it}) \sum_{k=1}^{t_i} \log\left[1 - \lambda(k|\mathbf{x}_{it})\right]\right], \tag{10}$$

which can be rewritten as

$$l = \sum_{i=1}^{n} y_{it} \log \frac{\lambda(t_i|\mathbf{x}_{it})}{[1 - \lambda(t_i|\mathbf{x}_{it})]} + \sum_{i=1}^{n} \sum_{k=1}^{t_i} \log\left[1 - \lambda(k|\mathbf{x}_{it})\right]. \tag{11}$$

Equation (11) can be solved using numerical methods such as Iteratively Reweighted Least Squares (IRLS) in R. [See, McCullagh and Nelder (1989)].

3.2. GLMMs

In this section, we explain random effects models. Let b_i be a random intercept specific to individual i that follows a mixing distribution with density $f(.)$. We define the discrete-time hazard function for individual i by

$$\lambda(t|\mathbf{x}_{it}, b_i) = \Pr(T_i = t|T_i \geq t, \mathbf{x}_{it}, b_i) = h(\eta'_{it}), \tag{12}$$

where η'_{it} is the linear predictor given by

$$\eta'_{it} = b_i + \gamma_{0t} + \mathbf{x}_{it}^T \boldsymbol{\gamma}. \tag{13}$$

Similar to GLMs we define the survival probability in discrete-time by

$$S(t|\mathbf{x}_{it}, b_i) = \Pr(T_i > t|\mathbf{x}_{it}, b_i) = \prod_{k=1}^{t}[1 - \lambda(k|\mathbf{x}_{it}, b_i)].$$

Since our model only includes a random intercept, we call it the random intercept model. Estimation of parameters in GLMMs is not as straightforward as the GLMs. To estimate the parameters, we need to integrate the likelihood function over all possible values of the random effects, i.e.

$$L = \prod_{i=1}^{n} \int \left[\lambda(t|\mathbf{x}_{it}, b_i) \prod_{k=1}^{t_i-1} [1 - \lambda(k|\mathbf{x}_{it}, b_i]\right]^{y_{it}} \left[\prod_{k=1}^{t_i}[1 - \lambda(k|\mathbf{x}_{it}, b_i)]\right]^{1-y_{it}} f(b_i)db_i. \tag{14}$$

To handle the integral, we need to use numerical integration methods. In this study, we use Gauss–Hermite quadrature which is a numerical method used by "glmer" in R package "lme4" to approximate the value of integrals of the following kind

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)\mathrm{d}x.$$

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

The idea is that we can approximate an integral as a sum of polynomials, i.e.

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} e^{-x^2}\left[e^{x^2} f(x)\right]\mathrm{d}x \approx \sum_{i=1}^{n} w_i f(x_i),$$

where w_i and x_i are the weights and quadrature points, respectively. Parameters are chosen so that this approximation is exact for polynomials $f(x)$ up to and including degree $2n-1$, where n is the number of points in the summation. Suppose $n=3$, then we have 3 unknown w, 3 unknown x and polynomials with degrees up to and including 5, i.e. $f(x)=1$, $f(x)=x$, …, $f(x)=x^5$. Therefore, we have 6 equations to find 6 unknowns. The values of w_i and x_i for different n can be obtained from Table 25.10 in Abramowitz and Stegun (1964).

3.3. Optimization

Parameter estimation in the above models is an optimization problem as we either need to maximize the likelihood function or minimize the negative likelihood function, i.e. $\min f(x)$, $x \in R^n$. The common approach to solving such problems is to take a derivate of the objective function (here, the likelihood function). However, when the objective function is not smooth and/or is complicated, we can use gradient-free optimizers and numerical methods to find the optimum value. In Table A2 in the Appendix, we provide a comparison of some of the gradient-free algorithms in terms of the estimated parameters, convergence and the time taken until convergence. In the following, we briefly explain the optimizers in this table.

Nelder-Mead simplex algorithm

This is a direct search method, which means that it calculates the value of a function and compares it to its values at other points. It uses a simplex which is a shape consisting of $n+1$ vertices in n dimensions, such as a triangle in 2 dimensions. To optimize a function in 2 dimensions, we need three arbitrary points and find the value of our function at those points, say, $f(a)<f(b)<f(c)$ in this step a is the best point as our function takes the minimum value at a and c is the worst point. Then we reflect c through the centroid of a and b. If the new point is better than b, we replace the old c with our new c. If the new c is better than a, then we extend c even further. The algorithm takes different steps such as reflection, extension and contraction until it converges and the optimum point is obtained (Nelder and Mead (1965).

Bound Optimization BY Quadratic Approximation (BOBYQA)

This is a trust region method, where a surrogate model is used. If f is a complicated function, we can use a simpler function like $\tilde{f}$ for optimization that approximates f well in the trust region. The common candidate for $\tilde{f}$ is a quadratic function obtained by Taylor series, i.e.

$$\tilde{f}(x_k+s) = f(x_k) + \nabla f_k^T s + \frac{1}{2} s^T H_k s,$$

where x_k is the point at the kth iteration, ∇f is the gradient and H is the Hessian. This algorithm, does not require the exact calculation of the gradient and Hessian matrix,

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

instead it uses interpolation equation $\tilde{f}_k(x_i) = f(x_i)$ for $i = 1, 2, \ldots, m$ to estimate the gradient and the Hessian matrix, where m is chosen from the interval $[n + 2, (n + 1)(n + 2)/2]$. We then minimize $\tilde{f}$ subject to a trust region with radius δ at each iteration

$$\|s\|_2 \leq \delta_k,$$

To determine the size of the δ at each iteration we define the ratio

$$r = \frac{f(x_k) - f(x^*)}{\tilde{f}(x_k) - \tilde{f}(x^*)},$$

where x^* is the solution to the subproblem in each iteration. If r is greater than a threshold, say, 1 then, the trust region is increased, i.e. δ increases as the algorithm is moving in the right direction towards the optimum point, otherwise, the trust region is decreased. The algorithm terminates when the radius of the trust region is less than a given tolerance level [Powell (2009); Rios and Sahinidis (2013)].

L-BFGS-B

This method is based on quasi-Newton optimization method. The idea is that as we can use Newton's method to find the roots of a function, we can also apply this method in order to find the roots of the gradient of a function, i.e.

$$x_{k+1} = x_k - H_k^{-1} \nabla f_k. \tag{15}$$

Since H^{-1} does not always exist and it is computationally expensive, it can be replaced by another function which is an approximation to H. Therefore, Equation (15) in quasi-Newton algorithms can be written as

$$x_{k+1} = x_k - B_k^{-1} \nabla f_k, \tag{16}$$

where B is an approximation to H. The algorithm starts with an arbitrary B, say, I an identity matrix and updates B until convergence. The difference between different classes of quasi-Newton methods such as Broyden-Fletcher-Goldfarb-Shanno (BFGS) and limited memory BFGS (L-BFGS-B) is on how B is determined. For example, L-BFGS-B is more suitable for large scale optimization problems with restricted memory [Liu and Nocedal (1989); Nash and Varadhan (2011)].

Other algorithms in Table A2 include nonlinear minimization with box constraints (nlminb) which is based on Newton optimization methods subject to a box constraint, alternative versions of Nelder-Mead and bobyqa which are provided via "nloptr" wrapper package.

3.4. Imputation algorithms

Let Y be a variable that contains the missing values. We denote the observed values by y_{obs} and the missing values by y_{mis}. Let also $\mathbf{X}_{obs}$ be a matrix of covariates for which y is observed and $\mathbf{X}_{mis}$ be a matrix of covariates for which y is missed. In this study, we use predictive mean matching and logistic regression to impute missing values.

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

Predictive mean matching

This method is based on Bayesian regression and we need to draw samples from the posterior distribution of the parameters. This is an important step in MICE algorithm to ensure uncertainty in the imputations [van Buuren (2018)].

Step 1 Let $\mathbf{y}_{obs} \sim N(\mathbf{X}_{obs}\beta, \sigma^2\mathbf{I})$, where $\mathbf{X}$ is a matrix with n rows and k columns which represent the number of covariates. Here we assume β and σ^2 are both unknown. Suppose the prior distribution of $\beta|\sigma^2$ is given by

$$\beta|\sigma^2 \sim N_k(0, \sigma^2\mathbf{V}_0) \tag{17}$$

and the prior distribution of σ^2 by

$$\begin{aligned}\sigma^2 &\sim \text{Inv} - \text{Gamma}\left(a_0 = \frac{\nu_0}{2}, b_0 = \frac{1}{2}\nu_0 s_0^2\right)\\ &= \text{Scale} - \text{inv}{-}\chi^2(\nu_0, s_0^2) = \nu_0 s_0^2 \chi_{\nu_0}^{-2}.\end{aligned} \tag{18}$$

Then, the posterior distribution has the following form [see, for example, Box and Tiao (1973); Murphy (2012)]

$$\begin{aligned}p(\beta, \sigma^2|\mathbf{X}_{obs}, \mathbf{y}_{obs}) &\propto p(\beta|\mathbf{X}_{obs}, \mathbf{y}_{obs}, \sigma^2)p(\sigma^2|\mathbf{X}_{obs}, \mathbf{y}_{obs})\\ &\propto N(\beta|\beta_n, \sigma^2\mathbf{V}_n)\text{Inv} - \text{Gamma}(a_n, b_n),\end{aligned}$$

where

- $\beta_n = \mathbf{V}_n\mathbf{X}_{obs}^T\mathbf{y}_{obs}$
- $\mathbf{V}_n = (\mathbf{X}_{obs}^T\mathbf{X}_{obs} + \mathbf{V}_0^{-1})^{-1}$
- $a_n = a_0 + n/2$
- $b_n = b_0 + (\mathbf{y}_{obs}^T\mathbf{y}_{obs} - \beta_n^T\mathbf{V}_n^{-1}\beta_n)/2.$

Here, we only need the first two expressions. (See, Appendix A.2 for details).

Step 2 Let [Box and Tiao (1973)]

$$s^2 = (\mathbf{y}_{obs} - \hat{\mathbf{y}})^T(\mathbf{y}_{obs} - \hat{\mathbf{y}})/\nu,$$

where $\hat{\mathbf{y}} = \mathbf{X}_{obs}\hat{\beta}$ is the vector of predicted values of $\mathbf{y}$. Then from (18) we have $\hat{\sigma}^2 \sim \nu s^2\chi_\nu^{-2}$. Draw a random variable $g \sim \chi_\nu^2$ with $\nu = n - k$, where n is the number of rows with observed values and k is the number of covariates for which y is observed. Calculate $\hat{\sigma}^2 = s^2/g$.

Step 3 Draw k random values $Z \sim N(0, 1)$.

Step 4 Calculate $\beta^* = \hat{\beta} + \hat{\sigma}\mathbf{z}_1\mathbf{V}^{1/2}$.

Step 5 Calculate $\mathbf{y}^\star = \mathbf{X}_{mis}\beta^\star$.

Step 6 For each missing value of $y^\star$, find the closest predicted values, i.e. $\hat{y}$.

Step 7 Choose d values of $\hat{y}$ which are close to $y^\star$ and randomly choose one of them. The values of d depend on the sample size. Small d may lead to repetition and large d may increase bias [van Buuren (2018)].

Step 8 Find the corresponding observed value of $\hat{y}$ and set $y_{mis} = y_{obs}$.

Example: Suppose ADL score for individual I is *NA* and the observed values of ADL, y, for individuals II, III, IV and V are 5, 6, 3 and 4, respectively.

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

Using $\hat{\beta}$ we can estimate ADL, $\hat{y}$, for these individuals, say, 7, 5, 4 and 3, respectively. Then suppose using $\beta^{\star}$, we find $y^{\star} = 6$. The 3 closest $\hat{y}$ to 6 are 7, 5 and 4 with the corresponding observed values of $y = 5$, 6 and 3. The algorithm then selects one value from the observed values of y randomly to impute the missing ADL for individual I.

Inverting the matrix in Step 1 may not be numerically stable. MICE applies Cholesky decomposition for matrix inversion [Murphy (2012); van Buuren (2018)].

Logistic regression

In this section, we explain imputation using logistic regression method.

Step 1 Consider the binary model:

$$p(y_{obs} = 1|\mathbf{X}_{obs}, \beta) = \frac{1}{1 + \exp(-\beta^T \mathbf{X}_{obs})}.$$

MICE uses Iterative Reweighted Least Squares (IRLS) to estimate β. Therefore, the negative log-likelihood function is given by

$$NLL := l = -\sum_{i=1}^{n} \left[y_{i,obs} \log p(x_{i,obs}; \beta) + (1 - y_{i,obs}) \log (1 - p(x_{i,obs}; \beta)) \right]. \quad (19)$$

Taking the partial derivative with respect to β yields

$$\mathbf{g} = \frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} [p(x_{i,obs}; \beta) - y_{i,obs}] \mathbf{x}_{i,obs} = \mathbf{X}_{obs}^T (\mathbf{p} - \mathbf{y}_{obs}), \quad (20)$$

and the second partial derivate gives

$$\mathbf{H} = \frac{\partial^2 l}{\partial \beta \partial \beta^T} = \sum_{i=1}^{n} p(x_{i,obs}; \beta)[1 - p(x_{i,obs}; \beta)] \mathbf{x}_{i,obs} \mathbf{x}_{i,obs}^T = \mathbf{X}_{obs}^T \mathbf{W} \mathbf{X}_{obs}, \quad (21)$$

where $\mathbf{W} = \text{diag}(p(x_{i,obs}; \beta)[1 - p(x_{i,obs}; \beta)])$. Beginning with Newton–Raphson algorithm we have

$$\beta_{new} = \beta_{old} - \mathbf{H}^{-1} \mathbf{g}.$$

Substituting (20) and (21) yields

$$\beta_{new} = (\mathbf{X}_{obs}^T \mathbf{W} \mathbf{X}_{obs})^{-1} \mathbf{X}_{obs}^T \mathbf{W} \mathbf{z}, \quad (22)$$

where $\mathbf{z} = \mathbf{X}\boldsymbol{\beta}_{old} + \mathbf{W}^{-1} (\mathbf{y} - \mathbf{p})$ (see Appendix A.3 for details.). Since each iteration solves the weighted least squares problem, i.e. $(\mathbf{z} - \mathbf{X}\beta)^T \mathbf{W} (\mathbf{z} - \mathbf{X}\beta)$, this method is known as IRLS. [See, for example, Hastie *et al.* (2009); Murphy (2012)]

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

Step 2 Calculate the estimated covariance matrix of $\hat{\beta}$, i.e. $\mathbf{V} = \mathbf{H}^{-1} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$ where $\mathbf{W}$ is as defined in (21). Murphy (2012)
Step 3 Draw k random values $Z \sim N(0, 1)$.
Step 4 Calculate $\beta^* = \hat{\beta} + \mathbf{z}_1\mathbf{V}^{1/2}$.
Step 5 Calculate the predicted probability based on missing values, i.e. $\mathbf{p}^\star = 1/[1 + \exp(-\mathbf{X}_{mis}\beta^\star)]$.
Step 6 Draw random variates from the uniform distribution $U(0, 1)$.
Step 7 Set imputations $y_{mis} = 1$ if $u \leq p^\star$, and $y_{mis} = 0$ otherwise.

4. Results

In this section, we discuss the models that we fit to our datasets I, II, III, IV and V. In each dataset, we have 59, 265 observations and 14, 964 unique individuals. We then divide our datasets into 70% training set and 30% validation set (test set). We will use the test set in Section 4.1 to compare the predictive power of our fitted models. In our training set, we have 41, 543 observations and 10, 475 unique individuals. In order to select our variables, we first build a "full model" by incorporating all variables and fitting a GLM with a clog-log link function. The results based on training set I are provided in Table 6. The variables "time", "gender", "age", "employment status", "IADL", "heart attack", "diabetes", "lung disease" and "cancer" are statistically significant at 0.1% and marked with three asterisks, which indicate that these variables have a significant impact on the hazard probability of death. We then use these significant variables from the "full model" and build our reduced GLM. We can see a similar pattern in the "reduced model" with greater log-likelihood at the expense of losing 10 variables. We test the null hypothesis that all these variables are identical. The difference between deviances of the full model and the reduced model is 9.4 which is less than 5% critical value of a χ^2 distribution with 10 degrees of freedom and therefore we do not have sufficient evidence against our null hypothesis. We consider the variables of our reduced model and fit a random effects model (GLMM) to our training set I. As we discussed in Section 3.2 fitting GLMMs involve numerical optimization which is computationally expensive with the possibility of convergence issues. We apply the adaptive Gauss–Hermite quadrature [Kabaila and Ranathunga (2019)] with 9 quadrature points, i.e. nAGQ = 9. The higher the quadrature points, the more exact the approximation is. However, this happens at the cost of speed. Also, according to Tutz and Schmid (2016) for simple models with random intercept, 8 to 15 quadrature points yield stable estimates. Since most of our variables are binary and the scale of age is different from other variables, we have convergence issues which can be resolved by scaling the variable age. This can be done by subtracting the column mean and dividing by the standard deviation. After that, we try several optimizers (see Table A2 in Appendix). As we can see, even after scaling age, we have convergence warnings for most of the optimizers. Bates *et al.* (2015) recommend "bobyqa". We use this optimizer with the maximum number of function evaluations before termination of 100, 000. As we can see both Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) have reduced. The random effect shows the individual-level effects, i.e. how much does an individual differ from the population? This variability is given by a variance of 1.346. The results of GLMM fitted to datasets II, III, IV and V are

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

Table 6. Coefficient and standard error estimates of models

	Full model (cloglog)		Reduced model (cloglog)		Random effects model	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Intercept	−11.31***	0.3746	−11.25***	0.3692	−5.4944***	0.3492
Time 2	−0.1698	0.1063	−0.1649	0.1060	−0.0488	0.1184
Time 3	−0.4357***	0.1153	−0.4141***	0.1139	−0.2683*	0.1313
Time 4	−0.6566***	0.1197	−0.6324***	0.1191	−0.4687***	0.1397
Time 5	−1.3313***	0.1495	−1.2934***	0.1486	−0.1371***	0.1661
Gender(F)	−0.4792***	0.0852	−0.4578***	0.0834	−0.5309***	0.0991
Age	0.0989***	0.0051	0.0981***	0.0050	1.2049***[1]	0.0883
Employment 2	−1.0983	1.0291	−1.0828	1.0290	−1.0803	1.0344
Employment 3	1.5226***	0.3279	1.5800***	0.3262	1.4866***	0.3399
Employment 4	0.3354	0.2773	0.3627	0.2767	0.0892	0.2951
Employment 5	1.0513*	0.4170	1.0525*	0.4169	1.0352*	0.4253
Marital status(S)	−0.1679+	0.0866	−0.1679+	0.0866	−0.1456	0.0985
Mobility	0.0574**	0.0187	0.0585**	0.0183	0.0623**	0.0200
ADL	0.0644+	0.0350	0.0658+	0.0348	0.0880*	0.0389
IADL	0.1210***	0.0301	0.1319***	0.0288	0.1591***	0.0331
High pressure(Y)	0.0722	0.0792				
Angina(Y)	−0.0648	0.1257				
Heart attack(Y)	0.6553***	0.1484	0.6468***	0.1420	0.6959***	0.1581
Heart failure(Y)	0.7199**	0.2245	0.7709***	0.2207	0.9739***	0.2768
Heart murmur(Y)	0.0116	0.1649				

(*Continued*)

Table 6. (*Continued.*)

	Full model (cloglog)		Reduced model (cloglog)		Random effects model	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Heart rhythm(Y)	0.1204	0.1206				
Diabetes(Y)	0.3548***	0.1066	0.3633***	0.1057	0.4037**	0.1228
Stroke(Y)	0.3158*	0.1389	0.3413*	0.1377	0.3811*	0.1540
Other heart D(Y)	0.1048	0.1808				
Lung(Y)	0.5881***	0.1170	0.5819***	0.1144	0.6829***	0.1362
Asthma(Y)	−0.0909	0.1252				
Arthritis(Y)	−0.1970*	0.0832	−0.1866*	0.0830	−0.2298*	0.0941
Osteoporosis(Y)	0.1859	0.1275				
Cancer(Y)	1.1763***	0.1068	1.1917***	0.1062	1.2938***	0.1276
Parkinson's(Y)	0.2525	0.2668				
Psychiatric(Y)	0.2331	0.1480				
Alzheimer(Y)	0.5334*	0.2649	0.5233*	0.2613	0.7538*	0.3179
Dementia(Y)	0.1689	0.1949				
AIC	5,623.5		5,612.8		5,599.8	
BIC	5,908.4		5,811.4		5,807.1	
Log-likelihood	−2, 778.7, df = 33		−2, 783.4, df = 23		−2, 775.9	
Random effect: id number						
Variance					1.346	
Number of obs:					41,543, groups: idauniq, 10,475	

$^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.
[1] Scaled.

provided in Table A1 in Appendix. As we can see the estimated parameters and the information criteria for these datasets are reasonably close.

To interpret the coefficients in a regression model with a clog-log link function, we note that $-\log(1-x) \approx x$. Therefore, from Equation (6), $\lambda(t|x) \approx e^{\eta}$, where η is the linear predictor. Negative coefficients correspond to a negative effect on the hazard probability and positive coefficients correspond to a positive effect on the hazard probability. The variable "time" has a negative coefficient which means that as time passes we expect improvement in hazard probability, i.e. mortality. The coefficient of gender is negative which means that the probability of mortality for females is less than the probability of mortality for males. The coefficient of age is positive, which means that death is more likely as people get older.

Mortality trends

Figure 3 shows the probability of hazard for males (solid lines) and females (dashed lines) who are healthy, retired and in partnership. This figure is based on GLMM and each curve represents the probability of death for the average population at a particular age during 5 waves. We can observe that the probability of death decreases over these 5 waves, which suggests an improvement in mortality. From wave 1 to wave 2 which corresponds to the years 2002–2003, the curves are relatively flat. From wave 2 to 4, we can see a steeper downward slope. This period corresponds to the years 2003–2009. The steepest slope is in wave 5 which is during 2010–2011. This suggests a greater mortality improvement during these years. We can also observe that this improvement is more pronounced among males than females. The mortality improvement happens faster in older ages. This mortality trend is in agreement with

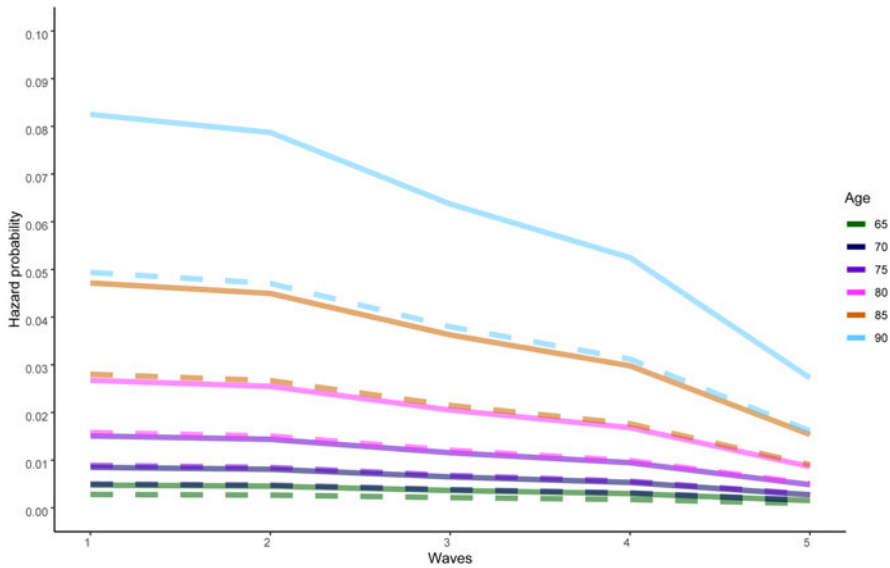

Figure 3. Hazard probability for males (solid lines) and females (dashed lines), retired, in relationship, with no disease (population).

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

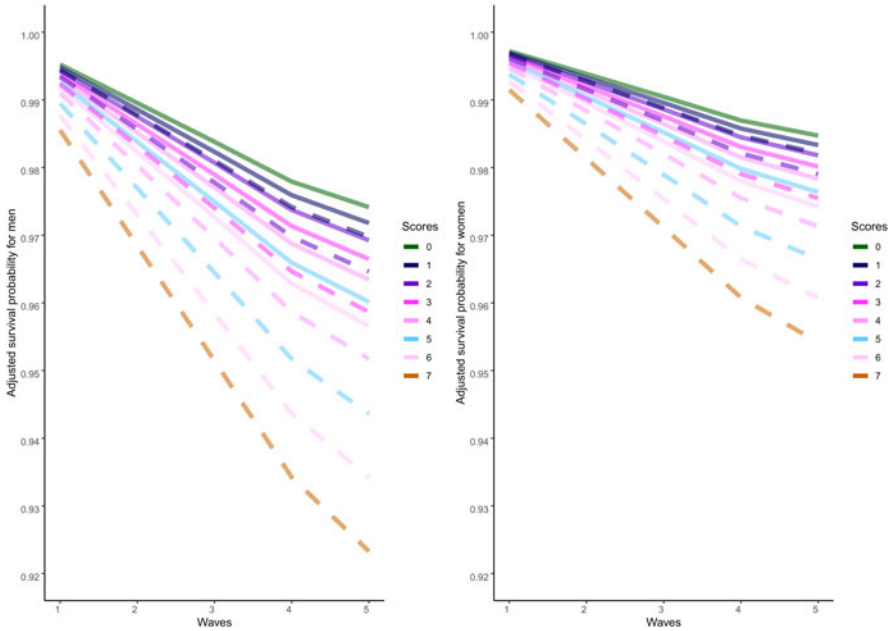

Figure 4. Adjusted survival probability for different ADL scores (solid line) and IADL scores (dashed line) for a male (left) and female (right) aged 65 in wave 1, retired, in relationship, with no disease (population).

Deaths Registered in England and Wales: 2013, Figure 1[2]. We can see similar trends in terms of survival probability in Figures 4, 5 and 6. In these figures, which are also based on GLMM and average population, the green curve represents the survival probability for a healthy individual. Comparing the green curve of the plots in Figure 4, we can see a sharper negative slope in survival probability among males (left plot) than among females (right plot) before wave 4. After wave 4, the slope is less steep on the left plot than on the right plot compared with before wave 4. In other words, between waves 4 and 5, the improvement in mortality among males is more than the improvement in mortality among females.

Socio-economic factors

Now, we consider the coefficient of "employment", where the reference category is "employed". Employment 3, 4 and 5 correspond to "permanently sick or disabled", "retired" and "self-employed", respectively. The positive coefficients, although not always statistically significant, suggest an increase in the risk of death for a retired, disabled and self-employed compared with an employed individual. Similar results have been found by Sorlie and Rogot (1990) for the US population. They find that the mortality rate is particularly higher for those who are unable to work. In our study, we can see that the coefficient of employment 3 (permanently sick or

[2]https://webarchive.nationalarchives.gov.uk/ukgwa/20160105181301/http://www.ons.gov.uk/ons/rel/vsob1/mortality-statistics–deaths-registered-in-england-and-wales–series-dr-/2013/stb-deaths-registered-in-england-and-wales-in-2013-by-cause.html

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

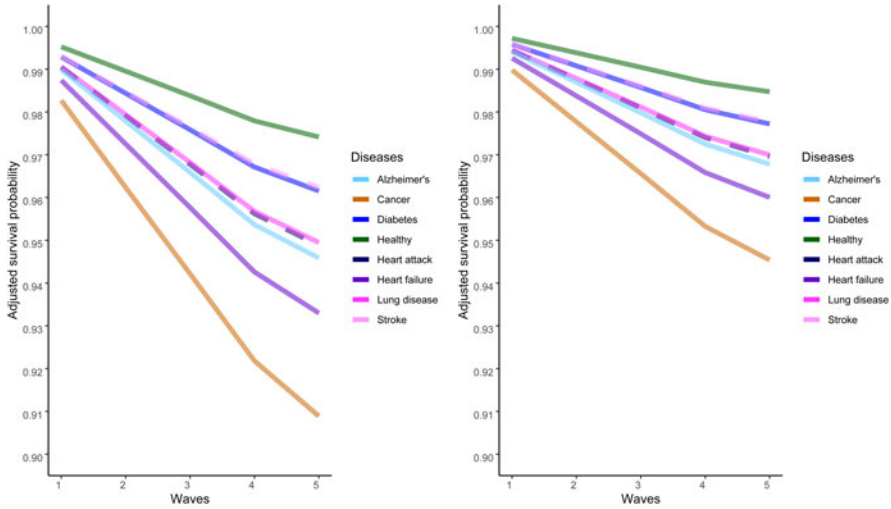

Figure 5. Adjusted survival probability for different diseases for a male (left) and female (right) aged 65 in wave 1, retired, in relationship (population).

disabled) is higher than the coefficient of other categories. In another study, Morris *et al.* (1994) find that the risk of death for British men who are unemployed is much higher than for those who are employed. They even report a higher risk of mortality

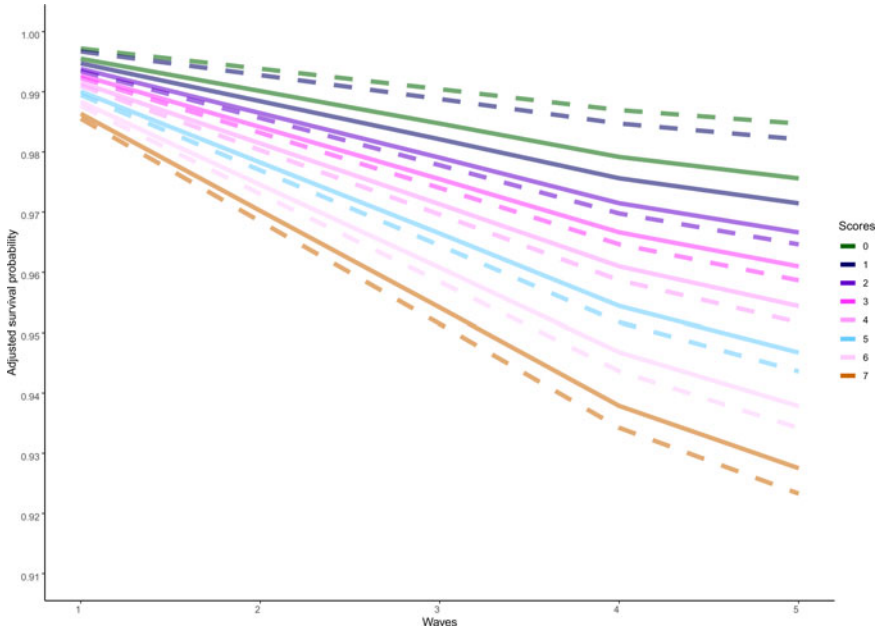

Figure 6. Adjusted survival probability for different IADL scores for a male, aged 65 in wave 1, retired, in relationship, individual (solid line) and average population (dashed line).

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

among people who retired early for reasons other than illness. The results for the relationship between mortality and self-employment are mixed. Our result shows that self-employed people have a higher probability of death. Toivanen *et al.* (2016) find that mortality from CVDs and suicide is lower among self-employed in Sweden, whereas Gauffin and Dunlavy (2021) find that poor health conditions are more common among self-employed than employed people in Sweden. Gonçalves and Martins (2018) report a low mortality rate among self-employed in Portugal. We categorize "marital status" into two groups: singles which include divorced, widowed and not married categories and couples which include any form of partnership. The coefficient of marital status is negative, i.e. the risk of death for a single individual is less than the risk of death for a couple. Ebrahim *et al.* (1995) look at the relationship between marital status and mortality among British men and find that although the risk of mortality is higher among single men, divorced men are not at increased risk of mortality. Johnson *et al.* (2000) show that the risk of mortality among single people aged $45-64$ is higher than older American people. Rendall *et al.* (2011) find that the survival probability among US married men is higher than the survival probability among married women, but they did not find mortality differences among never-married, divorced and widowed categories. In another study, Lindström and Rosvall (2019) also point to mortality differences among married men and women. They show that the risk of mortality is higher among unmarried, divorced and widowed men, but there are no significant differences in mortality among women with different marital status in Sweden. Ramezankhani *et al.* (2019) study all-cause and cause-specific mortality for different marital status among Iranian population and find that marital benefit is different among men and women. Their results show that the risk of mortality among never-married men is higher than among never-married women.

Disability factors

We then consider the coefficients of factors related to disability: mobility, ADLs and IADLs. All these disability factors are significant and have a positive impact on mortality. This is in agreement with other studies that see disability factors as predictors of mortality. For example, Scott *et al.* (1997) show that there is a direct relationship between mortality and disability factors among American population. Gobbens and van der Ploeng (2020) find similar results among Dutch population and Yang *et al.* (2021) among Chinese population. In Figure 4 we compare survival probability for different levels of ADLs and IADLs between females and males by controlling age, marital and employment status factors. We can observe that the slope of survival curves is steeper for males (left plot) than for females (right plot), which indicates survival differences due to disability factors between males and females. This is in agreement with Pongiglione *et al.* (2016) where they find a strong relationship between mortality and disability factors in ELSA participants and survival inequality among men and women. However, their findings are based on binary classifications of disability and they do not consider different levels of severity of disability. In another study, they point out that ordinal classifications of disability are more informative than binary classifications of disability [Pongiglione *et al.* (2017a)]. In Figure 4 we can also observe that IADLs contribute more to a decrease in survival probability than ADLs. Further, as IADL score increases, the survival probability falls faster. However, we can see that slope of the curves is less steep in the last wave which indicates an improvement in mortality due to disability during that

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

period. Pongiglione *et al.* (2017b) also look at disability trends between 2002–2012 by classifying disability into no disability, mild, moderate and severe groups and conclude that severe and moderate disability has improved among women, but moderate disability has increased and severe disability has been stable among men.

Disease trends

We expect all diseases to have a positive impact on the risk of death. However, in our full model, the coefficients of "angina" and "asthma" are negative, although not statistically significant and therefore they are not included in the reduced model and GLMM. The coefficient of "arthritis" is negative and significant at 5% which does not meet our expectations. However, this does not necessarily mean that arthritis has a negative effect on mortality. Table 7, shows the observed number and proportion of deaths and survivors for different diseases. For example, there are 288 cases that arthritis has been observed together with death and 14, 025 cases that arthritis has been reported without the occurrence of death. The proportion of reported arthritis without the occurrence of death is about 34% which is considerably higher than other diseases. Mendy *et al.* (2018) show that there is no relationship between self-reported arthritis and the risk of death. However, they find that knee arthritis is associated with a higher risk of CVDs and diabetes and therefore death. Similar results have been found by Turkiewicz *et al.* (2019) that arthritis does not lead to increased mortality, but knee and hip arthritis are associated with a higher risk of CVDs as a result of the inability to walk and/or be active. Figure 5 shows the adjusted survival probability for a male (left plot) and female (right plot) who is retired, in partnership, aged 65 in wave 1 and only suffers from one disease. The green curves show the survival probability for a healthy individual and the brown one shows the survival probability for an individual who only suffers from cancer. We can observe that the survival probability for some diseases is very close. In that case, one of the causes of death is shown by a dashed line. According to this figure, cancer contributes to a decrease in survival probability much more than other diseases. The second largest contribution to death is heart failure and the third one is Alzheimer's. This pattern is in line with UK government statistics for the leading causes of death amongst people aged 75 in 2007[3], where the largest proportion of death was reported to be due to cancer, chronic heart disease and dementia. In our full GLM, although the coefficient of dementia was positive, it was not statistically significant and therefore dementia was removed in our reduced GLM and GLMM. However, as we can see it is one of the causes of death that contributes to mortality among the old population even more than stroke, heart attack, lung disease and diabetes. In this figure, we can also observe that the survival probability for an individual with diabetes, illustrated by blue solid lines, and the survival probability for an individual who had a stroke attack, illustrated by pink dashed lines, is almost the same. In fact, the probability of death among diabetics is about 49.7% higher than healthy people and the probability of death among those who had a stroke attack is about 46% higher than those who did not have a stroke attack. Perhaps the reason that the probability of death for these two causes is so close is that according to research diabetics are exposed to a higher risk of death due to stroke. [See, for example, Hewitt *et al.* (2012); Lau *et al.* (2019), and the references therein]. Similarly,

[3]https://www.gov.uk/government/publications/death-in-people-aged-75-years-and-older-in-england-in-2017/death-in-people-aged-75-years-and-older-in-england-in-2017

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

Table 7. The reported number and proportion of deaths for different diseases in training set. Disease (1), no diseases (0)

	Heart attack		Heart failure		Diabetes		Stroke		Lung disease	
	0	1	0	1	0	1	0	1	0	1
Number of death	633	63	672	24	585	111	672	69	600	96
Proportion	(0.015)	(0.002)	(0.016)	(0.001)	(0.014)	(0.003)	(0.015)	(0.002)	(0.014)	(0.002)
Number of survivors	39, 983	864	40, 673	174	37, 218	3, 629	40, 004	843	38, 891	1, 956
Proportion	(0.962)	(0.021)	(0.979)	(0.004)	(0.896)	(0.087)	(0.963)	(0.020)	(0.936)	(0.047)
	Arthritis		Cancer		Alzheimer					
	0	1	0	1	0	1				
Number of death	408	288	586	110	677	19				
Proportion	(0.010)	(0.007)	(0.014)	(0.003)	(0.016)	(0.000)				
Number of survivors	26, 822	14, 025	39, 182	1, 665	40, 725	122				
Proportion	(0.646)	**(0.338)**	(0.943)	(0.040)	(0.980)	(0.003)				

we can observe that the survival probability for an individual with lung disease, illustrated by magenta solid lines, is very close to the survival probability for an individual who had a heart attack, illustrated by dark blue dashed lines. In fact, controlling for other factors, the hazard ratio for an individual with lung disease is 1.98 and for an individual with a heart attack is 2. Research shows that there is a relationship between lung diseases and CVDs and lung diseases can lead to a higher risk of mortality due to CVDs. [See, for example, Sin and Man (2005); Carter *et al.* (2019)]. Further, we can observe an improvement in mortality due to different causes of death in the last wave. This figure shows that cancer, followed by heart failure, is the leading cause of death which is in agreement with Deaths Registered in England and Wales: 2012 (see, Section 1)[4]. The difference in survival probability among males and females can also be detected by looking at the slope of the curves.

Population and individual effects

One of the advantages of GLMM is the ability of the model to consider observations belonging to nested or hierarchical subgroups within the population. In our study, the subgroups are individuals with unique id numbers who are repeatedly surveyed over the period 2002–2012. Therefore, when we investigate the impact of one factor on the survival probability of the population, we can also look at the variability of the impact of that factor for a unique individual over the same period. In other words, we consider the effects of factors both on population and individual levels. In this study, the repeated measures are therefore treated as random effects as there are some underlying factors which are unique to each individual that have not been considered by the model. Figure 6 compares the survival probability of a unique individual (solid lines) with a unique id number for different levels of disability due to IADL difficulties with an average population (dashed lines). In this figure, we controlled age, sex, employment and marital status. We can observe that the adjusted survival probability of this particular individual in healthy status is less than the adjusted survival probability of a healthy individual on average. The reason is that this particular individual may be subject to other factors such as financial problems or family health history which have not been considered by this model. On the other hand, we can see that all solid lines for different IADL scores are above the dashed lines, which indicates that the adjusted survival probability for this individual with different levels of disability is slightly higher than the adjusted survival probability for the average population with the same level of disability.

4.1. Discrimination measures

In this section, we look at the prediction accuracy of our 3 models: GLM, reduced GLM and GLMM based on the test set. For this, we use time-dependent ROC curves which are more suitable for follow-up studies than standard ROC curves. First, we explain standard ROC curves and then the time-dependent ROC curves.

Standard ROC curve

In classification problems, where we assign our observations to different classes such as death $y = 1$ and survival $y = 0$, we consider two types of error:

[4]https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsregistrationsummarytables/2013-07-10

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

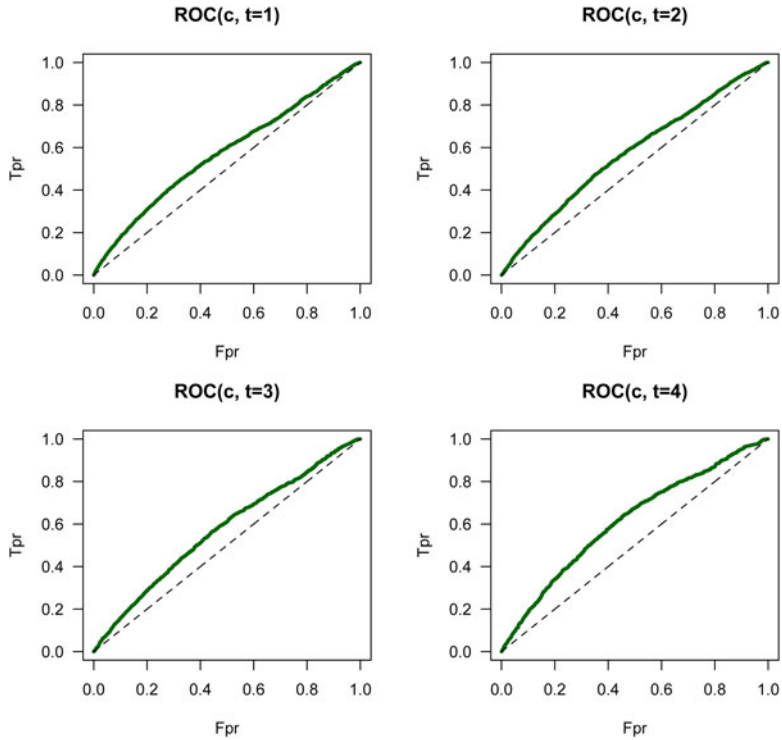

Figure 7. ROC curve for GLM based on dataset I: AUC(t = 1)=0.572, AUC(t = 2) = 0.574, AUC(t = 3) = 0.574, AUC(t = 4) = 0.614.

- False positive (Type I error): we predict that an individual will experience an event, say, death, but they did not. In other words, we estimate $\hat{y} = 1$, but the truth is $y = 0$.
- False negative (Type II error): we predict that an individual will be event-free and will survive, but they died. In other words, we estimate $\hat{y} = 0$, but the truth is $y = 1$.

We can illustrate these two types of error in a matrix which is known as the confusion matrix [see, Table 8]:

The in-diagonal values are related to predictions that are correct and the off-diagonal values are related to incorrect predictions. Using the information from a confusion matrix, we can define the following measures:

- True positive rate: Tpr (sensitivity) = $TP/(TP + FN)$
- False positive rate: Fpr (type I error) = $FP/(FP + TN)$
- False negative rate: Fnr (type II error) = $FN/(FN + TP)$
- True negative rate: Tnr (specificity) = $TN/(TN + FP)$

The ROC curve plots Tpr against Fpr [Murphy (2012)]. The more this curve is away from the 45° line, the better is our model at discriminating the positive ($y = 1$) and negative ($y = 0$) events. Another way to measure the predictability of the model is by

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

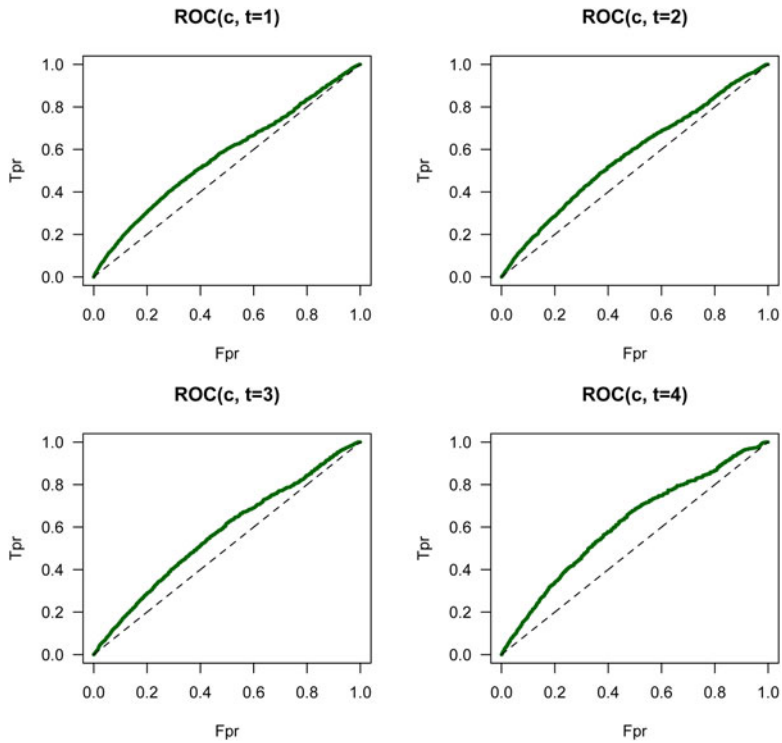

Figure 8. ROC curve for reduced GLM based on dataset I: AUC(t=1)=0.569, AUC(t = 2) =0.571, AUC(t = 3) = 0.573, AUC(t = 4) = 0.614.

looking at the area under the ROC curve, i.e. AUC. The larger AUC, the better the models at distinguishing between death and survival.

Time-dependent ROC curve

Let T_i be the time of death and η be the linear predictor, which represents risk score and is defined in Equations (4) and (13) for GLM and GLMM, respectively. Further, let y be the death status as defined by (2). The idea is that at each time t, each individual is classified as a case, i.e. the individual dies or control, i.e. the individual survives and we would like to see what percentage of cases and controls can be discriminated by our models for different values of thresholds c. In this case, each individual who had the role of control at the earlier time, may contribute as a case at a later time [see, for example, Tutz and Schmid (2016); Kamarudin *et al.* (2017)]. To plot the ROC curve, we need sensitivity and specificity rates. The time-dependent sensitivity is defined as

$$\text{sensitivity}(c, t) := \Pr(\eta > c | T = t, y = 1)$$

which is the probability that an individual who has died ($T = t$) is predicted to have the hazard probability η of greater than c, indicating that cases are correctly identified (TP).

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

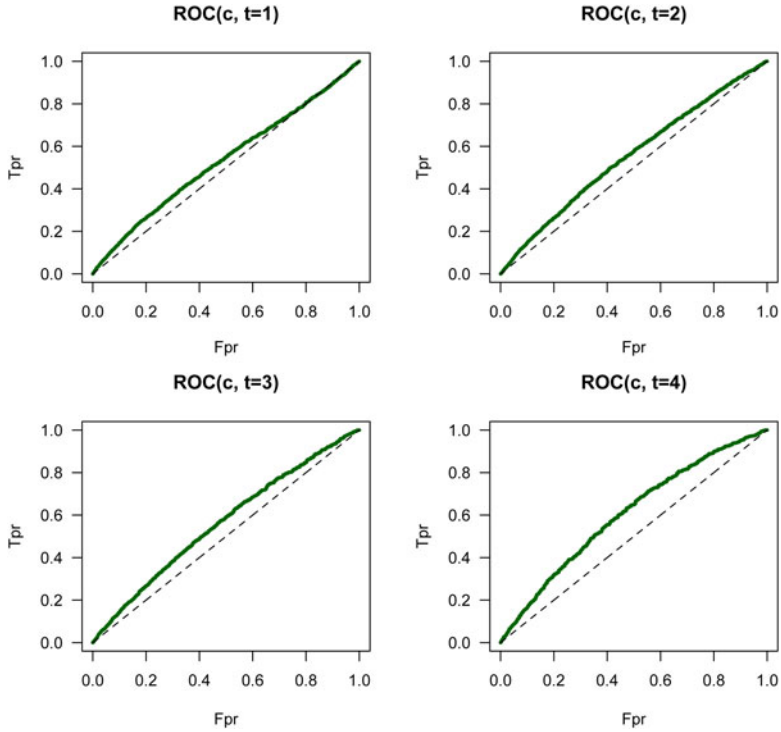

Figure 9. ROC curve for random effects model for the population average based on dataset I: AUC(t = 1) = 0.534, AUC(t = 2) = 0.554, AUC(t = 3) = 0.562, AUC(t = 4) = 0.605.

The time-dependent specificity is also defined by

$$\text{specificity}(c, t) := \Pr(\eta \leq c | T > t, y = 0),$$

which is the probability that an individual who has survived ($T > t$) is predicted to have the hazard probability η of less than or equal to c, indicating that controls are correctly identified (TN). These probabilities change as the value of threshold c changes. In a follow-up study, the status of the individuals changes over time and some observations may be censored before time t. Therefore to calculate sensitivity and specificity we need to apply the inverse probability of censoring weight (IPCW) in order to maintain consistency and to avoid bias. Let $G(.)$ be the survival function of the censoring times U as defined in Section 3. Then we define

$$G(t) = \Pr(U > t | \mathbf{x}). \tag{23}$$

If G is large, then we can conclude that with high probability censoring occurs after time t, whereas if G is small, then censoring occurs before time t and hence the number of observations fully observed up to time t is only a small part of the whole sample. To allow for this under-represented group, we use the inverse probability of censoring as the weight [Tutz and Schmid (2016)]. Therefore,

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

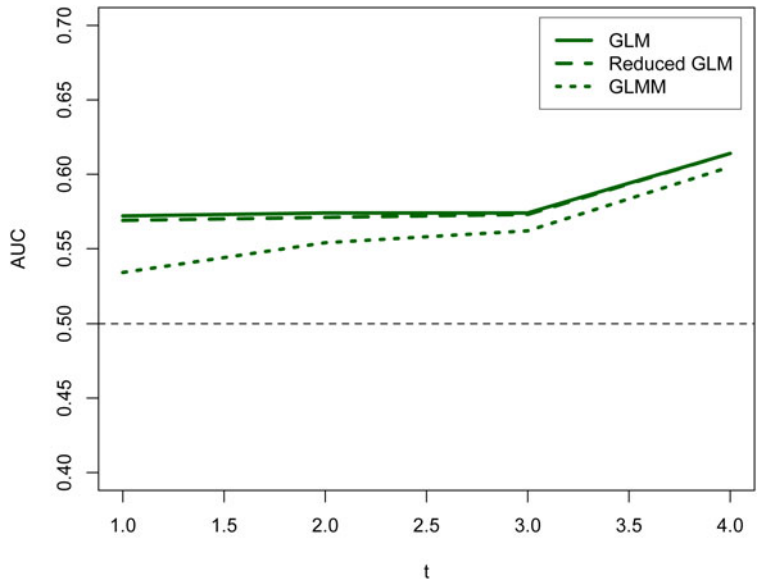

Figure 10. AUC for GLM, reduced GLM and GLMM at t = 1, 2, 3 and 4.

Table 8. Confusion matrix

	$y = 1$	$y = 0$
$\hat{y} = 1$	True positive (*TP*)	False positive (*FP*)
$\hat{y} = 0$	False negative (*FN*)	True negative (*TN*)

sensitivity is given by

$$\text{sensitivity}(c, t) = \frac{\sum_i \delta_i I\left(\eta_i > c \cap t_i = t\right)/G_i(t_i - 1)}{\sum_i \delta_i I(t_i = t)/G_i(t_i - 1)},$$

where δ is defined in Equation (1), I is the indicator function and G can be obtained from the training data similar to $S(t)$. The only difference is that this time, the event of interest is the censoring time rather than the time of death. Further, specificity is given by

$$\text{specificty}(c, t) = \frac{\sum_i I\left(\eta_i \leq c \cap t_i > t\right)}{\sum_i I(t_i > t)}.$$

We can plot the ROC curve for different points in time and compare the predictability of the models at different times. Similarly, we can define the

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

time-dependent AUC by

$$\mathrm{AUC}(t) = \Pr\Big(\{\eta_i > \eta_j\} | \{T_i = t\} \cap \{T_j > t\}\Big),$$

where η_i, η_j, T_i and T_j are the predictors and survival times of independent observations i and j [Tutz and Schmid (2016)]. We can use package "discSurv" in R to plot Figures 7, 8 and 9 which show the ROC curves for GLM, reduced GLM and GLMM, respectively. From these figures, we can observe that at t = 1, 2 and 3, the predictive power of GLM is better than the other 2 models and at t = 4, the reduced model performs as good as the GLM. Further, we can see that the GLMM has the worst predictive power. In fact, we know that we can use GLMM to predict the survival probability of the existing individuals, but we cannot use this to estimate the survival probability of a new individual and in the case of a new individual, it only gives the estimated survival probability of the average population. In other words, for a new individual, it only takes into account population level and not individual levels. Figure 10 compares the AUC for all 3 models at t = 1, 2, 3 and 4. We can observe that at t = 1, 2 and 3, the AUC is only slightly above 0.5. However, as t increases, the predictive power of all models represented by AUC improves. From these figures, we can conclude that GLM and reduced GLM can discriminate death and survival better than GLMM and are better at generalization than GLMM.

5. Conclusion

In this study, we applied survival analysis to the English Longitudinal Study of Ageing (ELSA). We looked at the impact of demographic and self-reported health conditions on the survival of the participants. We found that the survival probability for individuals who have difficulty in performing IADLs is less than the survival probability for individuals who have difficulty in performing ADLs. We also found that the survival probability for individuals with Alzheimer's disease is less than the survival probability for individuals with diabetes. Further, cancer was the deadliest of the disease that we considered in this study. We showed that a random effects model can distinguish between individual-level and population-level hazard probability. One of the problems with survey data is missing values and, in particular, temporary withdrawals of the participants. To address this issue we applied Last Observation Carried Forward (LOCF) method, which is a single imputation method, to fill in the missing values related to "employment status" and "marital status". For the rest of our covariates, we used MICE, which is a multiple imputation method. We produced 5 datasets and applied our analysis to each dataset. We found that the results under all these 5 datasets are very close and therefore, we performed our analysis based on only one dataset. The results of this study about the survival probability and factors that affect the survival probability can be used in areas such as life insurance and health insurance.

Acknowledgements. I wish to thank the referees and the editors for valuable comments and suggestions that led to a substantial improvement of this article.

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

References

Abramowitz, M. and I. A. Stegun (1964) *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. (Vol. 55). US Government printing office.

Aida, J., N. Cable, P. Zaninotto, T. Tsuboya, G. Tsakos, Y. Matsuyama, K. Ito, K. Osaka, K. Kondo, M. G. Marmot and R. G. Watt (2018) Social and behavioral determinants of the difference in survival among older adults in Japan and England. *Gerontology* 64(3), 266–277.

Allison, P. D. (1982) Discrete-time methods for the analysis of event histories. *Sociological Methodology* 1361–98.

Antonio, K. and E. A. Valdez (2012) Statistical aspects of a priori and a posteriori risk classification in insurance. *Advances in Statistical Analysis* 96(2), 187–224.

Antonio, K. and Y. Zhang (2014) *Predictive Modelling in Actuarial Science*. New York: Cambridge University Press.

Azur, M. J., E. A. Stuart, C. Frangakis and P. J. Leaf (2011) Multiple imputation by chained equations: what is it and how does it work?. *International Journal of Methods in Psychiatric Research* 20(1), 40–49.

Bates, D., M. Mächler, B. M. Bolker and S. C. Walker (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1–48.

Blake, M., S. Bridges, D. Hussey and D. Mandalia (2015) *The Dynamics of Ageing: The 2010 English Longitudinal Study of Ageing (wave 5)*. London: NatCen Social Research.

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens and J. S. S. White (2009) Generalised linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* 24(3), 127–135.

Box, G. E. P. and G. C. Tiao (1973) *Bayesian Inference In Statistical Analysis*. Philippines: Addison–Wesley publishing company.

Breslow, N. E. and D. G. Clayton (1993) Approximate inference in generalised linear mixed models. *Journal of the American Statistical Association* 88(421), 9–25.

Bridges, S., D. Hussey and M. Blake (2015) The dynamics of ageing: The 2012 English Longitudinal Study of Ageing (wave 6). *London: NatCen Social Research*.

Carter, P., J. Lagan, C. Fortune, D. L. Bhatt, J. Vestbo, R. Niven, N. Chaudhuri, E. B. Schelbert, R. Potluri and C. A. Miller (2019) Association of cardiovascular disease with respiratory disease. *Journal of the American College of Cardiology* 73(17), 2166–2177.

Cox, D. R (1972) Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 34(2), 187–202.

Davidian, M. and D. M. Giltinan (2003) Nonlinear models for repeated measurement data: An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics* 8(4), 387–419.

Davies, K, A. Maharani, T. Chandola, C. Todd and N. Pendleton (2021) The longitudinal relationship between loneliness, social isolation, and frailty in older adults in England: a prospective analysis. *The Lancet Healthy Longevity* 2(2), e70–e77.

Demakakos, P., J. P. Biddulph, M. Bobak and M. G. Marmot (2016) Wealth and mortality at older ages: a prospective cohort study. *Journal of Epidemiology and Community Health* 70(4), 346–353.

Demakakos, P., J. P. Biddulph, C. de Oliveira, G. Tsakos and M. G. Marmot (2018) Subjective social status and mortality: the English Longitudinal Study of Ageing. *European Journal of Epidemiology* 33(8), 729–739.

Donati, L., D. Fongo, L. Cattelani and F. Chesani (2019) Prediction of decline in activities of daily living through artificial neural networks and domain adaptation. In *International Conference of the Italian Association for Artificial Intelligence*, pp. 376–391. Springer, Cham.

d'Orsi, E., A. J. Xavier, A. Steptoe, C. de Oliveira, L. R. Ramos, M. Orrell, P. Demakakos and M. G. Marmot (2014) Socio-economic and lifestyle factors related to instrumental activity of daily living dynamics: results from the English Longitudinal Study of Ageing. *Journal of the American Geriatrics Society* 62 (9), 1630–1639.

Ebrahim, S., G. Wannamethee, A. McCallum, M. Walker and A. G. Shaper (1995) Marital status, change in marital status, and mortality in middle-aged British men. *American Journal of Epidemiology* 142(8), 834–842.

Fahrmeir, L. and L. Knorr-Held (1997) Discrete-time duration models: estimation via Markov Chain Monte Carlo. *Sociological Methodology* 27(1), 417–452.

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

Frees, E. W. (2004) *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. London: Cambridge University Press.
Friedman, M (1982) Piecewise exponential models for survival data with covariates. *The Annals of Statistics* 10(1), 101–113.
Gauffin, K. and A. Dunlavy (2021) Health inequalities in the diverse world of self-employment: A Swedish national cohort study. *International Journal of Environmental Research and Public Health* 18(23), 12301.
Gobbens, R. J. J. and T. van der Ploeng (2020) The prediction of mortality by disability among Dutch community-dwelling older people. *Clinical Interventions in Ageing* 151897–1906.
Gonçalves, J. and P. S. Martins (2018) The effect of self-employment on health: evidence from longitudinal social security data. IZA Discussion Papers Series, 11305. RePEc:iza:izadps:dp11305.
Guzman-Castillo, M., S. Ahmadi-Abhari, P. Bandosz, S. Capewell, A. Steptoe, A. Singh-Manoux, M. Kivimaki, M. J. Shipley, E. J. Brunner and M. O'Flaherty (2017) Forecasted trends in disability and life expectancy in England and Wales up to 2025: a modelling study. *The Lancet Public Health* 2 (7), e307–e313.
Ham, J. C. and S. A. Rea Jr (1987) Unemployment insurance and male unemployment duration in Canada. *Journal of Labor Economics* 5(3), 325–353.
Hanewald, K., H. Li and A. W. Shao (2019) Modelling multi-state health transitions in China: a generalised linear model with time trends. *Annals of Actuarial Science* 13(1), 145–165.
Hastie, T., R. Tibshirani and J. H. Friedman (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.
Heagerty, P. J. and Y. Zheng (2005) Survival model predictive accuracy and ROC curves. *Biometrics* 61(1), 92–105.
Hewitt, J., L. C. Guerra, M. del Carmen Fernändez-Moreno and C. Sierra (2012) Diabetes and stroke prevention: A review. *Stroke Research and Treatment* 20121–6.
Johnson, N. J., E. Backlund, P. D. Sorlie and C. A. Loveless (2000) Marital status and mortality: the national longitudinal mortality study. *Annals of Epidemiology* 10(4), 224–238.
Kabaila, P. and N. Ranathunga (2019) On adaptive Gauss–Hermite quadrature for estimation in GLMM's. In *Research School on Statistics and Data Science*, pp. 130–139. Springer, Singapore.
Kamarudin, A.N., T. Cox and R. Kolamunnage-Dona (2017) Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Medical Research Methodology* 17(1), 1–19.
Kessler, M., E. Thumë, S. Scholes, M. Marmot, L. A. Facchini, B. P. Nunes, K. P. Machado, M. U. Soares and C. de Oliveira (2020) Modifiable risk factors for 9-year mortality in older English and Brazilian adults: The ELSA and SIGa-Bagë ageing cohorts. *Scientific Reports* 10(1), 1–13.
Khondoker, M., S. B. Rafnsson, S. Morris, M. Orrel and A. Steptoe (2017) Positive and negative experiences of social support and risk of dementia in later life: An investigation using the English Longitudinal Study of Ageing. *Journal of Alzheimer's Disease* 58, 99–108.
Lau, L.-H., J. Lew, K. Borschmann, V. Thijs and E. I. Ekinci (2019) Prevalence of diabetes and its effects on stroke outcomes: A meta-analysis and literature review. *Journal of Diabetes Investigation* 10(3), 780–792.
Lee, K. J. and J. B. Carlin (2010) Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology* 171(5), 624–632.
Li, Z., A. W. Shao and M. Sherris (2017) The impact of systematic trend and uncertainty on mortality and disability in a multi-state latent factor model for transition rates. *North American Actuarial Journal* 21(4), 594–610.
Lindström, M. and M. Rosvall (2019) Marital status and 5-year mortality: A population-based prospective cohort study. *Public Health* 17045–48.
Liu, D. C. and J. Nocedal (1989) On the limited memory BFGS method for large scale optimisation. *Mathematical Programming* 45(1), 503–528.
McCullagh, P. and J. A. Nelder (1989) *Generalized Linear Models*, 2nd Ed., Chapman and Hall.
Mendy, A., J. Park and E. R. Vieira (2018) Osteoarthritis and risk of mortality in the USA: a population-based cohort study. *International Journal of Epidemiology* 47(6), 1821–1829.
Morris, J. K., D. G. Cook and A. G. Shaper (1994) Loss of employment and mortality. *The BMJ* 308(6937), 1135–1139.
Murphy, K. P. (2012) *Machine Learning: A Probabilistic Perspective*. England: The MIT Press.
Nash, J. C. and R. Varadhan (2011) Unifying optimisation algorithms to aid software system users: optimx for R. *Journal of Statistical Software* 43(9), 1–14.

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

Nelder, J. A. and R. Mead (1965) A simplex method for function minimisation. *The Computer Journal* 7(4), 308–313.

Petersen, T (1986) Fitting parametric survival models with time-dependent covariates. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 35(3), 281–288.

Pongiglione, B., B. L. De Stavola, H. Kuper and G. B. Ploubidis (2016) Disability and all-cause mortality in the older population: evidence from the English Longitudinal Study of Ageing. *European Journal of Epidemiology* 31(8), 735–746.

Pongiglione, B., G. B. Ploubidis and B. L. De Stavola (2017a) Levels of disability in the older population of England: Comparing binary and ordinal classifications. *Disability and Health Journal* 10(4), 509–517.

Pongiglione, B., G. Ploubidis and B. De Stavola (2017b) Disability-free life expectancy between 2002 and 2012 in England: trends differ across genders and levels of disability. *International Population Conference. IUSSP*.

Potente, C. and C. Monden (2016) Pathways to death by socio-economic status. *2016 Annual Meeting*. PAA.

Powell, M. J. D. (2009) The BOBYQA algorithm for bound constrained optimisation without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge, 26*.

Rafnsson, S.B., M. Orrell, E. d'Oris, E. Hogervorst and A. Steptoe (2020) Loneliness, social integration, and incident dementia over 6 years: Prospective findings from the English Longitudinal Study of Ageing. *Journals of Gerontology: Series B* 75(1), 114–124.

Ramezankhani, A., F. Azizi and F. Hadaegh (2019) Associations of marital status with diabetes, hypertension, cardiovascular disease and all-cause mortality: A long term follow-up study. *PloS one* 14(4), e0215593.

Rendall, M. S., M. M. Weden, M. M. Faveault and H. Waldron (2011) The protective effect of marriage for survival: a review and update. *Demography* 48(2), 481–506.

Renshaw, A. E. and S. Haberman (2000) Modelling the recent time trend in UK permanent health insurance recovery, mortality and claim inception transition intensities. *Insurance: Mathematics and Economics* 27(3), 365–396.

Richayzen, B. D. and D. E. P. Walsh (2002) A multi-state model of disability for the United Kingdom: implications for future need for long-term care for the elderly. *British Actuarial Journal* 8(2), 341–393.

Rios, L. M. and N. V. Sahinidis (2013) Derivative-free optimisation: a review of algorithms and comparison of software implementations. *Journal of Global Optimisation* 56(3), 1247–1293.

Scheike, T. H. and T. K. Jensen (1997) A discrete survival model with random effects: an application to time to pregnancy. *Biometrics* 1997318–329.

Scott, W. K., C. A. Macera, C. B. Cornman and P. A. Sharpe (1997) Fuctional health status as a predictor of mortality in men and women over 65. *Epidemiology* 50(3), 291–296.

Sin, D. D. and S. P. Man (2005) Chronic obstructive pulmonary disease as a risk factor for cardiovascular morbidity and mortality. *Proceedings of the American Thoracic Society* 2(1), 8–11.

Singer, J. D. and J. B. Willet (1993) It's about time: using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics* 18(2), 155–195.

Sorlie, P. D. and E. Rogot (1990) Mortality by employment status in the national longitudinal mortality study. *American Journal of Epidemiology* 132(5), 983–992.

Stamate, D., H. Musto, O. Ajnakina and D. Stahl (2022) Predicting risk of dementia with survival machine learning and statistical methods: results on the English longitudinal study of ageing cohort. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 436–447. Cham: Springer.

Steptoe, A., E. Breeze, J. Banks and J. Nazroo (2013) Cohort profile: the English longitudinal study of ageing. *International Journal of Epidemiology* 42(6), 1640–1648.

Steptoe, A., A. Deaton and A. A. Stone (2015) Psychological wellbeing, health and ageing. *Lancet* 385 (9968), 640.

Steptoe, A. and P. Zaninotto (2020) Lower socio-economic status and the acceleration of aging: An outcome-wide analysis. *Proceedings of the National Academy of Sciences* 117(26), 14911–14917.

Sullivan, D. F. (1971) A single index of mortality and morbidity. *HSMHA health reports* 86(4), 347.

Thompson Jr, W. A. (1977) On the treatment of grouped observations in life studies. *Biometrics* 33(3), 463–470.

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

Toivanen, S., R. H. Griep, C. Mellner, S. Vinberg and S. Eloranta (2016) Mortality differences between self-employed and paid employees: a 5-year follow-up study of the working population in Sweden. *Occupational and Environmental Medicine* 73(9), 627–636.

Torres, J. L., M. F. Lima-Costa, M. Marmot and C. de Oliveira (2016) Wealth and disability in later life: The English Longitudinal Study of Ageing (ELSA). *PloS ONE* 11(11), e0166825.

Turkiewicz, A., A. A. Kiadaliri and M. Englund (2019) Cause-specific mortality in osteoarthritis of peripheral joints. *Osteoarthritis and Cartilage* 27(6), 848–854.

Tutz, G. and M. Schmid (2016) *Modelling discrete time-to-event data*, Springer series in Statistics.

van Buuren, S. (2018) *Flexible Imputation of Missing Data*. 2nd Ed. ed. FL: Chapman & Hall/CRC.

van Buuren, S. and K. Groothuis-Oudshoorn (2011) mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 45(3), 1–67.

Yang, Y., Z. Du, Y. Liu, J. Lao, X. Sun and F. Tang (2021) Disability and the risk of subsequent mortality in elderly: a 12-year longitudinal population-based study. *BMC Geriatrics* 21(1), 1–9.

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

Appendix A: Appendix

A.1.

Table A1. Coefficient and standard error estimates of a random effects model based on datasets II, III, IV and V

	Model (II)		Model (III)		Model (IV)		Model (V)	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Intercept	−5.487***	0.3459	−5.500***	0.3522	−5.504***	0.3506	−5.4826***	0.3482
Time 2	−0.0556	0.1176	−0.0551	0.1186	−0.0556	0.1183	−0.0532	0.1182
Time 3	−0.2704*	0.1308	−0.2813*	0.1317	−0.2689*	0.1315	−0.2777*	0.1310
Time 4	−0.4751***	0.1391	−0.4694***	0.1403	−0.4709***	0.1400	−0.4742***	0.1394
Time 5	−1.142***	0.1653	−1.1415***	0.1665	−1.1411***	0.1663	−1.1423***	0.1657
Gender(F)	−0.5305***	0.0989	−0.5382***	0.0996	−0.5346***	0.0994	−0.5338***	0.0990
Age[1]	1.1996***	0.0873	1.2065***	0.0894	1.2071***	0.0889	1.2046***	0.0882
Emp. 2	−1.0755	1.0346	−1.0726	1.0346	−1.0834	1.0348	−1.0720	1.0346
Emp. 3	1.4899***	0.3394	1.4722***	0.3402	1.4867***	0.3400	1.4821***	0.3397
Emp. 4	0.1011	0.2946	0.0867	0.2954	0.0916	0.2950	0.0892	0.2948
Emp. 5	1.0356*	0.4253	1.0375*	0.4257	1.0522*	0.4256	1.0242*	0.4254
Mari. (S)	−0.1430	0.0984	−0.1423	0.0989	−0.1439	0.0987	−0.1436	0.0984
Mobility	0.0644**	0.0201	0.0639**	0.0202	0.0642**	0.0202	0.0633**	0.0201
ADL	0.0878*	0.0390	0.0871*	0.0392	0.0869*	0.0390	0.0890*	0.0391
IADL	0.1584***	0.0333	0.1636***	0.0334	0.1642***	0.0331	0.1600***	0.0332

(*Continued*)

Table A1. (*Continued.*)

	Model (II)		Model (III)		Model (IV)		Model (V)	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Heart A(Y)	0.7170***	0.1570	0.6740***	0.1579	0.6844***	0.1580	0.7013***	0.1580
Heart F(Y)	0.9796***	0.2765	0.9908***	0.2784	0.9908***	0.2783	0.9686***	0.2757
Diabetes(Y)	0.3992**	0.1225	0.4049**	0.1233	0.4094***	0.1235	0.4038**	0.1228
Stroke(Y)	0.3645*	0.1545	0.3549*	0.1550	0.3491*	0.1548	0.3463*	0.1542
Lung(Y)	0.6537***	0.1358	0.6606***	0.1369	0.6605***	0.1368	0.6551***	0.1363
Arthritis(Y)	−0.2357*	0.0942	−0.2319*	0.0945	−0.2456**	0.0948	−0.2330*	0.0941
Cancer(Y)	1.2610***	0.1268	1.2624***	0.1282	1.2710***	0.1278	1.2628***	0.1273
Alzheimer(Y)	0.7784*	0.3211	0.7797*	0.3236	0.7725*	0.3227	0.7584*	0.3202
AIC	5, 602.1		5, 603.8		5, 601.6		5, 605.1	
BIC	5, 809.3		5, 811		5, 808.9		5, 812.4	
Log-likelihood	−2, 777.0		−2, 777.9		−2, 776.8		−2, 778.6	
Random effect: id number								
Variance	1.322		1.386		1.379		1.341	
Number of obs: 41,543; groups: idauniq, 10,475								

$^{+}p < 0.1$, * $p < 0.05$,** $p < 0.01$, *** $p < 0.001$.
[1] Scaled.

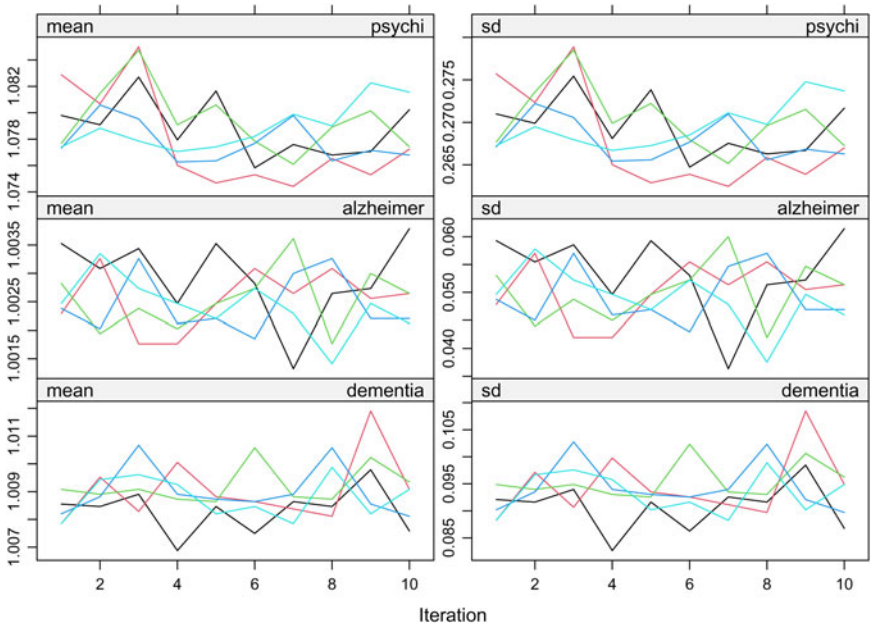

Figure A1. Trace lines do not show any trends after 10 iterations.

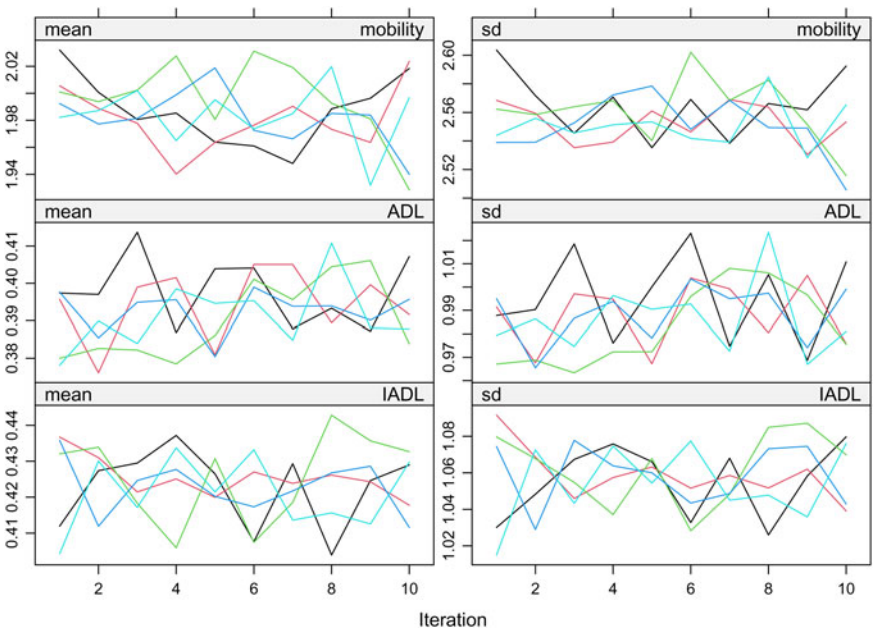

Figure A2. Trace lines do not show any trends after 10 iterations.

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

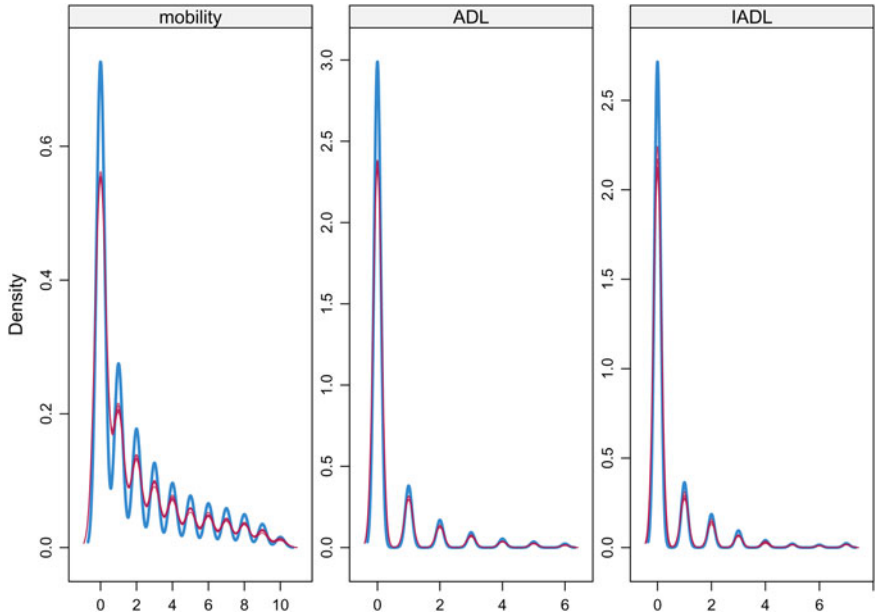

Figure A3. Comparison of the distribution of the generated dataset (red curve) with the distribution of the observed dataset (blue curve).

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

Table A2. Comparison of different optimizers

	bobyqa	Nelder Mead	nlminbwrap	optimx.L-BFGS-B	nloptwrap Nelder Mead	nloptwrap BOBYQA
Intercept	−5.4944	−5.1417	−4.9241	−5.4844	−5.4944	−5.4961
Time 2	−0.0488	−0.1797	−0.1650	−0.0489	−0.0488	−0.0478
Time 3	−0.2683	−0.4208	−0.4141	−0.2683	−0.2683	−0.2677
Time 4	−0.4687	−0.6108	−0.6324	−0.4687	−0.4686	−0.4681
Time 5	−1.1371	−1.3016	−1.2934	−1.1376	−1.1371	−1.1369
Gender(F)	−0.5309	−0.5358	−0.4578	−0.5309	−0.5309	−0.5309
Age[1]	1.2049	1.1117	1.0222	1.2046	1.2049	1.2040
Emp. 2	−1.0803	−1.7030	−1.0828	−1.0925	−1.0805	−1.1050
Emp. 3	1.4866	1.4334	1.5801	1.4777	1.4866	1.4879
Emp. 4	0.0892	0.1827	0.3627	0.0820	0.0892	0.0922
Emp. 5	1.0352	0.8771	1.0525	1.0231	1.0352	1.0388
Mari. (S)	−0.1456	−0.1603	−0.1679	−0.1458	−0.1456	−0.1449
Mobility	0.0623	0.0616	0.0585	0.0624	0.0624	0.0624
ADL	0.0880	0.0819	0.0657	0.0881	0.0880	0.0879
IADL	0.1591	0.1558	0.1319	0.1588	0.1591	0.1590
Heart A(Y)	0.6959	0.6214	0.6468	0.6966	0.6959	0.6969
Heart F(Y)	0.9739	0.8832	0.7709	0.9725	0.9739	0.9734
Diabetes(Y)	0.4037	0.3605	0.3633	0.4039	0.4037	0.4038
Stroke(Y)	0.3811	0.3223	0.3413	0.3810	0.3811	0.3803
Lung(Y)	0.6829	0.5993	0.5819	0.6828	0.6829	0.6826
Arthritis(Y)	−0.2298	−0.2014	−0.1866	−0.2298	−0.2297	−0.2296

(*Continued*)

Table A2. (*Continued.*)

	bobyqa	Nelder Mead	nlminbwrap	optimx.L-BFGS-B	nloptwrap Nelder Mead	nloptwrap BOBYQA
Cancer(Y)	1.2938	1.2162	1.1917	1.2940	1.2938	1.2938
Alzheimer(Y)	0.7538	0.6682	0.5231	0.7530	0.7538	0.7564
AIC	5, 599.8	5, 603.5	5, 614.8	5, 599.8	5, 599.8	5, 599.8
BIC	5, 807.1	5, 810.7	5, 822.1	5, 807.1	5, 807.1	5, 807.1
Log-likelihood	−2, 775.9	−2, 777.8	−2, 783.4	−2, 775.9	−2, 775.9	−2, 775.9
Convergence warning	No	Yes	Yes	Yes	No	Yes
Random effect: id number						
Std.Dev	1.16	0.873	0.0000	1.158	1.16	1.159
Number of obs: 41,543; groups: idauniq, 10,475						
Elapsed time	1, 646.517	1, 133.596	96.209	1, 005.173	1, 130.972	630.508

[1] Scaled.

A.2. Posterior distribution of β and σ^2

Let $Y \sim N(X\beta,\ \sigma^2\ I)$; $\beta|\sigma^2 \sim N_k(0,\ \sigma^2\ V_0)$ and $\sigma^2 \sim$ Inv-Gamma$(a_0 = \nu_0/2,\ b_0 = \nu_0 s_0^2/2)$. To find the posterior distribution of β and σ^2 we have

$$p(\beta,\ \sigma^2|X,\ y) \propto p(y|X,\ \beta,\ \sigma^2)p(\beta|\sigma^2)p(\sigma^2). \tag{A1}$$

The first term is the likelihood function which is given by

$$p(y|X,\ \beta,\ \sigma^2) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right\}. \tag{A2}$$

First we consider the exponential term:

$$\begin{aligned}
&(y - X\beta)^T(y - X\beta)\\
&= (y - X\hat{\beta} + X\hat{\beta} - X\beta)^T(y - X\hat{\beta} + X\hat{\beta} - X\beta)\\
&= [(y - X\hat{\beta}) + (X\hat{\beta} - X\beta)]^T[(y - X\hat{\beta}) + (X\hat{\beta} - X\beta)]\\
&= (y - X\hat{\beta})^T(y - X\hat{\beta}) + (X\hat{\beta} - X\beta)^T(X\hat{\beta} - X\beta) + 2(y - X\hat{\beta})^T(X\hat{\beta} - X\beta)\\
&= (y - X\hat{\beta})^T(y - X\hat{\beta}) + (\hat{\beta} - \beta)^T X^T X(\hat{\beta} - \beta),
\end{aligned} \tag{A3}$$

where $\hat{\beta} = (X^TX)^{-1}X^Ty$. The last term is equal to 0 since

$$(y - X\hat{\beta})^T X(\hat{\beta} - \beta) = (y^TX - \hat{\beta}^TX^TX)(\hat{\beta} - \beta).$$

Substituting for $\hat{\beta}$ in the first term, we have

$$y^TX - [(X^TX)^{-1}X^Ty]^TX^TX = y^TX - y^TX[(X^TX)^{-1}]^TX^TX,$$

which is equal to 0 since $[(X^TX)^{-1}]^TX^TX = [(X^TX)^T]^{-1}X^TX = (X^TX)^{-1}X^TX = I$. Then we consider the second term in Equation (A1) which is the prior density of $\beta|\sigma^2$ and is given by

$$p(\beta|\sigma^2) \propto (\sigma^2)^{-k/2} \exp\left\{-\frac{1}{2\sigma^2}\beta^T V_0^{-1}\beta\right\} \tag{A4}$$

and the third term in Equation (A1) is the prior density of σ^2 which is given by

$$p(\sigma^2) \propto (\sigma^2)^{-\nu_0/2-1} \exp\left\{-\frac{\nu_0 s_0^2}{2\sigma^2}\right\} \tag{A5}$$

Combining (A2), (A4) and (A5), posterior distribution is proportional to

$$\begin{aligned}
&(\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right\}(\sigma^2)^{-k/2} \exp\left\{-\frac{1}{2\sigma^2}\beta^T V_0^{-1}\beta\right\}(\sigma^2)^{-(\nu_0/2+1)}\\
&\exp\left\{-\frac{\nu_0 s_0^2}{2\sigma^2}\right\}.
\end{aligned} \tag{A6}$$

Next, we consider the first two exponential terms. Using the result in (A3), we have

$$\begin{aligned}
&(y - X\hat{\beta})^T(y - X\hat{\beta}) + (\beta - \hat{\beta})^T(X^TX)(\beta - \hat{\beta}) + \beta^T V_0^{-1}\beta\\
&= y^Ty + \hat{\beta}^TX^TX\hat{\beta} - 2y^TX\hat{\beta} + \beta^TX^TX\beta + \hat{\beta}^TX^TX\hat{\beta} - 2\beta^TX^TX\hat{\beta} + \beta^TV_0^{-1}\beta\\
&= y^Ty + \beta^T(X^TX + V_0^{-1})\beta - 2\beta^TX^TX\hat{\beta},
\end{aligned} \tag{A7}$$

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

where in the last term we use the fact that

$$2\hat{\beta}^T X^T X\hat{\beta} = 2[(X^T X)^{-1}X^T y]^T X^T X\hat{\beta} = y^T X[(X^T X)^{-1}]^T X^T X\hat{\beta} = 2y^T X\hat{\beta}.$$

Our aim is to write (A7) in a quadratic form. Consider

$$(\beta-\mu)^T S(\beta-\mu) = \beta^T S\beta + \mu^T S\mu - 2\beta^T S\mu.$$

Comparing the above expression with (A7), $\beta_n = \mu = S^{-1}(X^T X\hat{\beta})$ and $V_n^{-1} = S = X^T X + V_0^{-1}$. Therefore, we can write (A7) as

$$(\beta-\beta_n)^T V_n^{-1}(\beta-\beta_n) - \beta_n^T V_n^{-1}\beta_n + y^T y. \tag{A8}$$

Substituting in (A6), we have

$$\begin{aligned}(\sigma^2)^{-k/2}\exp\left\{-\frac{1}{2\sigma^2}(\beta-\beta_n)^T V_n^{-1}(\beta-\beta_n)\right\}(\sigma^2)^{-(n+\nu_0)/2-1}\\ \exp\left\{-\frac{1}{2\sigma^2}(y^T y - \beta_n^T V_n^{-1}\beta_n + \nu_0 s_0^2\right\}.\end{aligned} \tag{A9}$$

which is the posterior distribution of (β, σ^2).

A.3. Gradient and Hessian matrix for a logistic regression

Consider $x_i = (x_{i1}, \ldots, x_{id})^T$. In Equation (20), we first consider the first logarithm

$$\log p = \log\left(\frac{1}{1+e^{-\beta^T x}}\right) = -\log(1+e^{-\beta^T x}).$$

Taking the partial derivative with respect to β_j gives

$$\frac{\partial}{\partial\beta_j}\log p = x_j(1-p).$$

Then we take the second logarithm

$$\log(1-p) = -\beta^T x - \log(1+e^{-\beta^T x}).$$

Taking the partial derivative with respect to β_j gives

$$\frac{\partial}{\partial\beta_j}\log(1-p) = -x_j + x_j(1-p) = -px_j.$$

Hence the gradient is given by

$$g = \frac{\partial}{\partial\beta_j}l = -\sum_{i=1}^{n}[y_i x_{ij}(1-p_i) - (1-y_i)x_{ij}p_i] = \sum_{i=1}^{n}(p_i - y_i)x_{ij} = X^T(p-y),$$

https://doi.org/10.1017/dem.2023.3 Published online by Cambridge University Press

where X is a design matrix and is given by

$$X = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nd} \end{pmatrix}. \tag{A10}$$

Next we look at the Hessian matrix by taking the second partial derivative of l

$$H = \frac{\partial^2}{\partial\beta_j \partial\beta_k} l = \sum_{i=1}^{n} x_{ij} \frac{\partial}{\partial\beta_k} p_i = \sum_{i=1}^{n} x_{ij} x_{ik} p_i (1 - p_i) = X^T W X, \tag{A11}$$

where W is a diagonal matrix given by

$$W = \begin{pmatrix} p_1(1-p_1) & & \\ & \ddots & \\ & & p_n(1-p_n) \end{pmatrix}. \tag{A12}$$

We can now show that the application of Newton–Raphson method results in an equation which is the solution to a weighted least square problem.

$$\begin{aligned} \beta_{new} &= \beta_{old} - H^{-1} g \\ &= \beta_{old} - (X^T W X)^{-1} X^T (p - y) \\ &= (X^T W X)^{-1} X^T W [X\beta_{old} - W^{-1}(p - y)] \\ &= (X^T W X)^{-1} X^T W z, \end{aligned} \tag{A13}$$

which is the solution to the following problem

$$\arg\min_{\beta} \sum w_i (z_i - \beta^T x_i)^2.$$

Cite this article: Qazvini M (2023). Survival analysis of longitudinal data: the case of English population aged 50 and over. *Journal of Demographic Economics* **89**, 419–463. https://doi.org/10.1017/dem.2023.3