Autoregressive mixed effects models and an application to annual income of cancer survivors

by

Lisa McQuarrie

B.Sc., University of Victoria, 2019

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

© Lisa McQuarrie 2021

SIMON FRASER UNIVERSITY
Spring 2021

Copyright in this work rests with the author. Please ensure that any reproduction or re-use is done in accordance with the relevant national copyright legislation.

Declaration of Committee

Name:	**Lisa McQuarrie**
Degree:	**Master of Science (Statistics)**
Title:	**Autoregressive mixed effects models and an application to annual income of cancer survivors**
Committee:	**Chair:** Richard Lockhart Professor Statistics and Actuarial Science
	Rachel Altman Supervisor Associate Professor, Statistics and Actuarial Science
	Stuart Peacock Committee Member Professor, Faculty of Health Sciences
	Joan Hu Examiner Professor, Statistics and Actuarial Science

Abstract

Longitudinal observations of income are often strongly autocorrelated, even after adjusting for independent variables. We explore two common longitudinal models that allow for residual autocorrelation: 1. the autoregressive error model (a linear mixed effects model with an AR(1) covariance structure), and 2. the autoregressive response model (a linear mixed effects model that includes the first lag of the response variable as an independent variable). We explore the theoretical properties of these models and illustrate the behaviour of parameter estimates using a simulation study. Additionally, we apply the models to a data set containing repeated (annual) observations of income and sociodemographic variables on a sample of breast cancer survivors. Our preliminary results suggest that the autoregressive response model may severely overestimate the magnitude of the effect of cancer. Our findings will guide future, comprehensive study of the short- and long-term effects of a breast cancer diagnosis on a survivor's annual net income.

Keywords: Longitudinal data, mixed models, autoregressive models, breast cancer, annual net income

Acknowledgements

First and foremost, I would like to express my sincerest appreciation to my supervisor, Rachel, for her endless support and commitment. Thank you for helping me reach my goals, challenging me to think deeper, and giving me the confidence to find solutions on my own. I will truly miss our weekly zoom calls.

I would also like to thank the research team at the Canadian Centre for Applied Research in Cancer Control, who welcomed me into their team and supported this project with invaluable subject matter expertise. Specifically, thank you to Stuart Peacock, Kim van der Hoek, Paulos Teckle, Colene Bentley, and Shiraz El Adam. I have enjoyed working with all of you immensely.

Thank you to my examining committee for taking the time to provide thoughtful feedback and ask interesting questions; your insight is greatly appreciated.

I wish my fellow grad students best of luck in their studies and future endeavors. I would particularly like to express my gratitude to Renny Doig, for frequently reviewing my work (including this project) and providing thoughtful, kind feedback. I would also like to thank Matt Berkowitz, who was my guide through the program and would always happily answer my “quick” questions.

Lastly, thank you to my parents for feigning interest whenever I want to discuss the exciting new statistical concept I learned that day, or, when things go wrong, for patiently listening while I express my frustrations. To my sister Amy, thank you for being my best friend and partner in the vexations of grad student life.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this project, we discuss the application of longitudinal mixed effects models to incomes of breast cancer survivors. Using yearly observations of annual income, both pre- and post-diagnosis, on a sample of breast cancer survivors, we characterize and estimate the effect of a cancer diagnosis on a survivor's income.

1.1 Background

In Canada, cancer is expected to impact one out of every two people, and, in particular, one in four cancer cases in women will be breast cancer (Brenner et al., 2020). Thanks to improved screening procedures, diagnostic tests, and treatments, the mortality rate for female breast cancer has declined significantly in the last 35 years (Brenner et al., 2020), and now female breast cancer has a 5-year survival rate of 88% (Canadian Cancer Society, 2020). Consequently, the number of breast cancer survivors living for extended periods of time past their diagnosis and initial treatment has grown, and, as such, focus has expanded to caring for all aspects of survivors' health, including mental and physical health and quality of life. In this project, "survivor" refers to any person diagnosed with cancer, regardless of how long they live after diagnosis.

An essential contributor to quality of life, both for those with and without long-term health problems, is often employment. The effect of financial stress and hardship on quality of life has been extensively documented; individuals with insufficient income for food, medical care, and other necessities have lower quality of life, in general (Park and Look, 2018; Fitch and Longo, 2018; Jones et al., 2020; Kale and Carroll, 2016). However, employment provides more than just income. To many, the workplace is a venue that provides social stimulation and interaction, which has been shown to improve mental health in general, and in particular the mental health of cancer survivors (Rasmussen and Elverdam, 2008).

The routine, mental stimulation, and familiarity of work and the workplace can provide a sense of normalcy and purpose that also contribute to cancer survivors' quality of life (Duijts et al., 2017; Lundh et al., 2013; Rasmussen and Elverdam, 2008).

Continuing or returning to work after a cancer diagnosis and treatment can be a vital step in recovery. However, depending on the type and severity of the cancer or the type of work (e.g., manual labour versus office job), there can be many barriers in maintaining or returning to employment after cancer treatment. Cancer survivors may experience significant long-term health problems, such as chronic lymphedema, fatigue, and cognitive impairment, all of which can affect their ability to participate in the labour force (Ahles et al., 2002; Erickson et al., 2001; Ganz et al., 1998; Ganz, 1999; Ganz et al., 2002). In addition to the physical effects of treatment, cancer survivors often struggle with their mental health (Hewitt et al., 2003), which can influence their decision to return to work (Islam et al., 2014). Lastly, lack of support from employers and colleagues has been shown to have a significant negative impact on the success of a cancer survivor's return to work (Islam et al., 2014; Tamminga et al., 2012).

Female breast cancer survivors warrant in-depth study due to the large and growing population of such individuals who are pre-retirement age. Most cancers affect an older population; for example, the median age at diagnosis of survivors of prostate and lung cancers, two of the other most prevalent cancer types, are 66 and 70, respectively (National Cancer Institute, 2015). However, the median age at diagnosis for the female breast cancer survivor is 61 (National Cancer Institute, 2015). Many female breast cancer survivors are pre-retirement age when diagnosed, which, when combined with the high survival rate, indicates that many women diagnosed with breast cancer may be able and willing to return to employment post-treatment. Additionally, female cancer survivors have been found to have longer return-to-work times (van Muijen et al., 2013), higher likelihood of job loss, and lower likelihood of re-employment than male cancer survivors (Park et al., 2008). For these reasons, we focus our study on female breast cancer survivors. Henceforth, all references to breast cancer will refer to female breast cancers.

The literature on the impact of cancer on economic outcomes, such as income or employment status, is inconclusive. Different studies have reported negative (Hewitt et al., 2003; Yabroff et al., 2004; Drolet et al., 2005a,b; Chirikos et al., 2002a; Bradley et al., 2002a,b; Grunfeld et al., 2004; Maunsell et al., 1999, 2004; Syse et al., 2008; Lauzier et al., 2008; Chirikos et al., 2002b), neutral (Hensley et al., 2005; Peuckmann et al., 2009; Gudbergsson et al., 2009; Hakanen and Lindbohm, 2008), and positive (Bradley et al., 2002a,b; Chirikos et al., 2002a; Satariano and DeLorenze, 1996) consequences of cancer on various economic outcomes. However, our review of the existing literature on this subject identified five commonly occurring weaknesses:

1. Subjects are often recruited through advertisements or cancer support groups and centres, which can result in selection bias.

2. The sample sizes are generally small, which limits the ability to identify important effects.

3. The majority of studies use self-reported income data, which is systematically underreported and prone to recall bias (Fukuoka et al., 2007).

4. Few studies are conducted in Canada, and, due to differing health care systems (private versus public), results from studies in other countries may not be applicable to Canadian cancer survivors. For example, in the USA, the health insurance system is strongly tied to the workplace; therefore, survivors may be incentivised to return to work to have access to health insurance.

5. Similar studies have presented almost exclusively cross-sectional analyses of determinants of income of cancer survivors; longitudinal analysis of changes in income over time is usually not considered.

Economics researchers have found a strong autocorrelation structure in annual incomes. Specifically, income in year t is highly positively correlated with income in year $t + \ell$ when ℓ is small and decreases toward zero as ℓ grows (Blundell, 2014; Baker, 1997; Parsons, 1978). Research also suggests that this autocorrelation structure persists even after adjusting for common socio-demographic variables, such as education, gender, and race (Parsons, 1978; Blundell, 2014). Consequently, models with a marginal covariance matrix that allows for changing covariance with increasing lag are often used to model income (Berry et al., 1988; Browning and Ejrnæs, 2013; Zhang and Zhou, 2019; Lillard and Willis, 1978). We consider two of the most common such models, the autoregressive response (ARR) and autoregressive error (ARE) models. The ARR model is specified in terms of lagged outcomes and independent errors, while the ARE model is specified in terms of first-order autoregressive (AR(1)) errors. Both models include a random, subject-specific intercept to allow for unexplained heterogeneity in income across subjects.

These models are frequently used to model income in the longitudinal setting. For example, the ARR model is used to model income and related economic outcomes in numerous papers (e.g., Zhang and Zhou (2019); Cai et al. (2014); Bond (2002)). Additionally, the ARR model is taught in several popular econometrics textbooks (e.g., Hsiao (2014); Baltagi (2008); Greene (2003); Wooldridge (2002)), and a statistics textbook (Funatogawa and Funatogawa, 2019). The ARE model is also frequently used to describe income (e.g., Berry et al. (1988); Browning and Ejrnæs (2013)).

1.2 Motivation and objectives

The primary objective of this research is to describe breast cancer survivors' income over time, with special focus on the short- and long-term effects of breast cancer diagnosis. Our analysis, described in Chapter 4, of a large, longitudinal administrative data set addresses each of the weaknesses that we identified in our literature review in Section 1.1.

As income is directly related to quality of life, determining how a woman's breast cancer diagnosis impacts her income is of interest. Longitudinal models are used to capture changes in the income of breast cancer survivors over time and to characterize the autocorrelation structure of yearly income observations.

We investigate the ARR and ARE models because they are commonly used to model longitudinal income trends. In addition, we use a linear mixed effects model with a woman-specific intercept to inform our choice of mean, variance, and correlation structures in the ARR and ARE models.

The remainder of this report is organized as follows. We review the data set in Chapter 2 and introduce the methods used in the analysis in Chapter 3. We describe the analysis in Chapter 4 and conduct a simulation study to further explore the models and their properties in Chapter 5. We conclude with a discussion and suggestions for future work in Chapter 6.

Chapter 2

Description of data set

For this project, we use national administrative data files from Statistics Canada that consist of a linkage of the Canadian Cancer Registry (CCR) with the T1 Family File (T1FF). Access to the data files, which are not publicly available, was provided by Statistics Canada through their Research Data Centres. Due to the highly confidential nature of the data, we are limited in the descriptions and results we can provide by strict guidelines meant to prevent a breach of confidentiality. As a consequence, we are unable to present some of our results publicly.

The CCR data file contains diagnosis and treatment information about all patients diagnosed with cancer in Canada between the years 1992 and 2015. The T1FF data files contain variables related to the income of individuals and their families for all individuals who filed a tax return in Canada between 1992 and 2015.

In the CCR, primary tumours are classified using the International Classification of Diseases for Oncology (ICD-O) codes, which allows us to identify breast tumours. Sex, diagnosis date, and birth date are also supplied in the CCR. Our inclusion criteria are: female breast cancer survivor, diagnosed in Canada between January 1st, 1992, and December 31st, 2015, and between 18 and 60 years of age at diagnosis. As mentioned earlier, a key reason for focusing on breast cancer survivors is that they are often of pre-retirement age when diagnosed. Retired workers frequently have a fixed income from a pension plan, which would not vary due to a cancer diagnosis. We include only observations recorded on patients when they are between the ages of 18 and 60 to avoid having to incorporate the effect of retirement in the model.

From the T1FF data files, we extract our response variable, annual after-tax net income, defined as the total income received by the filer less all tax deductions. We use after-tax income (as opposed to other definitions of income, such as income generated by employment) as the response variable because after-tax income represents the consumption and saving opportunity of the individual. Therefore, after-tax income best represents the ability of the individual to support herself. Incomes are observed both pre- and post-diagnosis. In general, an individual's record ends either in 2015 or when she is lost to follow-up, whichever occurs first. An implication is that we do not consider the loss of income due to death.

Our main interest is in estimating the effect of cancer. However, the effect of cancer could manifest in many different ways. For instance, cancer could cause a change in income, on average, only in the year of diagnosis. Or, cancer could cause a change in income that lasts for many years post-diagnosis. Therefore, based on recommendations by our collaborators at the Canadian Centre for Applied Research in Cancer Control (ARCC), we define three indicator variables to allow for different possible manifestations of the effect of cancer: `cancerImmediate`, `cancerShort`, and `cancerLong`. We describe these variables below.

The variable `cancerImmediate` represents the immediate cancer effect that occurs in the year of diagnosis and the year after diagnosis; it is an indicator variable that takes the value 1 in these two years and 0 in all other years. Our rationale for this choice is as follows. We have data only on annual income, i.e., the income received in a calendar year. Therefore, we cannot investigate the effect of cancer in the 12 months following a cancer diagnosis; for example, if a woman is diagnosed in June 2010, we cannot identify the effect of the diagnosis on her income from June 2010 to June 2011. We may reasonably expect to find an effect of cancer on a woman's income in the year of diagnosis if she is diagnosed early in the year, but, if she is diagnosed, for example, in December 2010, then cancer would likely have a stronger effect on her income in 2011. We therefore define the immediate cancer effect as the effect in the year of diagnosis and the following year so that, at a minimum, we are capturing the effect of cancer in the first 12 months post-diagnosis for all survivors.

We define the short-term cancer effect, `cancerShort`, as the effect of cancer from the year of diagnosis to 5 years post-diagnosis, and we capture this effect using an indicator variable that takes the value 1 from the year of diagnosis to 5 years post-diagnosis and 0 in all other years.

The long-term cancer effect, represented by `cancerLong`, is defined as the effect of cancer in the year of diagnosis and all following years and is captured by an indicator variable taking the value of 1 in the year of diagnosis and all following years and 0 in the years before diagnosis. All three cancer indicator variables are derived from data provided in the CCR about diagnosis date.

In addition to cancerImmediate, cancerShort, and cancerLong, we include the explanatory variables yearsSince18, birthYear, selfEmploy, and spouseStatus in our models. We define these variables next.

The variable yearsSince18 is defined as age (in years) less 18 and is used to capture the effect of age on income. Justification for using years since 18 rather than age is provided in Chapter 3. This variable is derived from data on birth date provided in the CCR.

The variable birthYear is defined as the year of birth, and we use this variable to capture the cohort effect of birth year. This variable originates in the CCR. In the longitudinal context, birthYear and yearsSince18 are not correlated since, for each woman, birthYear is constant and yearsSince18 is time varying. Therefore, both variables can be included as explanatory variables in a model without introducing multicollinearity issues.

The variable selfEmploy is 1 if the woman receives some income from self-employment at a given time point and 0 otherwise. It originates in the T1FF.

The variable spouseStatus is 1 if the woman has a spouse at a given time point and 0 otherwise. This variable is derived from data about family composition provided in the T1FF.

Although not used as a explanatory variable, we also use the variable years since diagnosis in descriptions of the data. This variable is defined as the year of the observation less the year of diagnosis, so that it takes negative integer values in the years prior to diagnosis, 0 in the year of diagnosis, and positive integer values after diagnosis.

All residents of Canada must file a tax return if they have an income source. Additionally, medical practitioners are required to submit their patients' information for recording in the CCR. As a result, only a small proportion, $< 5\%$, of the total observations contained missing values for one or more of the variables considered in the analysis. Because this proportion is small, we chose to remove rows with missing values. After removing missing values, the number of women observed at one or more time points is 14,565 and the total number of observations is 243,475.

2.1 Descriptive statistics and data visualizations

The bar chart in Figure 2.1 shows the distribution of number of observations on each woman. Of all the women in the sample, 23% were observed in all 24 years. The median number of observations on a woman is 18.

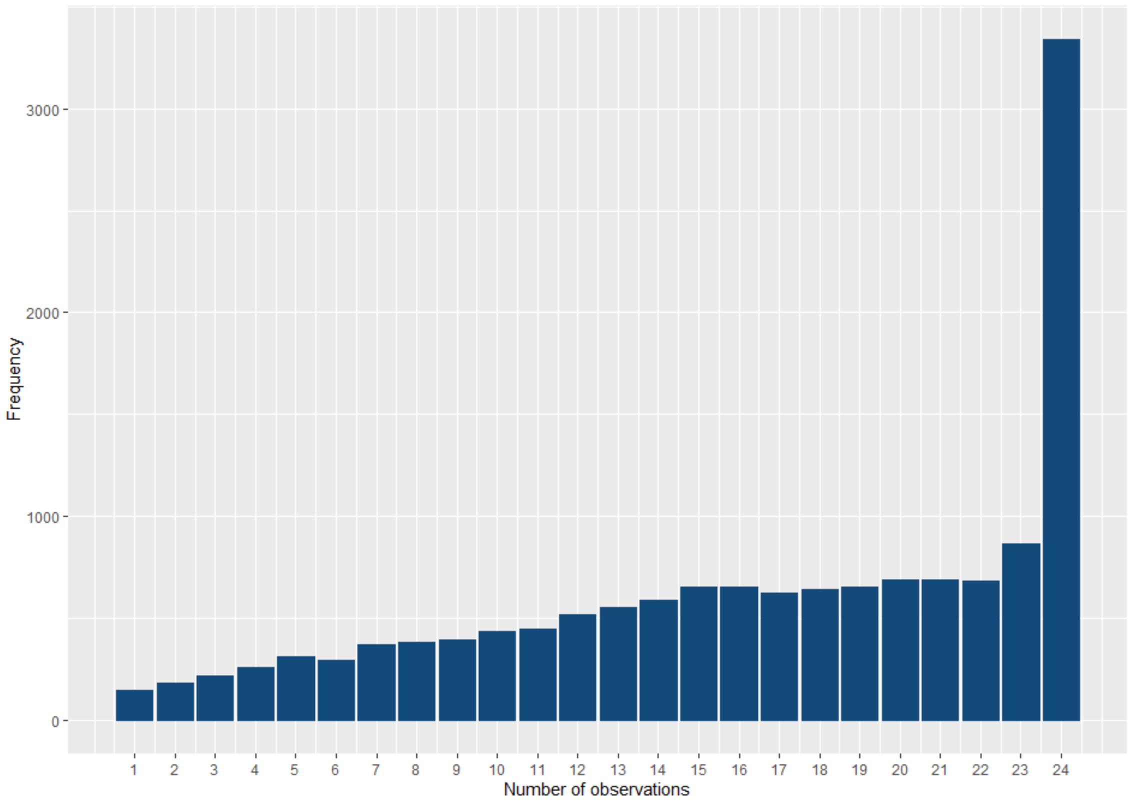

Figure 2.1: Bar chart of number of observations per woman.

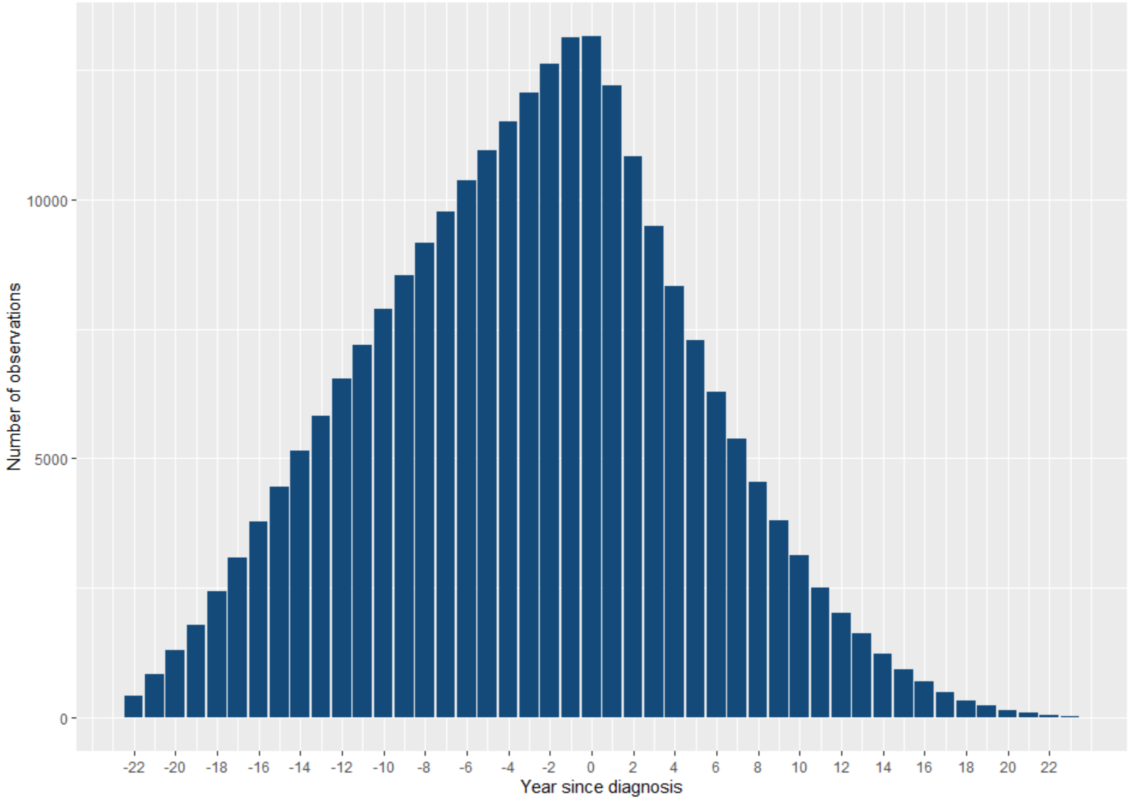

Figure 2.2: Bar chart of number of observations by year since diagnosis.

The distribution of number of observations by years since diagnosis is shown in Figure 2.2. Fewer records are observed on women post-diagnosis compared to pre-diagnosis.

The proportion of women who had a spouse in one or more years is 88%, and the proportion of women who received self-employment income in one or more years is 32%. The average year of birth in the sample is 1954. The average and median ages at diagnosis in the sample are 49 and 50, respectively. Since the sample includes only women diagnosed before the age of 60, the average and median ages at diagnosis in the sample are much younger than in the general population of breast cancer survivors.

Figure 2.3 is a plot showing the sample mean and standard deviation (SD) of income$^{0.25}$ at each age. The justification for the transformation income$^{0.25}$ is provided in Chapter 4. Only observations from years prior to a woman's diagnosis were used to create this plot, i.e., cancer effects are not confounding the trend with age. This plot clearly shows that the relationship between average income$^{0.25}$ and age is not linear; in particular, income$^{0.25}$ grows rapidly between the ages of 18 and 25, then tapers off, and then decreases gradually between the ages of 50 and 60 (perhaps women who are planning for early retirement begin

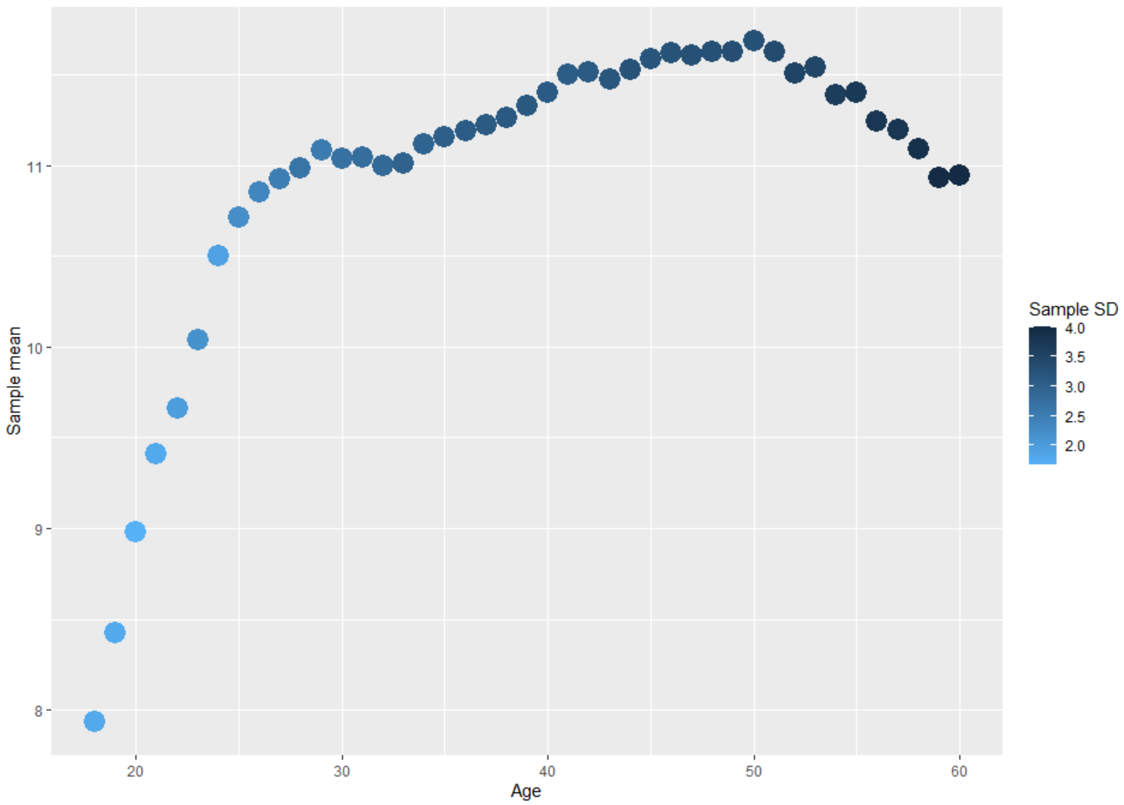

Figure 2.3: Sample mean income$^{0.25}$ by age (years).

to reduce their employment income after the age of 50). Also, as the scale on the right shows, the sample standard deviation of income$^{0.25}$ increases from ages 18 to 60.

2.1.1 Negative income observations

Some women in the sample have negative values of after-tax income. Negative after-tax income values can occur when a person's tax deductible expense claims (including, for example, medical expenses, pension plan contributions, and education costs) are greater than their income in that year. Additionally, self-employed individuals may report a negative after-tax income value if they do not make a profit in that year.

In total, 2,055 observations (0.844% of total observations) of after-tax income are negative, and 1,065 women (7.31% of sample) had one or more negative observations.

In each year relative to the year of diagnosis, the proportion of income observations that are negative is shown in Figure 2.4. Observations from the years -22 to -15 and 15 to 23 were binned together due to confidentiality concerns about the low counts in those years. The proportion of negative income values increases noticeably after diagnosis, and remains elevated

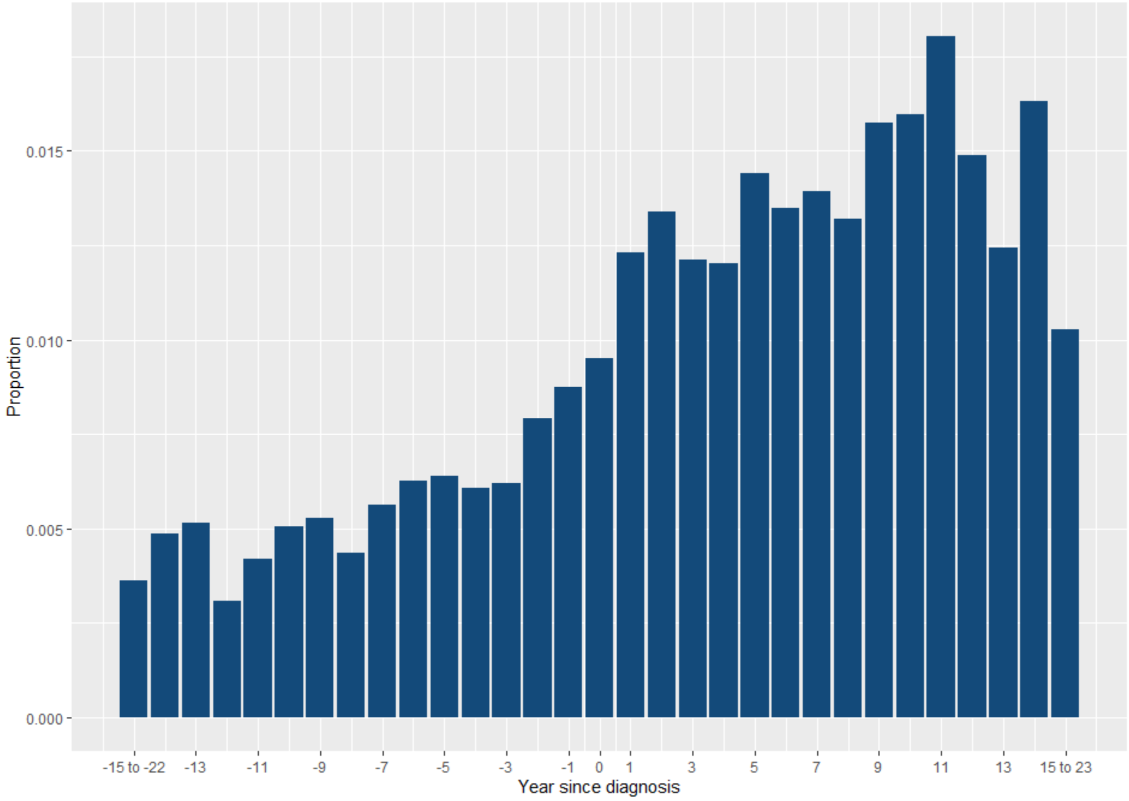

Figure 2.4: Proportion of total observations that are negative by years since diagnosis.

in all years post-diagnosis. In other words, a cancer diagnosis may increase the probability that a woman will file a tax return with a negative value of after-tax income.

We handle the negative observations by setting all negative values of after-tax income to 0. One advantage of this choice is that we can apply our chosen transformation, income$^{0.25}$, to all observations. A disadvantage is that we create a point mass at 0 in the distribution of income$^{0.25}$. We discuss the implications of our choice further in Section 6.1.

Chapter 3

Methods

As described in Section 1.1, this project explores two models, the ARR and ARE models. We also consider the linear mixed effects (LME) model—a special case of both models—as a benchmark.

Before specifying the models, we define some notation that will be used throughout the following sections. Let Y_{it} be a transformation of annual after-tax income for subject i at time t; justification for using a transformation of income as the response variable is given in Chapter 4, and henceforth we use the term "transformed income" to refer to our response variable. We define time as the number of years since subject i was age 18 (i.e., $t = 0$ corresponds to age 18, the typical age at which individuals begin filing tax returns). Consequently, $t \in \{0, 1, 2, \ldots\}$. Our chosen definition of time has important implications in the case of the ARR model, as we discuss in Section 3.3.1. Let $\mathbf{Y}_i = (Y_{i0}, Y_{i1}, \ldots, Y_{iT_i})$ be the vector of observations on subject i, $i = 1, 2, \ldots, N$. Let $\mathbf{x}_{it}$ be a vector of explanatory variables (with 1 as the first element to allow for an intercept) for subject i at time t, $i = 1, 2, \ldots, N$, $t = 0, 1, \ldots, T_i$.

We now describe the three models and their interpretations in detail. We follow with a discussion of estimation of the model parameters.

3.1 Linear mixed effects model

The LME model has a random intercept and independent errors. While we do not expect this model to be realistic, we use it as a basis for comparison with the ARR and ARE models, especially when choosing the components of these more complex models. Our LME model is defined as

$$
\begin{aligned}
Y_{it} &= \mathbf{x}_{it}'\boldsymbol{\alpha} + u_i + \delta_{it} \\
&= \alpha_0 + \alpha_1\texttt{birthYear}_{it} + \alpha_2\texttt{yearsSince18}_{it} + \alpha_3\texttt{yearsSince18}_{it}^2 + \\
&\quad \alpha_4\texttt{yearsSince18}_{it}^3 + \alpha_5\texttt{yearsSince18}_{it}^4 + \alpha_6\texttt{cancerImmediate}_{it} + \\
&\quad \alpha_7\texttt{cancerShort}_{it} + \alpha_8\texttt{cancerLong}_{it} + \alpha_9\texttt{spouseStatus}_{it} + \alpha_{10}\texttt{selfEmploy}_{it} + \\
&\quad \alpha_{11}\texttt{yearsSince18}_{it} \times \texttt{cancerLong}_{it} + \alpha_{12}\texttt{spouseStatus}_{it} \times \texttt{cancerLong}_{it} + \\
&\quad \alpha_{13}\texttt{selfEmploy}_{it} \times \texttt{cancerLong}_{it} + u_i + \delta_{it} \qquad (3.1)
\end{aligned}
$$

for $i = 1, 2, \ldots, N$, $t = 0, 1, \ldots, T_i$. Additionally, we assume that the u_i's are independent and identically distributed (IID) $N(0, \sigma_u^2)$ random variables, the δ_{it}'s are IID $N(0, \sigma^2)$, and the u_i's are independent of the δ_{it}'s. Justification for including the higher order `yearsSince18` terms and interaction terms is provided in Chapter 4.

Under this model, $\mathbf{Y}_i$ has a multivariate normal distribution with the following properties:

$$
\begin{aligned}
E(Y_{it}) &= \mathbf{x}_{it}'\boldsymbol{\alpha} && (3.2) \\
\mathrm{Var}(Y_{it}) &= \sigma_u^2 + \sigma^2 && (3.3) \\
\mathrm{Cov}(Y_{it}, Y_{i,t+\ell}) &= \sigma_u^2, \quad \ell \neq 0 && (3.4) \\
\mathrm{Cov}(Y_{it}, Y_{j,t+\ell}) &= 0, \quad i \neq j && (3.5)
\end{aligned}
$$

The interpretation of the regression coefficients in the LME model is straightforward. For example, if two women are identical except that one was born one year later than the other, then α_1 is their expected difference in transformed income at a given age—regardless of whether we have conditioned on u_i.

However, interpretation of the effect of cancer requires more care, as we use three indicator variables and several interaction effects to capture different aspects of this effect. Rather than interpreting the effects of `cancerImmediate`, `cancerShort`, and `cancerLong` individually, we define the "total cancer effect" as the expected difference between the transformed income of a woman at a given post-diagnosis year and her transformed income had she never been diagnosed with cancer. To be clear, this notion of total effect is unrelated to that of total effect in causal analysis. The size of the total cancer effect depends on the number of years since diagnosis. Table 3.1 shows these effect sizes. Here, the value of `yearsSince18` in the year of diagnosis is represented by t_d.

Years since diagnosis	**Total cancer effect**
0	$\alpha_6 + \alpha_7 + \alpha_8 + \alpha_{11} t_d$
1	$\alpha_6 + \alpha_7 + \alpha_8 + \alpha_{11}(t_d + 1)$
2	$\alpha_7 + \alpha_8 + \alpha_{11}(t_d + 2)$
3	$\alpha_7 + \alpha_8 + \alpha_{11}(t_d + 3)$
4	$\alpha_7 + \alpha_8 + \alpha_{11}(t_d + 4)$
5	$\alpha_7 + \alpha_8 + \alpha_{11}(t_d + 5)$
6	$\alpha_8 + \alpha_{11}(t_d + 6)$
$\vdots$	$\vdots$

Table 3.1: Total cancer effect by years since diagnosis under the LME model. The value of `yearsSince18` in the year of diagnosis is represented by t_d.

Under the LME model, the variance of the responses is constant, and the covariance structure does not depend on ℓ or t, i.e., the correlation between observations on the same subject at any two time points is constant.

We expect that the correlation between income observations on the same subject should, in fact, decrease as the time between the observations increases; therefore, we do not expect the LME model to be realistic. However, as we will show, the LME model is nested within both the ARE and ARR models, and, consequently, we can use the LME to guide our choice of simpler features of the autoregressive models (e.g., the mean structure) before making decisions about more complex features (e.g., the variance and covariance structures).

3.2 Autoregressive error model

The ARE model is a linear mixed effects model with AR(1) errors. For $i = 1, \dots, N$, the model is defined as

$$Y_{it} = \mathbf{x}_{it}' \boldsymbol{\alpha} + u_i + \epsilon_{it}, \qquad t = 0, 1, \dots, T_i \tag{3.6}$$

$$\epsilon_{it} = \gamma \epsilon_{i,t-1} + \delta_{it}, \qquad t = 1, 2, \dots, T_i, \tag{3.7}$$

$$\epsilon_{i0} \sim N(0, \frac{\sigma^2}{1 - \gamma^2}), \tag{3.8}$$

where u_i and δ_{it} are as defined in Section 3.1. We constrain γ to lie in $(-1, 1)$. All other variables are as previously defined.

We make the assumption that $\epsilon_{i0} \sim N(0, \frac{\sigma^2}{1-\gamma^2})$ so that the time series $\{\epsilon_{it}\}_{t=0}^{T_i}$ is stationary. In particular, the distribution of ϵ_{it} is the same for all t and, consequently, $\text{Var}(Y_{it})$ is constant for all t. This restriction can be relaxed; however, exploration of that extension is beyond the scope of this project.

Under this model, $\mathbf{Y}_i$ has a multivariate normal distribution with the following properties:

$$E(Y_{it}) = \mathbf{x}_{it}'\boldsymbol{\alpha} \tag{3.9}$$

$$\mathrm{Var}(Y_{it}) = \sigma_u^2 + \frac{\sigma^2}{1-\gamma^2} \tag{3.10}$$

$$\mathrm{Cov}(Y_{it}, Y_{i,t+\ell}) = \sigma_u^2 + \frac{\gamma^\ell}{1-\gamma^2}\sigma^2, \qquad \ell \neq 0 \tag{3.11}$$

$$\mathrm{Cov}(Y_{it}, Y_{j,t+\ell}) = 0, \qquad i \neq j. \tag{3.12}$$

If $\gamma = 0$, then this model reduces to the LME model.

The regression coefficients in the ARE model have the same interpretation as in the LME model; in particular, they represent both the marginal and conditional (on u_i) effects of the explanatory variables. The total cancer effect can also be interpreted as for the LME model.

The variance of the response is constant over time and is a monotonic, increasing function of γ. The covariance structure of the response varies depending on the parameter γ and the lag ℓ, but does not depend on t. If $\gamma > 0$, then the covariance decreases monotonically with increasing lag, and, if $\gamma < 0$, then the covariance is not monotonic with lag. Since the covariance does not depend on t, any two time points separated by the same lag will have the same covariance.

3.3 Autoregressive response model

In the literature, the ARR model is typically specified in a conditional form. For example, in our context, the mean transformed income of woman i at time t would be specified given her transformed income at earlier times as well as a vector of explanatory variables, $\tilde{\mathbf{x}}_{it}$ (which includes 1 as the first element). On this conditional level, lagged response variables are effectively treated as additional explanatory variables. In our setting, we use a lag of 1 so that a woman's mean transformed income at age t is specified as a function of her transformed income at age $t-1$. We also include a random, woman-specific intercept. Formally, the *conditional form of the ARR model* is

$$Y_{it} = \rho Y_{i,t-1} + \tilde{\mathbf{x}}_{it}'\boldsymbol{\beta} + u_i + \delta_{it}, \tag{3.13}$$

for $t \geq 1$. The random variables u_i are IID $N(0, \sigma_u^2)$ and δ_{it} are IID $N(0, \sigma^2)$. The δ_{it}'s are assumed to be independent of the u_i's. The parameter ρ is constrained to lie in the interval $(-1, 1)$.

The specific form of the ARR model that we use in this project is

$$\begin{aligned} Y_{it} = \rho Y_{i,t-1} &+ \beta_0 + \beta_1 \texttt{birthYear}_{it} + \beta_2 \texttt{yearsSince18}_{it} + \beta_3 \texttt{cancerImmediate}_{it} + \\ &\beta_4 \texttt{cancerShort}_{it} + \beta_5 \texttt{cancerLong}_{it} + \beta_6 \texttt{spouseStatus}_{it} + \beta_7 \texttt{selfEmploy}_{it} + \\ &\beta_8 \texttt{yearsSince18}_{it} \times \texttt{cancerLong}_{it} + \beta_9 \texttt{spouseStatus}_{it} \times \texttt{cancerLong}_{it} + \\ &\beta_{10} \texttt{selfEmploy}_{it} \times \texttt{cancerLong}_{it} + u_i + \delta_{it}, \end{aligned} \tag{3.14}$$

for $t \geq 1$. Justification for using only a linear `yearsSince18` term and for the interaction terms is provided in Section 4.3.

If $\rho = 0$, then the model reduces to the LME model.

The distribution of the first time point, Y_{i0}, must also be specified. Many different models can be assumed for the structure of Y_{i0}, and we must consider how the initial time point was chosen when selecting a model. In our case, we deliberately selected the age at which many people begin to earn a substantial income (age 18) as time point 0. Consequently, we follow Funatogawa and Funatogawa (2019) and assume that $Y_{i0} = \tilde{\mathbf{x}}_{i0}'\boldsymbol{\beta} + u_i + \delta_{i0}$, where $\delta_{i0} \sim N(0, \frac{\sigma^2}{1-\rho^2})$. A thorough discussion of different choices for the model for Y_{i0} can be found in Anderson and Hsiao (1982), and justification for our definition of the initial time point can be found in Section 3.3.1.

By applying (3.13) recursively, we can show that $Y_{it} = \sum_{k=0}^{t} \rho^k (\tilde{\mathbf{x}}_{i,t-k}'\boldsymbol{\beta} + u_i + \delta_{i,t-k})$. Therefore, the distribution of $\mathbf{Y}_i$ is multivariate normal. The marginal expectation of Y_{it} is

$$E(Y_{it}) = \sum_{k=0}^{t} \rho^k \tilde{\mathbf{x}}_{i,t-k}'\boldsymbol{\beta} \tag{3.15}$$

for $t \geq 0$.

To determine the marginal variance-covariance structure, we first need to determine $\mathrm{Cov}(Y_{it}, u_i)$. For $t = 0$, $\mathrm{Cov}(Y_{i0}, u_i) = \sigma_u^2$. For $t \geq 1$,

$$\begin{aligned}
\mathrm{Cov}(Y_{it}, u_i) &= \mathrm{Cov}(\rho Y_{i,t-1} + \tilde{\mathbf{x}}_{it}'\boldsymbol{\beta} + u_i + \delta_{it}, u_i) \\
&= \rho \mathrm{Cov}(Y_{i,t-1}, u_i) + \mathrm{Var}(u_i) \\
&= \rho^2 \mathrm{Cov}(Y_{i,t-2}, u_i) + \rho \mathrm{Var}(u_i) + \mathrm{Var}(u_i) \\
&\dots \\
&= \rho^t \mathrm{Cov}(Y_{i0}, u_i) + \sum_{k=0}^{t-1} \rho^k \mathrm{Var}(u_i) \\
&= \sigma_u^2(\rho^t + \sum_{k=0}^{t-1} \rho^k) \\
&= \sigma_u^2 \left(\frac{1-\rho^{t+1}}{1-\rho} \right). \qquad (3.16)
\end{aligned}$$

Next, the marginal variance of the response can be derived as

$$\begin{aligned}
\mathrm{Var}(Y_{i0}) &= \mathrm{Var}(\tilde{\mathbf{x}}_{i0}'\boldsymbol{\beta} + u_i + \delta_{i0}) \\
&= \sigma_u^2 + \frac{\sigma^2}{1-\rho^2} \\
\mathrm{Var}(Y_{i1}) &= \mathrm{Var}(\rho Y_{i0} + \tilde{\mathbf{x}}_{i1}'\boldsymbol{\beta} + u_i + \delta_{i1}) \\
&= \rho^2 \left(\sigma_u^2 + \frac{\sigma^2}{1-\rho^2} \right) + \sigma_u^2 + \sigma^2 + 2\rho \mathrm{Cov}(Y_{i0}, u_i) \\
\mathrm{Var}(Y_{i2}) &= \mathrm{Var}(\rho Y_{i1} + \tilde{\mathbf{x}}_{i2}'\boldsymbol{\beta} + u_i + \delta_{i2}) \\
&= \rho^2 \mathrm{Var}(Y_{i1}) + \sigma_u^2 + \sigma^2 + 2\rho \mathrm{Cov}(Y_{i1}, u_i) \\
&= \rho^4 \left(\sigma_u^2 + \frac{\sigma^2}{1-\rho^2} \right) + (\rho^2+1)(\sigma_u^2 + \sigma^2) + 2\rho^3 \mathrm{Cov}(Y_{i0}, u_i) + 2\rho \mathrm{Cov}(Y_{i1}, u_i) \\
&\dots \\
\mathrm{Var}(Y_{it}) &= \rho^{2t} \left(\sigma_u^2 + \frac{\sigma^2}{1-\rho^2} \right) + \sum_{k=0}^{t-1} \rho^{2k}(\sigma_u^2 + \sigma^2) + 2\sum_{k=0}^{t-1} \rho^{2k+1} \mathrm{Cov}(Y_{i,t-1-k}, u_i) \\
&= \rho^{2t} \left(\sigma_u^2 + \frac{\sigma^2}{1-\rho^2} \right) + \frac{1-\rho^{2t}}{1-\rho^2}(\sigma_u^2 + \sigma^2) + 2\sum_{k=0}^{t-1} \rho^{2k+1} \mathrm{Cov}(Y_{i,t-1-k}, u_i) \\
&= \rho^{2t} \left(\sigma_u^2 + \frac{\sigma^2}{1-\rho^2} \right) + \frac{1-\rho^{2t}}{1-\rho^2}(\sigma_u^2 + \sigma^2) + 2\sigma_u^2 \sum_{k=0}^{t-1} \rho^{2k+1} \left(\frac{1-\rho^{t-k}}{1-\rho} \right), \qquad (3.17)
\end{aligned}$$

for $t \geq 1$.

Lastly, we derive the marginal covariance, $\mathrm{Cov}(Y_{it}, Y_{j,t+\ell})$. Because the δ_{it}'s are independent, $\mathrm{Cov}(Y_{it}, Y_{j,t+\ell}) = 0$ when $i \neq j$. On the other hand, when $i = j$, for $\ell = 1, 2, \ldots$ and $t \geq 0$,

$$\begin{aligned}
\mathrm{Cov}(Y_{it}, Y_{i,t+1}) &= \mathrm{Cov}(Y_{it}, \rho Y_{it} + \tilde{\mathbf{x}}'_{i,t+1}\boldsymbol{\beta} + u_i + \delta_{i,t+1}) \\
&= \rho \mathrm{Var}(Y_{it}) + \mathrm{Cov}(Y_{it}, u_i) \\
\mathrm{Cov}(Y_{it}, Y_{i,t+2}) &= \mathrm{Cov}(Y_{it}, \rho Y_{i,t+1} + \tilde{\mathbf{x}}'_{i,t+2}\boldsymbol{\beta} + u_i + \delta_{i,t+2}) \\
&= \rho \mathrm{Cov}(Y_{it}, Y_{i,t+1}) + \mathrm{Cov}(Y_{it}, u_i) \\
&= \rho^2 \mathrm{Var}(Y_{it}) + \rho \mathrm{Cov}(Y_{it}, u_i) + \mathrm{Cov}(Y_{it}, u_i) \\
&\cdots \\
\mathrm{Cov}(Y_{it}, Y_{i,t+\ell}) &= \rho^\ell \mathrm{Var}(Y_{it}) + \sum_{k=0}^{\ell-1} \rho^k \mathrm{Cov}(Y_{it}, u_i) \\
&= \rho^\ell \mathrm{Var}(Y_{it}) + \left(\frac{1-\rho^\ell}{1-\rho}\right) \mathrm{Cov}(Y_{it}, u_i). \qquad (3.18)
\end{aligned}$$

Therefore, if $t = 0$, $\mathrm{Cov}(Y_{i0}, Y_{i\ell}) = \frac{1-\rho^{\ell+1}}{1-\rho}\sigma_u^2 + \frac{\rho^\ell}{1-\rho^2}\sigma^2$, and, if $t \geq 1$,

$$\begin{aligned}
\mathrm{Cov}(Y_{it}, Y_{i,t+\ell}) = \rho^\ell &\left[\rho^{2t}\left(\sigma_u^2 + \frac{\sigma^2}{1-\rho^2}\right) + \frac{1-\rho^{2t}}{1-\rho^2}(\sigma_u^2 + \sigma^2) + 2\sigma_u^2 \sum_{k=0}^{t-1} \rho^{2k+1}\left(\frac{1-\rho^{t-k}}{1-\rho}\right)\right] \\
&+ \left(\frac{1-\rho^\ell}{1-\rho}\right)\sigma_u^2\left(\frac{1-\rho^{t+1}}{1-\rho}\right). \qquad (3.19)
\end{aligned}$$

To summarize, $\mathbf{Y}_i$ is multivariate normal distributed with mean, variance, and covariance structures given by (3.15), (3.17), and (3.19), respectively. We will refer to this formulation of the model as the *marginal form of the ARR model.*

3.3.1 ARR model visualizations

Unlike in the case of the LME and ARE models, the mean and variance-covariance structures of the ARR model are complex and difficult to interpret. Therefore, in this section, we illustrate these structures using simulated data. Specifically, we simulate transformed income from a sample of women, half of which are breast cancer survivors and half of which are controls. All women are assigned the same year of birth (1954), and the survivors are assigned the same age at diagnosis (49). These values are based on the average birth year and age at diagnosis in the data set described in Chapter 2. Specifying the same birth year and year of diagnosis for all women—and considering controls as well as survivors—allow the clear illustration of the cancer effect.

Transformed income is simulated from an ARR model for the sample of women for all years between the ages 18 and 60. This model contains an intercept, a birth year effect (`birthYear`), a linear effect of time (`yearsSince18`), and a cancer effect (for survivors only) that begins at diagnosis and is constant in the years following. The values for the regression coefficients are informed by our data set (a detailed analysis of which is presented in Section 4.3) and are listed in Table 3.2. We use four different values of ρ (0.1, 0.3, 0.7, and 0.9) to illustrate how the mean and variance of transformed income change with ρ. These values were chosen to span a realistic range for ρ; negative values of ρ are theoretically possible but, in practice, have not been reported in the literature. The value of $\rho = 0.7$ is close to the estimated value we report in Section 4.3 and can be considered the most realistic for the context of this project.

Variable	β
Intercept	-58.572
`birthYear`	0.032
`yearsSince18`	0.030
Cancer	-0.350

Table 3.2: Regression coefficients used in the simulation of income data from an ARR model.

Figure 3.1 shows the average transformed income for ages 18 through 60, with transformed incomes simulated from the ARR model with various values of ρ, as described. All parameters other than ρ are held constant among the four plots. The year of the cancer diagnosis is marked by a vertical black line. Larger values of ρ lead to a larger yearly increase in transformed income (the scales on the vertical axes of the four graphs are different).

These four plots illustrate several properties of the ARR model that were introduced in Section 3.3. First, as we will demonstrate theoretically in Example 2 of Section 3.3.2, the overall trend of the mean transformed income involves a steep increase at small t. The rate of yearly change of mean transformed income then decreases, approaching a constant as t increases. The rate of convergence to this constant depends on the value of ρ. This feature of the transformed incomes can be explained by the fact that, as t increases, the coefficient of the time trend approaches an asymptote, so the rate of change in mean transformed income is approximately linear for large t.

We can clearly see that, although the conditional form of the ARR model (see (3.13)) could appear, at first glance, to imply a linear relationship between mean transformed income and time, this relationship is, in fact, non-linear, particularly for small t or large ρ.

The effect of cancer is visible in Figure 3.1 as the vertical distance between the control and survivor trends. As shown in (3.15) and explored more thoroughly in Section 3.3.2, the effect of cancer is a function of both ρ and t. However, due to the differing scales of the four

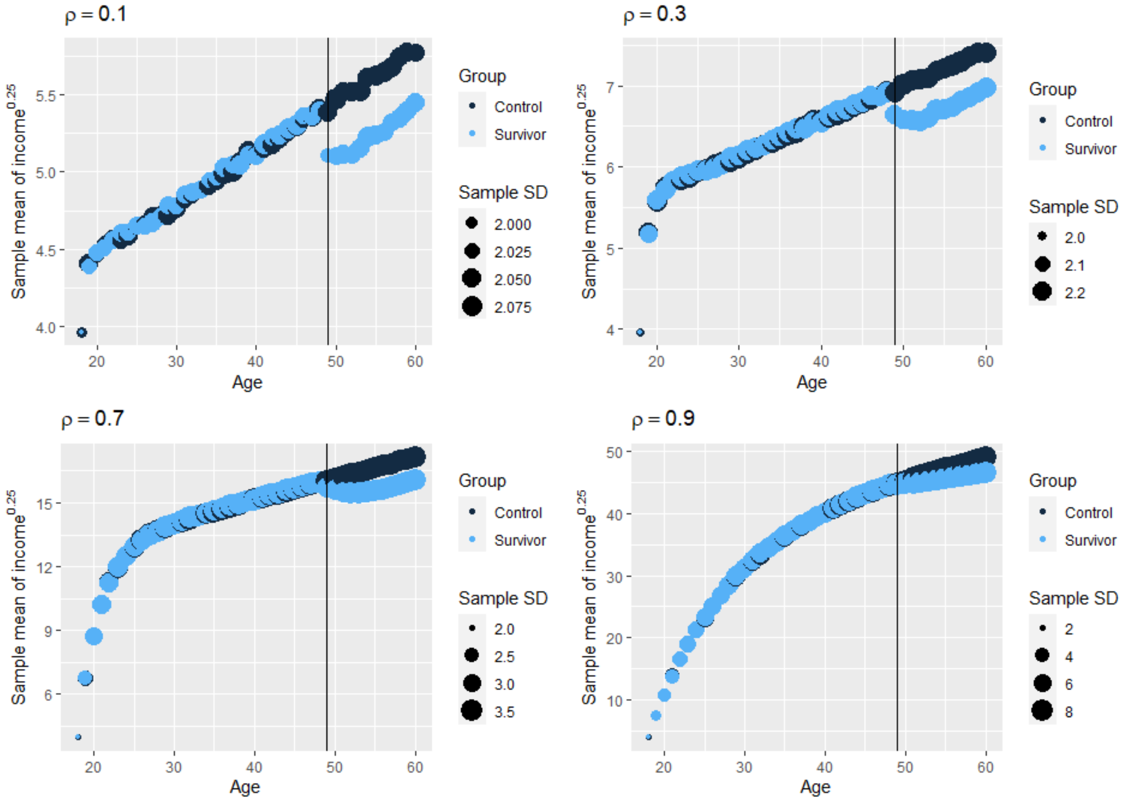

Figure 3.1: Sample mean transformed income by age for various values of ρ. The relative magnitudes of the sample SDs of transformed income for each age/group are reflected in the size of the points. The vertical black lines represent the year of cancer diagnosis.

plots, this feature is difficult to see. To show more clearly that the cancer effect is changing over time, Figure 3.2 is the same as the plot of $\rho = 0.7$ from Figure 3.1, but focuses on ages 47 (2 years prior to diagnosis) to 60. We can see that the effect of cancer increases as t increases, and, at age 60, the effect magnitude is larger than in the year of diagnosis. In the years just after the year of diagnosis, the effect magnitude clearly changes each year. In later years, the effect magnitude approaches an asymptote and changes very little with t; so, for example, the effect size at age 59 appears to be the same as the effect size at age 60.

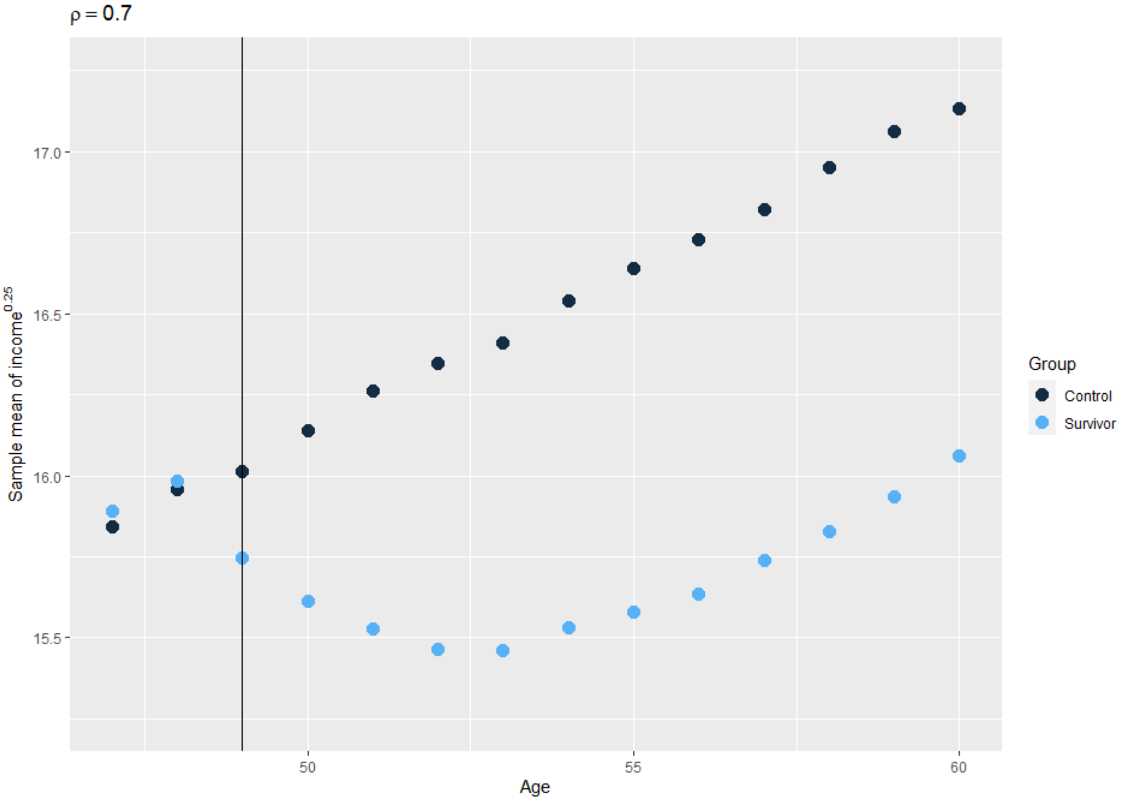

Figure 3.2: Sample mean transformed income by age for $\rho = 0.7$, focused on the years surrounding diagnosis. The vertical black line represents the year of cancer diagnosis.

Figure 3.1 also illustrates that the SD of transformed income increases with both time and ρ (σ_u^2 and σ^2 are held constant in all four plots). Additionally, larger values of ρ cause the SD to increase more over time; for example, when $\rho = 0.1$ the SD of transformed income grows from 1.99 at age 18 to 2.06 at age 60, but when $\rho = 0.9$, the SD grows from 1.99 to 8.16.

Figure 3.1 illustrates the important point that, for the ARR model to capture the trend in real income data with respect to age (see Figure 2.3), time 0 needs to correspond to the age at which the rate of change in income is greatest, i.e., age 18. In many applications, (e.g., Bond (2002), Cai et al. (2014), and Zhang and Zhou (2019)) time 0 is defined as the first observed time point, which may occur at different ages for different subjects. If we had followed these authors and chosen to define $t = 0$ as the first observed time point (which, in our application, tends to occur at ages much greater than 18), then we would have been assuming severe curvature in transformed income observations over ages when the trend is, in fact, approximately linear.

We also simulated data for multiple values of σ_u^2 (which governs the among-individual variation in income) and σ^2 (which governs the within-individual variation in income). The income trends (not shown) behaved as expected; larger values of both parameters lead to greater SD of income.

3.3.2 Interpretation of some features of the ARR model

In this section, we consider several perspectives on the interpretation of the ARR model (3.14). We first discuss the effects of the explanatory variables (both time invariant and time varying) on the mean response. We then describe the marginal variance-covariance structure.

The regression coefficients in the ARR model have different interpretations on the marginal and conditional levels. Consider a time-invariant, continuous variable, for example `birthYear`, which has the coefficient β_1. Then, when we condition on the previous year's income, β_1 has the usual interpretation as the expected change in transformed income that occurs when `birthYear` is increased by one unit and all other variables, including the transformed income in the preceding year, remain unchanged. However, in the marginal model, the coefficient of `birthYear` is $\phi(t) = \frac{1-\rho^{t+1}}{1-\rho}\beta_1$. Thus, the effect of `birthYear` varies over time as a function of ρ, and the conditional and marginal effects on mean transformed income, $\mathrm{E}(Y_{it})$, coincide only when $t = 0$. The coefficient $\phi(t)$ becomes more interpretable for large t. Specifically, as $t \to \infty$, then $\phi(t)$ converges to $\phi \equiv \beta_1 \frac{1}{1-\rho}$.

Figure 3.3 shows how $\phi(t)$ changes as a function of time for years 0 to 10 for various values of ρ and $\beta_1 = 0.5$. For all values of ρ, $\phi(0) = \beta_1$. When ρ is positive, the magnitude of $\phi(t)$ increases monotonically to its limit, ϕ, where $|\phi| > |\beta_1|$. When ρ is negative, $|\phi| < |\beta_1|$, i.e., the magnitude of $\phi(t)$ when t is large is assumed to be less than its magnitude at the initial time point. However, the decrease in the magnitude of $\phi(t)$ is not monotonic. To provide a specific example, suppose $\beta_1 = 0.5$ and $\rho = 0.7$. Then the effect of `birthYear` on mean transformed income is 0.5 when $t = 0$ but is approximately 1.63 at $t = 10$. However, if $\beta_1 = 0.5$ and $\rho = -0.7$, then the effect of `birthYear` at $t = 0$ is 0.5 and at $t = 10$ is approximately 0.3.

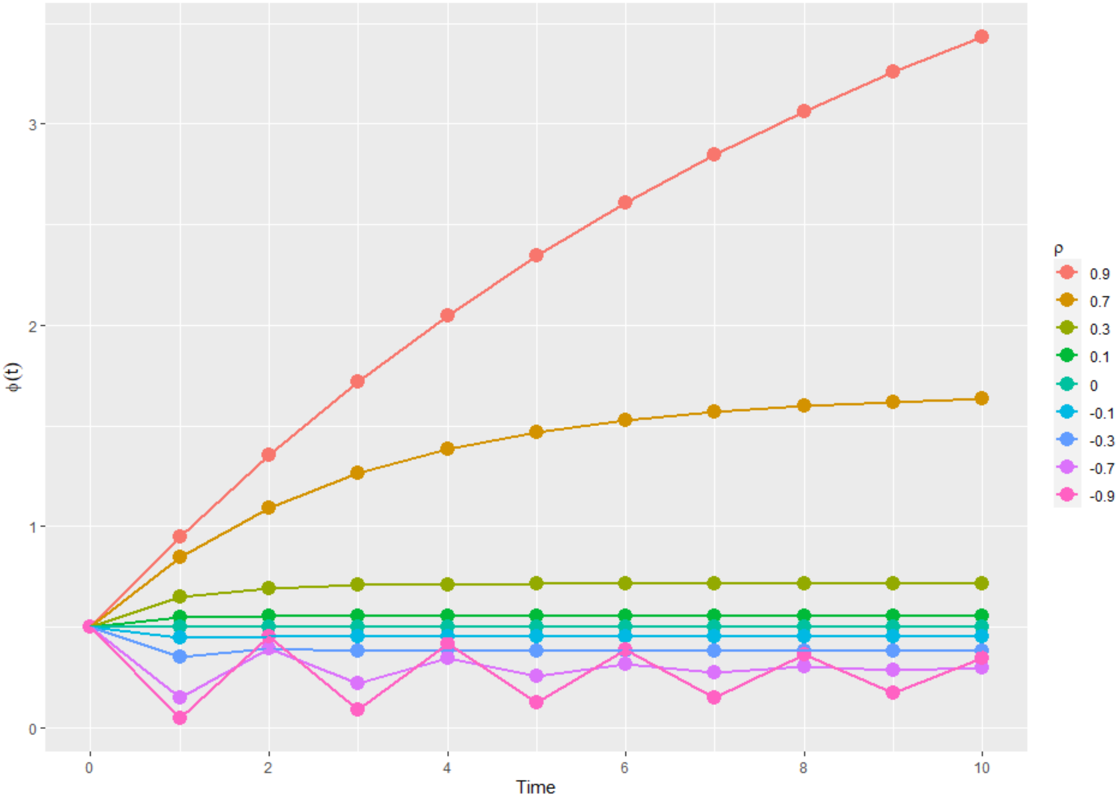

Figure 3.3: The coefficient, $\phi(t)$, of a time-invariant explanatory variable in the marginal form of the ARR model versus time for various ρ and $\beta_1 = 0.5$. Time is a discrete variable; however, curves are shown on the plot to emphasize the trends.

In the case of a time-varying variable, on the conditional level, its effect can be interpreted in the same way as the effect of a time-invariant variable. But the interpretation is more complicated on the marginal level since it depends on previous values of the variable and the particular way in which the variable changes over time. To illustrate, we provide interpretations of three effects of interest in our context.

Effect of `selfEmploy`

Consider the variable `selfEmploy`, represented by x_{it7}, which can be 0 or 1 at any time point. In (3.14), this variable has coefficient β_7. Suppose we have two subjects, i and j, identical up to and including time t except for their values of `selfEmploy` at time t ($x_{it7} = 1$, but $x_{jt7} = 0$). In other words, $x_{i07} = x_{j07}, \ldots, x_{i,t-1,7} = x_{j,t-1,7}$, but $x_{it7} = x_{jt7} + 1$. Under these conditions, $\mathrm{E}(Y_{it}) = \mathrm{E}(Y_{jt}) + \beta_7$. We can then interpret β_7 as the mean difference in transformed income in year t of a subject who is self-employed relative to a subject who is not, assuming that, prior to time t, the two subjects have the same value of all variables.

Effect of `yearsSince18`

Now consider the variable `yearsSince18`, which is represented by x_{it2} and has coefficient β_2. This variable differs from `selfEmploy` in that it changes systematically with time. Consider a woman, i, whose values of all variables except `yearsSince18` are constant from time 0 to time t. Using the notation introduced in Section 3.3, the change in mean transformed income between t and $t+1$ for woman i is

$$\begin{aligned}
\mathrm{E}(Y_{i,t+1}) - \mathrm{E}(Y_{it}) &= \sum_{k=0}^{t+1} \rho^k \tilde{\mathbf{x}}_{i,t+1-k}\boldsymbol{\beta} - \sum_{k=0}^{t} \rho^k \tilde{\mathbf{x}}_{i,t-k}\boldsymbol{\beta} \\
&= \rho^{t+1}\tilde{\mathbf{x}}_{i0}\boldsymbol{\beta} + \sum_{k=0}^{t} \rho^k \beta_2 (x_{i,t+1-k,2} - x_{i,t-k,2}) \\
&= \rho^{t+1}\tilde{\mathbf{x}}_{i0}\boldsymbol{\beta} + \sum_{k=0}^{t} \rho^k \beta_2 (1) \\
&= \rho^{t+1}\tilde{\mathbf{x}}_{i0}\boldsymbol{\beta} + \frac{1-\rho^{t+1}}{1-\rho}\beta_2
\end{aligned}$$

This yearly change in mean transformed income is a function of t and ρ, and results in the the curvature illustrated in Figure 3.1. However, as $t \to \infty$, $\mathrm{E}(Y_{i,t+1}) - \mathrm{E}(Y_{it}) \to \frac{1}{1-\rho}\beta_2$, i.e., the rate of change is approximately constant for large t.

Total cancer effect

Years since diagnosis	Total cancer effect Conditional	 Marginal
0	$\beta_3+\beta_4+\beta_5+\beta_8 t_d$	$\beta_3+\beta_4+\beta_5+\beta_8 t_d$
1	$\beta_3+\beta_4+\beta_5+\beta_8(t_d+1)$	$(\rho+1)(\beta_3+\beta_4+\beta_5)+$ $\beta_8((t_d+1)+\rho t_d)$
2	$\beta_4+\beta_5+\beta_8(t_d+2)$	$(\rho+\rho^2)\beta_3+(1+\rho+\rho^2)(\beta_4+\beta_5)$ $+\beta_8((t_d+2)+\rho(t_d+1)+\rho^2 t_d)$
3	$\beta_4+\beta_5+\beta_8(t_d+3)$	$(\rho^2+\rho^3)\beta_3+$ $(1+\rho+\rho^2+\rho^3)(\beta_4+\beta_5)+$ $\beta_8((t_d+3)+\rho(t_d+2)+\rho^2(t_d+1)+\rho^3 t_d)$
4	$\beta_4+\beta_5+\beta_8(t_d+4)$	$(\rho^3+\rho^4)\beta_3+$ $(1+\rho+\rho^2+\rho^3+\rho^4)(\beta_4+\beta_5)+$ $\beta_8((t_d+4)+\rho(t_d+3)+\rho^2(t_d+2)+$ $\rho^3(t_d+1)+\rho^4 t_d)$
5	$\beta_4+\beta_5+\beta_8(t_d+5)$	$(\rho^4+\rho^5)\beta_3+$ $(1+\rho+\rho^2+\rho^3+\rho^4+\rho^5)(\beta_4+\beta_5)+$ $\beta_8((t_d+5)+\rho(t_d+4)+\rho^2(t_d+3)+$ $\rho^3(t_d+2)+\rho^4(t_d+1)+\rho^5 t_d)$
6	$\beta_5+\beta_8(t_d+6)$	$(\rho^5+\rho^6)\beta_3+$ $(\rho+\rho^2+\rho^3+\rho^4+\rho^5+\rho^6)\beta_4+$ $(1+\rho+\rho^2+\rho^3+\rho^4+\rho^5+\rho^6)\beta_5+$ $\beta_8((t_d+6)+\rho(t_d+5)+\rho^2(t_d+4)+$ $\rho^3(t_d+3)+\rho^4(t_d+2)+\rho^5(t_d+1)+\rho^6 t_d)$
$\vdots$	$\vdots$	$\vdots$

Table 3.3: Conditional and marginal total cancer effects under the ARR model. The value of `yearsSince18` in the year of diagnosis is represented by t_d.

As mentioned in Section 3.1, the total cancer effect is multifaceted. Interpretation of this effect is even more complicated in the ARR setting due to the differences between the conditional and marginal forms of the model. Consider the ARR model in (3.14). The marginal total cancer effect at a given age (measured in years since 18) can be interpreted as the mean difference between the transformed incomes of two women, identical across all explanatory variables except for cancer status. The conditional total cancer effect has the same interpretation except that the two women's incomes in the year prior to their current age are also assumed to be identical. Table 3.3 gives the conditional and marginal effects for each year post-diagnosis, where t_d represents the value of `yearsSince18` in the year of diagnosis. Clearly, the marginal total effect of cancer is a complicated function of time (measured in years since diagnosis) that also depends strongly on ρ and age at diagnosis.

Variance-Covariance structure

In the marginal ARR model, the variance of the response varies over time according to a relationship determined by ρ, as shown in (3.17). Our preliminary work suggests that, if $\rho > 0$, then the variance increases monotonically over time, but, if $\rho < 0$, the variance does not change monotonically over time. The change in $\text{Var}(Y_{it})$ over time seems to be greater for values of ρ of larger magnitude. Likewise, $\text{Cov}(Y_{it}, Y_{i,t+\ell})$ depends on both t and ℓ. Our initial explorations suggest that, if $\rho > 0$, $\text{Cov}(Y_{it}, Y_{i,t+\ell})$ monotonically increases with t and decreases with ℓ. But, if $\rho < 0$, then $\text{Cov}(Y_{it}, Y_{i,t+\ell})$ is not monotonic in t or ℓ. $\text{Cov}(Y_{it}, Y_{i,t+\ell})$ seems to increase monotonically with ρ when $\rho > 0$ and decrease monotonically with ρ when $\rho < 0$. We leave a formal proof of these properties to future work.

If $\rho > 0$, as is found in practice in studies of income, then the general increasing/decreasing trends in the variance and covariance functions assumed by the ARR model may be reasonable. However, the reasonableness of the specific forms of the variance and covariance of Y_{it}, particularly their relationship with ρ, would be difficult to determine in a given setting.

3.4 Model comparison

In Section 3.1, we introduced three models: the LME, ARE, and ARR models. The LME model, which assumes that, given the woman-specific random effects, incomes are independent, provides a baseline for comparison to the ARE and ARR models. In contrast, the ARE and ARR models allow for annual income to be autocorrelated over time, even when conditioning on the random effect. But these models differ in how they incorporate autocorrelation.

The LME and ARE models have a marginal form that resembles their conditional form. The effects of the explanatory variables are assumed to be constant over time, and the variance and covariance of the transformed incomes are not functions of t. Interpreting these models is thus straightforward.

On the other hand, the ARR model has an unusual marginal form that reveals numerous strong underlying assumptions that may be difficult to justify in practice. In particular, the mean, variance, and covariance are assumed to be functions of ρ and t. The effects of all explanatory variables are functions of ρ and t, which makes interpretation difficult. We must consider whether these assumptions are realistic. In the case of transformed income and the data in Chapter 2, we did observe curvature in the mean at early ages, and carefully chose a definition of time so that the ARR model could match the curvature. We also observed that the variance of income increases over time. However, we question the reasonableness of the assumption that the effects of all explanatory variables (not just `yearsSince18`) would change as a function of time and ρ.

If the ARR model is to be used, given that the marginal and conditional effects differ in their interpretation, careful consideration of which effect is of interest is required. In our case, we would like to capture the cumulative effect of a cancer diagnosis at a time t. The conditional effect is the effect of cancer at time t, given the previous year's income; if cancer had an effect on income at time $t-1$, then that effect is not captured in the conditional effect. Therefore, the marginal effect is of greater interest in our setting. However, in the literature, focus tends to rest almost exclusively on the conditional form, while the marginal form and its interpretation are ignored, for example, Greene (2003), Baltagi (2008), and Wooldridge (2002). The conditional effect of the variable of interest is often presented in a way that could imply that it is a marginal effect, i.e., the estimated coefficient is often described as the variable's estimated effect without specifying that it is conditional on the previous observation (see, for example, Karaivanov et al. (2020) and Zhang and Zhou (2019)). We believe that the marginal model is integral to the interpretation of this model and that it should be considered whenever the ARR model is applied.

3.5 Estimation

We use the method of restricted maximum likelihood (REML) to estimate the parameters of the ARE and LME models and the method of maximum likelihood (ML) to estimate the parameters of the ARR model. In this section, we describe the implementation of these methods and the properties of the resulting parameter estimates.

3.5.1 Estimation of LME and ARE models

REML is used to estimate the parameters of both the LME and ARE models. REML estimates are obtained using the `nlme::lme` function in `R`, which implements the Expectation-Maximization (EM) algorithm as described in Laird and Ware (1982). The REML estimates of the regression coefficients, $\hat{\boldsymbol{\alpha}}$, are consistent and asymptotically efficient (Verbeke and Molenberghs, 2000). Complete data is not required; if either $\mathbf{x}_{it}$ or Y_{it} has a missing value, then `R`, by default, removes the entire record for woman i at time t.

3.5.2 Estimation of ARR model

In the economics literature, generalized method of moments (GMM) estimation is commonly used for fitting the autoregressive response model (for example, Bond (2002)). GMM estimation produces consistent estimates. However, the efficiency of the GMM estimators depends on the number of moment conditions used (Hsiao, 2014). We prefer the ML estimation procedure, which generates consistent and (presumably) more efficient estimates (Hsiao, 2014).

For efficient computation of the ML estimates, rewriting the marginal form of the model in matrix form is convenient. Suppose woman i is observed at time points $0, \ldots, T_i$. Let the number of observations on woman i be $n_i = T_i + 1$. As per Funatogawa and Funatogawa (2019), let $\mathbf{B}_i = (I_{n_i} - \rho \mathbf{F}_i)^{-1}$, where I_{n_i} is the $n_i \times n_i$ identity matrix and $\mathbf{F}_i$ is the $n_i \times n_i$ matrix

$$\mathbf{F}_i = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ 1 & 0 & \ldots & 0 \\ \cdot & \cdot & \ldots & \cdot \\ 0 & \ldots & 1 & 0 \end{bmatrix}.$$

Define $\tilde{\mathbf{X}}_i$ as the $n_i \times p$ matrix containing the observations of the explanatory variables and $\mathbf{z}_i$ as an n_i-dimensional vector of ones. Then

$$\mathbf{Y}_i = \mathbf{B}_i(\tilde{\mathbf{X}}_i \boldsymbol{\beta} + \mathbf{z}_i u_i + \boldsymbol{\delta}_i). \tag{3.20}$$

Letting $\boldsymbol{\delta}_i \sim MVN(\mathbf{0}, \sigma^2 I_{n_i})$, $\mathbf{Y}_i$ has a multivariate normal distribution with mean

$$\mathrm{E}(\mathbf{Y}_{\mathrm{i}}) = \mathbf{B}_i \tilde{\mathbf{X}}_i \boldsymbol{\beta} \tag{3.21}$$

and variance-covariance matrix

$$\boldsymbol{\Sigma}_i = \mathrm{Var}(\mathbf{Y}_i) = \mathbf{B}_i(\sigma_u^2 \mathbf{z}_i' \mathbf{z}_i + \sigma^2 I_{n_i}) \mathbf{B}_i'. \tag{3.22}$$

Equations (3.21) and (3.22) are equivalent to (3.15), (3.17), and (3.19). We emphasize that $\boldsymbol{\Sigma}_i$ depends on i only through its dimensions ($n_i \times n_i$), i.e., it does not depend on the explanatory variables.

Let $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N)$ be the vector of all responses. From (3.21) and (3.22), the log-likelihood can be determined:

$$\ell(\boldsymbol{\beta}, \rho, \sigma_u^2, \sigma^2 | \mathbf{Y}) = -\frac{1}{2} \sum_{i=1}^{N} \left\{ n_i \log(2\pi) + \log(|\boldsymbol{\Sigma}_i|) + (\mathbf{Y}_i - \mathbf{B}_i \tilde{\mathbf{X}}_i \boldsymbol{\beta})' \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \mathbf{B}_i \tilde{\mathbf{X}}_i \boldsymbol{\beta}) \right\}. \tag{3.23}$$

Due to the autoregressive response term in the conditional model, the mean of Y_{it} is a function of $\tilde{\mathbf{x}}_{i0}, \ldots, \tilde{\mathbf{x}}_{it}$, i.e., to calculate the likelihood, complete data on the explanatory variables is required at all time points. However, calculating the likelihood does not require complete data on $\mathbf{Y}_i$; if Y_{it} is missing, then the likelihood based on the observed responses can be calculated by deleting the row corresponding to time t in $\mathbf{B}_i\tilde{\mathbf{X}}_i$ and the row and column corresponding to time t in $\boldsymbol{\Sigma}_i$.

The parameters ρ, σ_u^2, and σ^2 have restricted ranges. Rather than using a constrained optimization method to maximize (3.23), we reparameterize the model so that all parameters have a range of $\mathbb{R}$. The following transformations were used:

$$\begin{aligned} \rho^* &= \log\left(\frac{1+\rho}{1-\rho}\right) \\ \lambda_u &= \log\sigma_u^2 \\ \lambda &= \log\sigma^2 \end{aligned}$$

We then maximize the likelihood numerically to find $\hat{\boldsymbol{\beta}}, \hat{\rho}^*, \hat{\lambda}_u,$ and $\hat{\lambda}$. For this purpose, we use the `optim` function in `R` and the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method, a quasi-Newton method.

An alternative to direct optimization is an iterative procedure. Define $\boldsymbol{\psi} = (\rho^*, \lambda_u, \lambda)$. Because a closed-form solution for $\hat{\boldsymbol{\beta}}$ exists if $\boldsymbol{\psi}$ is treated as fixed, we can assume a value for $\boldsymbol{\psi}$, compute a closed form estimate of $\boldsymbol{\beta}$ based on the assumed value of $\boldsymbol{\psi}$, and then maximize the likelihood with respect to $\boldsymbol{\psi}$ given our current estimate of $\boldsymbol{\beta}$. We can then repeat these steps, updating our estimate of $\boldsymbol{\psi}$ and $\boldsymbol{\beta}$ at each iteration.

Specifically, using the subscript (k) to denote the values of the parameters at iteration k, our iterative estimation algorithm is as follows:

1. Select initial values, $\boldsymbol{\psi}_{(0)}$.

2. At iteration k, use $\boldsymbol{\psi} = \boldsymbol{\psi}_{(k)}$ and solve for $\boldsymbol{\beta}_{(k)}$ as

$$\boldsymbol{\beta}_{(k)} = \left[\sum_{i=1}^{N}(\mathbf{B}_{i(k)}\tilde{\mathbf{X}}_i)'\boldsymbol{\Sigma}_{i(k)}^{-1}(\mathbf{B}_{i(k)}\tilde{\mathbf{X}}_i)\right]^{-1}\sum_{i=1}^{N}(\mathbf{B}_{i(k)}\tilde{\mathbf{X}}_i)'\boldsymbol{\Sigma}_{i(k)}^{-1}\mathbf{Y}_i. \tag{3.24}$$

3. Use a numerical optimization procedure to minimize $-\ell(\boldsymbol{\beta}_{(k)}, \boldsymbol{\psi}_{(k+1)}|\mathbf{Y})$ with respect to $\boldsymbol{\psi}_{(k+1)}$, treating $\boldsymbol{\beta}_{(k)}$ as fixed.

4. Repeat Steps 2 and 3 until $|\ell(\boldsymbol{\beta}_{(k+1)}, \boldsymbol{\psi}_{(k+1)}|\mathbf{Y}) - \ell(\boldsymbol{\beta}_{(k)}, \boldsymbol{\psi}_{(k)}|\mathbf{Y})| < \nu$, where ν is a pre-specified tolerance level.

We wrote our own code to implement this algorithm. The `optim` function with the BFGS method is used in Step 3.

In our experience, when the number of observations per subject is small and the sample size is large, the iterative procedure has a shorter computation time. However, when the number of observations per subject is large, the direct optimization approach is faster. Ultimately, we used direct optimization for the analysis in Section 4.3 due to the large number of observations per subject in that setting.

After the ML estimates of the parameters, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\psi}}$, are obtained (by either procedure), the Hessian, H, of the log-likelihood can be determined. Let $V = -H^{-1}$. Then the diagonal elements of $\sqrt{V}$ provide standard errors for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\psi}}$. The delta method can be used to find standard errors of $\hat{\rho}$, $\hat{\sigma}_u^2$, and $\hat{\sigma}^2$.

We use the following relationships in the derivation of the standard errors of $\hat{\rho}, \hat{\sigma}_u^2$, and $\hat{\sigma}^2$:

$$\begin{aligned}
\rho &= \frac{e^{\rho^*} - 1}{e^{\rho^*} + 1} \\
\frac{\partial \rho}{\partial \rho^*} &= \frac{2e^{\rho^*}}{(1 + e^{\rho^*})^2} \\
\sigma_u^2 &= e^{\lambda_u} \\
\frac{\partial \sigma_u^2}{\partial \lambda_u} &= e^{\lambda_u} \\
\sigma^2 &= e^{\lambda} \\
\frac{\partial \sigma^2}{\partial \lambda} &= e^{\lambda}
\end{aligned}$$

Let the matrix of derivatives of $\boldsymbol{\psi}$ be

$$\triangledown \boldsymbol{\psi} = \begin{pmatrix} \frac{\partial \rho}{\partial \rho^*} & \frac{\partial \rho}{\partial \lambda_u} & \frac{\partial \rho}{\partial \lambda} \\ \frac{\partial \sigma_u^2}{\partial \rho^*} & \frac{\partial \sigma_u^2}{\partial \lambda_u} & \frac{\partial \sigma_u^2}{\partial \lambda} \\ \frac{\partial \sigma^2}{\partial \rho^*} & \frac{\partial \sigma^2}{\partial \lambda_u} & \frac{\partial \sigma^2}{\partial \lambda} \end{pmatrix}.$$

Define $\widehat{\triangledown \boldsymbol{\psi}} = \triangledown \boldsymbol{\psi}|_{\psi = \hat{\psi}}$, and let V_ψ be the 3×3 submatrix of V containing only the elements of V corresponding to the estimated variances and covariances of $\hat{\rho}$, $\hat{\lambda}_u$, and $\hat{\lambda}$. Applying the delta method, the estimated variance-covariance matrix of $(\hat{\rho}, \hat{\sigma}_u^2, \hat{\sigma}^2)$ is $V_\psi^* = \widehat{\triangledown \boldsymbol{\psi}}' V_\psi \widehat{\triangledown \boldsymbol{\psi}}$, where

$$\widehat{\bigtriangledown\boldsymbol{\psi}} = \begin{pmatrix} \frac{2e^{\hat{\rho}^*}}{(1+e^{\hat{\rho}^*})^2} & 0 & 0 \\ 0 & e^{\hat{\lambda}_u} & 0 \\ 0 & 0 & e^{\hat{\lambda}} \end{pmatrix}.$$

The diagonal elements of V_{ψ}^* are the estimated variances of $\hat{\rho}, \hat{\sigma}_u^2$, and $\hat{\sigma}^2$.

The estimate of a marginal effect of a explanatory variable and of its SD can be similarly computed. For example, consider a time-invariant variable such as `birthYear` with coefficient β_1 in the conditional model. At time t, the coefficient of `birthYear` in the marginal model is

$$g = \frac{1-\rho^{t+1}}{1-\rho}\beta_1 = \frac{1-\left(\frac{e^{\rho^*}-1}{e^{\rho^*}+1}\right)^{t+1}}{1-\left(\frac{e^{\rho^*}-1}{e^{\rho^*}+1}\right)}\beta_1.$$

Let V_g be the 2×2 dimensional submatrix of V that contains only the elements of V corresponding to the variances of $\hat{\beta}_1$ and $\hat{\rho}$ and their covariance. Define $\widehat{\bigtriangledown g} = \bigtriangledown g|_{(\beta_1,\rho)=(\hat{\beta}_1,\hat{\rho})}$. Again applying the delta method, the estimated variance of g is $\widehat{\bigtriangledown g}' V_g \widehat{\bigtriangledown g}$, where

$$\begin{aligned} \bigtriangledown g &= \begin{pmatrix} \frac{\partial g}{\partial \beta_1} \\ \frac{\partial g}{\partial \rho^*} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial g}{\partial \beta_1} \\ \left(\frac{\partial g}{\partial \rho}\right)\left(\frac{\partial \rho}{\partial \rho^*}\right) \end{pmatrix} \\ &= \begin{pmatrix} \frac{1-\left(\frac{e^{\rho^*}-1}{e^{\rho^*}+1}\right)^{t+1}}{1-\left(\frac{e^{\rho^*}-1}{e^{\rho^*}+1}\right)} \\ \beta_1 \left(\frac{\left(1-\left(\frac{e^{\rho^*}-1}{e^{\rho^*}+1}\right)\right)^t - t\left(\frac{e^{\rho^*}-1}{e^{\rho^*}+1}\right)^t \left(1-\left(\frac{e^{\rho^*}-1}{e^{\rho^*}+1}\right)\right)}{\left(1-\left(\frac{e^{\rho^*}-1}{e^{\rho^*}+1}\right)\right)^2} \right) \left(\frac{2e^{\rho^*}}{(1+e^{\rho^*})^2}\right) \end{pmatrix}. \end{aligned}$$

Chapter 4

Application to breast cancer survivor data

In this section, we apply the models described in Chapter 3 to the data described in Chapter 2.

First, we discuss our choice to model income$^{0.25}$ rather than income. A well-known characteristic of income is that its distribution tends to be heavily right-skewed, and our sample is no exception. Since the models we use to analyze the income trends assume normality of the response, we opted to apply a transformation. Initially, we planned to use the natural log transformation, as is common in econometrics. However, after conducting initial analyses using this transformation, we observed that it did not completely remove the skewness in income. We proceeded to search for a different transformation and found that income$^{0.25}$ has an approximately normal distribution.

4.1 Linear mixed effects model

Although we do not expect the LME model to be realistic, it provides a simple way of determining some key features that a more realistic model should possess.

One important use of the LME model in this setting is as an aid in determining the mean structure, particularly the trend in transformed income over age. As shown in Figure 2.3, this trend is non-linear, particularly at younger ages. By considering plots of the residuals from various LME models, we determined that the trend of transformed income over age can be reasonably captured by including linear, quadratic, cubic, and quartic effects of age in the mean structure.

We can also use the LME model to visualize the effect of a cancer diagnosis. The effect of cancer is not visible in a plot of mean transformed income over age because the women are diagnosed at different ages, i.e., the effect of cancer does not begin at one particular point on the horizontal axis. As an alternative, we fit a LME model with all variables except the cancer effects and plot the residuals versus the number of years since diagnosis. Specifically, this model includes the following variables: `yearsSince18`, $\texttt{yearsSince18}^2$, $\texttt{yearsSince18}^3$ and $\texttt{yearsSince18}^4$, `birthYear`, `spouseStatus`, and `selfEmploy`. The residuals are calculated as

$$e_{it} = Y_{it} - (\mathbf{x}'_{it}\hat{\boldsymbol{\alpha}} + \hat{u}_i),$$

where the $\hat{u}_i$'s are the best linear unbiased predictions of the u_i's (Pinheiro et al.).

This plot appears in Figure 4.1. Due to reasons of confidentiality, we cannot show a plot of the individual residuals from this model. Figure 4.1 thus shows summary statistics of the residuals in each year since diagnosis. The vertical lines travel from the 90th percentile to the 3rd quartile and from the 1st quartile to the 10th percentile. The triangular and circular points represent the sample medians and means, respectively. Due to the low number of women observed 21 through 23 years post-diagnosis, the residuals from these years had to be combined (as labelled on the plot).

The sample mean and recorded percentiles of the residuals decrease noticeably at the year of diagnosis and remain lowered for approximately 5 years post-diagnosis. This effect of cancer, while visible, is small compared to the total variation in the residuals, which is large (despite having adjusted for several explanatory variables).

To incorporate the effect of cancer in the model, we add the `cancerImmediate`, `cancerShort`, and `cancerLong` variables to the mean structure. We also include the interactions `yearsSince18`×`cancerLong`, `spouseStatus`×`cancerLong`, and `selfEmploy`×`cancerLong` (this LME model is given in (3.1)). The interaction `yearsSince18`×`cancerLong` is included to allow the rate of change of transformed income to be different before and after a cancer diagnosis. The `spouseStatus`×`cancerLong` and `selfEmploy`×`cancerLong` interactions are included to allow women with spouses or women who receive income from self-employment, respectively, to experience a different change in mean income post-diagnosis. We refer to the LME model with this mean structure as fully adjusted. The estimates of the parameters of this model are shown in Table A.1 of Appendix A.

To evaluate the mean structure of the fully adjusted LME model, we plot the residuals against year since diagnosis and against `birthYear` (not shown). These plots showed no apparent trends and the residuals are approximately symmetric about 0. We find no evidence against the final chosen mean structure.

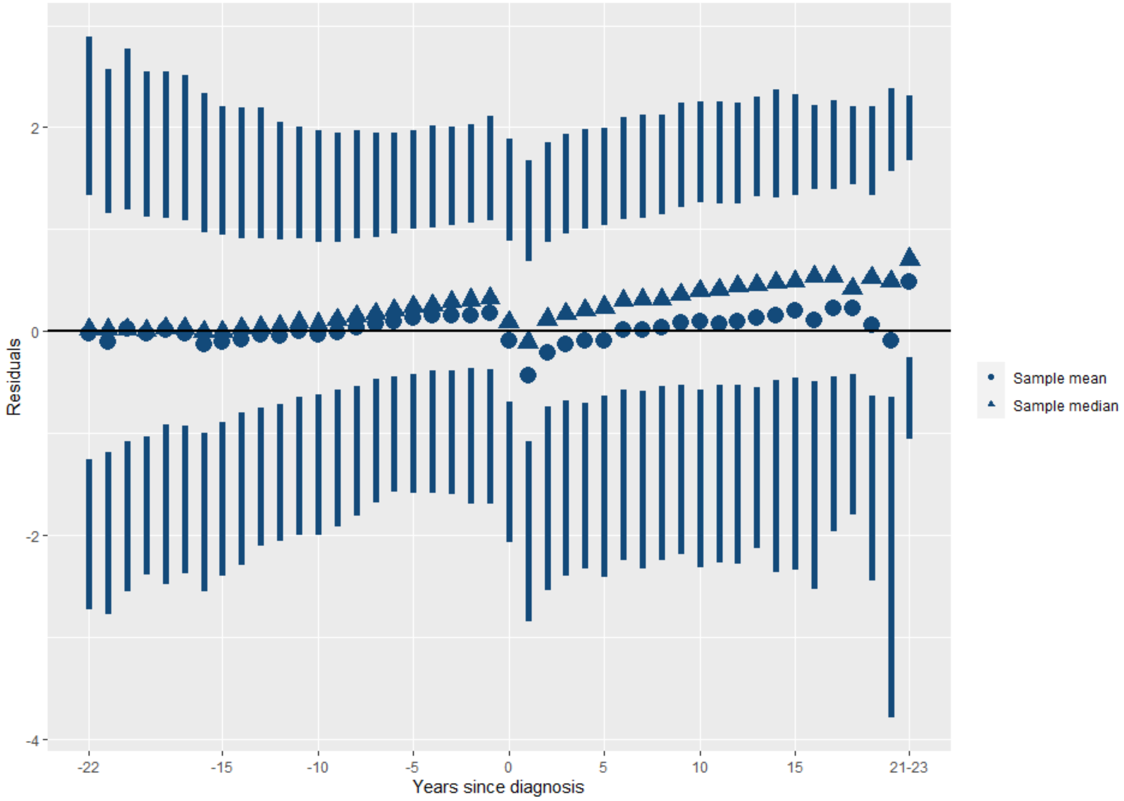

Figure 4.1: Residuals from the LME model without cancer effects. The vertical lines travel from the 90th percentile to the 3rd quartile and from the 1st quartile to the 10th percentile. The triangular and circular points represent the sample medians and means, respectively.

Lag	Estimated Autocorrelation
0	1.000
1	0.383
2	0.184
3	0.076
4	0.007
5	−0.044
6	−0.077
7	−0.099
8	−0.113
9	−0.122
10	−0.127
11	−0.131
12	−0.140
13	−0.144

Table 4.1: Average estimated ACF of residuals from fully adjusted LME model.

Finally, we can use the LME model to assess the autocorrelation in incomes observed on a given woman. Specifically, if the residuals of the fully adjusted model are autocorrelated, then the variance-covariance structure of the ARE or ARR models may be better suited to the data. The autocorrelation function (ACF) is commonly used in time series analysis and characterizes the correlation between the residuals for each lag. We estimate the ACF for the residuals of each woman individually, then average these estimates over all women. In this way, we can informally assess the correlation structure of the error term in the LME model and determine whether assuming δ_{it} are IID for all t is appropriate. The estimated ACF, shown in Table 4.1, suggests moderate autocorrelation in the residuals at lag 1 and (possibly) low autocorrelation at lag 2. The negative estimated autocorrelations at higher lags are unexpected and are presumably statistically insignificant. These findings suggest that the ARE or ARR model, which allow the covariance to vary with lag, may provide a better fit to the data.

To summarize, as we found no evidence that the mean structure of the LME model is misspecified, we expect that the estimators of the regression coefficients are consistent (Liang and Zeger, 1986). In contrast, Table 4.1 indicates that the LME model does not properly account for the autocorrelation in the repeated observations on each woman. Since the variance-covariance structure of this model is likely misspecified, the standard errors of the parameter estimates (shown in Table A.1) are likely not consistent (Liang and Zeger, 1986). Therefore, we use the LME model only as a guide for choosing the mean and variance-covariance structures of the ARE and ARR models, not as a basis for formal inference.

Variable	**Estimate**	**Standard error**
`Intercept`	-179.592	5.639
`birthYear`	0.095	0.003
`yearsSince18`	0.479	0.028
`yearsSince18`^{2}	-0.019	0.002
`yearsSince18`^{3}	0.000	0.000
`yearsSince18`^{4}	-0.000	0.000
`cancerImmediate`	-0.148	0.019
`cancerShort`	-0.087	0.023
`cancerLong`	0.138	0.088
`yearsSince18`×`cancerLong`	-0.005	0.002
`spouseStatus`	-0.466	0.020
`spouseStatus`×`cancerLong`	-0.191	0.028
`selfEmploy`	0.352	0.024
`selfEmploy`×`cancerLong`	0.191	0.034

Table 4.2: Estimated regression coefficients and standard errors from the fully adjusted ARE model.

4.2 Autoregressive error model

In this section, we discuss the results of fitting the ARE model with fully adjusted mean structure (described in Section 4.1). Parameter estimates and standard errors are shown in Table 4.2.

4.2.1 Cancer effect

The estimate of the `yearsSince18`×`cancerLong` interaction effect indicates that the growth rate of a woman's income after being diagnosed with cancer is lower than before being diagnosed. However, as explained in Section 3.1, we cannot easily interpret the immediate, short-term, and long-term cancer effects individually. We therefore consider their combined effect over the years post-diagnosis. Table 4.3 shows, for a woman who does not have a spouse and is not self-employed, the estimated expected difference in her mean transformed income if she was diagnosed with breast cancer at the average age of 49 years compared to her mean transformed income if she had never been diagnosed with cancer. Table 3.1 shows how the values in Table 4.3 are calculated.

A cancer diagnosis is associated with a decrease in transformed income that is estimated to be largest in the year following the diagnosis. The total effect of cancer is significant up to 5 years post-diagnosis, after which the effect of cancer is not significant. In the year of diagnosis, the estimated difference between the mean transformed income of a woman diagnosed with cancer at age 49 with no spouse and no self-employment income, and the same woman had she never been diagnosed with cancer, is -0.252.

Years since diagnosis	**Estimated total cancer effect**	**Standard error**
0	−0.252	0.0274
1	−0.257	0.0275
2	−0.114	0.0323
3	−0.119	0.0326
4	−0.124	0.0331
5	−0.129	0.0338
6	−0.047	0.0408
7	−0.052	0.0416
8	−0.057	0.0424
9	−0.062	0.0434
10	−0.067	0.0445

Table 4.3: Total estimated effect of cancer based on the ARE model for a woman who is diagnosed at age 49, does not have a spouse, and is not self-employed.

4.2.2 Other parameter estimates

Birth year has a positive effect on transformed income, meaning that women born later have higher incomes at a given age, on average. (This effect is sometimes called a *cohort effect.*) Transformed income is estimated to increase by 0.095 units when birth year is increased by 1 year, on average.

The cubic and quartic `yearsSince18` terms have non-zero and significant coefficients; however, we are restricted in the precision of the estimates that we present and are allowed to express the coefficients to only 3 decimal places. Therefore, these estimated effects appear to be 0. Due to the scale of the `yearsSince18` variable, $\texttt{yearsSince18}^3$ and $\texttt{yearsSince18}^4$ are very large; more than 3 decimal places are required to quantify the effects of these variables precisely. All we can say, then, about the trend in transformed income over `yearsSince18` is that the linear and cubic terms are positive and the quadratic and quartic terms are negative.

Having a spouse is associated with a decrease in a woman's transformed income, on average, and women with spouses experience a larger decrease in mean transformed income after a cancer diagnosis. If we compare two women, identical except that woman i has a spouse and woman j does not, then, prior to diagnosis, the mean transformed income of woman i is estimated to be 0.466 units less than the transformed income of woman j. After diagnosis, that difference increases to 0.657 units.

Receiving income from self-employment is associated with an increase in mean transformed income, and women with self-employment income experience a smaller change in mean transformed income after a cancer diagnosis. Specifically, if we assume woman i and woman

j differ only by their value of `selfEmploy`, where woman i receives self-employment income and woman j does not, then, prior to diagnosis, the transformed income of woman i is estimated to be 0.352 units greater than the transformed income of woman j, on average. After diagnosis, that difference increases to 0.543 units.

The estimate of the variance of the random effect, σ_u^2, is 5.847, and the estimate of the error variance, σ^2, is 6.345. The autoregressive parameter, γ, is estimated at 0.667. The estimates of the variance parameters are consistent with our previous observation that the variance of the responses is high, both among and within individuals. The estimate of γ is positive, consistent with the lag-1 sample autocorrelation reported in Table 4.1.

4.2.3 Diagnostics

The LME model is nested within the ARE model. Therefore, we can formally compare the two models using a likelihood ratio test. This test requires ML, not REML, estimates, so we refit the models using ML. The test statistic is 96,693 and is (approximately) a random draw from a $\chi^2_{(1)}$ distribution if $\gamma = 0$. The p-value is approximately 0, so we can conclude that the ARE model, which allows for autocorrelation via an AR(1) structure for the error terms, provides a significantly better fit to the data than does the LME model.

We examine several diagnostic plots based on the fully adjusted ARE model (not shown). The plot of the observed versus fitted values suggests that the model describes moderate incomes well. However, fitted values associated with observed values of zero are non-zero and often very large. This discrepancy is unsurprising given that we created a point mass at 0 by converting negative incomes to 0. The left tail of the resulting distribution of transformed income is thus not well described by a normal distribution. Similarly, for large observed values, the fitted values tend to underestimate the actual values. Overall, the plot suggests a lack of fit only in the case of extreme observed values.

We also plot the fitted values versus residuals based on the fully adjusted ARE model, where the residuals are calculated as in Section 4.1. This plot again indicates that points with an observed value of 0 are not handled well by the model. However, ignoring these points, the plot does not indicate lack of fit.

The plots of the residuals versus `birthYear` and the residuals versus `yearsSince18` do not reveal any obvious patterns; the points are approximately symmetric about 0 and have approximately constant vertical spread.

Based on these plots, we conclude that overall the fit of the ARE model is not problematic, but that more careful handling of the negative incomes would presumably lead to a model with improved fit.

4.3 Autoregressive response model

We now fit the ARR model (3.14) to the data. The explanatory variables included in this model are `birthYear, yearsSince18, cancerImmediate, cancerShort, cancerLong`, `spouseStatus`, and `selfEmploy` and the interaction terms `yearsSince18`×`cancerLong`, `spouseStatus`×`cancerLong`, and `selfEmploy`×`cancerLong`. Because the autoregressive response term in the ARR model introduces curvature into the mean structure of the model (as shown in Section 3.3.1), we opted to exclude higher order functions of `yearsSince18`.

The coding of the algorithm to estimate the parameters of the ARR model (see Section 3.5.2) requires that all women be observed at two or more time points. Therefore, women observed at only one time point are removed from the sample; 243,335 total observations on 14,425 women remain.

4.3.1 Missing data

Before fitting the model, we must first manage some missing data issues that, due to the autoregressive response term in this model, are a problem with this model but not with the LME or ARE models. Since time is defined as the number of years since age 18 and we do not observe the women in the sample prior to 1992, the observations from $t = 0$ until the first observed time point are unavailable. Additionally, some women have records with intermittent missingness, i.e., they did not file a tax return in one or more of the years between their first and last observations. The absence of a tax return may be informative (e.g., the woman may not have earned any income that year) or uninformative (e.g., the woman was living out of country that year). Differentiating between these two causes of intermittent missingness is impossible given the information we have available. We assume that values missing before the first observed time point are missing completely at random and that intermittent missing values are missing at random.

For the ARR model, as shown in Section 3.5.2, values of the explanatory variables at all time points between $t = 0$ and the last observation are necessary to calculate the likelihood. The variables `yearsSince18, birthYear, cancerImmediate, cancerShort,` and `cancerLong` are known for all time points. However, some values of `spouseStatus` and `selfEmploy` may be unknown. We use a simplistic imputation approach to fill in these values and acknowledge that a more sophisticated handling of the missing values could lead to improved parameter estimates. Specifically, for the years from age 18 to the first observation, if `spouseStatus` $= 1$ at the first observation, then we impute the value of 1 for all years in which the woman is older than 24 (the median age of first marriage for Canadian women according to Statistics Canada) and impute a value of 0 for the years between ages 18 and 24. If `spouseStatus` $= 0$ at the first observation, we impute a value of 0 for all earlier years. If `selfEmploy` $= 1$ at the first observation, then we impute the value of 1 at all

	Estimate	**Standard error**
`Intercept`	-58.636	3.402
`birthYear`	0.032	8.866×10^{-7}
`yearsSince18`	0.030	5.542×10^{-7}
`cancerImmediate`	-0.304	2.394×10^{-4}
`cancerShort`	0.008	2.339×10^{-4}
`cancerLong`	0.448	2.161×10^{-3}
`yearsSince18`×`cancerLong`	-0.016	1.376×10^{-6}
`spouseStatus`	-0.327	1.499×10^{-4}
`spouseStatus`×`cancerLong`	-0.083	3.206×10^{-4}
`selfEmploy`	0.105	2.602×10^{-4}
`selfEmploy`×`cancerLong`	0.115	6.011×10^{-4}
σ_u^2	0.592	4.62×10^{-4}
σ^2	3.427	8.82×10^{-6}
ρ	0.684	5.31×10^{-5}

Table 4.4: Parameter estimates and standard errors for the ARR model.

earlier years; otherwise we impute the value 0 for all earlier years. For imputing intermittent missing values, we use last value carried forward.

The total number of imputed values for years prior to the first observation is 303,680, and the total number of imputed intermittent missing values is 9,365.

4.3.2 Parameter estimates

We maximize the log-likelihood (3.23) based on the imputed data set. Starting values are required for the parameters; these values are chosen based on the estimates of the parameters of the ARE model. The estimated effects of the explanatory variables in the conditional form of the ARR model ($\hat{\boldsymbol{\beta}}$) are shown in Table 4.4.

Cancer effect

The estimated rate of change of transformed income is greater before a cancer diagnosis than afterwards. However, as discussed in Section 3.3.2, we cannot easily interpret the cancer indicators individually. The estimated total cancer effect, made up of the estimates of the effects of the three cancer indicators and the interactions, is shown in Table 4.7 from the year of diagnosis until 10 years post-diagnosis. Table 3.3 shows how to calculate these values. The total cancer effect can be calculated on the conditional and marginal levels, where the conditional cancer effect is conditional on the previous year's income. The marginal total cancer effect can be interpreted as the mean difference in transformed income between a woman who was diagnosed with cancer and the same woman had she never been diagnosed. The estimated effect sizes in Table 4.5 are calculated for a woman who was diagnosed at the

	Estimated total cancer effect			
Years since diagnosis	**Conditional**	**SE**	**Marginal**	**SE**
0	−0.344	0.0193	−0.344	0.0193
1	−0.360	0.0193	−0.595	0.0325
2	−0.072	0.0189	−0.479	0.0378
3	−0.088	0.0189	−0.415	0.0438
4	−0.104	0.0191	−0.388	0.0486
5	−0.120	0.0193	−0.387	0.0523
6	−0.144	0.0211	−0.408	0.0534
7	−0.160	0.0212	−0.439	0.0563
8	−0.176	0.0214	−0.476	0.0593
9	−0.192	0.0217	−0.518	0.0619
10	−0.208	0.0220	−0.562	0.0641

Table 4.5: Marginal and conditional estimated total effect of cancer based on the ARR model for a woman who is diagnosed at age 49, does not have a spouse, and is not self-employed.

average age of 49, who does not have a spouse, and who does not receive self-employment income.

Cancer has a negative effect on transformed income, on average, and the magnitudes of both the conditional and marginal effects are largest one year after diagnosis. In that year, the difference between the mean transformed income of a woman who was diagnosed at age 49, who does not have a spouse, and who does not receive self-employment income, compared to the mean transformed income of the same woman if she had never been diagnosed, is -0.595 units. Both the marginal and conditional total cancer effects are significantly negative in all years post-diagnosis.

To illustrate the effect of cancer on the income scale, we compare the predicted income for an average woman who is diagnosed with cancer (i.e., a woman who has a birth year of 1954 and who was diagnosed at age 49) and who does not have a spouse or self-employment income to the predicted income of the same woman had she not been diagnosed. Specifically, using the estimated parameters from Table 4.4, we can calculate the predicted transformed income of a woman for any year post-diagnosis and her predicted transformed income in the same year had she never been diagnosed. We can then raise these values to the fourth power to obtain predictions on the scale of dollars.

For the average woman, the total effect of cancer translates to a predicted loss of $4,536 in the year of diagnosis, compared to the same woman had she not been diagnosed with cancer. At 10 years post-diagnosis, the predicted difference increases to $8,741.

Years since diagnosis	Diagnosed	Not diagnosed	Difference
0	46,821	51,357	−4,536
1	44,862	52,665	−7,803
2	47,522	53,998	−6,476
3	49,594	55,356	−5,762
4	51,240	56,739	−5,499
5	52,584	58,148	−5,564
6	53,602	59,583	−5,981
7	54,507	61,045	−6,538
8	55,334	62,533	−7,199
9	56,109	64,048	−7,939
10	56,850	65,591	−8,741

Table 4.6: Predicted income (dollars) of a woman who was diagnosed versus a woman who was not diagnosed. Values are for a woman born in 1954, diagnosed at age 49, with no spouse or self-employment income.

Other parameter estimates

The cohort effect of year of birth is positive, i.e., women born in later years have greater income, on average. If we compare a woman, i, to a woman, j, who was born 1 year later, then, conditional on their previous year's incomes, the mean transformed income of woman j is estimated to be 0.032 units greater than that of woman i when they are the same age. Or, if we do not condition on their previous year's incomes, the mean transformed income of woman j is estimated to be, for example, 0.032 units greater than woman i at age 18 and 0.101 units greater than woman i at age 60.

We can interpret the effect of `yearsSince18` as described in Section 3.3.2. For example, the estimated change in mean transformed income from age 49 to 50 for a woman who was born in 1954, was not diagnosed with cancer, and neither has a spouse nor is self-employed in any year from ages 18 to 50, is 0.0950.

For `spouseStatus` and `selfEmploy`, we can interpret the parameter estimates as explained in Section 3.3.2. First, consider two women who, in all years prior to time t (where t is prior to diagnosis), had the same value of all variables, but in year t differ only in that woman i has a spouse and woman j does not. Then the estimated mean difference in transformed income in year t of woman i compared to woman j is -0.327. Similarly, if the two women differ in year t only in that woman i has self-employment income and woman j does not, then the estimated mean difference in transformed income in year t of woman i compared to woman j is 0.105.

Not only do women with spouses have lower values of mean transformed income pre-diagnosis, but they also experience stronger negative effects of a cancer diagnosis. If, at some post-diagnosis time t, we compare two women who were diagnosed with cancer in the same year, have the same values of all variables in all years prior to t, and at time t differ only in that woman i has a spouse and woman j does not, then the estimated mean difference in transformed income of woman i compared to woman j is -0.083.

In contrast, women with self-employment income experience smaller negative effects of a cancer diagnosis. Again, we can compare two women at some post-diagnosis time t who were diagnosed with cancer in the same year, have the same values of all variables in all years prior to time t, and at time t differ only in that woman i has self-employment income and woman j does not. The estimated mean difference in transformed income of woman i compared to woman j is 0.115.

The variance of the random effect, σ_u^2, is estimated at 0.592 and the error variance, σ^2, is estimated at 3.427. The autoregressive parameter, ρ, is estimated at 0.684. The estimate of σ^2 is consistent with our expectations, as we have observed that the within-subject variance is large. However, the variance of the random effect is surprisingly low.

4.3.3 Diagnostics

If $\rho = 0$ and the same linear predictor is used in the LME model and conditional form of the ARR model, then the LME model is nested within the ARR model. However, our versions of these two models used different mean structures (the LME model had higher order `yearsSince18` terms), so we cannot compare them using a likelihood ratio test. To enable some sort of comparison of the models, we refit the LME model (by ML) with the same explanatory variables as in the ARR model. The likelihood ratio test statistic is $103,628$ and is approximately a random draw from a $\chi^2_{(1)}$ distribution if $\rho = 0$. The p-value is approximately 0, so we can conclude that the ARR model provides a significantly better fit to the data than the LME model (with this simplified mean structure, at least).

To investigate the fit of the ARR model, we calculate the residuals conditional on $Y_{i,t-1}$ and u_i (which we will call the *conditional residuals*) as

$$e_{it} = Y_{it} - (\hat{\rho} Y_{i,t-1} + \mathbf{x}_{it}\hat{\beta} + \hat{u}_i), \tag{4.1}$$

where $\hat{u}_i$ is the value of u_i predicted using an empirical Bayes method (Verbeke and Molenberghs, 2000).

Lag	Estimated autocorrelation
0	1.000
1	−0.070
2	−0.026
3	−0.010
4	−0.014
5	−0.025
6	−0.027
7	−0.036
8	−0.032
9	−0.036
10	−0.038
11	−0.040
12	−0.042
13	−0.041

Table 4.7: Average estimated ACF of conditional residuals from the ARR model.

The errors in the conditional form of the ARR model in (3.13) are uncorrelated and have constant variance. Therefore, the conditional residuals should be approximately uncorrelated and have approximately constant variance if the fit of the model is reasonable.

First, we informally evaluate if the conditional residuals are uncorrelated. The estimated ACF of the conditional residuals is calculated as described in Section 4.1 and is shown in Table 4.7. The estimated autocorrelation is negligible at all lags, indicating that the residuals are indeed close to uncorrelated.

A plot of the conditional residuals versus fitted values (not shown) indicates, as was the case with the ARE model, that the points with an observed value of 0 are not handled well by the model. The residuals tend to be more positive at small fitted values and more negative at large fitted values, indicating lack of fit at the extremes. Plots of the conditional residuals versus the explanatory variables (not shown) did not reveal other problems with the model fit.

We can calculate *marginal residuals* using the marginal expectation of the responses as

$$\mathbf{e}_i = \mathbf{Y}_i - \hat{\mathbf{B}}_i \mathbf{X}_i \hat{\boldsymbol{\beta}},$$

where $\mathbf{e}_i = (e_{i0}, e_{i1}, \ldots, e_{iT_i})$ is vector of residuals at each time point, and the remaining notation is the same as in Section 3.5.2.

The marginal residuals are difficult to interpret due to the complicated properties of the marginal model; for example, we do not expect them to have constant variance over time or to be uncorrelated. However, the plot of the marginal residuals versus the fitted values (not shown) could potentially reveal problems with the assumed marginal mean structure. In our case, the residuals were slightly positive, on average, but did not otherwise reveal any patterns.

Overall, we did not find clear statistical evidence of lack of fit of the ARR model except at the extremes of the range of incomes.

4.4 Summary of application

Using diagnostic plots based on the LME and ARE models, we did not find strong evidence that the ARE model is misspecified. Likewise, we found little evidence of misspecification of the ARR model. The ARE and ARR models are not nested; we cannot formally compare them with a likelihood ratio test. Therefore, we have no statistical evidence with which to distinguish between the fit of the ARE and ARR models. However, we prefer the ARE model for its simplicity and interpretability and believe that inference based on the ARE model is more reliable than that based on the ARR model.

The ARR and ARE models describe the nature of the cancer effect very differently. The total cancer effect estimates based on the ARE model shown in Table 4.3 and the marginal total cancer effect estimates from the ARR model shown in Table 4.5 are estimates of the same quantity (the total difference between mean income of a woman who is diagnosed with cancer and that had she not been diagnosed). However, the estimates differ greatly. For instance, in the year of diagnosis, a 95% confidence interval based on the ARE estimate for the total cancer effect is $(-0.306, -0.198)$ and based on the marginal estimate from the ARR model is $(-0.382, -0.306)$. These confidence intervals do not overlap; therefore, the estimated effect sizes are significantly different. By 10 years post-diagnosis, this difference has grown dramatically: a 95% confidence interval based on the ARE model is $(-0.154, 0.0202)$ and from the ARR model is $(-0.688, -0.436)$. Notably, based on the ARE model, the total cancer effect is insignificant 5 years post-diagnosis, but, based on the ARR model, the cancer effect is still significant 10 years post-diagnosis.

Clearly, the two models suggest very different impacts of a cancer diagnosis. As discussed in Section 3.3.2, the ARR model assumes that the effects of all explanatory variables vary over time and depend on ρ. This assumption is likely unreasonable, and the dramatic difference between the total cancer effect estimates from the ARE and ARR models could suggest that this assumption is not met in the case of the cancer effect. Moreover, results from previous studies on the effect of cancer on income align with the results from the ARE model, i.e., they suggest that the effect of cancer decreases over time (Syse et al., 2008; Jeon, 2017).

We would have liked to present predicted income values from the ARE model as we did for the ARR model in Table 4.6. However, because we are not allowed to express the estimated regression coefficients to more than 3 decimal places, we lack the precision in the coefficients of the cubic and quartic `yearsSince18` terms necessary to generate reasonable predicted values for the ARE and LME models. As we have already expressed, we have reservations about the accuracy of the estimates based on the marginal ARR model and the same reservations apply to the predicted incomes in Table 4.6.

Chapter 5

Simulation study

The purpose of this simulation study is to investigate the performance of the three models when they are fit to data generated from an ARE model. LME, ARE, and ARR models are all fit to these simulated data, and the performance of the estimator of the effect of cancer in each model is judged. We chose to use the ARE model as the true model in this simulation because, of the three models we are studying, it seems to be the most justifiable in this context, according to the analysis in Chapter 4. We expect the ARE estimator to perform the best since it is based on the true model; we use it as a "control" against which we can compare the other estimators.

The simulated values of transformed income are generated from a simplified version of our ARE model with only `yearsSince18`, `cancerLong`, and `birthYear` as explanatory variables. We fix birth year at 1954 and age of diagnosis at 49. Setting the year of birth and age of diagnosis of all women to the same values allows for easier comparison between the ARR estimator and the other estimators.

The parameter values used in the simulation are shown in Table 5.1. Two different values for each of the coefficient of `cancerLong`, σ_u^2, σ^2, and the number of women, N, were considered. The larger magnitude value of the cancer effect was chosen to be similar to the total cancer effect estimated in Section 4.2, and the smaller magnitude value was chosen to assess performance when the effect size is smaller than what is observed in Section 4.2. The values of the variance parameters were chosen so that the larger values approximately match the large variance of the data from Chapter 4, and the smaller values were chosen to assess whether performance improves when the responses have lower variance. Lastly, the values of N were chosen to assess performance on a very small and a moderate size sample. A run is defined as a unique combination of the values of these parameters; we therefore consider 16 runs in total. The values of the other parameters (the intercept, `yearsSince18`

Parameter	**Value(s)**
Intercept	-179.592
Coefficient of `birthYear`	0.095
Coefficient of `yearsSince18`	0.479
Coefficient of `cancerLong`	-0.35, -0.1
γ	0.65
σ_u^2	1, 6.25
σ^2	1, 6.25
N	50, 1000

Table 5.1: Parameter values chosen for the simulation.

effect, `birthYear` effect, and γ) remain constant across all runs; they were chosen to reflect the parameter estimates from Section 4.2. We simulate 100 data sets (replicates) for each run.

When estimating the parameters of the ARR model using the method of direct optimization (see Section 3.5.2), starting values are required. For the coefficients and variance parameters, starting values were randomly selected from an interval around the true values. For the autoregressive parameter, ρ, which has no corresponding true value, the starting value was randomly selected from an interval around γ.

For the LME and ARE models, the estimated coefficient of the cancer term can be interpreted as both the conditional and marginal effects of cancer. For the ARR model, the conditional and marginal estimates are different. In particular, the estimated marginal effect of `cancerLong` at time t is calculated as $\frac{1-\hat{\rho}^{t-t_d+1}}{1-\hat{\rho}}\hat{\beta}$, where β is the conditional effect and t_d is the time of diagnosis. The marginal estimate is a function of the number of years since diagnosis; in the following discussion, the marginal estimate refers to the estimate at one year post-diagnosis, i.e., for $t - t_d = 1$.

We investigate three properties of the four estimators (ARE, LME, ARR conditional, and ARR marginal): bias, accuracy of the standard errors, and the coverage probability of a 95% confidence interval for the effect.

We first present results on the bias of the estimated effects. Figure 5.1 shows the estimated bias of each estimator. The estimated bias for each run is calculated as the average of the estimated effects from the 100 replicates less the true value of the cancer effect. The error bars indicate the 95% Wald confidence interval for bias, where the standard error of the estimated bias is defined as the sample standard deviation of the 100 bias estimates.

Figure 5.1 shows that the LME and ARE estimators are approximately unbiased across the 16 runs; the 95% confidence intervals for bias of both the ARE and LME estimators

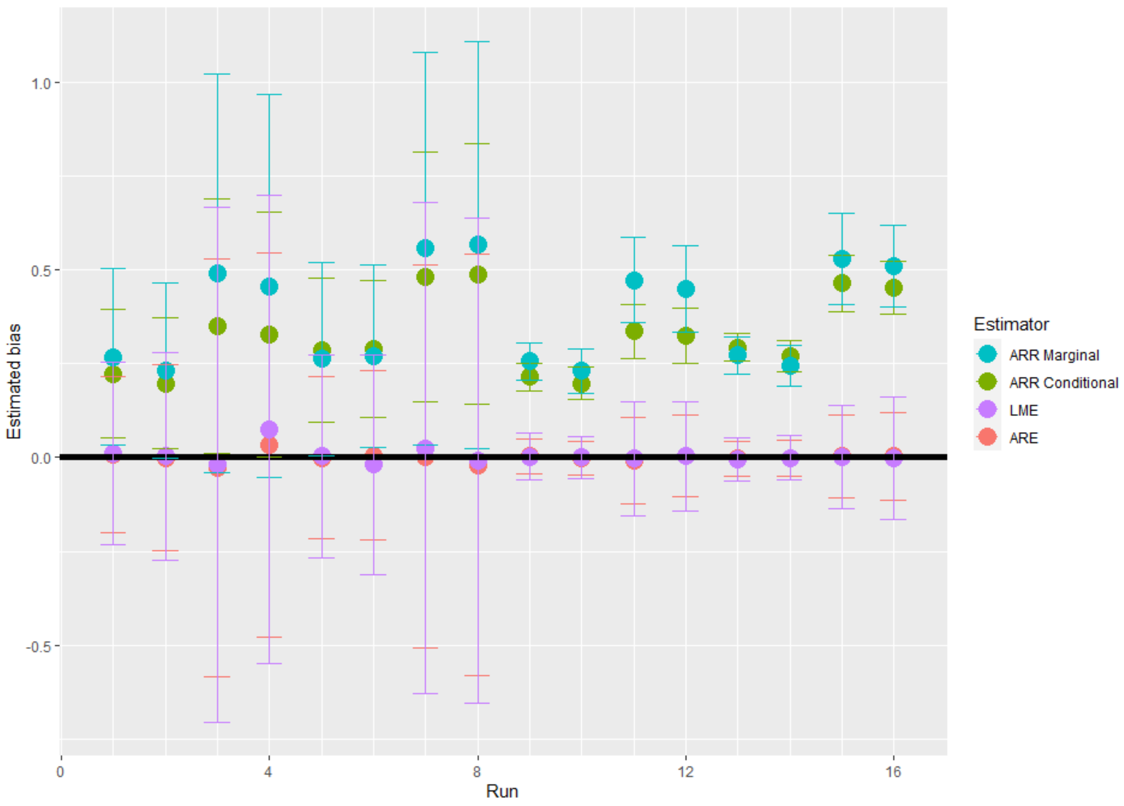

Figure 5.1: Estimated bias of estimators of the cancer effect. Error bars indicate the 95% Wald confidence interval for bias.

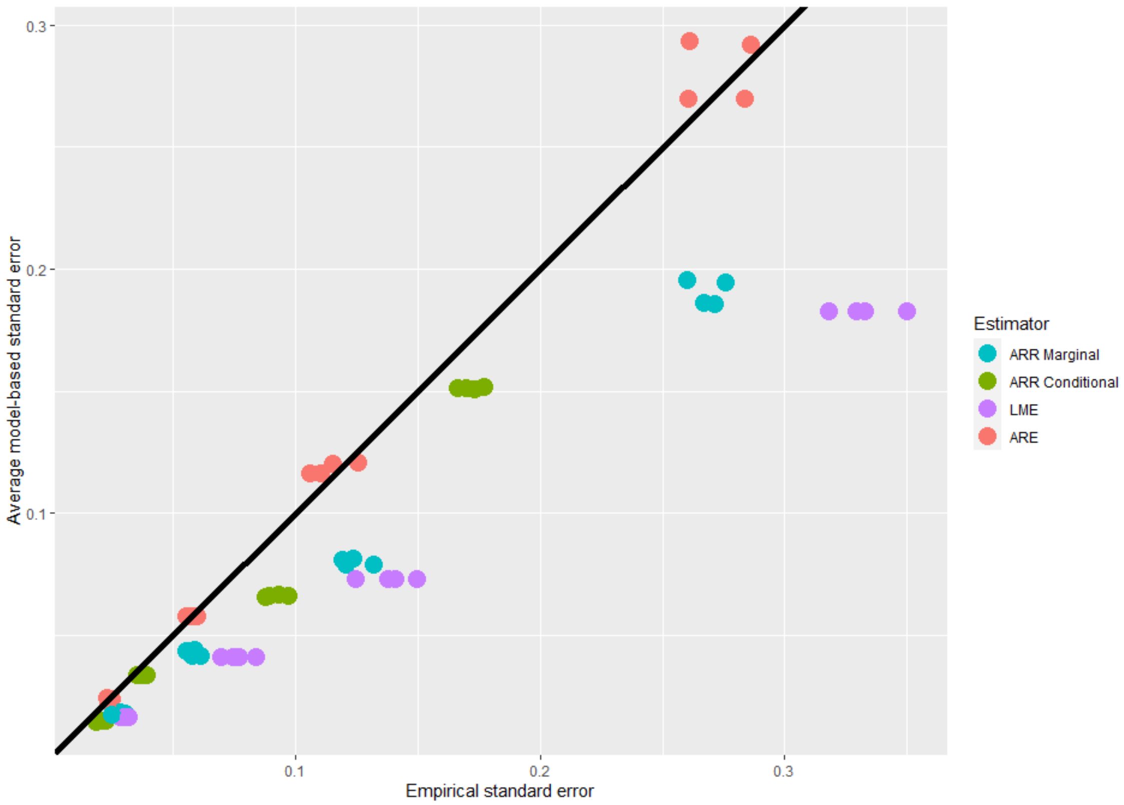

Figure 5.2: Model-based standard errors vs. empirical standard errors of the effect estimators for all runs.

include 0 in all runs. In contrast, the ARR marginal and conditional estimators have very large estimated bias across all runs and most of the associated confidence intervals exclude 0.

Next, we assess the accuracy of the standard errors associated with the estimators. Figure 5.2 is a plot of the average model-based standard error vs. the empirical standard error for each run and estimator. The former is the average of the standard errors associated with the 100 replicates. The latter is the sample SD of the 100 effect estimates. The empirical standard error approximates the true SD of the estimator. If the model-based standard errors are accurate, we would expect them to lie close to the $y = x$ line.

Figure 5.2 shows that, as expected, the standard errors of the ARE estimators are very accurate. The standard errors of the ARR conditional estimators are close to accurate; however, the standard errors of the ARR marginal estimators tend to underestimate the true SDs, especially larger SDs. Similarly, the standard errors of the LME estimators tend to underestimate the true SDs and become less accurate as the true SD increases. Unsurprisingly, the negative bias of the standard errors of the marginal ARR and LME estimators

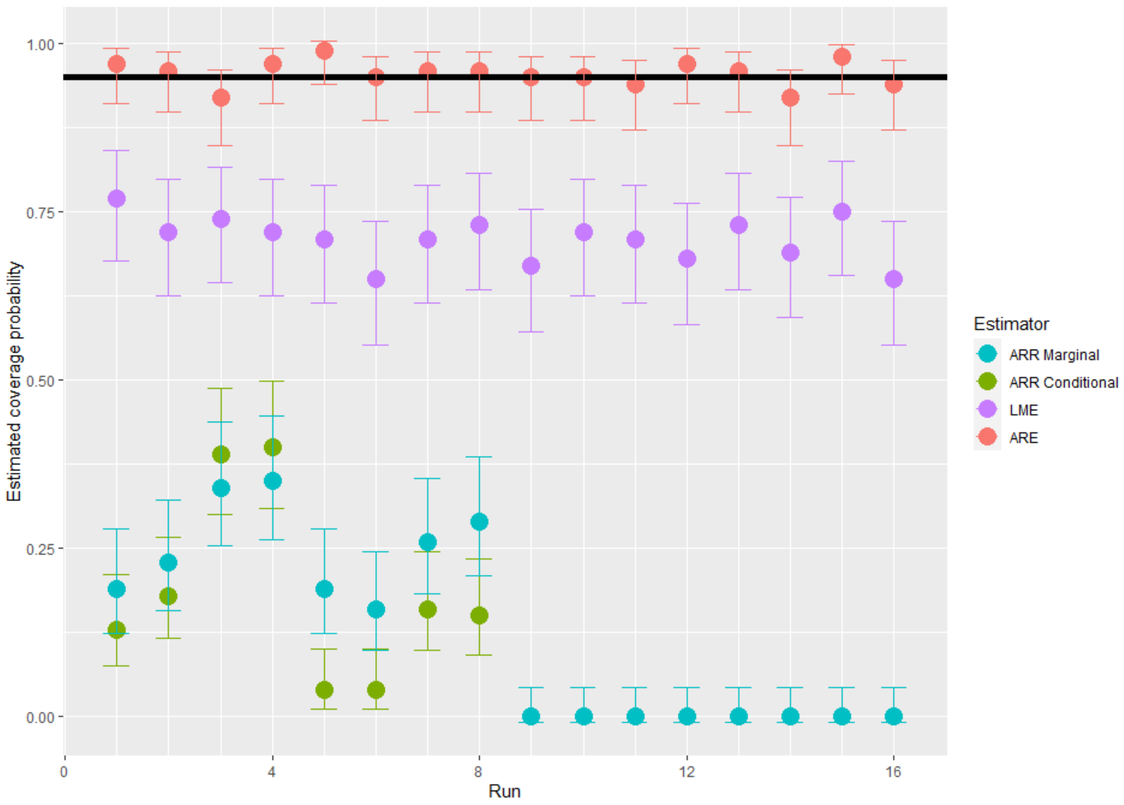

Figure 5.3: Estimated coverage probabilities of confidence intervals for the cancer effect. Error bars represent the 95% Agresti-Coull confidence interval for coverage probability.

is more pronounced when the error variance is large and the sample size is small (see points on the right that lie far below the $y = x$ line).

Lastly, we inspect the coverage probability of 95% confidence intervals for the cancer effect. For each replicate, the 95% Wald confidence interval is calculated. The coverage probability is estimated as the proportion of the 100 replicates from a run that contain the true effect size within the confidence interval. Figure 5.3 shows the estimated coverage probability for each estimator and each run, along with the 95% Agresti-Coull confidence interval for the coverage probability. Note that, for runs 9–16, the coverage probabilities based on the ARR marginal and conditional estimators are the same, so only the former are visible on the plot.

As anticipated, the nominal coverage probability of 0.95 is within all confidence intervals based on the ARE estimator. The coverage probability of confidence intervals based on the LME estimator is smaller than 0.95 for all runs, which is due to inaccurate standard errors, not bias (as per our discussion above).

The coverage probabilities of confidence intervals based on either the conditional or marginal ARR estimators are very low. This result is due to bias in both the effect estimators and their standard errors. In runs 9–16, the coverage probability is 0. Perhaps surprisingly, the coverage probabilities are slightly better in runs 1–8, where the sample size is 50, than in runs 9–16, where the sample size is 1000. If we were fitting the correct model, we would expect improved performance with increased sample size. However, under this misspecified model, the bias of the effect estimates remains high, even for larger N, while the standard error decreases. As a consequence, the confidence intervals are shorter and less likely to include the true effect.

To summarize, as expected, the performance of the ARE estimator, which is based on the true model, is excellent. Likewise, because the LME model has the same mean structure as the true model (but has a simpler variance-covariance structure), the LME estimator appears to be unbiased but its standard error is biased low, resulting in reduced coverage probability of confidence intervals for the cancer effect. In contrast, the ARR model has different mean and variance-covariance structures from the true model. The estimator based on this model is biased and its standard error is biased low, leading to very poor coverage probabilities of confidence intervals for the cancer effect.

Chapter 6

Conclusions

In this project, we explored the theoretical properties of three models, the LME, ARE, and ARR models, and the (RE)ML estimators of their parameters. We considered the conditional and marginal forms of the models, both of which are important for understanding and interpreting the models. We showed that the marginal form of the ARR model has a marginal mean structure that is non-linear in time. Moreover, this model is based on strong assumptions about the relationship between the mean response and the explanatory variables, namely, that the effects of these variables change over time and are a function of ρ and t. This feature of the ARR model leads to difficulty in interpreting the model parameters and in handling missing values of explanatory variables.

We also applied these models in an analysis of the income trends of breast cancer survivors. The simplest model, the LME model, provides insight into the mean and autocorrelation structure of the income observations. The ARE model appears to provide a reasonable fit to the data and is simple to interpret. On the other hand, while our diagnostics did not indicate severe lack of fit of the ARR model, the marginal mean structure of this model seems unrealistic, and the estimates of the cancer effects based on this model are thus questionable.

Using the ARE model, we detected a significant and negative effect of cancer that is largest in the year of diagnosis and the year following. The rate of growth of income was found to be lower post-diagnosis compared to pre-diagnosis, on average. Additionally, women with spouses tend to experience a greater negative effect of cancer, while women with self-employment income tend to experience a smaller negative effect of cancer.

Lastly, we investigated the performance of the three models in a simulation study where the true model was an ARE model with a linear age term. Under the settings considered, the ARE estimator of the effect of cancer on income performed very well. However, the

ARR estimator performed very poorly, raising questions about the performance of this estimator under model misspecification, particularly when the true mean trend over time is linear.

6.1 Future work

Many facets of this work are left unexplored, particularly regarding issues identified in the analysis in Chapter 4, that, due to time constraints and delays in data access, we were unable to address fully.

To describe different aspects of the effect of cancer, we used multiple indicator variables and interaction terms. While feasible, this approach led to difficulties in interpreting the cancer effect—even in the LME and ARE models. In the future, we would consider alternative definitions of the cancer effect; one possibility is to include only `cancerLong` but allow its coefficient to be time varying (e.g., Hastie and Tibshirani (1993)). With this approach, the differences in the effect of cancer in each year post-diagnosis could be handled in a more flexible way. Likewise, varying-coefficient models could provide a simpler and more accurate characterization of the age trend (which we presently capture using a fourth-order polynomial).

We were able to include the effects of spousal status and self-employment status in our models. Other explanatory variables, such as geographic location, children indicator, and employment sector, are also of interest for the future. In particular, we would like to determine if cancer has a disparate effect in groups defined by these variables. Including a children indicator could pose challenges; in our models, including the indicators for spouse, self-employment, and children together resulted in a multicollinearity issue. We are unsure as to the cause of this issue. Initially, we thought the cause could be a strong association between the spouse and children indicators; however, these two variables are not strongly associated and can be included in a model together without incident. This issue requires further investigation, and, for this analysis, we chose to include only the indicators for spouse status and self-employment in the model as those variables are of greater interest. However, in the future, we plan to investigate the multicollinearity issue further and develop a solution that allows all three effects to be estimated.

One of the most substantial issues we encountered in the application was the management of negative observed values of after-tax income. A number of women in the sample had a negative value of after-tax income in one or more years; since our chosen transformation can be applied only to non-negative incomes, we converted all negative incomes to 0. We acknowledge that this choice (in particular, ignoring the magnitude of the negative observations) may bias the estimate of the cancer effect. Moreover, our approach caused a point mass at 0 in the distribution of transformed income that violates the normality assumption and negatively affects the fit of the models. To improve the handling of the 0 values, one possible solution is a mixture model that assigns a probability, p, to 0 and the probability $1 - p$ to some positive, continuous distribution.

Alternatively, we could use employment income or gross income, rather than after-tax income, as the response variable, since negative values of these types of income cannot occur.

In general, more work is needed to understand the information that the negative income values provide about the effect of a cancer diagnosis. For example, negative values could arise if medical expenditures (resulting from a cancer diagnosis) claimed on a tax return are greater than income earned. This aspect of the effect of cancer would not be fully captured in our analysis. One option is to use a bivariate longitudinal model to describe how a cancer diagnosis affects medical expenditures and employment income simultaneously.

An alternative approach to the mixed effects models studied in this project is the modelling of the marginal effect of cancer and other explanatory variables directly. Given the complexity of the distribution of incomes (particularly when negative incomes are included), the method of generalized estimating equations (GEEs) (Liang and Zeger, 1986) could be desirable. With this approach, we would need to specify the marginal mean, variance, and covariance structure of income, but not its distribution. In addition to avoiding the specification of a distribution for income, GEEs have two advantages in this context. First, the simple form of the marginal mean structure (which is not conditional on random effects) is appealing when population-level effects, such as the effect of cancer, are of interest. Second, by using raw income as the response variable, the regression coefficients can be directly interpreted in terms of expected change in income (not transformed income). However, since the model is not fully specified, likelihood-based procedures, for example the likelihood ratio test, cannot be used. Additionally, GEEs require a stricter assumption about the missing data mechanism compared to likelihood-based estimation methods (Verbeke and Molenberghs, 2000). Nonetheless, this method, which allows modelling income without a transformation and assumption of normality, may provide a solution to several of the issues we encountered in this project. If a full model (and ensuing likelihood methods) are deemed necessary, Heagerty and Zeger (2000) presents a general marginal modelling approach.

Household income is the income generated by a family unit. The effect of cancer on household income is also of interest and may differ from the effect on individual income if, for example, the spouse of a breast cancer survivor increases their employment income to compensate for the survivor's decrease in income. A bivariate longitudinal model of the couple's individual incomes could be used to characterize such an effect.

The ARR model required complete data on the explanatory variables from age 18 until the last observation. We used a naive single imputation method to fill in any missing values, but, if we were to use this model in the future, we would explore different methods of creating complete data.

We observed in Chapter 2 that the SD of transformed income increases with age. One extension of the ARE model allows the model for the errors to be non-stationary, in which case, the variance of the response is time varying. Although we did not see strong evidence of changing variance in the residual plots, we would like to apply this model to the breast cancer data set to see if the model fit is improved.

The simulation study we conducted in Chapter 5 is limited in scope; we chose a simple mean structure that is linear in age and used a limited selection of parameter values. As shown in Chapter 2, the trend in transformed income over age from a real sample displays curvature. As shown in Section 3.3.1, the trend implied by the ARR model has curvature at small t (even when the model contains only a linear age term). Therefore, the ARR model would presumably perform better in situations more similar to that described in Chapter 4, where the true model contains curvature. Nevertheless, our simulation study raises questions about the performance of the ARR estimator of the cancer effect when the true model is not an ARR model. Further work is needed to identify scenarios where this model performs well and where it does not. Based on the results of our preliminary study, we would strongly caution against the application of the ARR model to data sets where the mean structure does not contain curvature in the time trend.

Further, more detailed study is required to confirm our results. However, our preliminary findings indicate that the income of breast cancer survivors decreases significantly immediately following diagnosis and that the growth in income of breast cancer survivors is slower after diagnosis. These results suggest that planning for the needs of breast cancer survivors should consider not only their health but also their financial well-being.

In the near future, we expect data on both breast cancer survivors and controls, as well as additional important explanatory variables, to become available. Our work provides a solid basis for a comprehensive study of the effect of cancer on the income of breast cancer survivors.

Bibliography

Tim A. Ahles, Andrew J. Saykin, Charlotte T. Furstenberg, Bernard Cole, Leila A. Mott, Karen Skalla, Marie B. Whedon, Sarah Bivens, Tara Mitchell, E. Robert Greenberg, and Peter M. Silberfarb. Neuropsychologic Impact of Standard-Dose Systemic Chemotherapy in Long-Term Survivors of Breast Cancer and Lymphoma. *Journal of Clinical Oncology*, 20(2):485–493, 2002.

Theodore W. Anderson and Cheng Hsiao. Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, 18(1):47–82, 1982.

Michael Baker. Growth-rate heterogeneity and the covariance structure of life-cycle earnings. *Journal of Labor Economics*, 15(2):338–375, 1997.

Badi H. Baltagi. *Econometric Analysis of Panel Data*. John Wiley & Sons, 2008.

Steve Berry, Peter Gottschalk, and Doug Wissoker. An Error Components Model of the Impact of Plant Closing on Earnings. *The Review of Economics and Statistics*, 70(4): 701–707, 1988.

Richard Blundell. Income dynamics and life-cycle inequality: mechanisms and controversies. *The Economic Journal*, 124(576):289–318, 2014.

Stephen R Bond. Dynamic panel data models: a guide to micro data methods and practice. *Portuguese Economic Journal*, 1(2):141–162, 2002.

Cathy J. Bradley, Heather L. Bednarek, and David Neumark. Breast cancer survival, work, and earnings. *Journal of Health Economics*, 21(5):757–779, 2002a.

Cathy J. Bradley, Heather L. Bednarek, and David Neumark. Breast cancer and women's labor supply. *Health Services Research*, 37(5):1309–1327, 2002b.

Darren R. Brenner, Hannah K. Weir, Alain A. Demers, Larry F. Ellison, Cheryl Louzado, Amanda Shaw, Donna Turner, Ryan R. Woods, and Leah M. Smith. Projected estimates of cancer in canada in 2020. *CMAJ*, 192(9):E199–E205, 2020.

Martin Browning and Mette Ejrnæs. Heterogeneity in the Dynamics of Labor Earnings. *Annual Review of Economics*, 5(1):219–245, 2013.

Lixin Cai, Kostas Mavromaras, and Umut Oguzoglu. The effects of health status and health shocks on hours worked. *Health Economics*, 23(5):516–528, 2014.

Statistics Canada. Table 39-10-0020-01 mean age and median age at divorce and at marriage, by sex. Accessed: 2021-03-23, `https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3910002001`.

Thomas N. Chirikos, Anita Russell-Jacobs, and Alan B. Cantor. Indirect economic effects of long-term breast cancer survival. *Cancer Practice*, 10(5):248–255, 2002a.

Thomas N. Chirikos, Anita Russell-Jacobs, and Paul B. Jacobsen. Functional impairment and the economic consequences of female breast cancer. *Women and Health*, 36(1):1–20, 2002b.

Mélanie Drolet, Elizabeth Maunsell, Jacques Brisson, Chantal Brisson, Benoît Mâsse, and Luc Deschênes. Not working 3 years after breast cancer: Predictors in a population-based study. *Journal of Clinical Oncology*, 23(33):8305–8312, 2005a.

Mélanie Drolet, Elizabeth Maunsell, Myrto Mondor, Chantal Brisson, Jacques Brisson, Benoît Mâsse, and Luc Deschênes. Work absence after breast cancer diagnosis: A population-based study. *CMAJ*, 173(7):765–769, 2005b.

Saskia F. A. Duijts, Jacobien M. Kieffer, Peter van Muijen, and Allard J. van der Beek. Sustained employability and health-related quality of life in cancer survivors up to four years after diagnosis. *Acta Oncologica*, 56(2):174–182, 2017.

Virginia S. Erickson, Marjorie L. Pearson, Patricia A. Ganz, John Adams, and Katherine L. Kahn. Arm edema in breast cancer patients. *Journal of the National Cancer Institute*, 93(2):96–111, 2001.

Margaret Fitch and Christopher J. Longo. Exploring the impact of out-of-pocket costs on the quality of life of canadian cancer patients. *Journal of Psychosocial Oncology*, 36(5):582–596, 2018.

Yoshimi Fukuoka, Sally H. Rankin, and Diane L. Carroll. Systematic bias in self-reported annual household incomes among unpartnered elderly cardiac patients. *Applied Nursing Research*, 20(4):205–209, 2007.

Ikuko Funatogawa and Takashi Funatogawa. *Longitudinal Data Analysis: Autoregressive Linear Mixed Effects Models*. SpringerBriefs in Statistics. Springer Singapore Pte. Limited, Singapore, 2019.

Patricia A. Ganz. The quality of life after breast cancer - Solving the problem of lymphedema. *New England Journal of Medicine*, 340(5):383–385, 1999.

Patricia A. Ganz, Julia H. Rowland, Katherine Desmond, Beth E. Meyerowitz, and Gail E. Wyatt. Life after breast cancer: Understanding women's health-related quality of life and sexual functioning. *Journal of Clinical Oncology*, 16(2):501–514, 1998.

Patricia A. Ganz, Katherine A. Desmond, Beth Leedham, Julia H. Rowland, Beth E. Meyerowitz, and Thomas R. Belin. Quality of life in long-term, disease-free survivors of breast cancer: A follow-up study. *Journal of the National Cancer Institute*, 94(1):39–49, 2002.

William H. Greene. *Econometric analysis*. Pearson, 5th ed. edition, 2003.

Eva Grunfeld, Doug Coyle, Timothy Whelan, Jennifer Clinch, Leonard Reyno, Craig C. Earle, Andrew Willan, Raymond Viola, Marjorie Coristine, Teresa Janz, and Robert Glossop. Family caregiver burden: Results of a longitudinal study of breast cancer patients and their principal caregivers. *CMAJ*, 170(12):1795–1801, 2004.

Sævar B. Gudbergsson, Sophie D. Fosså, Marja Liisa Lindbohm, and Alv A. Dahl. Received and needed social support at the workplace in Norwegian and Finnish stage 1 breast cancer survivors: A study from the Nordic Study Group of Cancer and Work (NOCWO). *Acta Oncologica*, 48(1):67–75, 2009.

Jari J. Hakanen and Marja Liisa Lindbohm. Work engagement among breast cancer survivors and the referents: The importance of optimism and social resources at work. *Journal of Cancer Survivorship*, 2(4):283–295, 2008.

Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757 – 796, 1993.

Patrick J. Heagerty and Scott L. Zeger. Marginalized multilevel models and likelihood inference. *Statistical Science*, 15(1):1–26, 2000.

Martee L. Hensley, Jeannette Dowell, James E. Herndon, Eric Winer, Nancy Stark, Jane C. Weeks, and Electra Paskett. Economic outcomes of breast cancer survivorship: CALGB study 79804. *Breast Cancer Research and Treatment*, 91(2):153–161, 2005.

Maria Hewitt, Julia H. Rowland, and Rosemary Yancik. Cancer survivors in the United States: Age, health, and disability. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, 58(1):82–91, 2003.

Cheng Hsiao. *Analysis of Panel Data*. Cambridge University Press, third edition, 2014.

National Cancer Institute. Age and cancer risk. `https://www.cancer.gov/about-cancer/causes-prevention/risk/age`, 2015. Accessed: Jan 6, 2021.

Tania Islam, Maznah Dahlui, Hazreen Majid, Azmi Nahar, Nur Mohd Taib, and Tin Su. Factors associated with return to work of breast cancer survivors: a systematic review. *BMC public health*, 14(Suppl 3):S8–S8, 2014.

Sung Hee Jeon. The Long-Term Effects of Cancer on Employment and Earnings. *Health Economics (United Kingdom)*, 26(5):671–684, 2017.

Salene M. W. Jones, Trung Nguyen, and Shasank Chennupati. Association of financial burden with self-rated and mental health in older adults with cancer. *Journal of Aging and Health*, 32(5-6):394–400, 2020.

Hrishikesh P. Kale and Norman V. Carroll. Self-reported financial burden of cancer care and its effect on physical and mental health-related quality of life among us cancer survivors. *Cancer*, 122(8):283–289, 2016.

Alexander Karaivanov, En Lu, Hitoshi Shigeoka, Cong Chen, Stephanie Pamplona, and Shih En Lu. Face Masks, Public Policies and Slowing the Spread of COVID-19: Evidence from Canada. Oct 2020. URL `https://github.com/C19-SFU-Econ`.

Nan M. Laird and James H. Ware. Random-Effects Models for Longitudinal Data. *Biometrics*, 38(4):963, 1982.

Sophie Lauzier, Elizabeth Maunsell, Mélanie Drolet, Douglas Coyle, Nicole Hébert-Croteau, Jacques Brisson, Benoît Mâsse, Belkacem Abdous, André Robidoux, and Jean Robert. Wage losses in the year after breast cancer: Extent and determinants among Canadian women. *Journal of the National Cancer Institute*, 100(5):321–332, 2008.

Kung-Yee Liang and Scott L. Zeger. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73(1):13, 1986.

Lee A. Lillard and Robert J. Willis. Dynamic Aspects of Earning Mobility. *Econometrica*, 46(5):985, 1978.

Marie Høyer Lundh, Claudia Lampic, Karin Nordin, Johan Ahlgren, Leif Bergkvist, Mats Lambe, Anders Berglund, and Birgitta Johansson. Changes in health-related quality of life by occupational status among women diagnosed with breast cancer—a population-based cohort study. *Psycho-Oncology*, 22(10):2321–2331, 2013.

Elizabeth Maunsell, Chantal Brisson, Lise Dubois, Sophie Lauzier, and Annie Fraser. Work problems after breast cancer: An exploratory qualitative study. *Psycho-Oncology*, 8(6): 467–473, 1999.

Elizabeth Maunsell, Mélanie Drolet, Jacques Brisson, Chantal Brisson, Benoit Mâsse, and Luc Deschênes. Work situation after breast cancer: Results from a population-based study. *Journal of the National Cancer Institute*, 96(24):1813–1822, 2004.

Jae-Hyun Park, Eun-Cheol Park, Jong-Hyock Park, Sung-Gyeong Kim, and Sang-Yi Lee. Job loss and re-employment of cancer patients in korean employees: A nationwide retrospective cohort study. *Journal of Clinical Oncology*, 26(8):1302–1309, 2008.

Joohyun Park and Kevin A. Look. Relationship between objective financial burden and the health-related quality of life and mental health of patients with cancer. *Journal of Oncology Practice*, 14(2):e113–e121, 2018.

Donald O Parsons. The Autocorrelation of Earnings, Human Wealth Inequality, and Income Contingent Loans. *The Quarterly Journal of Economics*, 92(4):551–569, 1978.

V. Peuckmann, O. Ekholm, P. Sjøgren, N. K. Rasmussen, P. Christiansen, S. Møller, and M. Groenvold. Health care utilisation and characteristics of long-term breast cancer survivors: Nationwide survey in Denmark. *European Journal of Cancer*, 45(4):625–633, 2009.

Jose Pinheiro, Douglas Bates, and Saikat DebRoy. Linear and nonlinear mixed effects models. `https://cran.r-project.org/web/packages/nlme/nlme.pdf`.

Dorte M. Rasmussen and Beth Elverdam. The meaning of work and working life after cancer: An interview study. *Psycho-Oncology*, 17(12):1232–1238, 2008.

William A. Satariano and Gerald N. DeLorenze. The likelihood of returning to work after breast cancer. *Public Health Reports*, 111(3):236–241, 1996.

Canadian Cancer Society. Breast cancer statistics. `https://www.cancer.ca/en/cancer-information/cancer-type/breast/statistics/?region=on`, 2020.

Astri Syse, Steinar Tretli, and Øystein Kravdal. Cancer's impact on employment and earnings-a population-based study from Norway. *Journal of Cancer Survivorship*, 2(3): 149–158, 2008.

Sietske J Tamminga, Angela GEM de Boer, Jos HAM Verbeek, and Monigue HW Frings-Dresen. Breast cancer survivors' views of factors that influence the return-to-work process - a qualitative study. *Scandinavian Journal of Work, Environment & Health*, 38(2):144–154, 2012.

P. van Muijen, N.L.E.C. Weevers, I.A.K. Snels, S.F.A. Duijts, D.J. Bruinvels, A.J.M. Schellart, and A.J. van der Beek. Predictors of return to work and employment in cancer survivors: a systematic review. *European Journal of Cancer Care*, 22(2):144–160, 2013.

Geert Verbeke and Geert Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York, 2000.

Jeffrey M Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, Mass, 2002.

K. Robin Yabroff, William F. Lawrence, Steven Clauser, William W. Davis, and Martin L. Brown. Burden of illness in cancer survivors: Findings from a population-based national sample. *Journal of the National Cancer Institute*, 96(17):1322–1330, 2004.

Yonghui Zhang and Qiankun Zhou. Estimation for time-invariant effects in dynamic panel data models with application to income dynamics. *Econometrics and Statistics*, 9:62–77, 2019.

Appendix A

Variable	**Estimate**	**Standard error**
`Intercept`	-204.572	5.200
`birthYear`	0.108	0.003
`yearsSince18`	0.549	0.018
`yearsSince18`^{2}	-0.024	0.001
`yearsSince18`^{3}	0.001	0.000
`yearsSince18`^{4}	-0.000	0.000
`cancerImmediate`	-0.068	0.019
`cancerShort`	-0.115	0.020
`cancerLong`	-0.113	0.078
`yearsSince18`×`cancerLong`	-0.001	0.002
`spouseStatus`	-0.756	0.019
`spouseStatus`×`cancerLong`	-0.217	0.024
`selfEmploy`	-0.024	0.024
`selfEmploy`×`cancerLong`	0.248	0.033

Table A.1: Estimated regression coefficients and standard errors from the fully adjusted LME model. The standard errors are likely inconsistent, as we suspect the variance-covariance structure of the LME model is misspecified.

Parameter	**Estimate**
σ_u^2	7.295
σ^2	4.937

Table A.2: Variance parameter estimates from the fully adjusted LME model. We expect these estimates to be biased due to misspecification of the variance-covariance structure in the LME model.

Years since diagnosis	**Estimated total cancer effect**	**Standard error**
0	−0.327	0.0253
1	−0.328	0.0255
2	−0.261	0.0260
3	−0.262	0.0264
4	−0.263	0.0270
5	−0.264	0.0278
6	−0.150	0.0318
7	−0.151	0.0325
8	−0.152	0.0333
9	−0.153	0.0342
10	−0.154	0.0352

Table A.3: Total estimated effect of cancer based on the LME model for a woman who is diagnosed at age 49, does not have a spouse, and is not self-employed. The standard errors are likely inconsistent, as we suspect the variance-covariance structure of the LME model is misspecified.