Age-period-cohort models for individual-level data:

An acceleration-based regression framework

Zoë Fannon

Somerville College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2020

Word count: 407 words on p.6 × 195 pages = 79, 365

Dedicated to Deirdre Bolger, who taught me Maths.

Acknowledgements

I would like to thank my supervisor, Bent Nielsen, for his advice and guidance over the past five years. This thesis would not have been possible without his support. I would also like to thank my final examiners, Steve Bond and Thomas Crossley, as well as my confirmation examiner, Jurgen Doornik, for their interest in and thoughtful commentary on the thesis.

Many people played a role in the development of this thesis. I would particularly like to thank Christiaan Monden, for his contribution to the analysis of BMI data in §3. I am grateful to all those who provided opportunities to discuss or present my research, including the members of the Econometrics group in the Oxford Economics Department; Francesco Billari, Rebecca Graziani, and the other members of the DisCont research group at Universit Bocconi; Concetta Rondinelli and the economists at the Banca dItalia; and the participants in the 2019 Royal Economic Society Junior Researcher Symposium. Further thanks are due to Jonas Harnau and Matthias Qian, who served as role models throughout my DPhil.

I am grateful to Somerville College for welcoming me back for my DPhil years. I would particularly like to thank Guido Ascari, Karen Nielsen, and Steve Rayner for taking the time to offer advice and support at critical points during the preparation of the thesis, and Seun Alabi for making my return to college accommodation in my final year an incredibly stress-free process. I am also grateful for the guidance of Sean Veitch, of the Oxford University Counselling Service.

I would like to express my gratitude to the many people who kept me grounded throughout my graduate studies at Oxford. First and

foremost, I want to thank my parents, Tom Fannon and Irene Lynch-Fannon, and my siblings, Tim Fannon and Grace Fannon, for their steady love and support. I am grateful for my partner, Cyrus Motashaw, whose patience and kindness sustained me through writing the thesis; and for Ritu Motashaw, who welcomed me into her home.

Thanks are due to my coach, Graham Nichols; my teammates, the Womes; and the members of Oxford University Basketball Club. I am so grateful for my Oxford family at 8 Bullingdon Road: Tetyana Vasylyeva, Arushi Garg, Ben Abraham, Mikesh Udani, Kalpana Sivabalah, Miriam Zachau Walker, and Yulia Ioffe. Thanks are due to the other students who shared the journey through the DPhil in Economics, in particular Genevieve Nelson. Thank you also to my exceptional friends Gill Geng, Heather Wilson, and Lena Naassana.

I owe a debt of gratitude to the Economic and Social Research Council (grant ES/J500112/1), to the DisCont European Research Council project (grant 694262), and to the Oxford Economics Department Graduate Teaching Assistant scheme, which together provided the financial support that enabled me to produce this thesis.

Abstract

I develop a framework to analyse the relationship between an outcome of interest and an individual's age, period of observation, or birth cohort, using individual-level data. The framework is suitable for any continuous or binary outcome. I have created a package for the statistical software R which implements this framework.

It is well-known that linear relationships between age, period, or cohort and an outcome of interest are not separately identified. My framework instead focuses on non-linear relationships, described by age, period, and cohort (APC) accelerations.

My framework embeds the age, period, and cohort (APC) accelerations as parameters in a regression model. The regression approach makes it easy to include covariates and to test restrictions on the model. I develop a regression-based test of the APC acceleration model against a more general model, the time-saturated model.

My APC acceleration framework is suitable for repeated cross section and panel data. For repeated cross section data, a generalized linear modelling strategy is used which accommodates continuous and binary outcomes. For panel data, a generalized least squares strategy is used which accommodates continuous outcomes. I consider three panel settings: pooled ordinary least squares, random effects, and fixed effects.

I give three examples of applying the APC acceleration framework. First, I identify a new stylized fact about the role of birth cohort in obesity among English men. Second, I use the framework as a diagnostic tool to evaluate control variables in a model linking commute time and hospital in-patient stays. Third, I show that the framework can detect well-known relationships between wages and age and period.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis develops a regression framework for identifying age, period, and cohort acceleration effects from individual-level data.

An individual's age, their year of observation (period), and their year of birth (cohort) can separately influence outcomes, such as health, wages, or savings. However, because of the exact relationship *age = period - cohort*, not all effects of the three variables can be separately identified. The effects that can be separately identified are accelerations. Accelerations in age capture how the effect of aging one year changes with age. A constant positive acceleration effect of age on savings means that the change in savings from one age to the next is more positive the older a person is. Accelerations in period and cohort are similarly defined.

There are at least four uses for acceleration effects of age, period, and cohort. First, accelerations can be used for policy evaluation, because they identify discontinuities in the relationship between age, period, or cohort and some outcome variable. An educational policy change to raise the school-leaving age would show up as a cohort acceleration effect on years of schooling. Second, accelerations can be used to evaluate the shape of the relationship between some outcome and age, period, or cohort, leading to the identification of new stylized facts; an example is given in §3. Third, accelerations can be used to test functional form restrictions on the relationship between an outcome and age, period, or cohort. Such restrictions might arise from economic theory; for example, the Life Cycle Hypothesis implies a flat relationship between consumption and age, so all age accelerations should be zero. They may also be imposed for convenience, such as the quadratic in age

that I test in §4. Fourth, accelerations may be used for forecasting.

The key contribution of this thesis is the development of a framework to embed these age, period, and cohort (APC) accelerations in regressions which can be estimated from individual-level data. Accelerations have been estimated from individual-level data before, but using a cumbersome, two-step procedure (Van Landeghem, 2012). The framework I develop in this thesis allows accelerations to be estimated as parameters from regressions. I have also created an R package which performs these regressions in a single step.

The framework and R package I develop have roots in an existing regression framework and R package for aggregate data (Nielsen, 2015). Since age, period, and cohort are frequently of interest in analyses of individual-level data in economics and other social sciences, the extension of both the framework and the package to allow for individual-level data is valuable.

The structure of the thesis is as follows. In §2, I provide an overview of the literature on age, period, and cohort effects. I discuss some of the many papers in economics that have relied on age, period, and cohort as explanatory variables. I introduce the analytical framework used to identify APC accelerations in both existing work for aggregate data, and in this thesis which deals with individual-level data. I discuss some of the other approaches to identifying APC effects and their strengths and weaknesses relative to the framework developed here.

In §3 of the thesis, I develop the APC acceleration framework for the analysis of repeated cross section data. Repeated cross section data is individual-level data collected in multiple time periods, where new individuals are recorded at each period. Examples of repeated cross section data include the American CPS (Current Population Survey) and the English LFS (Labour Force Survey). My framework permits analysis of continuous and binary outcomes using a generalized linear modelling approach. I provide for estimation of a model with accelerations in age, period, and cohort. I also provide for testing of that model against both sub-models that it nests, which omit some of the accelerations, and a more general model which nests it. I apply the framework to identify a new stylized fact regarding the evolution of obesity by cohort among English men. This chapter is co-authored with Bent Nielsen and Christiaan Monden, and a modified version

has been accepted for publication at the *Journal of the Royal Statistical Society Series A*.

In §4, I develop the APC acceleration framework for the analysis of panel data. Panel data is individual-level data where the same people are recorded in multiple time periods. Examples include the American NLSY (National Longitudinal Survey of Youth) and the BHPS (British Household Panel Survey). My framework permits analysis of continuous outcomes only but considers three standard panel data settings: pooled OLS, random effects, and fixed effects. The statistical approach used here is generalized least squares, because that is the standard approach used by economists for panel data. I provide for estimation of a model with accelerations in age, period, and cohort, as well as testing of that model against sub-models that it nests. I apply the framework to evaluate the use of a standard set of controls in a study of the impact of commute time on hospital in-patient stays. I find that the standard quadratic model for age is rejected, and a more careful analysis of the non-linear shape of the age effect leads me to adapt the model in a way that improves the fit and strengthens the result regarding the effect of commute time on hospital in-patient stays.

In §5, I develop an R package which implements the APC acceleration framework developed in the preceding chapters. The package is designed for ease of use: there are three central commands which each operate on individual-level data presented in the standard "long" format. The first command is for estimation of a single model with age, period, and cohort accelerations; the second command is for model selection; and the third command creates plots of the estimated APC accelerations which facilitate their interpretation. The package is also designed to be compatible with standard testing tools in R, such as `waldtest()` and `linearHypothesis()`. The use of the package is demonstrated with models for log wages, using two datasets already available in the R environment.

Chapter 2

Identification of age, period, and cohort effects: a literature review

2.1 Introduction

This chapter provides the background to the contributions made in the three substantive chapters of the thesis. I explain the relevance of age, period, and cohort effects to economics and outline the analytical framework which will be used throughout the thesis.

I motivate the thesis in §2.2 with a discussion of existing papers in economics which use age, period, or cohort as explanatory variables. Examples are particularly prevalent in labour economics, health economics, and industrial organisation. I suggest ways in which accelerations in age, period, and cohort could be used to address the research questions of these applications.

The bulk of this chapter is devoted to an outline of the approach used in this thesis to identify age, period, and cohort (APC) accelerations from regressions, given in §2.3. I discuss the general APC identification problem and the fact that accelerations are robust to this. I then introduce the framework used to identify APC accelerations from regression. This framework is based on a reparametrization of what is known as the classical APC model. This approach was developed in a series of papers working with aggregate data, starting with Kuang et al. (2008).

In the final sections of this chapter I describe the advantages of the acceleration-based regression framework as an approach to the analysis of APC effects, as compared to a number of other common approaches. Many of these approaches have the undesirable property that their estimates are sensitive to untestable identifying constraints, which is avoided by the APC acceleration framework.

2.2 Age, period, and cohort (APC) effects in economics

In this section, I provide a brief tour of the applied literature that uses individual-level data in which age, period, or cohort are explanatory variables. This highlights the range of potential use cases for the framework developed in this thesis.

Many of the papers cited here use what I call "constraint methods" to estimate age, period, and cohort effects. These methods may be favoured by researchers because they give estimates of slopes in age, period, and cohort, and researchers may be more accustomed to working with slopes than the accelerations used in this thesis. The problem with constraint methods is that they only obtain estimates of slopes by imposing untestable constraints on the age, period, and cohort effects. The choice of constraints impacts the resulting estimates of the slopes in age, period, and cohort. This is explored further in §2.5.1. The framework I develop, in terms of accelerations, does not require these constraints.

I argue that the use of these constraint methods to estimate age, period, and cohort effects is unnecessary and can lead to incorrect inference. Many of the underlying questions of interest in the applications described below could be answered using the accelerations, which can be identified without imposing such constraints. Where the question of interest can truly only be answered using slopes, it seems unwise to rely on an answer to that question which depends on an untestable constraint imposed by the researcher. Even in the best examples of the application of constraint methods, such as Lagakos et al. (2018), where the constraints are carefully selected and argued for based on the context, it is impossible to formally test the constraints. It is therefore preferable to use accelerations, which are identified without the need for constraints. In the following, I describe how accelerations could be used to address research questions in economics.

In life-cycle models, age effects are naturally the primary object of interest. Life-cycle models may be estimated from individual-level data to better understand the shape of the relationship between age and an outcome of interest, controlling for other factors. For example, Hanoch & Honig (1985) estimate age profiles of employment and earnings for white Americans using social security data, and Browning et al. (2016) use British survey data to estimate an age profile of purchases of durable goods in later life. Accelerations are very suitable for exploring age (or period or cohort) profiles, as I demonstrate in my exploration of the age and cohort profiles of obesity in §3.5. Life-cycle models may be estimated in order to evaluate structural models, as is the case in Low et al. (2010) where a constraint method is used to estimate the age profile of employment rates from the American SIPP dataset. It would be straightforward to compare accelerations in the data with accelerations implied by the structural model, rather than comparing estimates of levels which rely on an untestable assumption. Other applications involve comparing life-cycle profiles between groups, which could easily be done in terms of accelerations. For example, Fitzenberger et al. (2004) use German micro-census data to compare the male and female age profiles of labour force participation and employment, and Lagakos et al. (2018) harmonized repeated cross section wage surveys to study variation in the experience-wage profile across countries.

Several studies have investigated whether a policy change or exogenous shock creates a discontinuity in one of age, period, or cohort. Krueger & Pischke (1992) use data from the repeated cross section Current Population Survey to investigate cohort discontinuities in American labor supply due to social security reforms. Almond (2006) looks for cohort discontinuities in various outcomes which could be linked to the 1918 influenza pandemic, using individual-level US census data. Discontinuities can be readily identified using APC accelerations; an example with aggregate data is McKenzie (2006), in which discontinuities in consumption by period are detected and attributed to the Mexican peso crisis.

Age, period, and cohort often appear as controls, rather than direct objects of interest, in studies which use individual-level data. An example is the literature on the economics of health and well-being, where it is essential to control for the impact of age (see for example Künn-Nelen, 2016; Roberts et al., 2011; Van Landeghem, 2012; Dickerson et al., 2014). Functional form restrictions are

often imposed on age, period, and cohort controls. The APC acceleration framework can be used to determine whether functional form restrictions on age, period, and cohort controls are appropriate. An example is given in §4.

Exploratory analyses of the importance of age, period, or cohort effects for some outcome of interest are seen in many sub-disciplines of economics. They are common in labour economics (in addition to the studies cited in the disussion of life-cycle models, see Méndez & Sepúlveda, 2012; Meghir & Whitehouse, 1996) and in the study of consumption and savings (in addition to the studies cited in the disussion of life-cycle models, see Kapteyn et al., 2005; Bíró, 2017; Attanasio, 1998). There is a literature on the relative importance of firm vintage, age, and year of observation for outcomes such as firm productivity (Jensen et al., 2001; Fukuda, 2013). In energy economics, Bardazzi & Pazienza (2018) seek to separate the contributions of age and cohort to energy demand.

A particularly concerning phenomenon is the existence of studies where one of age, period, or cohort is the main explanatory variable of interest, but the author seems unaware of the risk of confounding with the effects of the other two. An example is Rosenthal (2014), published in the *American Economic Review*. This paper examines the relationship between the age of a house and the income of the occupier, to examine whether the "filtering" process - where houses are passed down the income distribution as they age - provides a sufficient supply of housing for low-income families. There is no provision for cohort effects, although the year in which a house was built may also be related to the income of its occupiers.

The APC identification problem is not unique to economics; it appears in in applied work across the social sciences. Yang & Land (2013) and O'Brien (2015) describe examples in criminology, epidemiology, and sociology. The use of constraint methods is also common in these literatures. I discuss the constraint methods used in economics and other social sciences in more detail in §2.5.

The above discussion shows that age, period, and cohort are ubiquitous as explanatory variables. The use of regression-based constraint methods in applications with individual-level data is widespread, despite the general knowledge that accelerations can be identified and the ease with which accelerations could be used to answer the questions of interest. One of the factors preventing the wider use of accelerations may be the lack of a straightforward, regression-based framework

from which to estimate them. It is also possible that there is uncertainty around the use and interpretation of accelerations. This thesis attempts to fill both of these gaps, by developing a framework and software for estimating APC accelerations from individual-data regressions, and by providing applications and examples which show how accelerations may be used.

2.3 The reparametrized classical APC model: A framework to estimate APC accelerations from aggregate data regressions

The framework I develop to estimate APC accelerations from individual-level data regressions is based on an existing framework for estimation of APC accelerations from aggregate data regressions (Kuang et al., 2008; Nielsen, 2015). In this section, I explain that aggregate data framework. At the heart of the framework is a reparametrization of a model known as the classical APC model. The classical APC model is not identified; the reparametrization is identified, and permits isolation of the APC accelerations as regression parameters.

The remainder of this section is as follows. In §2.3.1, I provide a very brief overview of the aggregate data framework, which is explained in detail in the rest of the section. In §2.3.2, I explain the structure of the aggregate data, introduce the classical APC model, and explain why the classical APC model is not identified. In §2.3.3 I first outline the key ideas of the reparametrization approach to the classical APC model, and then provide a detailed derivation of the reparametrization. In §2.3.4 I explain the four use cases of the APC acceleration parameters as estimated from the reparametrized model.

2.3.1 Overview

The data for which the framework is developed is an array of aggregate outcome outcomes indexed by two of the three of age, period, and cohort. These are denoted by a, p, c respectively, and related by $p = a + c - 1$; this is explained in §2.3.2.

The analysis of this data begins with a model called the classical APC model, which is defined in terms of age, period, and cohort fixed effects:

$$\begin{aligned}\mu_{ac} =& \delta + \alpha_a + \beta_p + \gamma_c \\ =& d'_{ac}\theta \end{aligned} \tag{2.1}$$

Here μ_{ac} is the linear predictor for the outcome indexed by a, c; δ is a general intercept; and α_a, β_p, and γ_c are age, period, and cohort fixed effects. These are collected in the parameter vector θ which is associated with a design vector d_{ac}. The setting up of this classical APC model is detailed in §2.3.2.1.

The parameter vector of the classical APC model,

$$\theta = \{\delta, \alpha_1, \ldots, \alpha_A, \beta_1, \ldots, \beta_P, \gamma_1, \ldots, \gamma_C\} \tag{2.2}$$

is not identified. It has dimension $A + P + C + 1$, which is greater than can be identified from the data. This identification problem is well-known, see §2.3.2.2.

To address the identification problem, the model in (2.1) is reparametrized in terms of a parameter vector ξ, which is of lower dimension than θ and is therefore just-identified. The reparametrized model is

$$\mu_{ac} = x'_{ac}\xi, \tag{2.3}$$

for x_{ac} the design vector associated with ξ. The parameter vector ξ is unique with respect to the linear predictors μ_{ac}, and is invariant to the set of θ consistent with μ_{ac}. The details of the reparametrization are given in §2.3.3.

The APC accelerations that are the primary object of interest appear as elements of the new parameter vector ξ. The parameter vector is defined

$$\xi = \{v_o, v_a, v_c, \Delta^2\alpha_3, \ldots, \Delta^2\alpha_A, \Delta^2\beta_3, \ldots, \Delta^2\beta_P, \Delta^2\gamma_3, \ldots, \Delta^2\gamma_C\}. \tag{2.4}$$

Here age acceleration parameters are denoted $\Delta^2\alpha_a$, while period and cohort acceleration parameters are denoted $\Delta_2\beta_p$ and $\Delta^2\gamma_c$ respectively. There is also a linear plane defined by v_o, v_a, v_c. These parameters are explained in detail in §2.3.3.4.

There are at least four use cases of the accelerations, which are discussed in §2.3.4. First, the accelerations can be used independently to identify discontinuities due to policy changes or exogenous shocks; see §2.3.4.1. Second, accelerations

can be cumulated and detrended to construct a visual representation of the non-linear part of the relationship between an outcome of interest and age, period, or cohort; see §2.3.4.2. Third, restrictions on the model in terms of ξ can be imposed and tested using standard hypothesis testing frameworks; such restrictions may be implied by economic theories or simply desired for the sake of parsimony, see §2.3.4.3. Finally, the accelerations and other elements of ξ can be used in forecasting, see §2.3.4.4.

2.3.2 The classical APC model

The APC acceleration framework has its roots in the classical APC model. The classical APC model includes an indicator for each age, period, and cohort appearing in the dataset. This model is agnostic about functional form. It is well-known that the classical APC model is not identified due to the relationship $period = cohort + age - 1$[1], and there is an extensive literature addressing this lack of identification. The APC acceleration framework for aggregate data developed in Kuang et al. (2008) and Nielsen (2015) inherits features from this literature, as will be seen in this section. Other branches of the literature which do not make use of accelerations will be examined in §2.5.

2.3.2.1 Setting up the classical APC model

The classical APC model includes an indicator for each age, period, and cohort appearing in the dataset. Therefore, the first step in setting up the model is to determine the set of required indicators from the dataset.

Any aggregate dataset used for age-period-cohort analysis should be aggregated at the level of age-cohort combinations. An example is a dataset recording the unemployment rate at a given age for each of several cohorts, as seen in Table 2.1. This dataset is shown as an age-cohort array, where each row is an age and each column is a cohort. Each cell is referred to as an age-cohort cell.

[1]Why -1? This relation holds if age, period, and cohort are indexed such that cohort 1 is aged 1 in period 1. This indexing, referred to as time stamp accounting, is preferred because it simplifies later calculations. However, human ages are counted differently: babies are "zero years old" for all of their first year, only turning one in the beginning of their second year, yielding a relation of $period = age + cohort$. This is referred to as calendar accounting.

Table 2.1: US unemployment rate in an age-cohort array

age, cohort	1936-1940	1941-1945	1946-1950	1951-1955	1956-1960
20-24	0.076	0.060	0.082	0.136	0.115
25-29	0.038	0.046	0.086	0.080	0.075
30-34	0.037	0.067	0.057	0.064	0.051
35-39	0.060	0.049	0.054	0.044	0.046
40-44	0.042	0.049	0.039	0.040	0.029

Data from the OECD online database, copied from Fannon & Nielsen (2019)
Example: Top-left cell records unemployment in 1960 for the cohort 1936-1940.
In 1960 those born in 1936 were 24 and those born in 1940 were 20.

For an aggregate dataset represented as an age-cohort array, the classical model will include an indicator for each cohort (column), $c = 1 \dots C$, and an indicator for each age (row), $a = 1 \dots A$. In the example dataset in Table 2.1, $A = 5$ and $C = 5$. Indicators for period also need to be included. In this age-cohort array, the periods are given by the diagonals. In period 1 (1960, in the example), the first cohort is at the first age. In period 2, the first cohort is at the second age and the second cohort is at the first age. This continues for all periods, $p = 1 \dots P$. The relationship between the age, period, and cohort indices is thus given by the formula $p = a + c - 1$.

The linear predictor of the classical APC model for a dataset of this form is

$$\begin{aligned} \mu_{ac} = \delta &+ \alpha_1 \mathbb{1}(a = 1) + \cdots + \alpha_A \mathbb{1}(a = A) \\ &+ \beta_1 \mathbb{1}(p = 1) + \cdots + \beta_P \mathbb{1}(p = P) \\ &+ \gamma_1 \mathbb{1}(c = 1) + \cdots + \gamma_C \mathbb{1}(c = C). \end{aligned} \tag{2.5}$$

The left-hand side term, μ_{ac} is the value of the linear predictor implied by the classical APC model at the age-cohort cell $\{a, c\}$. This linear predictor is related to the outcome value y_{ac} at cell $\{a, c\}$ via a statistical model. For example, a linear model that can be estimated by OLS is

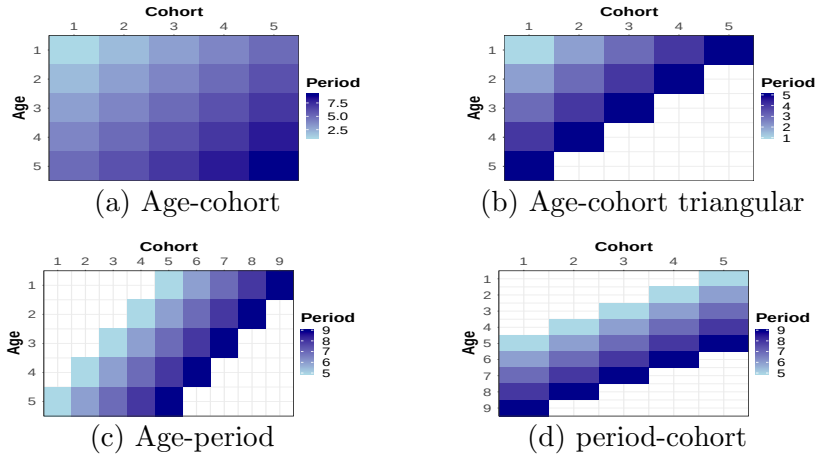
$$y_{ac} = \mu_{ac} + \varepsilon_{ac}, \tag{2.6}$$

for ε_{ac} a well-behaved noise term. On the right-hand side of equation (2.5) there are indicators for each age, period, and cohort appearing in the data. For example,

$\mathbb{1}(a = 1)$ is an indicator which takes the value 1 if the condition in brackets is satisfied for cell $\{a, c\}$ and 0 otherwise. Greek letters represent parameters: δ is a generalized intercept, α_a is the coefficient on the indicator for being at age a, β_p is the coefficient on the indicator for being in period p, and γ_c is the coefficient on the indicator for being a member of cohort c. This model is not identified, as will be explained in §2.3.2.2.

The dataset described above has an "age-cohort" structure, because it records values at given ages for a set of cohorts, but other structures exist. Three other structures are of interest. The first structure is age-cohort triangular, which is half of an age-cohort dataset (later periods have not been observed yet). The second structure is period-cohort, in which a set of cohorts are recorded for a fixed number of periods. The third structure is age-period, in which observations are recorded over a fixed range of ages for a fixed number of periods (a labour force survey is a good example). Each of these datasets can be represented as an portion of an age-cohort array, as seen in Figure 2.1.

Figure 2.1: Data structures as age-cohort arrays



(a) Age-cohort

(b) Age-cohort triangular

(c) Age-period

(d) period-cohort

For these three alternative structures, the classical APC model is set up in a very similar way as it was for data with an age-cohort structure. The only difference is in the period indexation. I now present a unified indexation which can account for all four data structures. Consider the age-period data in Figure 2.1c. The first periods in the age-cohort array - those periods in which the first

cohorts passed through the first ages - are not recorded in the dataset. Therefore the period index starts not at 1, but at $L+1$, where L is the number of periods in the age-cohort array not observed in the dataset. The general form of the classical APC model accounts for this by allowing for L in the indexing of period parameters:

$$\begin{aligned}\mu_{ac} = \delta &+ \alpha_1 \mathbb{1}(a=1) + \cdots + \alpha_A \mathbb{1}(a=A) \\ &+ \beta_{L+1} \mathbb{1}(p=L+1) + \cdots + \beta_{L+P} \mathbb{1}(p=L+P) \\ &+ \gamma_1 \mathbb{1}(c=1) + \cdots + \gamma_C \mathbb{1}(c=C). \end{aligned} \tag{2.7}$$

This encompasses the model for age-cohort data in equation (2.5), which is obtained by setting $L=0$.

The parameters of the classical model are collectively referred to as θ, which is of dimension $q = 1 + A + P + C$:

$$\theta = \{\delta, \alpha_1, \ldots, \alpha_A, \beta_{L+1}, \ldots, \beta_{L+P}, \gamma_1, \ldots, \gamma_C\}. \tag{2.8}$$

Two final considerations should be noted before we address the question of identification in the classical APC model. First, a feature of the classical APC model is that it is linear in age, period, and cohort; that is, it assumes no interaction effects between age, period, and cohort. Provision for testing this assumption, with repeated cross section data, is made in §3.4. It is also possible to test this assumption in aggregate data. Second, it is a feature of all the datasets presented above that they are contiguous, i.e. there are no empty age-cohort cells within the boundaries of the portion of the age-cohort array occupied by the data. Throughout this thesis, I assume all datasets worked with are contiguous. Further research is necessary to determine what violations of contiguity can be permitted without compromising identification.

2.3.2.2 The classical APC model is not identified

The full parameter vector θ in the classical APC model (2.7) cannot be identified because the model is over-parametrized due to the relationship $p = a + c - 1$.

To explain the identification problem it is useful to define a shorthand for model (2.7). Let model (2.7) be written as

$$\mu_{ac} = \delta + \alpha_a + \beta_p + \gamma_c. \tag{2.9}$$

Only those parameters associated with indicator variables which equal 1 in age-cohort cell $\{a, c\}$ appear.

The identification problem can be characterized by a group of transformations of model (2.9), defined by Carstensen (2007) as follows. For any constants $\{\mathfrak{a}, \mathfrak{b}, \mathfrak{c}, \mathfrak{d}\} \in \mathbb{R}$ the predictor μ_{ac} satisfies

$$\begin{aligned}\mu_{ac} = \{\alpha_a + \mathfrak{a} + (a-1)\mathfrak{d}\} + \{\beta_p + \mathfrak{b} - (p-1)\mathfrak{d}\} \\ + \{\gamma_c + \mathfrak{c} + (c-1)\mathfrak{d}\} + \{\delta - \mathfrak{a} - \mathfrak{b} - \mathfrak{c}\}.\end{aligned} \tag{2.10}$$

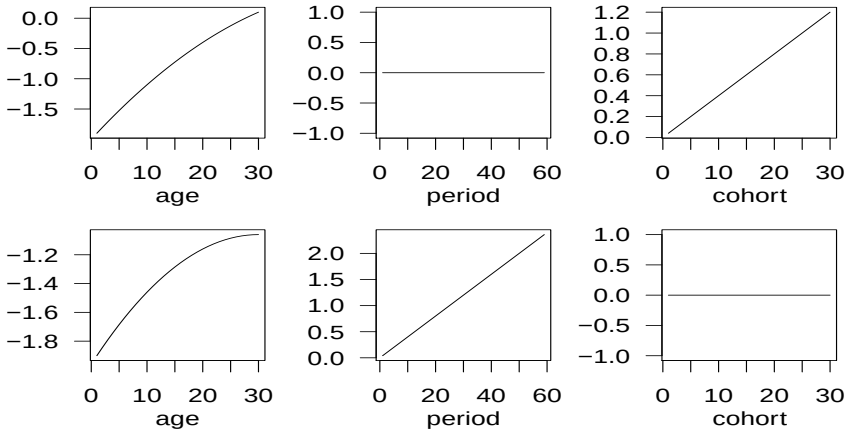
The constants $\{\mathfrak{a}, \mathfrak{b}, \mathfrak{c}, \mathfrak{d}\}$ cancel on the right hand side of (2.10), so that μ_{ac} does not depend on $\{\mathfrak{a}, \mathfrak{b}, \mathfrak{c}, \mathfrak{d}\}$. We say that μ_{ac} is invariant with respect to the transformations in (2.10). The individual effects α_a, β_p, γ_c, are not invariant to the transformation; they are only identified up to linear trends. For instance, the age effect α_a is observationally equivalent to $\alpha_a + \mathfrak{a} + (a-1)\mathfrak{d}$ for any $\{\mathfrak{a}, \mathfrak{d}\}$.

The implication of the identification problem is that the fit of model (2.9) with parameter vector θ is equal to that of model (2.9) with parameter vector θ^*, defined

$$\begin{aligned}\theta^* &= \{\delta^*, \alpha_1^*, \ldots, \alpha_A^*, \beta_{L+1}^*, \ldots, \beta_{L+P}^*, \gamma_1^*, \ldots, \gamma_C^*\} \\ \alpha_a^* &= \alpha_a + \mathfrak{a} + \mathfrak{d} \times a \\ \beta_p^* &= \beta_p + \mathfrak{b} - \mathfrak{d} \times p \\ \gamma_c^* &= \gamma_c + \mathfrak{c} + \mathfrak{d} \times c \\ \delta^* &= \delta - \mathfrak{a} - \mathfrak{b} - \mathfrak{c} - \mathfrak{d},\end{aligned} \tag{2.11}$$

As the fit of θ and θ^* are equal there is no unique θ which solves the model, i.e. the model is not identified. The following example is given in Fannon & Nielsen (2019): let $\mu_{ac} = -2 + 0.1a - 0.001a^2 + 0.04c$ as in Bell & Jones (2014). This could arise from θ defined by $\alpha_a = -2 + 0.1a - 0.001a^2$, $\gamma_c = 0.04c$, $\beta_p = \delta = 0$; or equally from θ^* defined by $\alpha_a^* = -1.96 + 0.06a - 0.001a^2$, $\beta_p^* = 0.04p$, $\gamma_c^* = \delta^* = 0$. In this example, θ^* arises from θ when choosing $\mathfrak{a} = 0$, $\mathfrak{b} = -\mathfrak{c} = -\mathfrak{d} = 0.04$. Thus, the vector θ is not identified.

Figure 2.2: APC effects of equal-likelihood classical APC model parameters



The severity of the identification problem is clear from Figure 2.2. The top row plots the age, period, and cohort effects implied by the parameter vector θ in the example from Bell & Jones (2014), while the bottom row plots the age, period, and cohort effects implied by the parameter vector θ^*. Under θ, there is a linear effect in cohort but not in period; under θ^* it is the opposite. Switching that linear effect from cohort, under θ, to period, under θ^*, also influences the age effect; the slope in age is much less steep in the bottom row. Clearly the choice between θ and θ^* will substantially impact interpretation, but both parameter vectors are an equally good fit to the data.

Kuang et al. (2008) show that the Carstensen transformation fully characterises the lack of identification in the model. There are not more than these four dependencies. Equation (2.10) therefore summarizes the $q - 4$ dimensional variation of the linear function from θ to μ.

2.3.3 Towards identification: Reparametrization of the classical APC model in terms of accelerations

The APC acceleration framework for aggregate data relies on a reparametrization of the classical APC model which addresses the lack of identification outlined in §2.3.2.2. The idea of reparametrization is to combine parameters in the original model (2.7) such that the number of parameters is reduced by the amount needed to resolve the under-identification. In the context of the classical APC model, it is

desirable to find a reparametrization that is invariant to the constants $\{\mathfrak{a}, \mathfrak{b}, \mathfrak{c}, \mathfrak{d}\}$ defined in §2.3.2.2. Such a reparametrization does not constrain the original parameters θ of the classical APC model, thus avoiding the problems associated with constraint-based approaches to the APC identification problem. Ideally, the reparametrized model should also be easy to interpret.

The reparametrization used in the APC acceleration framework for aggregate data is expressed in terms of a parameter vector ξ, which has four fewer elements than the original parameter vector θ, exactly addressing the under-identification. The reparametrization in terms of ξ is invariant to the constants $\{\mathfrak{a}, \mathfrak{b}, \mathfrak{c}, \mathfrak{d}\}$ defined in §2.3.2.2. The parameter ξ is also straightforward to interpret: it combines accelerations in age, period, and cohort, which have long been recognised as identified in the classical APC model, with a single linear plane.

2.3.3.1 Defining accelerations in the classical model

Accelerations in age, period, and cohort can be defined in terms of the parameters of the classical APC model. Recall that an acceleration in age captures how the effect of aging one year changes with age. In terms of the classical APC model the effect of aging one year at age a is given by $\Delta\alpha_a = \alpha_a - \alpha_{a-1}$. The change in that effect from the previous age is given by

$$\Delta^2\alpha_a = \Delta\alpha_a - \Delta\alpha_{a-1} = (\alpha_a - \alpha_{a-1}) - (\alpha_{a-1} - \alpha_{a-2}). \tag{2.12}$$

This term $\Delta^2\alpha_a$ is the acceleration at age a.

To show that the acceleration $\Delta^2\alpha_a$ is identified in the classical model, it must be shown to be invariant to the arbitrary constants $\{\mathfrak{a}, \mathfrak{b}, \mathfrak{c}, \mathfrak{d}\}$. That is, we must show that $\Delta^2\alpha_a = \Delta^2\alpha^*_a$ for α^*_a defined in (2.11). The proof is as follows:

$$\begin{aligned}
\Delta^2\alpha^*_a &= \alpha^*_a - 2\alpha^*_{a-1} + \alpha^*_{a-2} \\
&= \alpha_a + \mathfrak{a} + \mathfrak{d} \times a - 2[\alpha_{a-1} + \mathfrak{a} + \mathfrak{d} \times (a-1)] + \alpha_{a-2} + \mathfrak{a} + \mathfrak{d} \times (a-2) \\
&= \alpha_a - 2\alpha_{a-1} + \alpha_{a-2} + \underbrace{\mathfrak{a} - 2\mathfrak{a} + \mathfrak{a}}_{=0} + \mathfrak{d} \times [\underbrace{a - 2(a-1) + a - 2}_{=0}] \\
&= \Delta^2\alpha_a.
\end{aligned}$$

Similar proofs exist for period and cohort accelerations, $\Delta^2\beta_p$ and $\Delta^2\gamma_c$.

These accelerations have long been known to be identified from the classical model, and have been used in several applications. An early example is Clayton & Schifflers (1987a). McKenzie (2006) identified period accelerations in consumption from aggregate data and attributed the discontinuities they revealed to the Mexican peso crisis. He also identified age accelerations in consumption and used them to test the life cycle consumption hypothesis. Van Landeghem (2012) identified age accelerations in self-reported well-being from individual-level data and used them to evaluate the claim that well-being is U-shaped in age.

In previous applications using accelerations, the accelerations were constructed in multi-step procedures rather than being identified from a regression. The reliance on such cumbersome procedures may explain the limited adoption of the use of accelerations by researchers. The alternative approach of embedding the accelerations in a reparametrized version of the classical model, which can be estimated using standard linear regression, is expected to expand the use of accelerations. This is the approach of the framework developed for aggregate data in Nielsen (2015), and for individual-level data in this thesis.

2.3.3.2 The idea of reparametrization

Reparametrization is a general approach to dealing with overparametrized models which preserves all variation present in the model while still permitting the identification of a unique set of parameters. The number of parameters to be estimated is reduced by combining parameters of the original model into a new parameter vector which is identified and freely varying. In the context of the classical APC model, the original parameter vector θ, of dimension q, is combined to produce a new parameter vector ξ, which is selected to have the following properties. First, ξ should be of dimension $q - 4$. Second, it should be uniquely identified from the data. Third, ξ should be invariant to the arbitrary constants $\{\mathfrak{a}, \mathfrak{b}, \mathfrak{c}, \mathfrak{d}\}$. This invariance indicates that the variation present in the original model is preserved; the reparametrization to ξ does not constrain θ in any way.

To illustrate the idea of reparametrization, consider a simple example of estimating the effect of eye colour on vitamin D absorption for a population of

individuals indexed by i. The proposed model is

$$D_i = \psi_0 + \psi_1 \mathbb{1}(eye_i = blue) + \psi_2 \mathbb{1}(eye_i = brown) + \psi_3 \mathbb{1}(eye_i = green). \quad (2.13)$$

Here ψ_0 is the baseline level of vitamin D absorption and the coefficients ψ_1 through ψ_3 capture the additional effect on vitamin D absorption of having blue, brown, or green eyes respectively.

Unfortunately it is not possible to identify all four parameters. The model is over-parametrized by one degree. The vector $\{\psi_0, \psi_1, \psi_2, \psi_3\}$ is observationally equivalent to an alternative parameter vector defined in terms of an arbitrary constant $\mathfrak{g}$: $\{\psi_0 + \mathfrak{g}, \psi_1 - \mathfrak{g}, \psi_2 - \mathfrak{g}, \psi_3 - \mathfrak{g}\}$.

To address the under-identification the model can be reparametrized:

$$D_i = \phi_1 \mathbb{1}(eye_i = blue) + \phi_2 \mathbb{1}(eye_i = brown) + \phi_3 \mathbb{1}(eye_i = green)$$
$$\phi_1 = (\psi_0 + \psi_1) \quad ; \quad \phi_2 = (\psi_0 + \psi_2) \quad ; \quad \phi_3 = (\psi_0 + \psi_3)$$

The parameter ϕ_1 represents the combined, identified effect of baseline plus blue-eye-specific vitamin D absorption. The separate effects ψ_0 and ψ_1 cannot be identified, but the reparametrization does not limit their variation; that, is the parameter vector $\{\phi_1, \phi_2, \phi_3\}$ is invariant to the choice of $\mathfrak{a}$. This is a very simple example of the reparametrization approach taken by Nielsen (2015) to address the under-identification in the classical APC model.

An alternative approach to such identification problems that is sometimes used involves the imposition of untestable identifying constraints. For example, in model (2.13), I could impose the constraint that $\psi_1 = 0$, i.e. there is no additional effect of having blue eyes on vitamin D absorption. Under the assumption that this constraint is valid, the remaining original parameters are identified: $\phi_1 = \psi_0, \phi_2 = \psi_2, \phi_3 = \psi_3$. However this approach is not invariant to $\mathfrak{g}$; it relies on $\mathfrak{g} = 0$. The implication is that this approach constrains the variation of the underlying parameter. This creates two problems. First, there is no way to test if the constraint is valid. Second, there is a risk that in interpretation it may be forgotten that $\phi_1 = \psi_0, \phi_2 = \psi_2, \phi_3 = \psi_3$ if and only if the constraint is valid. In the example above, the risk would be that a large estimated value of ϕ_1 is taken

to imply a large baseline level of vitamin D absorption, where in fact this interpretation relies on the assumption that $\psi_1 = 0$. Constraints-based approaches to identification of the classical APC model are considered in §2.5.1.

2.3.3.3 APC reparametrization: overview

Kuang et al. (2008) developed a reparametrization of the classical APC model in (2.7) which combines the original q parameters of θ into a new parameter vector ξ of dimension $q - 4$, which contains the accelerations defined in §2.3.3.1. This reparametrization has two advantages. First, it preserves the full variation present in the classical model but is uniquely identified. Second, it incorporates the accelerations, which had been discussed in previous literature.

The first insight used in developing this reparametrization is that any single age, period, or cohort parameter in the classical APC model can be represented as a telescopic sum of accelerations as well as an initial parameter and slope parameter. Consider for example the age parameter, α_5, which can be represented as follows

$$\alpha_5 = \alpha_1 + 4\Delta\alpha_2 + 3\Delta^2\alpha_3 + 2\Delta^2\alpha_4 + \Delta^2\alpha_5. \tag{2.14}$$

Here α_1 is the initial parameter, $\Delta\alpha_2 = \alpha_2 - \alpha_1$ is a slope in age, and the parameters $\{\Delta^2\alpha_3, \Delta^2\alpha_4, \Delta^2\alpha_5\}$ are accelerations. In fact each age indicator α_a can be represented in terms of the same initial parameter α_1, slope parameter $\Delta\alpha_2$, and a telescopic sum of accelerations, as follows:

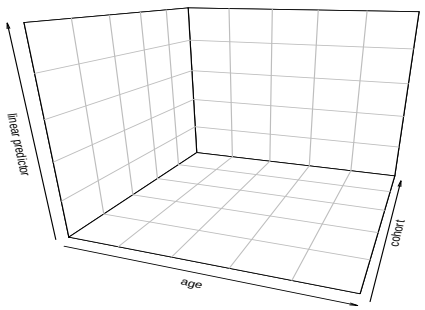
$$\alpha_a = \alpha_1 + \sum_{r=2}^{a} \Delta\alpha_r \qquad\qquad \Delta\alpha_r = \Delta\alpha_2 + \sum_{s=3}^{r} \Delta^2\alpha_s. \tag{2.15}$$

It is straightforward to see that the application of this decomposition to α_5 results in equation (2.14). Similar representations could be created for an alternative choice of slope or initial parameter; the telescopic sum would then be replaced by an alternative procedure for summation of accelerations. Such representations, in terms of a single initial parameter, slope, and accelerations, can also be created for period and cohort parameters.

The second insight is that the initial parameters and slope parameters in each of age, period, and cohort can be combined into three parameters which define a single linear plane in age-cohort space. Age-cohort space is a three-dimensional

space as seen in Figure 2.3. The horizontal axes of the space describe an age-cohort array like those seen in Figure 2.1. The vertical axis records the value of the linear predictor μ_{ac} associated with a given age-cohort cell. The single linear plane in age-cohort space, constructed by combining the initial parameters and slope parameters, can be used in conjunction with the sums of accelerations to study the shape of the relationship between the linear predictor and age, period, and cohort. It can also be used for forecasting. The initial parameters in age, period, and cohort are combined with the general intercept δ from parameter vector θ to form the origin point of the linear plane. The three slopes in age, period, and cohort can be reduced to the two slopes of a linear plane due to the relationship $p = a + c - 1$.

Figure 2.3: Age-cohort space



There are many possible linear planes and associated procedures for summing accelerations, because the sums in (2.15) could be constructed for any choice of initial parameter and slope. Indeed it is a general property of reparametrization as a technique that each model admits multiple reparametrizations. However, each reparametrization is unique in the sense that once a reparametrization is chosen, only one solution for the reparametrized vector is consistent with the data.

Kuang et al. (2008) and Nielsen (2015) focus on a particular choice of linear plane which is useful for estimation. This linear plane allows age and cohort accelerations to be summed symmetrically, facilitating construction of a design matrix. It is described in §2.3.3.4.

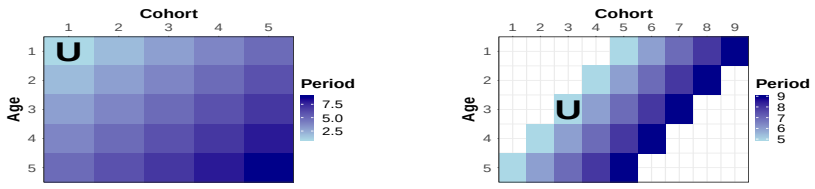
2.3.3.4 APC reparametrization: detail

The procedure for constructing the reparametrized classical model, in terms of the parameter vector ξ, has five steps. The first step is to select an appropriate linear plane. The second step is to express all parameters in the classical APC model in terms of the initial parameters and slopes of that linear plane, plus accelerations, using the approach in (2.15). The third step is to combine these re-expressed parameters to get an expression for the linear predictor μ_{ac} in terms of the linear plane and accelerations. The fourth step is to verify that the new parameter ξ has dimension $q - 4$. In the fifth step, a concise representation of both the full parameter vector ξ and associated design vector x_{ac} is defined.

The first step is to select a linear plane which results in a simple design matrix. Simplicity is achieved by selecting a plane in which age and cohort are treated symmetrically and the origin point of the linear plane is close to the origin corner of the age-cohort space. The linear plane is defined by three parameters: an origin point and two slopes. I first consider the origin point, then the slopes.

The origin point of the linear plane is chosen to be the first point where the age and cohort indices are equal in the portion of the age-cohort space occupied by the dataset. This origin point is shown for two data structures displayed in age-cohort space in Figure 2.4. The origin point is marked by the letter U. Figure 2.4a shows data that has an age-cohort structure, while Figure 2.4b shows data that has an age-period structure; the different data structures were defined in §2.3.2.1.

Figure 2.4: Location of origin point in data in age-cohort space



(a) Age-cohort (b) Age-period

Visualisations like those in Figure 2.4 are not necessary to find the origin point of the linear plane; it can be calculated from the data-determined constant L. Recall from §2.3.2.1 that L is the number of periods in the age-cohort space that predate data collection, and is used to index the period parameters in the classical

APC model: $\{\beta_{L+1}, \ldots, \beta_{L+P}\}$. For data with an age-cohort structure as in Figure 2.4a, $L = 0$. Define U to be the integer part of $(L+3)/2$, that is, $U = \lfloor (L+3)/2 \rfloor$. The origin of the linear plane is the point where $a = c = U$. The period in which that point lies is $p = 2U - 1$. The value of the linear predictor μ_{ac} from the classical APC model at the origin point is given by

$$v_o = \delta + \alpha_U + \beta_{2U-1} + \gamma_U. \tag{2.16}$$

This value v_o is the intercept, or origin point, parameter of the linear plane.

The slopes of the linear plane are constructed symmetrically along the age and cohort axes of the age-cohort space, with respect to the origin point of the linear plane $\{U, U\}$. The first slope is obtained by taking a unit increment along the age axis of the age-cohort space away from this point. The slope associated with this unit increment along the age axis is the difference between the value of the linear predictor at $\{U, U\}$, and the value of the linear predictor at the adjacent cell $\{U+1, U\}$. This difference,

$$\begin{aligned} v_a =& \mu_{U+1,U} - \mu_{U,U} \\ =& \alpha_{U+1} + \beta_{2U} - \alpha_U - \beta_{2U-1} \\ =& \Delta\alpha_{U+1} + \Delta\beta_{2U}, \end{aligned} \tag{2.17}$$

is the first of the two slopes of the linear plane. The second slope is based on taking a unit increment away from the origin point along the cohort axis of the age-cohort space, i.e. the difference between the value of the linear predictor at $\{U, U\}$, and the value of the linear predictor at the adjacent cell $\{U, U+1\}$:

$$\begin{aligned} v_c =& \mu_{U+1,U} - \mu_{U,U} \\ =& \beta_{2U} + \gamma_{U+1} - \gamma_U - \beta_{2U-1} \\ =& \Delta\gamma_{U+1} + \Delta\beta_{2U}. \end{aligned} \tag{2.18}$$

Having defined the linear plane, the second step is to express each of the parameters of the classical APC model (2.7) in terms of APC accelerations plus the parameters of the linear plane. This is done using telescopic sums, such as the following for age parameters of the form α_a

$$\alpha_a = \alpha_U + \sum_{r=U+1}^{a} \Delta\alpha_r \qquad \Delta\alpha_r = \Delta\alpha_{U+1} + \sum_{s=U+2}^{r} \Delta^2\alpha_s. \tag{2.19}$$

For example, α_{U+4} can be decomposed into

$$\alpha_{U+4} = \alpha_U + 4\Delta\alpha_{U+1} + 3\Delta^2\alpha_{U+2} + 2\Delta^2\alpha_{U+3} + \Delta^2\alpha_{U+4}. \tag{2.20}$$

Note the similarity to equations (2.14) and (2.15). Similar decompositions can be defined for period and cohort parameters.

The third step is to use these decompositions to express the linear predictor of the classical APC model, μ_{ac}, in terms of accelerations and parameters of the linear plane. Insert the decompositions for the age, period, and cohort parameters into the linear predictor $\mu_{ac} = \alpha_a + \beta_p + \gamma_c + \delta$. The result is

$$\begin{aligned}\mu_{ac} =& (\delta + \alpha_U + \gamma_U + \beta_{2U-1}) \\ &+ \Delta\alpha_{U+1}(a - U) + \Delta\gamma_{U+1}(c - U) + \Delta\beta_{2U}[p - (2U - 1)] \\ &+ \mathbb{A}_a + \mathbb{P}_p + \mathbb{C}_c\end{aligned} \tag{2.21}$$

where $\mathbb{A}_a, \mathbb{P}_p$, and $\mathbb{C}_c$ are sums of accelerations in each of age, period, and cohort:

$$\begin{aligned}\mathbb{A}_a &= \mathbb{1}(a < U) \sum_{r=a+2}^{U+1} \sum_{s=r}^{U+1} \Delta^2\alpha_s + \mathbb{1}(a > U + 1) \sum_{r=U+2}^{a} \sum_{s=U+2}^{r} \Delta^2\alpha_s \\ \mathbb{P}_p &= \mathbb{1}(L \textit{ odd } \& \; p = 2U - 2)\Delta^2\beta_{2U} + \mathbb{1}(p > 2U) \sum_{r=2U+1}^{p} \sum_{s=2U+1}^{r} \Delta^2\beta_s \\ \mathbb{C}_c &= \mathbb{1}(c < U) \sum_{r=c+2}^{U+1} \sum_{s=r}^{U+1} \Delta^2\gamma_s + \mathbb{1}(c > U + 1) \sum_{r=U+2}^{c} \sum_{s=U+2}^{r} \Delta^2\gamma_s\end{aligned}$$

These definitions are drawn from p.62 of Nielsen (2015). Recall that for data with an age-cohort structure, $L = 0$ so $U = 1$ and

$$\begin{aligned}\mathbb{A}_a =& \mathbb{1}(a > 2) \sum_{r=3}^{a} \sum_{s=3}^{r} \Delta^2\alpha_s \\ \mathbb{P}_p =& \mathbb{1}(p > 2) \sum_{r=3}^{p} \sum_{s=3}^{r} \Delta^2\beta_s \\ \mathbb{C}_c =& \mathbb{1}(c > 2) \sum_{r=3}^{c} \sum_{s=3}^{r} \Delta^2\beta_s.\end{aligned}$$

The three difference terms in equation (2.21), $\Delta\alpha_{U+1}$, $\Delta\beta_{2U}$, and $\Delta\gamma_{U+1}$, are combined to produce the two slopes of the linear plane, v_a and v_c. This is done using the relation $a + c = p + 1$. Line 2 of (2.21) becomes

$$(\Delta\alpha_{U+1} + \Delta\beta_{2U})(a - U) + (\Delta\gamma_{U+1} + \Delta\beta_{2U})(c - U) \tag{2.22}$$

where $(\Delta\alpha_{U+1} + \Delta\beta_{2U}) = v_a$ and $(\Delta\alpha_{U+1} + \Delta\beta_{2U}) = v_c$. Therefore the new parameter vector ξ is

$$\xi = \{v_o, v_a, v_c, \Delta^2\alpha_3, \ldots, \Delta^2\alpha_A, \Delta^2\beta_{L+3}, \ldots, \Delta^2\beta_{L+P}, \Delta^2\gamma_3, \ldots, \Delta^2\gamma_C\}. \tag{2.23}$$

The fourth step is to show that this new representation in terms of ξ contains exactly $q - 4$ parameters. Recall that q is the dimension of θ, i.e. one general intercept plus an indicator for each age, period, and cohort: $q = 1 + A + P + C$. The new representation has three parameters, v_o, v_a, and v_c, describing a single linear plane. It also has a complete set of accelerations in age, period, and cohort. The total number of accelerations is $A - 2 + P - 2 + C - 2$ (subtract 2 because accelerations cannot be defined for the first two ages, periods, or cohorts in the sample). Thus the total number of parameters in ξ is $3 + A + P + C - 6 = q - 4$.

The final step of the procedure is to write the reparametrized linear predictor in terms of a design vector x_{ac} and a parameter vector ξ: $\mu_{ac} = x'_{ac}\xi$. The design vector x_{ac} summarizes the linear plane and all cumulations of accelerations performed in $\mathbb{A}_a, \mathbb{P}_p$, and $\mathbb{C}_c$. It is defined as follows, with $m(r, s) = max(r - s + 1, 0)$:

$$x_{ac} = \{1, (a - U), (c - U), x^{\mathbb{A}}_{ac}, x^{\mathbb{P}}_{ac}, x^{\mathbb{C}}_{ac}\} \tag{2.24}$$

where $x^{\mathbb{A}}_{ac}$, $x^{\mathbb{P}}_{ac}$, and $x^{\mathbb{C}}_{ac}$ cumulate accelerations in age, period, and cohort:

$$x^{\mathbb{A}}_{ac} = m(1, a), \ldots, m(U - 1, a), m(a, U + 2), \ldots, m(a, A), \tag{2.25}$$

$$x^{\mathbb{P}}_{ac} = \begin{cases} \mathbb{1}(p = 2U - 2), m(p, 2U + 1), \ldots, m(p, 2U - 3 + P) & \text{for } L \text{ odd} \\ m(p, 2U + 1), \ldots, m(p, 2U - 2 + P) & \text{for } L \text{ even} \end{cases} \tag{2.26}$$

$$x^{\mathbb{C}}_{ac} = m(1, c), \ldots, m(U - 1, c), m(c, U + 2), \ldots, m(c, C) \tag{2.27}$$

The parameter vector is

$$\xi = \{v_o,\ v_a,\ v_c, \xi^{\mathbb{C}}, \xi^{\mathbb{P}}, \xi^{\mathbb{C}}\}, \tag{2.28}$$

where $\xi^{\mathbb{A}}$, $\xi^{\mathbb{P}}$, and $\xi^{\mathbb{C}}$ collect accelerations in age, period, and cohort respectively:

$$\begin{aligned}\xi^{\mathbb{A}} =& \Delta^2\alpha_3, \ ..., \Delta^2\alpha_A \\ \xi^{\mathbb{P}} =& \Delta^2\beta_{L+3}, \ ..., \Delta^2\beta_{L+P} \\ \xi^{\mathbb{C}} =& \Delta^2\gamma_3, \ ..., \Delta^2\gamma_C. \end{aligned} \tag{2.29}$$

The APC acceleration framework for aggregate data embeds this reparametrized linear predictor $\mu_{ac} = x'_{ac}\xi$ in a regression model. The regression model links the linear predictor to the outcome of interest y_{ac} and permits estimation of the parameters in ξ. Thus, it is possible to obtain estimates of the accelerations without needing to use the multi-stage procedures of previous studies.

2.3.4 Analysis with the reparametrized classical APC model

The APC acceleration framework, which estimates the APC accelerations as elements of the parameter vector ξ, can be used in at least four ways.

First, estimates of the accelerations alone can be used to evaluate policy changes or shocks. I explain this usage in §2.3.4.1.

Second, the sum of accelerations provided for in the design vector x_{ac} can be used to examine the shape of the non-linear relationship between an outcome of interest and age, period, or cohort. To isolate the non-linear part of the relationship, the sum of accelerations must be detrended. §2.3.4.2 presents the procedure for summing and detrending accelerations, and subsequently constructing a visual representation of the shape of the non-linear relationship between an outcome of interest and age, period, or cohort.

Third, restrictions on the reparametrized classical model can be tested, either with the goal of achieving a more parsimonious representation of the data or with the goal of testing some functional form restriction. The functional form restriction might be implied by economic theory or imposed for convenience. I discuss some common restrictions in §2.3.4.3.

Fourth, projections based on the reparametrized model can be used for forecasting. I give examples in §2.3.4.4.

2.3.4.1 Accelerations identify discontinuities

Estimates of the accelerations can be used to identify discontinuities in the relationship between an outcome of interest and age, period, or cohort. For example, McKenzie (2006) identifies a negative period acceleration in Mexican consumption in 1996, using data recorded at two-year intervals. This means that between 1994 and 1996, consumption either grew more slowly than it had between 1992 and 1994, or declined more rapidly. Either one indicates depressed consumption, which McKenzie attributes to the 1995 peso crisis.

McKenzie (2006) use a multi-step procedure to estimate accelerations, but the general framework described in §2.3.3 allows all accelerations to be estimated from any generalized linear model with linear predictor $\mu_{ac} = x'_{ac}\xi$. McKenzie's procedure is restricted to linear models of the form $y_{ac} = \mu_{ac} + \varepsilon_{ac}$ and is not generalizable to other linear models such as the logit. McKenzie's procedure for estimating accelerations is as follows: first, differences of the outcome variable between adjacent cells by age, period, and cohort are constructed. Second, differences of those differences along age, period, and cohort are constructed. An average of the appropriate sub-set of these differences is constructed to estimate a single acceleration. This procedure is repeated for each acceleration. In the general framework described in §2.3.3, each acceleration is an element of the parameter vector ξ and so is estimated by regression. Relative to the procedure used by McKenzie, this framework is faster and reduces the potential for error. Furthermore, this framework is suitable for all generalized linear models, whereas McKenzie's framework is suitable only for linear models that can be estimated by ordinary least squares.

2.3.4.2 Sums of accelerations describe relationships

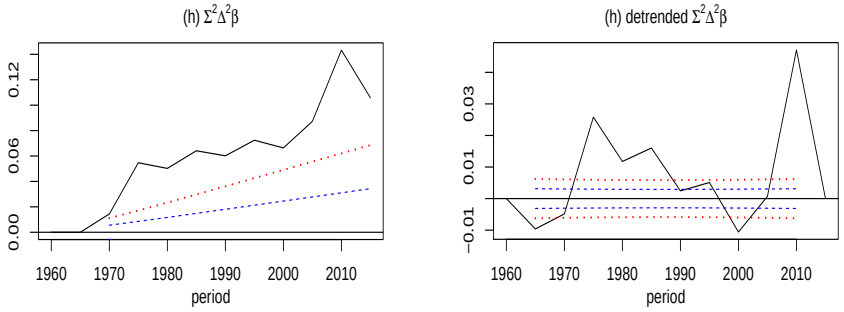
The sums of accelerations implied by the combination of the design vector x_{ac} and the parameter ξ can be used to study the shape of the relationship between an outcome and age, period, and cohort. Consider the sum of age accelerations at age a, given by $\mathbb{A}_a = x^{\mathbb{A}}_{ac}\xi^{\mathbb{A}}$. By plotting the value of this sum on the Y-axis against age a on the X-axis, one can visualise how the relationship between age and the outcome of interest evolves with age. One can determine whether the relationship is uniformly concave or convex, or some other shape. For example, Nielsen (2015)

finds a concave relationship between age and the number of mesothelioma deaths in Belgium, whereas Fannon & Nielsen (2019) find a convex relationship between age and the proportion of the US labour force that is unemployed. Similarly the shape of the relationship between an outcome and cohort can be studied by plotting $\mathbb{C}_c = x^{\mathbb{C}}_{ac}\xi^{\mathbb{C}}$ against cohort, and the shape of the relationship between an outcome and period can be studied by plotting $\mathbb{P}_p = x^{\mathbb{P}}_{ac}\xi^{\mathbb{P}}$ against period.

It is important here to only consider the non-linear part of the shape. The non-linear part of the shape, which describes deviations from linearity such as convexity or concavity, is identified through the accelerations. The linear part of the shape, indicating whether an outcome increases or decreases with age, period, or cohort, is not identified.

The reparametrization in terms of ξ, with sums of accelerations given by $\{\mathbb{A}_a, \mathbb{P}_p, \mathbb{C}_c\}$, as described in §2.3.3.3, is not optimal for visualising the non-linear part of the shape. Instead, it is optimized for ease of construction of the design vector by allowing age and cohort to be treated symmetrically. As a consequence, the estimated linear plane is highly dependent on the age, period, and cohort effects in the vicinity of U. This may lead to a problem where there is the appearance of a slope in the summed accelerations, as seen in Figure 2.5a.

Figure 2.5: US unemployment rate, summed period accelerations



(a) $\mathbb{P}_p$ as in §2.3.3.3 (b) Detrended representation $\mathbb{P}^{dt}_p$

Data for these figures is from the OECD online database, see Fannon & Nielsen (2019).
Blue dashed, red dotted lines = 1, 2 standard deviations from zero.

A slightly different parametrization of the classical APC model, which selects a different linear plane and summation of the accelerations, is preferred for isolating

the non-linear part of the shape. It is important to note that the accelerations are not changed by this different parametrization; it is only the design vector x_{ac} and the linear plane parameters $\{v_o, v_a, v_c\}$ that change. The linear plane in this parametrization is chosen to ensure that the sum of accelerations in each of age, period, and cohort is anchored to begin and end in zero. The parameter vector associated with this parametrization is ξ^{dt}, with the dt referring to the "detrending" implied by anchoring the sums of accelerations at zero. This is defined

$$\xi^{dt} = \{v_o^{dt}, v_a^{dt}, v_c^{dt}, \xi^{\mathbb{A}}, \xi^{\mathbb{B}}, \xi^{\mathbb{C}}\}. \tag{2.30}$$

The sum of period accelerations resulting from this detrended representation, $\mathbb{P}_p^{dt}$, is seen in Figure 2.5b. It uses the same data and accelerations as Figure 2.5a but clearly isolates the non-linear part of the shape.

The APC detrended acceleration parametrization in terms of ξ^{dt} can be constructed as a bijective mapping from the ξ parametrization. This is useful because the design matrix associated with ξ^{dt} is difficult to construct from scratch. The new linear plane parameters are constructed as follows:

$$\begin{aligned}
v_o^{dt} =& v_o - (U-1)(v_a + v_c) - \mathbb{A}_1 - \mathbb{P}_{P+1} - \mathbb{C}_1 - \frac{L}{P-1}(\mathbb{P}_{L+P} - \mathbb{P}_{L+1}) \\
v_a^{dt} =& v_a + \frac{1}{A-1}(\mathbb{A}_A - \mathbb{A}_1) + \frac{1}{P-1}(\mathbb{P}_{L+P} - \mathbb{P}_{L+1}) \\
v_c^{dt} =& v_c + \frac{1}{C-1}(\mathbb{C}_C - \mathbb{C}_1) + \frac{1}{P-1}(\mathbb{P}_{L+P} - \mathbb{P}_{L+1}).
\end{aligned}$$

The sums of accelerations are given by

$$\begin{aligned}
\mathbb{A}_a^{dt} &= (x_{ac}^{\mathbb{A}dt})'\xi^{\mathbb{A}} = \mathbb{A}_a - \mathbb{A}_1 - \frac{a-1}{A-1}(\mathbb{A}_A - \mathbb{A}_1) \\
\mathbb{P}_p^{dt} &= (x_{ac}^{\mathbb{P}dt})'\xi^{\mathbb{P}} = \mathbb{P}_p - \mathbb{P}_1 - \frac{p-L-1}{P-1}(\mathbb{P}_{L+P} - \mathbb{P}_{L+1}) \\
\mathbb{C}_c^{dt} &= (x_{ac}^{\mathbb{C}dt})'\xi^{\mathbb{C}} = \mathbb{C}_c - \mathbb{C}_1 - \frac{c-1}{C-1}(\mathbb{C}_C - \mathbb{C}_1).
\end{aligned}$$

The model with the new ξ^{dt} parametrization and design vector x_{ac}^{dt} is then

$$\mu_{ac} = (x_{ac}^{dt})'\xi^{dt} = v_o^{dt} + (a-1)v_a^{dt} + (c-1)v_c^{dt} + \mathbb{A}_a^{dt} + \mathbb{P}_p^{dt} + \mathbb{C}_c^{dt}. \tag{2.31}$$

These relationships are derived in Nielsen (2015).

This detrended parametrization is used exclusively for exploring the shape of the relationship between age, period, or cohort and some outcome. It was used to identify concavity and convexity in Nielsen (2015) and Fannon & Nielsen (2019), described above. For other applications using accelerations, the parametrization in terms of ξ is used because it is easier to estimate. I provide illustrations of how the detrended parametrization adapted to individual-level data may be used in §3 and §4 of this thesis.

2.3.4.3 Testing restrictions on the APC acceleration model

The regression framework makes it easy to test restrictions on the APC model, via likelihood ratio or Wald tests. The main APC model can be compared against sub-models, which restrict certain elements of the parameter vector ξ to zero. Seven categories of sub-model are considered (Nielsen, 2015; Oh & Holford, 2015).

AC/PC/AP First, the absence of accelerations in one of the APC effects can be tested. For instance, the absence of period accelerations is tested by imposing $\Delta^2\beta_{L+3} = \cdots = \Delta^2\beta_{L+P} = 0$. This gives an Age-Cohort (AC) model. In terms of the unidentified original parametrization θ, this is written as $\beta_{L+1} = \ldots \beta_{L+P} = 0$. The two formulations of the hypothesis are in fact equivalent (Nielsen & Nielsen, 2014). The latter formulation obscures the degrees of freedom and hides that the hypothesis does not constrain the linear period effects. Period-Cohort (PC) and Age-Period (AP) models are analogous to AC models.

Ad/Pd/Cd Second, the absence of accelerations in two components can be tested. For example, the absence of period and cohort accelerations is tested by imposing $\Delta^2\beta_{L+3} = \cdots = \Delta^2\beta_{L+P} = 0$ and $\Delta^2\gamma_3 = \cdots = \Delta^2\gamma_C = 0$, while leaving the linear plane unrestricted. This is the Age-drift (Ad) model (Clayton & Schifflers, 1987a). Analogous models are Cohort-drift (Cd) and Period-drift (Pd).

A/C/P Third, a model with only one slope in the linear plane can be tested. For instance, the Age (A) model is obtained by imposing $v_c = 0$ on the Ad model. The constraint is that $v_c = \Delta\gamma_{U+1} + \Delta\beta_{2U} = 0$. In this case, $v_a = \Delta\alpha_{U+1} + \Delta\beta_{2U}$ identifies a combined age-period slope. The linear

slope of the age effect, $\Delta\alpha_{U+1}$, remains unidentifiable. Analogous models are Cohort (C) and Period (P).

t Fourth, a pure linear plane model can be tested, where all double differences $\Delta^2\alpha_a$, $\Delta^2\beta_p$, $\Delta^2\gamma_c$ are set to zero. This is the trend (t) model.

tA/tP/tC Fifth, the trend can be further restricted. The tA model is constrained to have $v_c = 0$, while the tC model is constrained to have $v_a = 0$. The tP model is constrained to have $v_a = v_c$.

1 Sixth, the intercept (1) model has neither slope nor accelerations. All parameters are constrained to equal zero except v_o.

Finally, functional form restrictions may be of interest. For instance, a quadratic age effect $\alpha_a = \lambda_0 + \lambda_1 a + \lambda_2 a^2$ arises when $\Delta^2\alpha_3 = \cdots = \Delta^2\alpha_A = 2\lambda_2$ is imposed (Fannon & Nielsen, 2019; McKenzie, 2006).

An example of testing model restrictions is seen in Martínez Miranda et al. (2015), where restriction to the "AC" model is considered. Testing of model restrictions in the framework for individual-level data is explored in §3 and §4.

2.3.4.4 Forecasting

Estimates of the acceleration parameters can be used in conjunction with the linear plane to produce forecasts. First, the age-cohort space is extended to cover the future periods which it is desired to forecast. Then, the APC acceleration model is extended to incorporate the new age, period, and cohort accelerations required. These additional acceleration terms will need to be forecast, perhaps via an AR(1) or other time-series model for the accelerations in that time dimension. The linear plane is projected forward to cover the new age-cohort cells. Note that an implicit assumption of no structural change underpins any forecasts generated in this fashion. An example is the forecasting of mesothelioma mortality in Martínez Miranda et al. (2015). Further examples of epidemiological forecasting using the ξ reparametrization framework are Zhao et al. (2019), Ji et al. (2019), and Oddone et al. (2020).

2.4 Advantages of the reparametrized classical APC model

There are at least three benefits to using the APC acceleration framework to study the effects of age, period, and cohort. First, the parameter ξ is uniquely identified. Second, the parametrization is invariant to θ, meaning that it does not impose any constraints on the classical APC model. Third, the reparametrization is in terms of accelerations, which are familiar objects in the APC literature and have a natural interpretation. The APC acceleration framework permits these accelerations to be estimated from regression.

The parameter ξ is unique with respect to a particular set of linear predictors μ_{ac}. This is in contrast to the parametrization in terms of θ, where multiple values of θ fit the data equally well. For a given design vector x_{ac} and set of linear predictors μ_{ac}, there is a single ξ which provides the best fit.

The parameter ξ is invariant to the transformations in terms of $\{\mathfrak{a}, \mathfrak{b}, \mathfrak{c}, \mathfrak{d}\}$ described in equation (2.10) of §2.3.2.2. Corollary 2 in Kuang et al. (2008) shows that ξ is a maximal invariant function of θ with respect to (2.10). This means that for any θ which satisfies the linear predictors μ_{ac}, when it is reparametrized in terms of ξ, the same values of ξ will be obtained. The implication is that the reparametrization in terms of ξ does not limit the variation of θ. This stands in contrast to other approaches to identification that are discussed in §2.5. Requiring invariance draws focus to the identifiable part of θ, avoiding the part of θ that cannot be identified. While not exactly solving the APC identification problem, this approach makes the problem ignorable. Here, ξ represents what can be learned in an APC model, while the transformations (2.10) describe what cannot be learned.

The fact that the ξ parametrization is built around accelerations, which are well-known to be identified, means that it can be more readily used and interpreted. There are previous examples of successful use of accelerations in forecasting (Martínez Miranda et al., 2015); evaluating the impact of shocks (McKenzie, 2006); and testing theories (McKenzie, 2006; Van Landeghem, 2012). Two additional use cases, finding of new stylized facts and testing convenience functional form restrictions, are illustrated in §3 and §4 respectively.

The embedding of the accelerations in a linear regression framework via the ξ reparametrization is an improvement on the previous literature using accelerations. It makes the accelerations more accessible to researchers who are familiar with regression but may be put off by the multi-step procedures previously used to estimate accelerations. It simplifies model selection, because it is easy to test restrictions on the model. Finally it facilitates additional applications, including forecasting and using the non-linear shape of the APC effects to find new stylized facts. For forecasting, the combined linear plane provided for in ξ is essential. The design matrix x_{ac}^{dt} described in §2.3.4.2 provides for summation of the accelerations in a way that visually isolates the non-linear part of the relationship between an outcome of interest and age, period, and cohort.

2.5 Other approaches to analysis of APC effects

Several other approaches to APC analysis are commonly used in economics and other social sciences. The most common approach in economics uses the classical APC model in a regression framework; but instead of reparametrization, constraints are imposed to identify the model. In the best examples of this approach, the identifying constraints are carefully chosen based on the application at hand (e.g. Lagakos et al., 2018); but in many instances little justification is provided for the choice of constraints (Bardazzi & Pazienza, 2018; Bíró, 2017). The imposition of untestable and weakly-justified constraints creates problems for interpretation, since the estimated APC effects from these models depend upon the constraints.

A second approach common in economics is to replace the APC terms in the classical model with latent variables that reflect the phenomena the APC effects are expected to capture (Heckman & Robb, 1985). Again, problems of interpretation arise when this approach is not used carefully.

A third approach, less common in economics, is to use a model that is not derived from the classical APC model. A well-known example is the Lee-Carter model (Lee & Carter, 1992), used in actuarial science. A thorough analysis of such models is beyond the scope of this thesis, but they are also subject to the APC identification problem and often implicitly impose identifying constraints,

creating the potential for problems with interpretation analagous to those arising when constraints are imposed on the classical APC model.

A fourth approach is to abandon the effort to get precise estimates of age, period, and cohort effects and instead focus on careful inspection of plots of the data. This is useful in cases where the data is such that the effect of interest is clearly visible from the plots, as in Almond (2006), but that is not always the case.

A final phenomenon seen in economics, although not a solution to the APC problem, is the failure to recognise the APC identification problem at all, for example in Rosenthal (2014). In the following I consider the relationship between these approaches and the reparametrization framework outlined in §2.3.

2.5.1 Constraints on the classical APC model

The imposition of constraints to identify the classical APC model is common in economics. These constraints may be just-identifying or over-identifying. All constraints approaches carry a risk of mis-interpretation if the sensitivity of the estimated effects to the constraints is not fully appreciated.

Just-identifying constraints impose four constraints on the original parameter vector θ of the classical APC model. The new constrained parameter vector θ_c has a dimension of variation equal to $q - 4$ and is therefore identifiable.

The use of just-identifying constraints can result in mis-interpretation if the dependence of the resulting estimates on the constraints is not fully understood. Essentially, a problem arises if the solution for θ_c is mistaken as the solution for θ. It is only true that $\theta_c = \theta$ when the imposed just-identifying constraints are true; and it is not possible to test these constraints. A failure to recognise this may lead the applied researcher to make statements about θ when in fact those statements only apply to θ_c. A typical example would be a claim about the linear relationship between age, period, or cohort and some outcome of interest implied by the estimated θ_c, which fails to recognise that the linear relationship is only identified by the constraints.

A simple example of a just-identifying constraint is the approach suggested by Mason et al. (1973) and used by Ejrnæs & Hochguertel (2013). In this approach, four elements of θ are constrained to equal zero: one each in age, period, and

cohort, plus one further parameter. For example, one could set $\alpha_1 = \beta_{L+1} = \gamma_1 = \beta_{L+2} = 0$. The first three constraints (one parameter in each of age, period, and cohort) are standard normalizing constraints used when working with complete sets of dummy variables. The constraint on the additional parameter is needed as a consequence of the relationship $p = a + c - 1$, and it is this constraint that creates problems for interpretation. This constrains the linear slope in one of the three of age, period, and cohort, so that any remaining linear effect due to that variable must be distributed between the other two.

Another set of just-identifying constraints, common in economics, is that used by Deaton & Paxson (1994). Again, there are four constraints in total: three normalizing constraints and one linear constraint. The normalizing constraints restrict the sums of age, period, and cohort parameters to zero:

$$\sum_a \alpha_a = \sum_p \beta_{L+p} = \sum_c \gamma_c = 0. \tag{2.32}$$

The additional constraint is that there is no linear effect of period on consumption. The constraint takes the form

$$\sum_p p\beta_p = 0. \tag{2.33}$$

One could alternatively set the linear effect in age or cohort to zero. The sensitivity of the estimated effects to the fourth constraint is illustrated by Figure 4 of Lagakos et al. (2018), where one can see how the estimated trends in wages depend on whether the linear effects are constrained to be zero in period or in cohort.

A range of other just-identifying constraints exist. One class of just-identified models impose the fourth constraint by selecting a generalized inverse with particular properties. This class includes the intrinsic estimator of Yang et al. (2004) and the procedure described in Schulhofer-Wohl (2018). O'Brien (2015) provides a discussion of these models, and a refutation of the claim that the constraint implied by the generalized inverse results in $\theta_c = \theta$. There are also Bayesian versions of the constraints-based approach, see the discussion in Fannon & Nielsen (2019).

It is difficult to envision a situation where the use of just-identifying constraints could be preferred to the reparametrization outlined in §2.3. Interpretation of a just-identified model requires the researcher to keep track of how the imposed constraints affect the estimates, which is not straightforward (see the discussion in

Nielsen & Nielsen, 2014). The APC acceleration model does not have this requirement. The just-identified model may be preferred where the researcher has complete confidence in the imposed constraints, so that $\theta_c = \theta$; then the just-identified model permits inference on the linear effects, which the APC acceleration model does not. However, in most applications in economics the underlying phenomena are too complex, and our understanding of them too limited, for such confidence in untestable identifying constraints to be warranted.

Models which involve over-identifying constraints are also common in economics. These approaches share the problem of the just-identifying constraints approach, that the impact of the constraints on the estimated effects is often not fully appreciated. There are many types of over-identifying constraint. One is to assume that there are no effects attributable to one of the three of age, period, and cohort; for example Méndez & Sepúlveda (2012) assume that there are no cohort effects on skill acquisition. A second type involves restricting large numbers of age, period, or cohort effects to be equal, by creating "bands" within which there is no change in the relationship to the outcome of interest. For example, Bíró (2017) uses 10-year cohort bands in his analysis of durable good consumption.

Another over-identifying constraints approach, more common in sociology and demography than economics, is the hierarchical APC (HAPC) model (Yang & Land, 2006). This is model includes a polynomial in age, while period and cohort are modelled as normal, mean-zero random effects. The random effects approach effectively constraints the period and cohort effects, as is discussed in §3.1.2.2 of Fosse & Winship (2019) and §5.4.6 of Nielsen & Nielsen (2014).

The reparametrization framework is preferable to these constraints approaches because it can yield useful insights but is not vulnerable to mis-interpretation. While some of those who use constraints-based methods do so with appropriate care and recognise the sensitivity of their estimates to the identifying assumption (an excellent example is Lagakos et al., 2018), studies which do not take appropriate care are common. There is a particular risk if the constraints are believed to "solve" the identification problem, i.e. if θ_c is confused with θ. It is therefore important to have a general method for APC analysis which does not carry risks of mis-interpretation. The APC acceleration framework satisfies this need.

2.5.2 Latent variables

The latent variables approach treats age, period, and cohort as "proxies" for underlying latent variables which affect the outcome of interest (Heckman & Robb, 1985). A good latent variables analysis is time-consuming but is more informative about the determinants of the outcome of interest than an APC model. The APC acceleration framework is complementary to such an analysis; it can be used to guide the selection of latent variables. However, mis-application of the latent variables approach shares many of the problems that affect constraints-based analysis.

Good latent variables analysis adheres to the original idea of Heckman & Robb (1985), that all three of age, period, and cohort should be replaced by the latent variables and thus a better understanding of the outcome variable would be gained. An example is Kapteyn et al. (2005), where the focus of the paper is on determining which latent variables drive the cohort effects. The authors use economic theory to identify two candidate latent variables, and then test for the additional explanatory power of non-linear cohort effects while also accounting for age and period non-linearities. The APC acceleration framework can be used for this sort of testing. It can also be used to identify candidate latent variables; for example, in §4 my examination of the sums of age accelerations reveals the importance of accounting for childbirth in the model for hospital in-patient stays.

Problems arise when it is assumed that replacing one of the three of age, period, or cohort with a latent variable permits identification of the effects of the remaining two of age, period, and cohort. For example, in Jensen et al. (2001) industry-wide labour productivity and output variables are used to replace period so that the firm age and vintage effects on productivity can be estimated. This is in effect similar to the constraints approaches: the period effect is constrained to be equal to the effect of the latent variable. The estimated effects of age and cohort will be sensitive to this constraint, which cannot be tested.

2.5.3 Models other than the classical APC model

A range of alternative APC models exist, which do not inherit their structure from the classical APC model. Extensive analysis of these is beyond the scope of this paper. However, it is worth noting that they frequently suffer from the same issues

that arise with constraints on the classical APC model. A well-known example is the Lee-Carter model (Lee & Carter, 1992), typically used for mortality modelling. This differs from the classical APC model because it is not linear in age, period, and cohort effects. Instead, the cohort and period effects are interacted, while the age effects enter additively. Nielsen & Nielsen (2014) discuss how the identification of this model is subject to *a priori* constraints like those imposed on the classical APC model in §2.5.1.

2.5.4 Visual analysis

In some studies, no effort is made to statistically separate the effects of age, period and cohort. Instead the focus is on graphical analysis. This works well when the goal is simply to know whether APC effects exist, rather than to quantify them. An example is the work of Almond (2006), which examines the long-term effects of being *in utero* during the 1918 influenza pandemic. There are clear discontinuities in the data associated with the 1918 cohort, see for example his Figures 2 and 3. Another example is Voas & Chaves (2016), who consider the relationship between religious affiliation and the passage of time. They find that for each cohort, the relationship between the passage of time (reflecting either age or period or both) and religious affiliation is flat, but is shifted down compared to the previous cohort. Such a graph implies either pure cohort effects or perfectly balanced age and period effects; the authors argue that the latter is unlikely.

The limitation of this approach is that it is not widely applicable. The data often does not exhibit patterns that are clearly attributable to one of age, period, or cohort. Visual analysis breaks down in such situations. Additionally, it is often important to quantify the magnitude of effects, which is not possible with visual analysis. This is why Almond (2006) follows up his graphical analysis with estimation of deviations from linearity in cohort.

2.6 Conclusion

By reparametrizing the classical age-period-cohort (APC) model, Kuang et al. (2008) and Nielsen (2015) developed a general regression framework for estimating the identifiable effects of age, period, and cohort from aggregate data. The

identifiable effects are accelerations in each of age, period, and cohort, as well as a combined linear plane. These effects have been used in economic applications for data exploration, forecasting, evaluating the impact of policy changes and exogenous shocks, and testing theoretical claims. This thesis develops a complementary general regression framework for estimating age, period, and cohort accelerations, and a combined linear plane, from individual-level data.

The APC acceleration framework has several advantages relative to other approaches which are used to estimate the effects of age, period, and cohort. Other approaches which focus on the identifiable effects, in particular those focused on estimating accelerations, rely on multi-step procedures which are more cumbersome to implement than the APC acceleration framework. Approaches which do not focus on the identifiable effects rely on untestable identifying constraints. The extent to which the estimated APC effects are dependent on the constraints in these models is often not appreciated, leading to inaccurate interpretation of the estimates. The APC acceleration framework focuses on effects which are invariant to these constraints and so is less vulnerable to inaccurate interpretation.

The parameters identified from the APC acceleration framework have been used in a range of applications. The identified parameters are a single linear plane, which combines the linear effects of age, period, and cohort, and accelerations in each of age, period, and cohort. An acceleration in age, for example, captures how the effect of aging one year changes as a person ages. Accelerations have been used with aggregate data to estimate discontinuities arising from currency shocks and to test economic theories such as the life-cycle hypothesis (McKenzie, 2006). Accelerations in conjunction with the linear plane have been used to explore the age, period, and cohort effects in unemployment data (Fannon & Nielsen, 2019) and to forecast mesothelioma mortality (Martínez Miranda et al., 2015).

Despite the range of applications for which the APC acceleration framework is suitable, it has been relatively under-utilized in economics. I suspect four main factors are driving this. First, there may be a lack of awareness of the problems associated with other approaches to APC identification. This hypothesis is supported by the continued use of these approaches in recent papers, such as Bardazzi & Pazienza (2018). Second, there may be an incompatibility issue: many of the applied studies in economics where APC effects are of interest rely on individual-level

data (see §2.2), but the reparametrization framework was developed for aggregate data. Third, there may be a lack of understanding of how accelerations can be effectively used in economic applications. This hypothesis is supported by the fact that applied papers often cite methodological papers on accelerations, but do not discuss them in detail (e.g. Fukuda, 2013). Fourth, researchers may be deterred by the cumbersome multi-step procedures previously used to estimate accelerations.

In this thesis, I address each of the four factors that I suspect are driving the under-utilization of the APC acceleration framework. First, I have already drawn attention to the problems associated with other approaches to APC identification in §2.5. Second, in next two chapters of the thesis I develop a theoretical framework for using the APC acceleration framework with individual-level data, considering repeated cross section data in §3 and panel data in §4. Third, in each of §3, §4, and §5 I provide a different application or example which illustrates how accelerations may be used. Fourth, in §5 I develop an R package which implements the theoretical framework developed in §3 and §4, enabling accelerations to be estimated from regression in a single command.

Chapter 3

The framework for repeated cross section data, with an application to obesity in England 2001-2014

3.1 Introduction

This paper develops a new framework for the analysis of repeated cross section data where an individual's age, birth cohort, and period of observation are explanatory variables of interest. The framework is suitable for both continuous and binary variables. The age, period, and cohort (APC) effects are represented by an acceleration-based model which separates the identifiable non-linear parts of the APC effects from the unidentifiable linear parts of the APC effects. This parameters of this APC acceleration model are freely varying, and invariant to the well-known APC identification problem. We develop a test of the APC acceleration model against a more general "time-saturated" model. The test resembles a deviance test but can be used for both continuous and binary outcomes. The new framework is applied to an analysis of obesity in England, using both continuous and binary measures of obesity and accounting for the effect of covariates such as socio-economic status and alcohol consumption. Men and women are analysed separately. We find that the main deviations from linearity present in English obesity data are age-related among women and cohort-related among men.

The proposed framework uses generalized linear models (GLMs) for repeated cross section data, with both APC effects and individual covariates appearing in the linear predictor. GLMs are used because they can accommodate several types of dependent variables, including continuous and binary variables. The well-known APC identification problem, which is discussed at some length below, is addressed using the APC acceleration model of Kuang et al. (2008). This model has previously only been used with aggregate data. The combination of the acceleration model with GLMs gives simple likelihood functions. This makes it easy to formulate model restrictions as well as more general models, and to pose associated statistical tests. In this spirit, we propose to test the APC acceleration model against a more general time-saturated model, with a fixed effect for each age-cohort combination.

The APC identification problem is well-known, see §2.3.2 of this thesis as well as Holford (1983), Clayton & Schifflers (1987b), Glenn (2005), Carstensen (2007), O'Brien (2011), and Fannon & Nielsen (2019). We focus on the APC identification problem as it appears in the classical APC model; other models were mentioned briefly in §2.5.3, but are outside the scope of this thesis. In the classical APC model, the linear predictor contains an additive combination of the age, period, and cohort effects. The parameters are fixed effects for each age, period, and cohort appearing in the data. Knowing the linear predictor only allows partial identification of these parameters, due to the relationship $period = age + cohort - 1$[1]. This lack of complete identification is the APC identification problem.

A common approach to the identification problem as it appears in the classical APC model is to impose constraints on the APC effects. The problem with this constraints approach is that it is difficult to separate what is learned about the APC effects from the data and what is learned from the constraints, as discussed in §2.5.1. This in turn generates challenges with respect to interpretation, inference and recursive analysis; see Nielsen & Nielsen (2014) for a formal analysis.

[1]Why -1? This relation holds if we allow cohort 1 to be aged 1 in period 1, called time stamp accounting. We prefer this relationship because it greatly simplifies later calculations. However, human ages are counted differently: babies are "zero years old" for all of their first year, only turning one in the beginning of their second year, yielding a relation of $period = age + cohort$. This is called calendar accounting.

An alternative is the estimable functions approach, which focuses on those parts of the APC effects which are identifiable. Here, constraints are used to estimate the classical APC model, but only functions of the estimated APC effects that are invariant to the identification problem are interpreted. Accelerations in age, period, and cohort are examples of estimable functions. Theoretical discussion of the estimable functions approach can be found in Holford (1983) and Clayton & Schifflers (1987b). Accelerations have been estimated from individual-level data using the estimable functions approach by Van Landeghem (2012). The problem of the estimable functions approach is that it is cumbersome to implement; first the constrained classical APC model must be estimated, and then the accelerations constructed as functions of the estimated parameters of that model.

The framework developed in this chapter addresses the APC identification problem by focusing on APC accelerations, which are identified, rather than the unidentified APC slopes. Since it focuses on the identifiable parts of the APC effects, the framework is similar to the estimable functions approach. We improve upon the estimable functions approach by estimating the accelerations directly from regression, which is less cumbersome. This is achieved by reparametrizing the linear predictor of the classical APC model, following Kuang et al. (2008), and embedding that reparametrized linear predictor in a regression model. The reparametrization has two important features: the parameter vector is freely varying, and has desirable invariance properties. First, since the parameter vector is freely varying, it is canonical in a generalized linear model sense. This eases imposition and interpretation of hypotheses as well as counting the associated degrees of freedom. Second, the parameter vector is invariant to the identification problem. It is also invariant to incorporating additional data waves and other variations of the APC time horizons. Specifically, the parameter vector consists of: accelerations in each of age, period, and cohort, and a linear plane combining the inseparable APC slopes. We call this model the APC acceleration model.

The asymptotic analysis of the APC acceleration model can be understood by thinking of the data as a two-way age-cohort array with individual information accumulating in each cell. In the asymptotic analysis we keep the dimension of the age-cohort array fixed and exploit the individual level information for inference. The statistical analysis is then fairly simple. This asymptotic approach

resembles earlier work on two-way arrays of aggregate data where each cell entry is large, including a Poisson model for counts of cancer deaths (Martínez Miranda et al., 2015) and an over-dispersed Poisson model for insurance claims (Harnau & Nielsen, 2018). Asymptotics for aggregate data with a large period dimension have been considered by Fu (2016). That approach would be inappropriate for most repeated cross section data, including the obesity data used in the application in this chapter, due to the small period dimension.

We develop a test of the APC acceleration model against a more general model, where each cell in the age-cohort array has its own parameter. We call this more general model the time saturated (TS) model. The TS model nests the APC model. The proposed test of the APC acceleration model against the TS model resembles a deviance test, but is suitable for both continuous and binary outcomes. Inference is standard, but we address some computational challenges.

Applications with repeated cross section data where age, period, and cohort are of interest fall into two categories: those in which age, period, and cohort appear as controls, and those in which age, period, and cohort are the explanatory variables of interest. First, in some studies researchers are primarily interested in the effect of individual level covariates, but include APC variables as controls. An example is the study of unemployment insurance in Ejrnæs & Hochguertel (2013). In that case the same results for the effects of interest will be obtained from either the APC acceleration model or a set of just-identifying constraints, as we show in Appendix 3.A.2. Second, in the application to obesity in this chapter and in other studies the APC effects are of primary interest.

In our application we consider the evolution of adult obesity in England, which is a major public health concern. UK obesity rates almost tripled between 1980 and 2011, with over a quarter of adults estimated to be obese by 2016 (Department of Health, 2011; Moody, 2016). An individual is obese if their body mass index (BMI) exceeds 30. Obesity is linked to immediate and long term health risks, such as type II diabetes. The direct healthcare costs of obesity in 2006-07 were estimated to be £5.1 billion, 6% of the National Health Service budget (Scarborough et al., 2011). Reducing obesity has thus been a policy goal for many years, with specific government directives issued in 2007, 2011, and 2016. Policies to reduce obesity will have the greatest effect if they target the most at-risk subpopulations. The

analysis in this paper helps to identify such subpopulations by showing how obesity evolves with age and cohort, independent of population level period effects.

We use data from the 2001 through 2014 waves of the Health Survey for England. Our dependent variable is either the continuous measure, log BMI, or a binary obesity indicator. The explanatory variables include age, period, and cohort, as well as socio-demographic covariates. It is the age, period, and cohort effects that are of primary interest.

When applying our methods to the data from the Health Survey for England, 2001-2014, we find significant age accelerations for women and significant cohort accelerations for men. For women, the age accelerations result in a concave relationship between age and obesity. This concavity is consistent with previous research (Lean et al., 2013; Wang et al., 2011; Howel, 2011). For men, the cohort accelerations result in a a concave relationship between cohort and obesity. This relationship was not detected by previous studies. In previous studies using the Health Survey for England, cohort effects were omitted entirely (Zaninotto et al., 2009; Howel, 2011; Wang et al., 2011); while international studies typically imposed constraints on the cohort effects (Allman-Farinelli et al., 2008; Reither et al., 2009; Peeters et al., 2015; An & Xiang, 2016). The effects of covariates are broadly consistent with existing literature.

The paper is outlined as follows: §3.2 introduces the data and its APC structure, as well as the APC identification problem. It also reviews the reparametrization developed by Kuang et al. (2008), which we use to address the APC identification problem. §3.3 and §3.4 contain the main theoretical contributions of this paper. In §3.3, we discuss the conditions for standard inference in normal and logit models (continuous and binary outcomes, respectively) using the APC acceleration model and repeated cross section data. In §3.4 a new test is proposed which compares the APC acceleration model to a more general model, and an algorithm for this test is developed. §3.5 contains the application of the APC acceleration framework to analyse obesity dynamics in England, while §3.6 concludes.

3.2 Preliminaries: data and APC acceleration model

In this section we introduce the English obesity data and the standardized APC indices that we use to discuss the APC acceleration model. We give an overview of the classical APC model and the APC identification problem. We then describe how the APC acceleration model is constructed as a reparametrization of the classical APC model that avoids the APC identification problem.

3.2.1 Obesity data

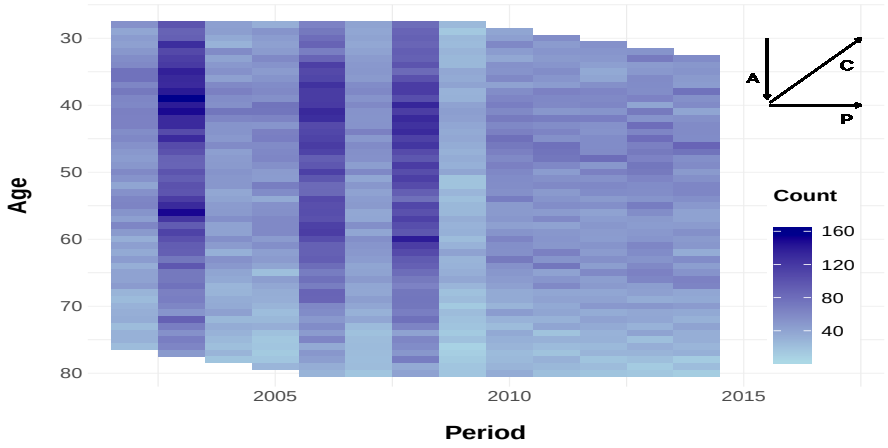
The data is a repeated cross section composed of representative samples of the English population taken from the Health Survey for England (HSE). We use waves from 2001-2014. The waves prior to 2001 do not include the National Statistics Socio-Economic Classification (NSSEC), one of our explanatory variables, while age is only recorded in five-year bands after 2014. For the analysis, we used the `apc.indiv` software described in §5, which is implemented in R (R Core Team, 2019). We separately analyse data for 43,077 women and 38,316 men. The sample size is denoted N and individuals are indexed $i = 1, \ldots, N$.

For each individual we have information on weight and height, measured by the interviewer. The body mass index (BMI) is defined as weight in kilograms divided by the square of height in metres. A few observations with BMI outside the range 12 to 60 were presumed to be subject to measurement error and excluded. In addition to BMI, age, and period of observation, we observe the following covariates: ethnicity, level of education, NSSEC, smoking history, and alcohol consumption. Tables 3.3 and 3.4 in Appendix 3.B report descriptive statistics.

We consider two choices of dependent variable: either log BMI or an indicator for obesity defined as $BMI \geq 30$. For each individual i then y_i is the dependent variable, a_i is the individual's age, p_i indicates the period in which the individual is observed, and c_i is the cohort of the individual constructed through $c_i = p_i - a_i + 1$.

In this dataset age and period vary in a rectangular age-period array, where age is between 28 and 80 and period is between 2001 and 2014 (see Figure 3.1). We therefore have $A = 53$ age groups and $P = 14$ period groups. Cohort then

Figure 3.1: Within-cell observation counts women



varies between 1921 and 1986. However, we exclude the first and last five cohorts as they are sparsely observed. This leaves $C = 56$ cohort groups. The final data, as an age-period array, is shown in Figure 3.1. The shading in that figure reflects the variation in survey size across waves.

To construct the standardized APC indices, we switch to an age-cohort coordinate system. This is preferred because of the age-cohort symmetry in the relation $a + c = p + 1$. Thus, throughout the paper we consider ages $a = 1, \ldots, A$ and cohorts $c = 1, \ldots, C$. We define the period index through $p = a + c - 1$ and get an index set $\mathcal{I}$ of the form

$$1 \leq a \leq A, \qquad 1 \leq c \leq C, \qquad L + 1 \leq p \leq L + P. \tag{3.1}$$

Here L is the necessary offset in the period index due to beginning the age and cohort indices at 1. With the present data, we then have that $A = 53$, $P = 14$, $C = 56$, $L = 48$ so that $age = 28$, $per = 2001$, $coh = 1921$ correspond to $a = 1$, $p = L + 1$, $c = 1$.

3.2.2 APC acceleration model

We account for the effects of age, period, and cohort in our analysis by embedding the APC acceleration model in a individual-specific linear predictor which also includes covariates. In this section, we review the APC acceleration model, which was discussed in detail in §2.3.

The APC acceleration model is derived from the classical APC model. The classical APC model is defined at the level of the age-cohort cell, as follows:

$$\mu_{ac} = \alpha_a + \beta_p + \gamma_c + \delta. \tag{3.2}$$

On the left-hand side of equation (3.2), μ_{ac} captures the predicted value from the classical APC model at cell $\{a, c\}$. Since μ_{ac} is determined by the age a and cohort c, it could be thought of as a function of age a and cohort c: $\mu_{ac} = \mu(a, c)$. We will use the notation μ_i to refer to the μ_{ac} associated with the age and cohort of individual i; using the function notation just defined, we can say $\mu_i = \mu(a_i, c_i)$. This μ_i will be later embedded in an individual-specific linear predictor η_i.

On the right-hand side of the classical APC model in equation (3.2), the terms α_a, β_p, and γ_c are fixed effects for age a, period p, and cohort c respectively. The full set of parameters of the classical APC model is θ, which is defined

$$\theta = \{\delta, \alpha_1, \ldots, \alpha_A, \beta_{L+1}, \ldots, \beta_{L+P}, \gamma_1, \ldots, \gamma_{L+c}\}. \tag{3.3}$$

In addition to the APC fixed effects, θ includes the general intercept δ. The overall dimension of θ is $q = A + P + C + 1$. We can then write

$$\mu_i = d_i'\theta, \tag{3.4}$$

where d_i is a design vector indicating which elements of θ appear in μ_i, the predicted value from the classical APC model for individual i.

The classical APC model in (3.2) is not identified, for the reasons outlined in §2.3.2.2. The vector θ contains four more parameters than can be identified from the set $\{\mu_i\}_{i \in N}$.

We resolve the identification problem by reparametrizing the classical APC model in the manner introduced by Kuang et al. (2008) and outlined in §2.3.3 of this thesis. The reparametrization is a mapping from $\mu_i = d_i'\theta$ to $\mu_i = x_i'\xi$, where x_i is a design vector and ξ is a parameter vector. Both x_i and ξ are of dimension $q - 4$. The new parametrization $\mu_i = x_i'\xi$ is referred to as the APC acceleration model, because the majority of the parameters it identifies are accelerations in each of age, period, and cohort.

The new parameter vector ξ is

$$\xi = (v_o, v_a, v_c, \Delta^2\alpha_3, \ldots, \Delta^2\alpha_A, \Delta^2\beta_{L+3}, \ldots, \Delta^2\beta_{L+P}, \Delta^2\gamma_3, \ldots, \Delta^2\gamma_C)'. \quad (3.5)$$

Here, v_o, v_a, v_c, are the level and two slopes of a linear plane while parameters of the form $\Delta^2\alpha_a, \Delta^2\beta_p, \Delta^2\gamma_c$ are accelerations in each of age, period, and cohort. The linear plane could be chosen in various ways, as discussed in §2.3.4.2, but here the plane parameters are those used in §2.3.3.4, which are defined

$$\begin{aligned} v_o &= \alpha_U + \beta_{2U-1} + \gamma_U + \delta \\ v_a &= \Delta\alpha_{U+1} + \Delta\beta_{2U} \\ v_c &= \Delta\gamma_{U+1} + \Delta\beta_{2U}. \end{aligned} \quad (3.6)$$

The acccelerations in ξ represent the non-linear parts of the APC effects. Corollary 2 in KNN shows that ξ is identifiable from a set of linear predictors $\{\mu_i\}_{i\in N}$ associated with a dataset that occupies a contiguous portion of an age-cohort array.

The new design vector x_i can be defined as a function of the age a_i and cohort c_i of individual i, so $x_i = x(a_i, c_i)$. The precise definition of the function $x_{ac} = x(a, c)$ is given by Nielsen (2015) and repeated in §2.3.3.4.

3.3 The APC acceleration model for repeated cross section data

We have described the data and the APC acceleration model. We now introduce generalized linear models for repeated cross section data which incorporate both the APC acceleration model and other explanatory variables of interest, called covariates. We consider a normal model for continuous outcomes and a logistic model for binary outcomes. We discuss estimation and inference for both cases.

3.3.1 The generalized linear model

We use generalized linear models for the dependent variable, y_i. The linear predictor η_i is a function of the APC acceleration model and covariates through

$$\eta_i = z_i'\zeta + \mu_i, \qquad \mu_i = x_i'\xi. \quad (3.7)$$

Here, ζ is a d_z-vector of parameters and z_i is a d_z-vector of covariates. These covariates can include regressors varying at the individual level or APC interaction effects. The term μ_i describes the APC acceleration model for an individual i with age a_i and cohort c_i observed in period $p_i = a_i + c_i - 1$, as described in §3.2.2.

In matrix notation, we stack η_i, μ_i, z_i, x_i over individuals to get η, μ, X, and Z. The linear predictor can then be written as

$$\eta = Z\zeta + \mu, \qquad \mu = X\xi. \tag{3.8}$$

The matrix X will have full column rank as long as $A, P, C \geq 2$. We will require that the combined design matrix (Z, X) also has full column rank.

When age, period, and cohort are the main explanatory variables of interest we recommend using the APC acceleration model, rather than any of the constraints methods commonly used to estimate the classical APC model. However, if the primary interest lies in the covariate coefficients ζ and it is only desired to control for the effects of age, period, and cohort, then any set of just-identifying constraints on the classical APC model will result in the same estimates of ζ as the APC acceleration model. This is demonstrated in Appendix 3.A.2. Ejrnæs & Hochguertel (2013) give an empirical example, where z_i includes both individual level regressors and a non-linear interaction effect.

3.3.2 The normal model

For continuous dependent variables a normal model is employed of the form

$$y_i = \eta_i + \varepsilon_i \qquad \text{for } i = 1, \ldots, N. \tag{3.9}$$

Conditional on the linear predictor, the errors ε_i are assumed independently $\mathsf{N}(0, \sigma^2)$ distributed. The model is estimated by ordinary least squares.

When it comes to inference, some parameters of the APC acceleration model may be unnecessary. Achieving a more parsimonious APC model is desirable for interpretation and for forecasting purposes. The standard sub-models of the APC acceleration model were described in §2.3.4.3. Likewise, some elements of covariate vector z_i may be redundant.

Exact inference on the parameters $\{\zeta, \xi\}$ can be performed using t- and F-tests by appealing to the classical results for analysis of variance under normality. Standard asymptotic inference can be conducted under weaker assumptions. In particular, the likelihood ratio test statistic will be asymptotically χ^2. For this we must be clear about the repetitive structure. We treat the dimensions A, P, C of the age-cohort array as fixed, but assume the number of individual observations N is large. To justify asymptotic inference we assume

1. The triplets y_i, x_i, z_i are i.i.d. across individuals i.
2. The covariance matrix of (z_i, x_i) is positive definite.
3. The errors ε_i satisfy $\mathsf{E}(\varepsilon_i|x_i, z_i) = 0$ and $\mathsf{Var}(\varepsilon_i|x_i, z_i) = \sigma^2$.

These assumptions imply that the relative frequencies of different age-cohort combinations are independent of the sample size N. In our data, the annual variation in sample size is due to financial constraints of the HSE and thus unrelated to the distribution of BMI. Similarly, it is plausible that selection into the survey is independent of the covariates z_i. It is therefore quite likely that inferences can be extrapolated beyond the sample (Wooldridge, 2010, §19.4).

3.3.3 The logit model

For binary dependent variables, a logistic model is employed with

$$\log\{\mathsf{P}(y_i = 1)/\mathsf{P}(y_i = 0)\} = \eta_i \qquad \text{for } i = 1, \ldots, N. \tag{3.10}$$

The corresponding logit log-likelihood is

$$\ell(\zeta, \xi) = \sum_{i=1}^{N} \eta_i y_i - \sum_{i=1}^{N} \ln(1 + \exp \eta_i). \tag{3.11}$$

It is strictly concave when the design matrix has full rank so the maximum likelihood estimator is unique (Wedderburn, 1976). It is finite in the absence of separation or quasi-separation (Agresti, 2013, §6.5). Under these conditions the maximum likelihood estimator can be found by Newton iteration.

The asymptotic theory of the logit estimator is outlined by Fahrmeir & Kaufmann (1986). Their Theorem 2 shows consistency and asymptotic normality under the assumptions 1 and 2 listed for the normal model in §3.3.2. The asymptotic

variance-covariance matrix of the logit estimator is given by $J = -\ddot{\ell}$, where $\ddot{\ell}$ is the second derivative of the log-likelihood. Theorem 3 of Fahrmeir & Kaufmann (1986) shows that likelihood ratio test statistics on the covariate parameter ζ and the APC parameter ξ are asymptotically χ^2.

3.4 A more general model: the time-saturated model

In practice, the APC acceleration model may be too parsimonious. There could be multiplicative relationships between age, period, and cohort, as in the Lee & Carter (1992) model, which are ruled out by the additive classical APC model from which the APC acceleration model is derived. Such multiplicative relationships could capture, for instance, the effect on obesity of a time- and age-limited government programme promoting healthy eating in schools. Alternatively, the values of the APC acceleration parameters could chnge over time. To address these possibilities, we provide a mis-specification test of the APC model against a time-saturated (TS) model, where μ_i, the portion of the linear predictor that captures the effects of age, period, and cohort, is unconstrained. We use a likelihood ratio test of the APC model against the TS model. For aggregate data, where there is only one observation per age-cohort cell, the test of APC against TS is a deviance test for the logistic model and not feasible for a normal model. However, with repeated cross section data the test applies for both the logistic and the normal model.

In the time-saturated model we replace the APC acceleration model $\mu_{ac} = \alpha_a + \beta_p + \gamma_c + \delta$ in (3.2) with a complete unstructured specification of μ_{ac}, using an indicator for each age-cohort combination. We replace the design vector x_i, of dimension $q - 4$, with an unit vector t_i, of dimension n, indicating the age-cohort cell to which individual i belongs. Here $q - 4 \ll n$ because n is the number of unique age-cohort combinations in the data. Thus, we compare the APC predictor $\eta_i = z_i'\zeta + x_i'\xi$ in (3.7) with the TS predictor

$$\eta_i = z_i'\zeta + t_i'\kappa. \tag{3.12}$$

We stack x_i and t_i in design matrices X and T, where X belongs to the linear span of T. We stack z_i in the design matrix Z.

Under classical normality assumptions, F-tests can be used to compare the models. Under asymptotic assumptions, likelihood ratio tests will be asymptotically χ^2 both for normal and logit models.

Computational issues arise with estimating the TS model when n is large. For example, in the obesity data n is 712 and storage issues arise when trying to estimate this model. We address these computational issues in the remainder of this section.

3.4.1 Estimation of the normal time-saturated model

The large size of the TS model creates computational challenges. The combined design matrix $M = (Z, T)$ has $d_z + n$ columns. In the obesity example, $d_z = 15$ and $n = 712$. Consequently, the standard tools in R run into storage limitations when trying to invert $M'M$ directly.

We address the computational problem by orthogonalizing the regressors and exploiting the fact that, since each row of T is a unit vector, $T'T$ is diagonal. Instead of regressing Y directly on M, we evaluate the partitioned regression

$$Y = \{Z - T(T'T)^{-1}T'Z\}\zeta + T\rho + \varepsilon. \tag{3.13}$$

Here $\{Z - T(T'T)^{-1}T'Z\} = \upsilon$ is the residual of a first-stage regression of Z on T. Since $T'T$ is diagonal it can be inverted by inverting the diagonal elements, avoiding general matrix inversion routines. It is therefore easy to compute υ. Since υ and T are orthogonal by construction, ζ and ρ are estimated by regressing Y on υ and T, respectively. This poses no computational challenge since υ has d_z columns, and d_z is small. If Z includes non-linear APC interactions, these can be expressed as functions of T. Then υ has reduced rank, so degrees of freedom calculations will be affected.

The model with orthogonalized regressors does not provide an estimate of κ. We can retrieve κ from $\hat{\kappa} = \hat{\rho} - (T'T)^{-1}T'Z\hat{\zeta}$. Note that (3.13) and the equation in terms of M, ζ, and κ give equivalent models with the same fit and residual variance. As a consequence we are normally not interested in the value of $\hat{\kappa}$.

3.4.2 Estimation of the logit time-saturated model

In the logit model as in the normal model, the many regressors in the TS model can cause computational problems in inverting the information matrix. The solution is similar to that used for the normal model.

The TS logit model has a linear predictor $\eta_i = z_i\zeta + t_i\kappa$ as in (3.12). The corresponding success probability is $\pi_i = \exp(\eta_i)/\{1+\exp(\eta_i)\}$. Thus, the score is

$$\dot{\ell} = (Y - \Pi)' \begin{pmatrix} Z & T \end{pmatrix}, \tag{3.14}$$

where Π is a H-length vector of probabilities π_i. The information matrix J is

$$J = -\ddot{\ell} = \begin{pmatrix} Z' \\ T' \end{pmatrix} W \begin{pmatrix} Z & T \end{pmatrix} = \begin{pmatrix} J_{ZZ} & J_{ZT} \\ J_{TZ} & J_{TT} \end{pmatrix}, \tag{3.15}$$

where W is a diagonal matrix of Bernoulli variances, $\pi_i(1-\pi_i)$. Using partitioned inversion we then find the inverse information as

$$J^{-1} = \begin{pmatrix} J_{ZZ\cdot T}^{-1} & -J_{ZZ\cdot T}^{-1} J_{ZT} J_{TT}^{-1} \\ -J_{TT}^{-1} J_{TZ} J_{ZZ\cdot T}^{-1} & J_{TT}^{-1} + J_{TT}^{-1} J_{TZ} J_{ZZ\cdot T}^{-1} J_{ZT} J_{TT}^{-1} \end{pmatrix}. \tag{3.16}$$

Here, the matrix $J_{TT} = T'WT$ is large, but diagonal, since the rows of T are unit vectors and W is diagonal. Thus, as before, the inverse of J_{TT} can be found simply by inverting the diagonal elements. Further, the matrices J_{ZZ} and $J_{ZZ\cdot T} = J_{ZZ} - J_{ZT}J_{TT}^{-1}J_{TZ}$ have low dimension and can be inverted by standard matrix inversion algorithms. The logit model can thus be estimated by Newton iteration using the above calculation of the inverse information.

3.5 Empirical application to obesity in England

We use the APC acceleration framework for repeated cross section data, developed in this chapter, to examine the dynamics of obesity in England. We examine a continuous outcome, log BMI, and a binary outcome, an indicator for obesity. Women and men are analysed separately. For women, we find that both log BMI and obesity are well-described by an age-drift (Ad) model. This model consists of a combined linear plane and accelerations in age. Detrended sums of the age accelerations reveal that the non-linear part of the relationship between age and

both log BMI and obesity is concave for women. For men, we find that log BMI is well-described by an age-cohort (AC) model, with a linear plane and accelerations in both age and cohort. Obesity is well-described by a cohort-drift (Cd) model, which omits age accelerations. Detrended sums of the cohort accelerations reveal that the non-linear part of the relationship exhibits concavity among later cohorts. Previous studies had detected non-linearities in age but had not detected the non-linearity in cohort among men.

3.5.1 Review of the APC literature on obesity

A substantial body of literature has tried to separate the effects of age, period, and cohort on obesity. This literature typically relies on descriptive statistics, or uses constraint methods to get estimates of age, period, and cohort effects.

The existing literature on adult obesity in England has largely failed to consider cohort effects. Howel (2011) uses data from the Health Survey for England to estimate models for obesity and overweight (BMI $\geq$ 25). She addresses the APC identification problem by omitting cohort entirely, constraining the effects of period to be equal within four-year bands, and constraining age to be polynomial. Wang et al. (2011) also use HSE data, and use a statistical model which omits cohort effects. Zaninotto et al. (2009) constructed within-period descriptive statistics by age, but did not consider cohort effects.

The literature on obesity outside of England has been more attentive to the possibility of cohort effects, but still relies on descriptive studies or constraint methods to obtain APC estimates. Lean et al. (2013) conducted a descriptive study of age patterns of obesity in Scotland, considering within-cohort patterns as well as within-period patterns. Allman-Farinelli et al. (2008) use the classical APC model in their analysis of Australian data, and impose four just-identifying constraints in the manner of Mason et al. (1973). The just-identifying constraints approach of Carstensen (2007), where the linear effect in one of the three of age, period, or cohort is constrained to equal zero, has been applied to both Australian obesity data (Peeters et al., 2015) and French obesity data (Diouf et al., 2010). A third commonly-used constraints method is the hierarchical APC (HAPC) model, developed by Yang & Land (2006) and Yang (2008). In the HAPC model, age is

constrained to be quadratic while period and cohort are assumed to be zero mean random effects. Reither et al. (2009) apply this model to US data and estimate period and cohort effects which deviate systematically from zero, indicating model misspecification. Separately, An & Xiang (2016) use a fixed-effects version of the HAPC model to assess US obesity, also constraining the effect of cohort to be equal within five-year bands.

There is consensus in the literature regarding the relationship between age and obesity, but not regarding the relationship between cohort and obesity. Most studies of obesity, in England and internationally, have detected a concave relationship between age and obesity (Howel, 2011; Zaninotto et al., 2009; Lean et al., 2013; Allman-Farinelli et al., 2008; Peeters et al., 2015; An & Xiang, 2016). However, the findings regarding the relationship between cohort and obesity vary across studies. It is likely that this is at least in part due to the fact that the various constraint methods restrict the cohort effects in different ways. In England, the existing studies of obesity do not attempt to isolate the effects of cohort.

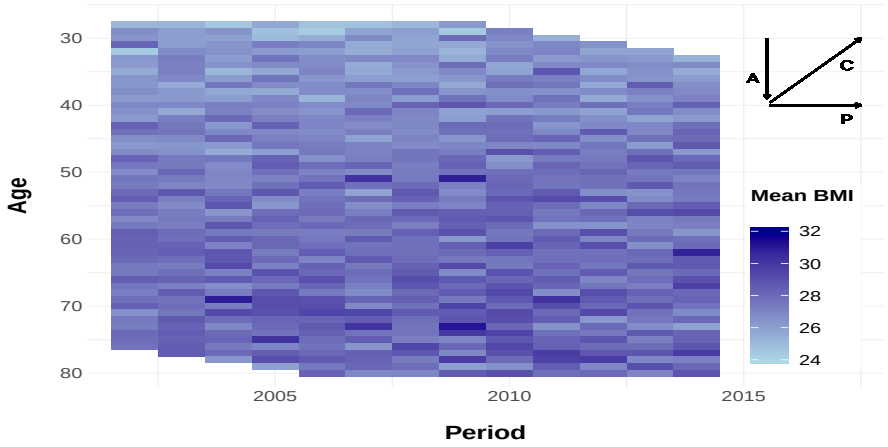
In this application, I find further support for a concave relationship between age and obesity, and also find evidence of a concave relationship between cohort and obesity among English men. The detection of a relationship between cohort and obesity in England, which is not an artefact of constraints imposed on the classical APC model, is a new stylized fact in the obesity literature.

3.5.2 Preliminary data analysis

We begin by visually inspecting the data, which was described in §3.2.1. The heatmap in Figure 3.2 shows the mean values of BMI in each age-cohort cell for women. The heatmap for men is similar, and therefore is not shown. We see a pattern of darker shading concentrated towards the right and centre of the graph and lighter shading concentrated towards the top, left, and perhaps bottom-left. This suggests that there are relationships between BMI and one or more of age, period, and cohort. The APC acceleration framework that we use cannot identify the linear components of such relationships, but it can identify non-linear components through the estimated accelerations. For example, the area of the array around ages 30-45 and periods 2005-2010 seems lighter in colour than the

central area below it (ages 45-60, periods 2005-2010). However, the central area is similar in colour to the area below it (ages 60-80, periods 2005-2010). This suggests curvature, i.e. non-linearity, in the relationship between age and BMI.

Figure 3.2: Within-cell BMI means women



In our analysis we use log BMI as the continuous outcome, rather than BMI. This is because the distribution of BMI has a right skew. The log-transformed data is closer to having the normal distribution required by the generalized linear modelling framework. This is discussed further in Appendix 3.B.1.

3.5.3 Covariates

The covariates are ethnicity, level of education, NSSEC, smoking history, and alcohol consumption as described below. Tables 3.3 and 3.4 in Appendix 3.B report descriptive statistics.

The reference ethnicity was taken to be white, with indicators for self-identification as black, Asian, of mixed ethnicity, or of "other" ethnicity (including e.g. Arab).

For education, the reference group are those who left school after attaining a GCSE, the minimum school-leaving qualification obtained around age 16, or equivalent qualification. We include three indicators: education below GCSE level, holders of a university degree, and education beyond GCSE but below degree level.

The three-class version of the National Statistics Socio-economic Classification (NSSEC) is used. The reference category is "Routine and Manual" occupations.

Indicators are included for "Intermediate", "Managerial and Professional", and "Other" occupation groups. The "Other" group includes students, those permanently outside the labour force, the long-term unemployed, and anyone whose employment could not be satisfactorily classified.

Smoking behaviour is classified into three groups. The reference category is individuals who have never smoked. One indicator records whether an individual currently smokes, while another captures former regular smokers. For alcohol consumption, the casual drinking population (those drinking one to four times a week) was taken to be the reference and indicators were introduced classifying individuals as not drinking at all, drinking rarely (less than once a week), and drinking frequently (five or more times a week). Note that the alcohol categories do not account for the quantity of alcohol consumed per drinking event.

3.5.4 Model for a continuous outcome variable: log BMI

We apply the APC acceleration model for continuous outcome variables, developed in §3.3.2, to analyse log BMI. We use the covariates described in §3.5.3. The analysis is performed separately for men and women. All analysis is performed using the `apc.indiv` software developed in §5.

3.5.4.1 Women

We begin by analysing the women. The first step in the analysis is model selection, performed using the Akaike Information Criterion (AIC) and F-tests. The selected model is the age-drift (Ad) model, which we estimate. We plot detrended sums of the estimated accelerations to explore the shape of the non-linear part of the relationship between age and log BMI among women. We also consider the estimated linear plane and covariate coefficients.

Model selection is performed using an analysis of variance of the TS model, the APC model and the sub-models described in §2.3.4.3. This analysis of variance is shown in the left-hand portion of Table 3.1. Recall that the TS model has a indicator for each age-cohort combination, as described in §3.4. Considering the sub-models, the AP, AC, and PC models each drop one set of accelerations; the Ad, Pd, and Cd models each drop two sets of accelerations; and the A, P, C models

Table 3.1: Model selection tables: log BMI

		Women				Men			
	df	F	p	AIC	ℓ	F	p	AIC	ℓ
TS				-22747.33	12101.67			-36449.64	18952.82
APC				-23321.91	11796.95			-37043.86	18657.93
AP	54	0.73	0.93	-23390.28	11777.14	1.85	0.00	-37051.87	18607.93
AC	12	1.38	0.17	-23329.33	11788.67	1.22	0.26	-37053.16	18650.58
PC	51	2.16	0.00	-23313.70	11741.85	1.57	0.01	-37065.67	18617.83
Ad	66	0.85	0.80	*-23397.46*	11768.73	1.73	0.00	-37061.31	18600.65
Pd	105	2.35	0.00	-23284.82	11673.41	4.80	0.00	-36750.82	18406.41
Cd	63	2.00	0.00	-23321.56	11733.78	1.50	0.01	*-37075.39*	18610.70
A	67	1.29	0.06	-23369.27	11753.64	2.63	0.00	-37001.23	18569.62
P	106	4.25	0.00	-23084.91	11572.46	5.05	0.00	-36722.71	18391.35
C	64	3.74	0.00	-23210.61	11677.31	3.33	0.00	-36958.31	18551.16

Degrees of freedom (df), F-statistics and p-values are for tests against APC model. The sub-models are defined in §2.3.4.3. Italics indicate the minimum AIC values.

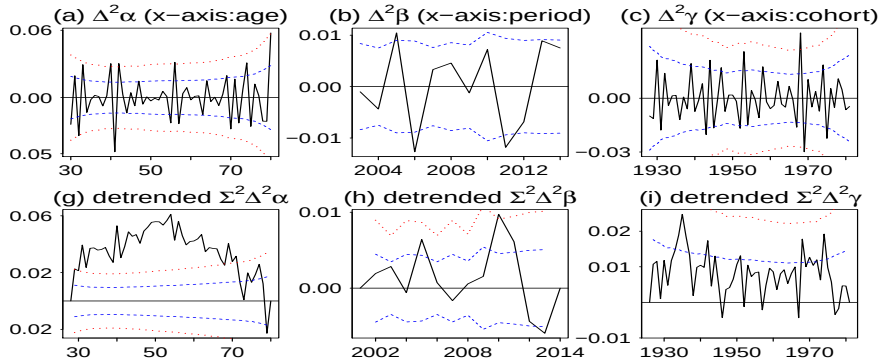
drop one slope of the linear plane in addition to two sets of accelerations. Further details of these submodels can be found in §2.3.4.3. Each of these models appears as a row in Table 3.1. Columns 2-4 of Table 3.1 report F-tests of the sub-models against the APC model. Column 6 is the log-likelihood. Column 5 contains the Akaike Information Criterion (AIC), which is a likelihood-based statistic that also incorporates a penalty for the number of parameters in the model.

We start by checking the APC specification against the TS model. From Table 3.1, we see that the APC acceleration model is preferred by the AIC. An F-test between the two models was also performed. The $F(592, 42942)$ statistic 1.02 has a p-value of 0.36, indicating that the restriction of the TS model to the APC acceleration model is not rejected.

The APC acceleration model can be reduced further to an Ad model, which we recall from §2.3.4.3 is the age-drift model with a linear plane and age accelerations. The reduction proceeds by the following logic. Minimizing the AIC in Table 3.1 points to the Ad model. The Ad model is rejected by neither an F-test against the TS model (not shown) nor an F-test against the APC model (columns 2-4).

A more careful analysis follows the model reduction paths TS–APC–AP–Ad and TS–APC–AC–Ad, using F-tests to compare each model to the model directly

Figure 3.3: Detrended summed accelerations, APC model of log BMI, women



Blue dashed, red dotted lines = 1, 2 standard deviations from zero.

preceding it. Neither path leads to rejection. The p-values for F-tests of Ad against AP and AC are 0.93 and 0.16, respectively. There is no support for a further reduction to the A model. The p-value for testing A against Ad is negligible.

Figure 3.3 displays the estimated accelerations for the APC model. The estimates for the preferred Ad model are very similar. Recall that the APC acceleration model contains a linear plane, described by the level v_o and two slopes v_a, v_c, and three sets of accelerations, $\Delta^2\alpha_a$, $\Delta^2\beta_p$, and $\Delta^2\gamma_c$. The accelerations are plotted in the top three sub-plots of Figure 3.3, marked (a), (b), (c).

The accelerations can also be interpreted as differences-in-differences. Recall that $\Delta^2\alpha_a = \mu_{ac} - \mu_{a-1,c} - \mu_{a-1,c+1} + \mu_{a-2,c+1}$ for any value of the cohort c. Thus, $\Delta^2\alpha_a$ captures the difference between the change of log BMI from age $a-1$ to age a, averaged over cohorts c, and the change of log BMI from age $a-2$ to age $a-1$, averaged over cohorts $c+1$. In sub-plot (a) of Figure 3.3, there is a large negative acceleration at age 41. This indicates that on average across cohorts and periods, the change in log BMI from age 40 to 41 is more negative than that from age 39 to 40. In this case the large negative acceleration may be due to sampling variation, since it is preceded and followed by large positive accelerations. However, had it been surrounded by near-zero accelerations, then the differences-in-differences interpretation would make sense and the large negative acceleration at age 41 could indicate a structural shift. The accelerations are not sensitive to the index array $\mathcal{I}$ defined in (3.1). If, for instance, those over 70 were truncated from the sample, the remaining accelerations would be unchanged apart from sampling error.

In each of the sub-plots (a)-(c), the reported pointwise confidence bands are not that informative about joint significance. However there is a slight tendency to more variation in (a) than in (b) or (c), in line with the above reduction to the Ad model.

To evaluate the overall non-linearity in the shape of the relationship between log BMI and age, period, and cohort, detrended sums of the accelerations are constructed as described in §2.3.4.2. the bottom sub-plots of Figure 3.3, marked (g), (h), and (i), show detrended sums of accelerations, anchored to start and end in 0. This cumulation gives us $\sum_{r=3}^{a}\sum_{s=3}^{r}\Delta^2\alpha_s$ which is the same as the original age effect α_a in (3.2) *apart* from an unidentifiable linear trend. By detrending the sums so that they start and end in zero, we isolate the deviation of the relationship between age and log BMI from a linear trend, i.e. how it accelerates or decelerates. This is the non-linear part of the relationship. The detrended sum of age accelerations in panel (g) has a concave appearance in line with many epidemiological studies (Nielsen, 2015). The detrended sums of period and cohort accelerations in panels (h) and (i) show no particular signal, in line with their formal insignificance as found by the reduction to the Ad model.

The detrended sums of accelerations have the same degrees of freedom as the accelerations. Recall that there are $A-2+C-2+P-2$ accelerations in ξ. There are $A+C+P$ sums of accelerations, of which six are set to zero by the detrending procedure (two each in age, period, and cohort). Therefore the detrended sums of accelerations have the same degrees of freedom as the accelerations. Due to the summation, the pointwise confidence bands in (g)-(i) give a better indication of the significance of the age non-linearity in (g) and the insignificance of the period and cohort non-linearities in (h),(i), in line with the Ad model.

Figure 3.3 could be repeated for the Ad model. That model excludes the accelerations in period and cohort, so the sub-plots (b), (c), (h), and (i) fall away. The corresponding figure for the Ad model has nearly the exact same sub-plots (a) and (g). This is unsurprising given that the Ad model is not rejected against the APC model.

The linear plane from the model with detrended sums of accelerations can also be studied. When detrending the sums of accelerations, the removed trends are added to the linear plane that is defined by v_o, v_a, v_c, see §2.3.4.2 and Nielsen (2015)

and for details. In the Ad model for log BMI among women, the resulting linear plane which combines both $\upsilon_o, \upsilon_a, \upsilon_c$ and the trends removed from the cumulated double-differences is

$$\underset{(0.02)}{3.16} + \underset{(0.0004)}{0.0025}(age - 28) - \underset{(0.0002)}{0.0013}(cohort - 1921).$$

We see that the slope along the cohort axis is significant. This matches the earlier finding that the Ad model cannot be reduced to an A model. This slope coefficient is interpreted as the change in the linear plane when increasing the cohort by one while keeping the age fixed. This of course means that period is also increased by one. Therefore, the interpretation of the cohort coefficient is that it is the sum of the cohort and the period slopes. Similarly, the age coefficient is interpreted as the sum of the age and period slopes.

The coefficients on the covariates of the Ad model are seen in column 1 of Table 3.2. Interpretation of these is deferred to §3.5.4.3, where they are discussed in conjunction with the model for men.

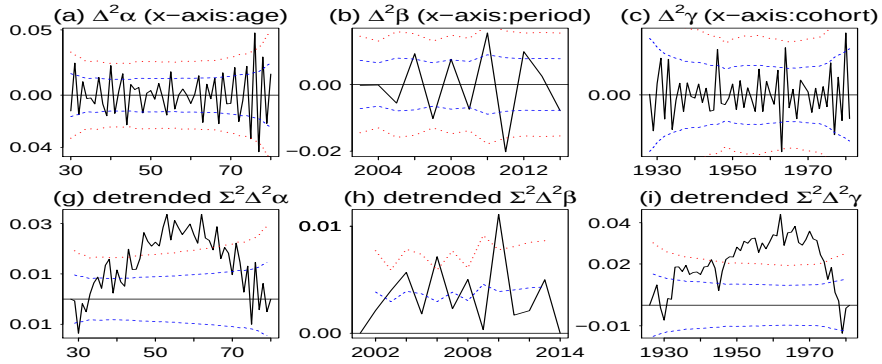
In Appendix 3.B, we report further mis-specification tests along with some comments on multiple testing issues and choice of significance level. Some further robustness checks are reported. We conclude that the overall outcomes are stable across changes to the specification.

3.5.4.2 Men

For the men, a similar approach is followed. The right-hand portion of Table 3.1 provides an analysis of variance. As before, both the AIC favours the APC model over the TS model. An F-test was also conducted to compare the two models. The $F(592, 37589)$ statistic is 0.98 with a p-value of 0.59, indicating support for the APC model against the TS model.

The APC model can be reduced further to an AC model. Minimizing the AIC in Table 3.1 points to the Cd model. However, the F-test for the Cd model against the APC model rejects. Likewise, the p-value for testing the Cd against AC model is 0.006. Since the AIC value for the AC model is also relatively low, and there is no formal way to compare the significance of the difference between the two AIC values, we select the AC model to describe this data.

Figure 3.4: Detrended summed accelerations, APC model of log BMI, men



Blue dashed, red dotted lines = 1, 2 standard deviations from zero.

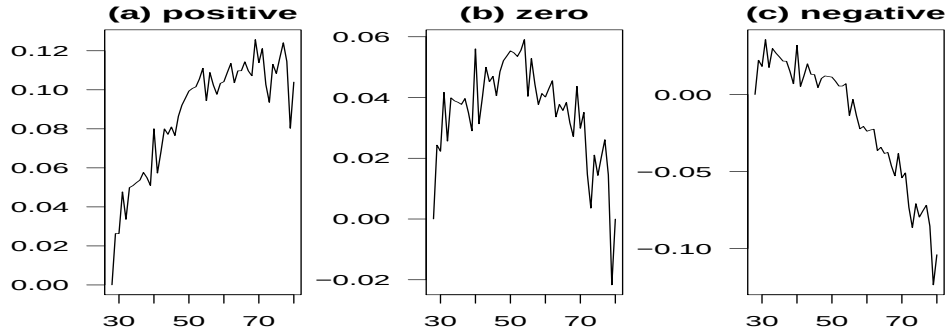
Figure 3.4 shows the estimates for the APC model. The estimates for the AC model are similar apart from omission of sub-plots (b) and (h). Looking at the plots of detrended sums of accelerations in sub-plots (g) through (i), there is some curvature in each of age and cohort, while the period non-linearity is driven by an anomalous spike in 2010. The fact that the only non-linearity was due to this spike, for which no clear explanation could be found, lent support to our decision to exclude the PC model and focus on the AC model.

The estimated coefficients on the covariates are seen in column 2 of Table 3.2. Mis-specification tests on the residuals are similar to those for women reported in Appendix 3.B and therefore are not shown.

3.5.4.3 Interpretation

We select the Ad model for women and the AC model for men. In both cases, the detrended sums of accelerations reveal significant non-linear relationships between log BMI and some subset of age, period, and cohort. For women, there is concavity in log BMI with age. The concavity may be consistent with general metabolic effects or selection effects towards the end of life, as those with higher BMI may die sooner (Hruby et al., 2016). Children may also be a factor, both due to the biological effect of child-bearing on the body and the impact of child-rearing on free time for personal healthcare. For men, there is non-linearity in log BMI with both age and cohort. The age non-linearity is not as significant as that for women, and

Figure 3.5: Age accelerations over various slopes, Ad model of women's log BMI



it begins later, suggesting that child-bearing may be an important factor among women. The cohort non-linearity among men is more difficult to explain, but may be related to generational shifts in the nature of employment. We hypothesize that men from the central cohorts may have similar dietary habits to men of earlier cohorts, but have a more sedentary lifestyle and do less physical labour; whereas more recent cohorts eat a more varied diet with less heavy, traditional British fare. Such factors could affect men more than women due to the long-standing social pressure on women to moderate their diets to "keep their figure". Further targeted research would be required to validate any of these hypotheses.

It should be recognised that overall effect of age or cohort on log BMI combines the identified concave non-linear effect with an unidentified linear effect. For example, the concavity in age among women may be combined with a positive, zero, or negative age slope, as seen in Figure 3.5. This slope is unidentifiable. Recognising the range of possible values for this slopes helps to think of explanations for the identified concavity; for example, an explanation of the age concavity based on a slowing metabolism is more consistent with Figure 3.5(a), while an explanation based on selection effects is more consistent with Figure 3.5(c).

The left-hand side of Table 3.2 shows the estimated effects of the covariates for both men and women. These are largely consistent with the literature. Black individuals have higher BMI than white individuals, on average, while those of other ethnicities have lower BMI (Ogden et al., 2015; An & Xiang, 2016). BMI and social class are negatively correlated (McPherson et al., 2007). Those with more education have lower BMI on average (Baum II & Ruhm, 2009; An & Xiang,

2016). Those who currently smoke have lower BMI on average, while those with a history of smoking have higher BMI on average, than those who have never smoked (Akbartabartoori et al., 2005). Non-drinkers and rare drinkers have higher BMI than casual drinkers (the reference group), who in turn have higher BMI than frequent drinkers. This may be explained by the mismatch between frequency and quantity of consumption (O'Donovan et al., 2018).

There are some sex differences in the covariate effects, primarily that some covariates are significant for women but not for men. Black and mixed ethnicity women have significantly higher and lower BMI, respectively, than white women; whereas for men these ethnicities are not significant. (Sproston & Mindell, 2006; Agyemang et al., 2015). Non-drinking men do not differ significantly from occasional drinkers. Social class is not significant for men and effects are larger in women than in men (Devaux & Sassi, 2011).

3.5.5 Model for a binary outcome variable: obesity

We apply the APC acceleration model for binary outcomes, developed in §3.3.3, to analyse an obesity indicator taking the value 1 for $BMI \geq 30$ and 0 otherwise. We use the same covariates as before and analyse women and men separately. All analysis is performed using the `apc.indiv` software developed in §5. We summarize the findings here, leaving details to Appendix 3.B.2.

For women we choose the Ad model. The detrended sums of age accelerations can be seen in Appendix 3.B.2 as Figure 3.7(a). They are similar to those estimated from the model for log BMI among women, seen in Figure 3.3(g). However, the concave shape is more right-skewed in the obesity model. This could be explained if the pattern of weight gain among women differs between early adulthood and middle age. If women who gain weight in middle age are less likely to pass the threshold into obesity than women who gain weight in early adulthood, this would produce the observed difference between the log BMI and obesity models.

For men we choose the Cd model, noting that for log BMI we favoured the AC model with weak support for the Cd model. The detrended sums of cohort accelerations are seen in Appendix 3.B.2 as Figure 3.7(b). They have a less skewed appearance than the detrended sums of cohort accelerations from the log BMI

Table 3.2: Estimated covariate effects: log BMI and obesity

	normal models for log BMI				logit models for obesity			
	Women, Ad		Men, AC		Women, Ad		Men, Cd	
	$\hat{\zeta}$	*se*	$\hat{\zeta}$	*se*	$\hat{\zeta}$	*se*	$\hat{\zeta}$	*se*
Ethnicity								
Black	0.068*	0.007	-0.008	0.006	0.658*	0.077	-0.034	0.095
Asian	-0.045*	0.007	-0.039*	0.005	-0.593*	0.105	-0.588*	0.089
Mixed	-0.023‡	0.011	-0.008	0.011	-0.190	0.149	-0.170	0.172
Other	-0.069*	0.012	-0.033*	0.011	-0.628*	0.178	-0.343†	0.188
Smoker								
Former	0.024*	0.002	0.026*	0.002	0.201*	0.027	0.310*	0.027
Current	-0.030*	0.002	-0.045*	0.002	-0.216*	0.030	-0.356*	0.033
Drinking frequency								
Never	0.043*	0.008	0.001	0.010	0.507*	0.097	0.347‡	0.143
Rarely	0.037*	0.002	0.014*	0.002	0.428*	0.025	0.201*	0.029
Frequently	-0.032*	0.003	-0.017*	0.002	-0.332*	0.035	-0.158*	0.029
Education								
Below GCSE	0.016*	0.003	0.010*	0.002	0.194*	0.031	0.160*	0.035
Some higher	-0.012*	0.003	0.002	0.002	-0.109*	0.033	-0.008	0.034
University	-0.048*	0.003	-0.026*	0.003	-0.452*	0.040	-0.280*	0.040
3 level NSSEC								
Intermediate	-0.021*	0.002	-0.001	0.002	-0.239*	0.029	-0.036	0.033
Managerial	-0.009*	0.003	-0.001	0.002	-0.084*	0.032	-0.117*	0.031
Other	-0.013†	0.007	-0.020†	0.011	-0.078	0.086	0.015	0.162

All variables are indicators with the following reference catagories: white ethnicity; never smoker; occasionally drink alcohol; GCSE education level; routine work for NSSEC. p-values: * $p \leq 0.01$, ‡ $0.01 < p \leq 0.05$, † $0.05 < p \leq 0.10$.

model for men, with more acceleration among earlier cohorts. This may be due to picking up some of the effects attributed to age accelerations in the log BMI model. This could be explained as follows: the group of men aged 40-60 in 2001-2014 have higher mean BMI than those of other ages. These men belong to cohorts 1940-1980. Because of the limited period range of this dataset, we do not observe middle-aged men from other cohorts, and we do not observe these cohorts at anything other than middle age. It is therefore impossible to separate the cohort and age influences for this group with this data.

Table 3.2 reports the estimated covariate coefficients for both men and women.

For women, significance (at a 5% level) and signs are exactly the same as for log BMI. For men, the logit coefficients are broadly speaking in line with those reported for log BMI.

3.6 Conclusion

In this chapter, we developed a framework for estimating accelerations in age, period, and cohort from repeated cross section data. We embedded the APC acceleration parametrization of Kuang et al. (2008) in generalized linear models for repeated cross section data. The generalized linear modelling approach allowed us to consider both continuous and binary outcomes. The APC acceleration parametrization facilitated estimation, inference, and identification of the non-linear APC effects, while avoiding issues relating to the unidentifiable linear trends. By embedding the APC acceleration parametrization in a generalized linear model, we were able to incorporate covariates as additional explanatory variables. To assess the adequacy of our APC acceleration model, we presented tests against a time-saturated model.

Our analysis of obesity data for England using this APC acceleration framework demonstrated clear non-linear age and cohort effects as well as covariate effects that were robust to a range of specifications. We considered both a continuous outcome, log BMI, and a binary outcome, the obesity indicator. For women, the only significant deviation from linearity identified from the APC acceleration model is concavity in age. We suggest metabolic changes, child-bearing, and child-rearing as potential explanations for this. These different explanations would likely imply different linear trends in age; for example, biological effects due to a slowing metabolism and child-bearing are persistent and so would be consistent with a positive linear age trend, whereas time constraints due to child-rearing are temporary and would be consistent with zero linear age trend. The linear trend in age cannot be identified, but alternative strategies (such as comparing women with and without children) may be used to evaluate the competing explanations of the identified non-linear effect. For men, there is significant concavity in cohort for both obesity and log BMI, and in age, for log BMI only. We suggest that the cohort non-linearities may be linked to generational shifts in lifestyle factors, such

as diet. For both sexes, the impact of covariates is largely consistent with existing literature, although more covariates are significant in the models for women.

The APC acceleration framework developed here could be expanded to allow for mixture models, interaction terms between APC effects and covariates, and heteroskedastic errors. This would enable us to address some of the mis-specification concerns in the log BMI analysis. It would be of interest to analyse the consequence of missing age-cohort cells within the dataset. Indeed, in the logit application we dropped all observations with ages below 28 to avoid perfect separation in a single age-cohort cell; it would be preferable to avoid this.

3.A Details on APC acceleration model

3.A.1 Properties of ad hoc identification schemes

Suppose θ is ad hoc identified by the constraint $\alpha_1 = \alpha_2 = \beta_P = \gamma_C = 0$ as in Mason et al. (1973). Another example is the analysis of Ejrnæs & Hochguertel (2013), where one age, one cohort, and two period indicators are constrained to zero. These are examples of constraints of the type $L'\theta = 0$, for L an appropriate selection matrix. We show that the APC predictor thus constrained has the form $\mu = XQ\phi$, for X the design matrix associated with the APC acceleration model, Q an invertible $q-4 \times q-4$ matrix, and ϕ a parameter vector. We describe how the constrained APC effects are found from ϕ.

Suppose, that $L'L$ is invertible. Then L has orthogonal complement $L_\perp$ of dimension $q \times q-4$, so that $L'_\perp L = 0$ and $(L, L_\perp)$ is invertible. For simplicity, suppose $L'L = I_4$ and $L'_\perp L_\perp = I_{q-4}$. Then, the orthogonal projection identity is $I_q = LL' + L_\perp L'_\perp$.

The identification problem is that μ as a function of θ is invariant to the transformations $\theta \mapsto \theta + A_\perp \mathfrak{v}$, where $\mathfrak{v} = (\mathfrak{a}, \mathfrak{b}, \mathfrak{c}, \mathfrak{d})'$ and $A_\perp$ is a $q \times 4$ matrix (Nielsen & Nielsen, 2014). Thus, μ only depends on θ through $\xi = A'\theta$ where A is a $q \times p$ matrix so that $A'A_\perp = 0$. Further, the collinearity of the d_{ac} vectors is given by $d_{ac} = x_{ac}A'$.

The original design matrix D, composed of stacked d_{ac}, has reduced column rank. Since $I_q = LL' + L_\perp L'_\perp$ we get $D\theta = DLL'\theta + DL_\perp L'_\perp\theta$. The constraint $L'\theta = 0$ implies $D\theta = DL_\perp L'_\perp\theta$. Here, $DL_\perp$ is a $q \times q-4$ matrix and it must have full column rank. The reason is that Corollary 2 in KNN implies that D has rank p. In other words, the constraint $L'\theta = 0$ is identifying when $DL_\perp$ has full column rank and the constrained θ is estimated by dropping the columns DL from D.

The identity $D = XA'$ shows $D\theta = XA'L_\perp L'_\perp\theta$. Let $Q = A'L_\perp$ and $\phi = L'_\perp\theta$, so that $D\theta = XQ\phi$. From above, $DL_\perp = XA'L_\perp = XQ$ has full column rank, so that Q is invertible.

To find the constrained parameter, satisfying $L'\theta = 0$ say, use $I_q = LL' + L_\perp L'_\perp$ to get $\theta = LL'\theta + L_\perp L'_\perp\theta$. Since $L'\theta = 0$ and $L'_\perp\theta = \phi$ then $\theta = L_\perp\phi$.

We note that in general ϕ is not invariant. Indeed, applying the transformation $\theta \mapsto \theta + A_\perp \mathfrak{v}$ to $\phi = L'_\perp \theta$ gives $L'_\perp \theta + L'_\perp A_\perp \mathfrak{v}$. In general, this depends on $\mathfrak{v}$, unless we choose $L_\perp = A$ so that $L'_\perp A_\perp \mathfrak{v} = 0$. With the choice $L_\perp = A$ we get $\phi = \xi$.

3.A.2 Covariates and identification of the APC model

Recall the model $\eta = Z\zeta + \mu$ for the linear predictor in equation (3.8). We show that when only the covariate effects ζ are of interest it does not matter whether the APC effects in μ are parametrized through the parametrization $\mu = X\xi$ as in (3.8), or by ad hoc identification.

Suppose that we identify the APC structure by a constraint such as $\alpha_1 = \alpha_2 = \beta_P = \gamma_C = 0$ or some other constraint on the form $L'\theta = 0$ as analyzed in Appendix 3.A.1. Following Appendix 3.A.1, we can just as well suppose the predictor satisfies $\mu = XQ\phi$, where Q is a known, invertible $p \times p$-matrix and $\phi = Q^{-1}\xi$. Thus, we have two parametrizations: $\eta = Z\zeta + X\xi$ and $\eta = Z\zeta + XQ\phi$. The mapping between the two parametrizations is one-one, since Q is invertible. Due to the equivariance of maximum likelihood estimators (Cox & Hinkley, 1974, §1) the maximum likelihood estimators of ζ are the same under the two parametrizations.

3.B Further data analysis

The data is available through the UK Data Service[2]. Tables 3.3 and 3.4 give descriptive statistics.

3.B.1 Robustness checks for normal models

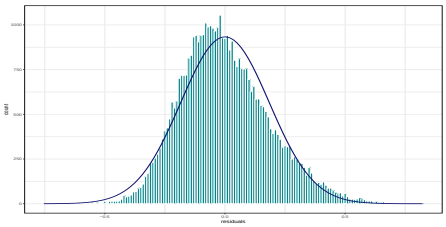
Formal mis-specification tests for the Ad model are reported in Table 3.5 in situations with and without log transformation of the dependent variable. The tests include a cumulant based test for normality of residuals and tests for functional form mis-specification and heteroskedasticity (Ramsey, 1969; White, 1980). The log transformation clearly improves the specification. Yet, given the large sample size, $N = 43,077$, it is difficult to avoid very small p-values. The histogram of the residuals in Figure 3.6 suggests a systematic, if modest skewness in the

[2]https://discover.ukdataservice.ac.uk/series/?sn=2000021

Table 3.3: Descriptive statistics: Continuous variables

	Women				Men			
	min	mean	median	max	min	mean	median	max
Age	28	51	50	80	28	52	51	80
Period	2001	2007	2006	2014	2001	2007	2006	2014
Cohort	1926	1956	1957	1981	1926	1955	1956	1981
BMI	13.2	27.4	26.4	58.9	13.6	27.9	27.4	59.5
Height (cm)	124	162	162	202	138	175	175	203
Weight (kg)	28.4	71.5	69	164	34.2	85.5	84	203

Figure 3.6: Residuals from Ad model of log BMI, women



solid line = normal distribution with mean and standard deviation from the data

residuals that could be addressed by adopting regression based on a non-normal distribution.

In the analysis in this chapter, we have conducted many tests, yet taken an informal approach to the choice of significance level. This is guided by the following ideas. The APC acceleration model is the central model, which we test using a handful of mis-specification tests: the test against the TS model, and those reported in Table 3.5. If each of these tests is conducted at a 1% level the overall level for mis-specification tests is about 5%. We then reduce the APC model by testing a number of nested sub-models. If each sub-model is compared both with the model immediately above and with the APC model at a 5% level the overall level for the reduction is likely to be in the neighbourhood of 5%. The reduction tests and the mis-specification tests are likely to be independent, which gives an overall size of about 10%. We refrain from a detailed analysis here, noting that it is more important to record marginal decisions with p-values in the range from 1% to 10% than to get the size calculation exactly right.

A range of alternative specifications of the APC acceleration model for log

Table 3.4: Descriptive statistics: Indicators

Variable	1	2	3	4	5
Women, 43077 observations					
Ethnicity	762	41071	702	288	254
Class	16306	11721	14302	748	
Education	13138	11847	9644	8448	
Alcohol	497	15703	19534	7343	
Smoker	23037	10724	9316		
Men, 38316 observations					
Ethnicity	600	36363	972	201	180
Class	14362	7390	16363	201	
Education	10927	7865	10456	9068	
Alcohol	228	8330	19489	10269	
Smoker	16692	13084	8540		

Classification. Ethnicity: 1 black, 2 white, 3 Asian, 4 mixed, 5 other.
Social class: 1 routine & manual, 2 intermediate, 3 managerial & professional, 4 other.
Education level: 1 below GCSE, 2 GCSE, 3 some higher, 4 university degree.
Alcohol drinking (events per week): 1 never, 2 rare (<1), 3 occasional (1-4), 4 frequent ($\geq$ 5).
Smoker: 1 never, 2 former, 3 current.

BMI were examined as robustness checks. We considered a model replacing the three-class NSSEC with the eight-class version. We considered using BMI instead of log BMI as the dependent variable. We also considered a model with log weight as the dependent variable and log height as an explanatory variable; a model with log BMI as the dependent variable implicitly imposes a coefficient of two in this regression. We also re-estimated the log BMI and obesity models using the interview weights provided by the HSE; these weights account for differences in participation by age, sex, region, household type, and social class. These models did not change our substantive findings.

We also considered different subsets of the original HSE data. To examine whether income yielded different results to the NSSEC, we tried a specification which replaced the NSSEC with inflation-adjusted household income (quadratic in logs) using two samples: first with all observations where income information

Table 3.5: Specification tests for Ad models, women

	BMI			Log BMI			
Test	value	statistic	p	value	statistic	p	distribution
Skewness	1.02	7488.99	0.00	0.45	1445.07	0.00	$\chi^2(1)$
Excess kurtosis	1.59	4524.01	0.00	0.24	102.92	0.00	$\chi^2(1)$
Normality test		12012.99	0.00		1547.99	0.00	$\chi^2(2)$
RESET test		23.07	0.00		18.93	0.00	F(2, 43006)
hetero test		5.20	0.00		4.94	0.00	F(120, 42956)

was available, then for only observations where both income and NSSEC information was available. There was no substantial change to the detrended sums of accelerations or the covariates. Given the apparent insensitivity of the estimated covariate coefficients to whether the TS model, APC model, or a sub-model was used, we decided that the APC and covariate effects were largely orthogonal to one another and tested a model which excluded the covariates. This gave us a much larger sample size due to less missing information. The substantive results were unchanged. Finally, to check whether the differences in sample size across years affected our results we randomly selected 2000 observations from each year and ran the original analysis on this smaller sample, using three different random seeds. The detrended sums of age accelerations were robust to this check for both men and women.

In our final set of robustness checks we tested extensions of the age-cohort space. We considered the original model but with the age range extended to be from 20-80, and the cohort range extended accordingly. This incorporated some cells in which perfect separation was present, but that is not a problem in the normal model. The main consequence of this was a strengthening of the significance of age non-linearities for men, with curvature in the early twenties that could be explained by particularly rapid growth in log BMI over those ages. The NSSEC was not recorded prior to 2001, but we have income information back to 1997, so we were able to consider the model with income over a longer period horizon. We were also able to evaluate a no-covariates model with data back to 1992. The estimated age effects remained similar to the original models throughout. With an extended period range, the period non-linearities become

significant and exhibit concavity, which could be explained by a reduction in the rate of growth in log BMI after the 1990s.

In addition to the robustness checks above, we have the mis-specification tests (normality, functional form, heteroskedasticity) on the estimated models. While our mis-specification tests show imperfections in our models, they do not invalidate our results. Fat tails mean that our standard errors may be incorrect, but the estimators will still be consistent. The functional form and heteroskedasticity results might be resolved with a more careful choice of covariates. We also intend to consider heteroskedasticity arising from the APC structure in future work. The lack of variation in the main substantive findings across all robustness checks is encouraging.

3.B.2 Details of binary analysis

The main findings of the binary analysis are summarized in §3.5.5. Here, we present details of the model selection. This follows the approach for continuous outcomes in §3.5.4.

Model comparison statistics are presented in Table 3.6. For the women, the APC model is preferred to the TS specification according to the AIC in column 5. An LR test between the two models yielded an LR statistic of 598.10, which has $p = 0.42$ when compared with a $\chi^2(592)$ distribution. This indicates that the reduction from the TS model to the APC model is not rejected. The APC model can be further reduced to the Ad model for the following reasons. First, the AIC favours the Ad model. Second, the Ad model is not rejected by a LR test against the APC model. Third, the Ad model is not rejected by LR tests against the AC and the AP models, which have p-values of 0.98 and 0.51, respectively. We note that although the A model is not rejected against the APC model, it is rejected when tested against the Ad model ($p \ll 0.00$). Figure 3.7(a) shows the detrended sums of age accelerations for women from the Ad model.

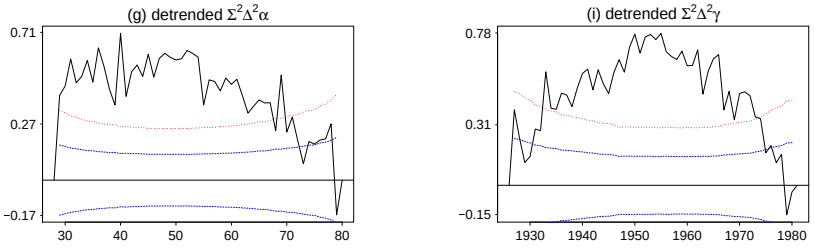
For the men, the APC model is preferred to the TS specification according to the AIC in column 9 of Table 3.6. An LR test between the two models yielded an LR statistic of 556.26, which has $p = 0.85$ when compared to a $\chi^2(592)$ distribution. This indicates that the reduction from the TS model to the APC model is not

Table 3.6: Model selection tables: obesity

		Women				Men			
	df	LR	p	AIC	ℓ	LR	p	AIC	ℓ
TS				48708.92	-23627.46			44563.50	-21554.75
APC				48123.02	-23926.51			43935.76	-21832.88
AP	54	34.49	0.98	48049.51	-23943.76	73.86	0.04	43901.62	-21869.81
AC	12	11.19	0.51	48110.21	-23932.10	20.81	0.05	43932.57	-21843.29
PC	51	73.75	0.02	48094.76	-23963.38	58.97	0.21	43892.74	-21862.37
Ad	66	45.72	0.97	*48036.74*	-23949.37	94.81	0.01	43898.57	-21880.28
Pd	105	154.69	0.00	48067.71	-24003.85	305.24	0.00	44031.00	-21985.50
Cd	63	85.31	0.03	48082.32	-23969.16	79.14	0.08	*43888.90*	-21872.45
A	67	71.48	0.33	48060.50	-23962.25	153.20	0.00	43954.96	-21909.48
P	106	211.26	0.00	48122.28	-24032.14	307.12	0.00	44030.88	-21986.44
C	64	147.62	0.00	48142.63	-24000.32	151.02	0.00	43958.78	-21908.39

Degrees of freedom (df), LR statistics and p-values are for tests against APC model.

Figure 3.7: Detrended summed accelerations, models of obesity indicator



(a) women, Ad model, age effect (b) men, Cd model, cohort effect

Blue dashed, red dotted lines = 1, 2 standard deviations from zero.

rejected. The APC model can be further reduced to the Cd model for the following reasons. First, the AIC favours the Cd model. Second, the Cd model is not rejected by a LR test against the APC model. Third, the Cd model is not rejected by LR tests against the AC and the PC models with p-values of 0.22 and 0.06, respectively, noting that the decision against the PC model is marginal. There is no support for reduction to the smaller C model. Figure 3.7(b) shows the detrended sum of cohort accelerations for men from the Cd model.

Chapter 4

The framework for panel data, with an application evaluating the treatment of control variables in a UK health study

4.1 Introduction

This paper proposes a general framework for identification of the separate effects of age, period, and cohort from panel data. The caveat is that the identified effects are accelerations, rather than the more customary slopes. This is because the well-known age-period-cohort problem makes identification of the separate slopes of age, period, and cohort impossible.

Many economists seek to understand the relationship between age, period, or cohort and some outcome variable. For example, life-cycle analysts study how wages, savings, and health evolve with age (e.g. Van Landeghem, 2012). Other researchers are interested in how these outcomes relate to the business cycle, which is composed of year-on-year (period-on-period) changes in macroeconomic variables. Still others are interested in whether these outcomes depend on generational experiences: did the cohort *in utero* during the 1918 influenza pandemic experience a lasting negative effect on socioeconomic outcomes (Almond, 2006)?

Studies of age, period, and cohort (APC) effects often rely on panel data. Panel data is individual-level data, with the distinguishing feature that each individual

is recorded in multiple periods. For example, the British Household Panel Survey (BHPS) interviewed the same households every year from 1991 to 2008. Other commonly-used panel datasets include the American National Longitudinal Survey of Youth (NLSY) and Panel Survey of Income Dynamics (PSID), and the German Socio-Economic Panel (GSOEP). Panel data confers advantages relative to repeated cross section data, where different individuals are surveyed each year. One is the ability to conduct comparisons across time free of noise from interpersonal variation. Another is the ability to control for time-invariant, unobserved, individual-specific factors influencing the outcome variable.

All studies of APC effects must contend with the well-known APC identification problem, which limits the separability of the three effects (Glenn, 2005). This can be explained intuitively using the example of the evolution of a person's self-reported health over their lifetime. Observing that individual, it is impossible to know whether they rated their health lower in a particular year because of life-cycle age effects or environmental period effects. Adding data on other people who experience that age at different time periods does not resolve the confusion, as it introduces the possibility of cohort effects: generation-specific preferences, beliefs, or behaviours that influence either the individual's health or their perception of it. Formally, this is a collinearity problem, due to the exact relation $period = age + cohort - 1$[1]. This is the APC identification problem.

In this paper, I propose a framework which enables separate identification of some parts of the APC effects from panel data: accelerations in each of age, period, and cohort. These accelerations are useful because they provide information about the shape of the response of the outcome variable of interest to age, period, and cohort. For example, predominantly negative accelerations of subjective well-being with age imply a concave relationship between well-being and age. This piece of evidence can be used to evaluate the claim that the relationship between age and well-being is U-shaped (Blanchflower & Oswald, 2008). It is not possible to separately identify the slopes, using my framework or any other method, unless

[1]Why -1? This relation holds if we allow cohort 1 to be aged 1 in period 1. This indexing, referred to as time stamp accounting, is preferred because it simplifies later calculations. However, human ages are counted differently: babies are "zero years old" for all of their first year, only turning one in the beginning of their second year, yielding a relation of $period = age + cohort$. This is referred to as calendar accounting.

strong untestable constraints are imposed (see Nielsen & Nielsen, 2014; Fosse & Winship, 2019). This means that it is not possible to determine whether subjective well-being increases or decreases with age. It is only possible to study the non-linear part of the shape of the relationship between an outcome of interest and age, period, or cohort, such as whether the relationship is concave or convex.

The fact that accelerations in age, period, and cohort are separately identified is not new; see for example Clayton & Schifflers (1987a). Accelerations have been exploited in economics and epidemiological studies with aggregate data (McKenzie, 2006; Martínez Miranda et al., 2015) and panel data (Van Landeghem, 2012). The limitation of the existing literature using accelerations in panel data is that the accelerations are not incorporated into a regression framework. In Van Landeghem (2012), the accelerations are constructed *post hoc* using estimates from a first-stage regression in differences. The first-stage regression is only identified by imposing an untestable constraint on the age, period, and cohort effects (the accelerations constructed in the second stage are robust to this constraint). There are three advantages to using a regression framework, as I do, instead of this approach. First, regression is a single-step procedure; Van Landeghem's two-step procedure is more cumbersome and increases the potential for errors. Second, the treatment of covariates in a regression is straightforward. In contrast, Van Landeghem includes the covariates in the first-stage regression and so their estimated effects may be sensitive to the untestable identifying constraint. Third, in a regression framework it is easy to perform model selection over both APC effects and covariates, by testing restrictions on both and then estimating reduced forms of the model. This is not straightforward in Van Landeghem's framework. A separate strand of literature has embedded the APC accelerations in a regression framework (see Kuang et al. (2008); Martínez Miranda et al. (2015) and §3 of this thesis), but has not to date considered panel data.

The first contribution of this paper is to explain how APC accelerations may be identified and estimated as parameters in panel data regressions. I use the APC acceleration parametrization of Kuang et al. (2008), originally developed for aggregate data, described in detail in §2.3 of this thesis. This parametrization includes accelerations in each of age, period, and cohort, plus a single linear plane that combines the inseparable slopes of all three. I embed the parametrization in

a panel data model which also includes covariates. I show that this model, which I refer to as the APC acceleration model, can be estimated in three common panel data settings: pooled OLS, random effects, and fixed effects. In pooled OLS and random effects settings, all three sets of accelerations are identified, while in the fixed effects setting, cohort accelerations are not identified.

The second contribution is to provide guidance on factors influencing the choice between the three panel settings in the context of the APC acceleration model. In particular, I show that the correlation between the APC variables and unobserved components is not a factor. Typically the correlation between explanatory variables, like the APC variables, and unobservable components does need to be taken into consideration, since the three panel settings make different assumptions about this correlation. I show that the deterministic nature of the APC variables renders redundant the differences between the three settings in terms of the assumptions they make about correlation between the APC variables and unobserved components, except when the data is unbalanced in a particular way that I discuss. Therefore this correlation need not be considered when choosing between the three panel settings. Instead, the choice should be based on other factors, which I discuss. Since the random effects setting imposes the strongest assumptions on the correlation between unobservables and explanatory variables, this finding strengthens the case for the use of the random effects setting.

The third contribution of this paper is an application which demonstrates how the APC acceleration model can be used as a diagnostic tool to evaluate standard approaches to accounting for age, period, and cohort in panel data regressions. In the application, which is health-related, age is an essential control variable. A standard approach is to assume a particular functional form, such as quadratic, for age. The APC acceleration model can be used to test such functional form restrictions. The fact that the APC accelerations are embedded in a panel data regression framework allows the appropriate functional form to be investigated while accounting for covariates.

The application considers the impact of commuting on hospital in-patient stays. By using the APC acceleration model as a diagnostic tool, I identify an amendment to the model. This improves the model fit and strengthens the previously weak

finding that commuting does not affect the probability of being a hospital in-patient. I closely follow the analysis of Künn-Nelen (2016), who examined the relationship between commute time and hospital in-patient stays. She assumed a quadratic functional form for the age control, as is common in the literature on health and commuting (Dickerson et al., 2014; Gimenez-Nadal et al., 2018). Using the APC acceleration model I test this quadratic functional form restriction and find that it is rejected. By closely examining the estimated shape of the age effect, I find that the model can be improved by the addition of a childbirth indicator. The addition of this indicator improves the fit of the model (as measured by the R^2), eliminates all remaining age non-linearities, and renders the coefficients on commute time and its square insignificant at all conventional significance levels.

The paper proceeds as follows: §4.2 provides an overview of the two elements combined in this paper: panel data, and the APC acceleration parametrization due to Kuang et al. (2008). The main theoretical contributions appear in §4.3, where I analyse the introduction of the APC acceleration parametrization to three panel data settings. The application in which the APC acceleration model is used as a diagnostic tool appears in §4.4, and §4.5 concludes.

4.2 Overview of panel data and APC acceleration model

The objective of this paper is to analyse a model of the following form:

$$y_{ip} = x'_{ip}\xi + z'_{ip}\zeta + \omega_i + \epsilon_{ip}. \tag{4.1}$$

In this model there are observations on individuals i at various time periods p. The observed scalar y_{ip} is the outcome of interest, the observed vector x_{ip} is the design vector associated with the APC acceleration parametrization, and the observed vector z_{ip} captures other explanatory variables of interest (covariates). The Greek letters $\{\xi, \zeta\}$ are parameter vectors, and the elements ω_i and ϵ_{ip} are unobserved scalars.

.

4.2.1 Types of panel data considered

The model in equation (4.1) is suitable for panel datasets, which contain information about a given set of individuals, observed over multiple time periods. For example, a dataset containing the health records of all children at a GP surgery over several years is a panel dataset. Formally, a panel dataset for which model (4.1) is relevant must have the following features:

- Collected for individuals $i = 1, \ldots, N$.
- Each individual is observed for time periods indexed by p.
- The recorded variables are
  - y_{ip}: a continuous outcome variable (scalar)
  - At least two of the following three variables
    - a_{ip}: the age of i at p, varies over i and p
    - p_{ip}: the period at p, varies over p, and
    - c_{ip}: the cohort of individual i, varies over i but not p.
  - A vector of other explanatory variables z_{ip} that can be split into
    - $z_{1,ip}$: varies over i and p, and
    - $z_{2,ip}$: varies over i but not p.

The period index is $p = L + 1, \ldots, L + P$, where L is a constant that depends on the structure of the dataset, as in §2.3.2.1. The total number of periods is P.

The size of the dataset depends on whether it is balanced or unbalanced. If the dataset is balanced, i.e. no individuals enter or leave the sample, the total number of observations will be NP. In practice many panel datasets exhibit entry and exit, either due to attrition or by design; this creates an unbalanced panel, where the total number of observations is fewer than NP.

Having at least two of age, period, and cohort is sufficient to calculate the third, via the relation $p_{ip} = a_{ip} + c_{ip} - 1$. This information is then used to construct the design vector x_{ip} that appears in equation (4.1). The design vector x_{ip} can be split into two parts: $x_{1,ip}$, which depends on both i and p, and $x_{2,ip}$, which depends on i alone. The construction and split of x_{ip} is discussed further in §4.2.2.

For subsequent discussion, additional notation is needed. Let Y_i and X_i refer to a vector and a matrix, respectively, which stack y_{ip} and x_{ip} over p for a given

i. Let $X_{1,i}, X_{2,i}, Z_i, Z_{1,i}, Z_{2,i}$, and ϵ_i be similarly defined. Further, let Y and X refer to a vector and a matrix, respectively, which stack Y_i and X_i over i. Let $X_1, X_2, Z, Z_1, Z_2, \omega$, and ϵ be similarly defined.

4.2.2 The APC acceleration model

The age, period, and cohort (APC) effects in model (4.1) are captured by the design vector x_{ip} and the parameter vector ξ. In this section I explain how this representation is constructed from information on a_{ip}, p_{ip}, and c_{ip}. This representation is the APC acceleration parametrization, which permits direct estimation of APC accelerations from the regression model.

The representation of APC effects in terms of of x_{ip} and ξ is derived from the classical APC model, which includes an indicator for each unique age, period, and cohort observed in the data. This classical APC model nests many of the common specifications used for age, period, and cohort variables in panel regressions, such as the quadratic age effect with year fixed effects that is common in the health literature described in §4.4.1. In what follows I explain how the classical APC model is constructed from panel data of the form described in §4.2.1 and then how to progress from the classical APC model to the representation in terms of $x_{ip}'\xi$.

For panel data the classical APC model takes the form

$$\begin{aligned} y_{ip} =& \mu_{ip} + z_{ip}'\zeta + \omega_i + \epsilon_{ip} \\ \mu_{ip} =& d_{ip}'\theta \\ =& \delta + \alpha_1 \mathbb{1}(a_{ip} = 1) + \cdots + \alpha_A \mathbb{1}(a_{ip} = A) \\ & + \beta_{L+1} \mathbb{1}(p_{ip} = L+1) + \cdots + \beta_{L+P} \mathbb{1}(p_{ip} = L+P) \\ & + \gamma_1 \mathbb{1}(c_{ip} = 1) + \cdots + \gamma_C \mathbb{1}(c_{ip} = C) \end{aligned} \tag{4.2}$$

Here μ_{ip} is the APC part of the model. The indicators for unique ages, periods, and cohorts are collected in the design vector d_{ip}. Each indicator has the form $\mathbb{1}(a_{ip} = 1)$ and takes the value 1 if the condition in brackets is satisfied for individual i in period p, and 0 otherwise. The associated parameters are $\{\alpha_1, \ldots, \alpha_A\}$ for age effects, $\{\beta_{L+1}, \ldots, \beta_{L+P}\}$ for period effects, and $\{\gamma_1, \ldots, \gamma_C\}$ for cohort effects. The age index is $a = 1, \ldots, A$, the period index is $p = L+1, \ldots, L+P$, and the cohort index is $c = 1, \ldots, C$. It is easy to see how the indicators can be constructed

from the variables a_{ip}, p_{ip}, c_{ip} which are present in the data as per §4.2.1. The parameters are combined in

$$\theta = \{\delta, \alpha_1, \ldots, \alpha_A, \beta_{L+1}, \ldots, \beta_{L+P}, \gamma_1, \ldots, \gamma_C\}. \tag{4.3}$$

As explained in §2.3.2.2, the classical APC model is overparametrized by four parameters, and so cannot be estimated. The exact relationship $p = a + c - 1$ means that there are many parameter vectors θ which are consistent with a set of d_{ip} and μ_{ip}. Therefore it is not possible to estimate a unique θ from model (4.2).

It is common in panel studies of age, period, and cohort to make model (4.2) estimable by imposing constraints, such as functional form constraints. Recurring examples include: omitting one of the three sets of indicators; restricting the age effect to be quadratic and constraining groups of cohort effects to be equal (often five-year bands); and constraining the linear effect in one dimension to zero (Deaton & Paxson, 1994). These constraint methods all share the limitation outlined in §2.5.1: they amount to choosing one of the many equally-well-fitting θ parameters on an *ad hoc* basis, and the validity of that choice - and the constraints used to make it - cannot be tested.

Instead, in this paper I resolve the overparametrization of model (4.2) by reparametrizing it following Kuang et al. (2008) and Nielsen (2015) as outlined in §2.3. The reparametrization is a mapping from the parameter vector θ, which is overparametrized by four, to a uniquely identified parameter vector ξ, which has four fewer elements. There is a complementary mapping from the design vector d_{ip} to a shorter design vector x_{ip}. The interpretable elements of the new parameter vector ξ are accelerations in each of age, period, and cohort; these are the "non-linear" APC effects. The vector ξ also includes a linear plane in age-cohort space, but this is chosen by the researcher and so is not interpretable in isolation. See Figure 2.3 in §2.3.3.3 for an illustration of age-cohort space.

The full expression of the reparametrized model is

$$\begin{aligned} y_{ip} =& \mu_{ip} + z'_{ip}\zeta + \omega_i + \epsilon_{ip} \\ \mu_{ip} =& x'_{ip}\xi \\ =& v_o + (a_{ip} - U)v_a + (c_{ip} - U)v_c \\ &+ x_{ip}^{\mathbb{A}\prime}\xi^{\mathbb{A}} + x_{ip}^{\mathbb{P}\prime}\xi^{\mathbb{P}} + x_{ip}^{\mathbb{C}\prime}\xi^{\mathbb{C}}; \end{aligned} \tag{4.4}$$

this is the model introduced in equation (4.1). The parameter vector ξ here is

$$\begin{aligned}
\xi =& \{v_o, v_a, v_c, \xi^{\mathbb{A}}, \xi^{\mathbb{P}}, \xi^{\mathbb{C}}\} \\
\xi^{\mathbb{A}} =& \{\Delta^2\alpha_3, \ldots, \Delta^2\alpha_A\} \\
\xi^{\mathbb{P}} =& \{\Delta^2\beta_{L+1}, \ldots, \Delta^2\beta_{L+P}\} \\
\xi^{\mathbb{C}} =& \{\Delta^2\gamma_3, \ldots, \Delta^2\gamma_C\}
\end{aligned} \tag{4.5}$$

The parameters of most interest are accelerations of the form $\Delta^2\alpha_a = \alpha_a - 2\alpha_{a-1} + \alpha_{a-2}$, which are collected in $\{\xi^{\mathbb{A}}, \xi^{\mathbb{P}}, \xi^{\mathbb{C}}\}$ for age, period, and cohort respectively. The design vectors $\{x_{ip}^{\mathbb{A}}, x_{ip}^{\mathbb{P}}, x_{ip}^{\mathbb{C}}\}$ are sums of accelerations in age, period, and cohort respectively. They are defined analogously to $\{x_{ac}^{\mathbb{A}}, x_{ac}^{\mathbb{P}}, x_{ac}^{\mathbb{C}}\}$ in §2.3.3.4, using a_{ip} and c_{ip} to replace a and c in the definitions. See also Appendix 4.A. The first three elements of the parameter vector ξ are $\{v_o, v_a, v_c\}$, which are the parameters of the linear plane. The plane is defined on the basis of a reference point in age-cohort space at which $age = coh = U$ for U a data-determined constant, and slopes in the age and cohort dimensions away from that origin point. See §2.3.3.4 for details on the derivation of ξ.

The design vector x_{ip} can be split into a time-varying component $x_{1,ip}$ and a time-invariant component $x_{2,ip}$. Since an individual's cohort is fixed at birth, elements defined as functions of c_{ip} do not change over time; therefore define $x_{2,ip} = \{1, (c_{ip} - U), x_{ip}^{\mathbb{C}}\}$. Age and period do change over time, in a deterministic fashion, so elements defined as functions of a_{ip}, p_{ip} will be time-varying. Thus define $x_{1,ip} = \{(a_{ip} - U), x_{ip}^{\mathbb{A}}, x_{ip}^{\mathbb{P}}\}$.

Previous work has shown that all parameters in ξ are identified in certain non-panel settings. First, Kuang et al. (2008) and Nielsen (2015) showed that ξ is identified from aggregate data, subject to some conditions on the data structure. Then in §3, I showed that ξ is identified from repeated cross section data, again subject to conditions on the data structure. The conditions on the data structure relate to the set of age-cohort combinations appearing in the data, referred to as the generalized trapezoid. It is sufficient for identification that the generalized trapezoid be contiguous, that is, that there be no age-cohort cells within the dataset in which no observations are recorded.

The object of this chapter is to determine which elements of ξ can be identified in panel settings. I maintain the requirement of a contiguous generalized trapezoid. Identification depends on two considerations: first, the relationship between the elements of the design vector; and second, the relationship between the design vector and the noise terms. I consider three different panel settings, each of which makes different assumptions about the noise terms. These three settings are pooled OLS, random effects, and fixed effects.

4.3 The APC acceleration model in three panel settings

I account for the identification of the age-period-cohort model in equation (4.1) in three panel data settings, distinguished by the different assumptions they make about unobserved components. The three settings are: pooled ordinary least squares (pooled OLS), random effects, and fixed effects. The pooled OLS setting, discussed in §4.3.1, permits all elements of the APC acceleration parameter vector ξ to be identified and makes the least restrictive assumptions of the three settings regarding unobserved components; however, it yields inefficient estimates. The random effects setting, discussed in §4.3.2, also permits all elements of the APC acceleration parameter vector ξ to be identified; it makes stronger assumptions than the pooled OLS framework, but gains efficiency as a result. The fixed effects setting, discussed in §4.3.3, yields consistent estimates in some cases where the other two settings do not, but at a cost: only the elements of ξ associated with age and period are identified.

I consider the impact of the APC variables, x_{ip}, on the choice between the three settings. Typically, in panel data analysis, this choice depends on several factors. One important factor is the assumptions made about the correlation between the explanatory variables and unobserved components. The three panel settings make different assumptions in this regard. I show in §4.3.4 that the deterministic nature of the APC variables, x_{ip}, renders redundant the differences between the three panel settings in terms of the assumptions they make about the correlation between x_{ip} and unobserved components. The only exception is when the dataset is unbalanced in a way that is linked to age and period, so that the unobservable

component may be correlated with age or period elements of x_{ip}; this is described in §4.3.4.2. Apart from this exceptional case, assumptions about the correlation between x_{ip} and unobserved components are not a factor that needs to be considered in the choice between the three panel settings. Instead, the choice should be determined by other considerations, such as correlation between covariates and the unobserved components, assumptions about the within-individual correlation of the unobserved components, identification, and efficiency.

4.3.1 The pooled OLS setting

In the pooled OLS (POLS) setting, the individual-specific unobserved component ω_i is constrained to equal zero. Any unexplained variation is attributed to the individual- and time-specific unobserved component ϵ_{ip}. The ϵ_{ip} terms are permitted to have a general variance-covariance structure within an individual i (they must be independent across individuals). This generality makes POLS inference quite robust, but also inefficient.

The pooled OLS estimator for model (4.1) is the standard OLS estimator:

$$\begin{pmatrix} \hat{\xi} \\ \hat{\zeta} \end{pmatrix} = \Big[\begin{pmatrix} X & Z \end{pmatrix}' \begin{pmatrix} X & Z \end{pmatrix} \Big]^{-1} \begin{pmatrix} X & Z \end{pmatrix}' Y. \tag{4.6}$$

However, inference differs slightly from the standard OLS case. Standard OLS inference relies on an assumption that the error terms ϵ_{ip} are IID across both individuals i and time periods p. However, it is not plausible that the unobserved components for the same individual in different time periods would be independent. It is likely that there is some unexplained correlation in observations on the same individual over time. We need a form of inference which allows for within-individual correlation in ϵ_{ip} (Cameron & Trivedi, 2005, see for example p.702 of).

Inference in the pooled OLS model is conducted by treating the individual i as the base unit of analysis. Thus there are N independent observations, with $\{\epsilon_i, X_i, Z_i\}$ assumed IID across individuals i. It is also assumed that

1. $\omega_i = 0$
2. $\mathsf{E}(\epsilon_{ip}|x_{ip}, z_{ip}) = 0$
3. (X, Z) has full rank.

I will argue that under these assumptions, all elements of $\{\xi, \zeta\}$ are identified. I will then discuss the asymptotic properties of the pooled OLS estimator for $\{\xi, \zeta\}$.

It is straightforward to make the case for identification of all elements of $\{\xi, \zeta\}$ in the pooled OLS model, because the relevant assumptions are the same as those for repeated cross section data, where $\{\xi, \zeta\}$ is known to be identified. Identification in an econometric model is determined by two things: the correlation among explanatory variables, and the correlation between explanatory variables and the unobserved components. Regarding the correlation among explanatory variables, we know from the existing literature on the APC acceleration parametrization that the correlation among the APC variables in x_{ip} does not inhibit identification. In §3, I showed that the inclusion of covariates z_{ip} does not inhibit identification, provided that those covariates are not linear functions of x_{ip}. Regarding the correlation between explanatory variables and the unobserved components, this is avoided by the pooled OLS assumptions $\omega_i = 0$ and $\mathsf{E}(\epsilon_{ip}|x_{ip}, z_{ip}) = 0$. The latter assumption of contemporaneous exogeneity is the same as that used in §3.3.2. Essentially, because POLS treats panel data as though it were repeated cross section, and because §3 shows that $\{\xi, \zeta\}$ is identified from repeated-cross section data, it follows that $\{\xi, \zeta\}$ is identified from panel data in the POLS setting.

Having established identification under pooled OLS, I now consider the asymptotic behaviour of the estimator (4.6). The estimator is consistent and asymptotically normal under the assumptions listed above, in conjunction with some technical existence and finiteness conditions (Wooldridge, 2010, §7.3.2). It converges at a rate governed by N. Correct inference depends on the appropriate selection of variance-covariance matrix for ϵ_{ip}. Tests are available to determine whether heteroskedasticity or within-individual correlation is present (Croissant & Millo, 2008). Typically a variance-covariance matrix which allows for within-individual correlation, such as that of Arellano (1987), is advisable. Allowing for this general variance-covariance matrix causes OLS procedures like POLS to be relatively inefficient compared to an FGLS procedure such as that used in the random effects setting.

4.3.1.1 APC sub-models and pooled OLS

One of the advantages of embedding the APC accelerations in a regression model is the ease with which reductions of the model can be tested. Standard Wald tests can be used on any combination of the parameters $\{\xi, \zeta\}$. These tests are asymptotically χ^2, or approximately F distributed if ϵ_{ip} are approximately normal.

In the POLS setting, all of the standard sub-models of the APC acceleration model are identified and can be tested as restrictions of the APC acceleration model using Wald tests. I give a brief account of these sub-models here, with more detail available from Nielsen (2015) and §2.3.4.3 of this thesis.

The full model is the APC model, which includes all elements of ξ. There are five sets of reductions to this model which we consider. The first set of reductions eliminate one set of accelerations and the associated letter is dropped from the name: for example, the "AP" model omits all cohort accelerations $\xi^{\mathbb{C}} = \{\Delta^2\gamma_3, \ldots, \Delta^2\gamma_C\}$. The "PC" and "AC" models are similarly defined. The second set of reductions eliminate a second set of accelerations: for example, the "Ad" model omits all cohort and period accelerations. The "Pd" and "Cd" models are similarly defined. The "d" in the names stands for "drift", indicating that the linear plane is unrestricted. The third set of reductions restricts the linear plane: the "A", "P", and "C" models are analogous to the "Ad", "Pd", and "Cd" models but with the additional restriction that the linear plane has only one slope. The "A" model omits the slope in the cohort dimension, leaving only the slope in the age dimension; the "C" model does the opposite; and the "P" model constrains the slopes in the age and cohort dimensions to be equal, effectively producing a slope in the period dimension. There is also a "d" model which omits all accelerations but does not constrain the linear plane. The fourth set of reductions allow a single slope and no accelerations. These models are "tA", "tP", and "tC"; the "t" refers to the single slope or trend, while the "A", "P", or "C" indicates which slope is retained. The fifth and final reduction is to a simple intercept model, named "1". All of these models are feasible in the POLS setting.

4.3.2 The random effects setting

The random effects setting improves on POLS by efficiently accounting for correlation of the unobservable component within individuals while also permitting all elements of $\{\xi, \zeta\}$ to be identified and estimated, except in some exceptional situations. The exceptional situations arise where there are no time-varying variables in the model, i.e. $x_{1,ip}$ and $z_{1,ip}$ are omitted; here the parameters are identified under the random effects assumptions, but cannot be estimated using the random effects estimator. The efficiency comes at the cost of a restriction on the the generality of the variance-covariance matrix of ϵ_{ip}, and a stronger assumption on the relationship between the unobserved components and any covariates z_{ip}. There is no stronger assumption on the relationship between unobserved components and x_{ip}, as will be explained in §4.3.4.

The random effects (RE) estimator is an FGLS (feasible generalized least squares) estimator. FGLS is a two-step procedure. In step one, the variance-covariance matrix Ω of the unobserved component $\varepsilon_{ip} = \omega_i + \epsilon_{ip}$ is estimated. In step two, the estimate of Ω is used as a weight to produce estimates of $\{\xi, \zeta\}$.

In the RE setting, structure is imposed on $\{\omega_i, \epsilon_{ip}\}$, and thereby on Ω, to efficiently account for within-individual correlation in the unobservables. The individual-specific noise terms ω_i are taken to be independent draws from a distribution with mean zero and variance σ^2_ω, conditional on $\{X_i, Z_i\}$. The observation-specific noise terms ϵ_{ip} are taken to be independent draws from a distribution with mean zero and variance σ^2_ϵ, conditional on the explanatory variables $\{X_i, Z_i\}$ and on ω_i. Then Ω has every element on the main diagonal equal to $\sigma^2_\omega + \sigma^2_\epsilon$ and every element on an off-diagonal equal to σ^2_ω. The fact that Ω contains only two parameters makes RE more efficient and less computationally costly than general FGLS, where Ω is unrestricted.

To estimate model (4.1) in the random effects setting, a standard FGLS estimator is used. Define the design matrix, $\Pi_i = \begin{pmatrix} X_i & Z_i \end{pmatrix}$. The estimator is

$$\begin{pmatrix} \check{\xi} \\ \check{\zeta} \end{pmatrix} = \left[N^{-1} \sum_{i=1}^{N} \Pi_i' \hat{\Omega}^{-1} \Pi_i \right]^{-1} \left[N^{-1} \sum_{i=1}^{N} \Pi_i' \hat{\Omega}^{-1} Y_i \right] \tag{4.7}$$

Due to the structure imposed on ω_i, ϵ_{ip}, all elements of the main diagonal of $\hat{\Omega}$ can be constructed from consistent estimates of σ^2_ω and σ^2_ϵ. Such estimates can

be obtained using residuals from simple "between" and "within" OLS regressions; further details are available in §2.3 of Baltagi (2005).

A drawback of FGLS, and thus of random effects, is the need to assume strict exogeneity i.e. $\mathsf{E}(\epsilon_{ip}|X_i, Z_i) = 0$. This is stronger than the analogous pooled OLS assumption of contemporaneous exogeneity, $\mathsf{E}(\epsilon_{ip}|x_{ip}, z_{ip}) = 0$. The full set of assumptions for the random effects estimator is:

1. $\{\epsilon_i, \omega_i, X_i, Z_i\}$ are IID across individuals i
2. $\{X, Z\}$ has full rank
3. $\mathsf{E}(\omega_i|X_i, Z_i) = E(\epsilon_{ip}|X_i, Z_i, \omega_i) = 0$
4. $\mathsf{E}(\epsilon_i\epsilon_i'|X_i, Z_i, \omega_i) = \sigma_\epsilon^2 I_P$ and $E(\omega_i^2|X_i, Z_i) = \sigma_\omega^2$.

Under these assumptions and some technical existence and finiteness conditions the random effects estimator converges to the true parameter value at rate N and is asymptotically normal.

We know that $\{\xi, \zeta\}$ is identified in the RE setting because it is identified in the POLS setting, and moving from the POLS to the RE setting does not introduce correlation that would inhibit identification. In §4.3.1, I argued that ξ is identified in the POLS setting. Recall that only correlation among explanatory variables or correlation between the explanatory variables and unobservables could inhibit identification. Moving from POLS to RE, the correlation among explanatory variables is unchanged. Considering the correlation between explanatory variables and unobservables, the POLS assumptions

$$\omega_i = 0 \quad ; \quad \mathsf{E}(\epsilon_{ip}|x_{ip}, z_{ip}) = 0 \tag{4.8}$$

are replaced with the RE assumptions

$$\begin{aligned} \mathsf{E}(\omega_i|x_{i,L+1}, \ldots, x_{i,L+P}, z_{i,L+1}, \ldots, z_{i,L+P}) &= 0; \\ \mathsf{E}(\epsilon_{ip}|x_{i,L+1}, \ldots, x_{i,L+P}, z_{i,L+1}, \ldots, z_{i,L+P}, \omega_i) &= 0. \end{aligned} \tag{4.9}$$

These RE assumptions rule out correlation between unobservables and explanatory variables and therefore ensure identification.

4.3.2.1 APC sub-models and random effects

Some of the sub-models of the full APC acceleration model listed in §4.3.1.1, are identified under RE assumptions but cannot be estimated using the RE estimator. These are models where there are no time-varying explanatory variables, i.e. where $x_{1,ip}$ and $z_{1,ip}$ are omitted. To understand why they cannot be estimated, recall that construction of the RE estimator requires estimates of the variances σ^2_ω and σ^2_ϵ, to construct the weight matrix $\hat{\Omega}$. These estimates are generated from "between" and "within" regressions. "Between" regressions consider only the variation between individuals, while "within" regressions consider only the variation within individuals. However if there are no time-varying variables in the model, there is no within-individual variation, so it is impossible to perform a "within" regression. Without the "within" regression is impossible to generate separate estimates of σ^2_ω and σ^2_ϵ, and thus it is impossible to construct $\hat{\Omega}$.

The sub-models which cannot be estimated using the random effects estimator are those which contain only elements of $x_{2,ip} = \{1, (c_{ip} - U), x^{\mathbb{C}}_{ip}\}$ and $z_{2,ip}$. Considering the set of APC sub-models discussed in §4.3.1.1, those which only contain elements of $x_{2,ip}$ are the C, tC, and 1 models. Inefficient estimates of these models can still be obtained under random effects assumptions using the POLS estimator with an unrestricted variance-covariance matrix for ϵ_i.

4.3.3 The fixed effects setting

The fixed effects setting permits only age and period, but not cohort, accelerations to be estimated. The fixed effects (FE) setting accounts for correlation between unobservable components within an individual, and is robust to correlation of the individual-specific unobservable component ω_i with explanatory variables. This robustness comes at a cost: the effects of time-invariant explanatory variables are not identified. In the APC acceleration model, the loss of time-invariant variables means the loss of all cohort accelerations, $\xi^{\mathbb{C}}$, as well as the intercept v_o and the slope in the cohort dimension v_c. This makes FE a relatively undesirable setting, especially since correlation between unobserved components and the remaining APC variables is impossible in most panel settings, as discussed in §4.3.4.2.

The FE estimator uses the repetition in panel data to control for unobserved, individual-specific, time-invariant factors that influence the outcome variable. These factors produce correlation in the unobservable components within an individual over time, represented by ω_i. They may also be correlated with the explanatory variables in the regression. Where ω_i is correlated with the explanatory variables in the regression, both POLS and RE estimators are are biased and inconsistent. FE estimation avoids this problem.

The cost of FE estimation is that the ability to identify coefficients on time-invariant explanatory variables $\{x_{2,i}, z_{2,i}\}$ is lost. Recall that $x_{2,i} = \{1, (c_{ip} - U), x_{ip}^{\mathbb{C}}\}$, so the associated coefficients are $\xi_2 = \{v_o, v_c, \xi^{\mathbb{C}}\}$; all cohort accelerations, as well as the origin point of the linear plane and one slope, are not identified in the FE setting. Similarly ζ_2, the coefficients associated with any time-invariant covariates $z_{2,ip}$, are not identified.

The fact that one of the two slopes of the APC acceleration model is time-invariant and therefore not identified in the FE setting is not simply an artefact of the choice made in §2.3.3.4 to define one of the slopes in the cohort dimension. Rather it reflects a fundamental limitation on the variation of the data, such that only one slope is identified. This is discussed further in Appendix 4.C.

In the FE setting ω_i are parameters to be estimated. One way to implement this is to run pooled OLS on a model which includes an indicator for each individual in the sample; this is the least squares dummy variable (LSDV) model. An equivalent implementation is the time-demeaned model,

$$\tilde{y}_{ip} = (\tilde{a}_{ip} - U)v_a + \tilde{x}_{ip}^{\mathbb{A}'}\xi^{\mathbb{A}} + \tilde{x}_{ip}^{\mathbb{P}'}\xi^{\mathbb{P}} + \tilde{z}_{1,ip}'\zeta_1 + \epsilon_{ip}. \tag{4.10}$$

where

$$\tilde{y}_{ip} = y_{ip} - P^{-1}\sum_p y_{ip} \tag{4.11}$$

and $\tilde{z}_{1,ip}, \tilde{x}_{ip}^{\mathbb{A}'}$, and $\tilde{x}_{ip}^{\mathbb{P}'}$ are constructed similarly (note that if an individual is observed for fewer than P periods then the sum should be premultiplied by the inverse of the number of periods for which that individual is observed rather than P^{-1}). The value and asymptotic properties of the estimates of $\{\xi_1, \zeta_1\}$ from this model are identical to those from the LSDV model (Wooldridge, 2010, §10.5.3).

The FE assumptions are:

1. $\{\epsilon_i, X_i, Z_i\}$ are IID across individuals i
2. $\mathsf{E}(\epsilon_{ip}|X_i, Z_i) = 0$ (strict exogeneity)
3. $\tilde{X}, \tilde{Z}$ has full rank.

I first discuss identification under these assumptions, and then the asymptotic properties of the FE estimator.

While it is already clear that the parameters $\{\xi_1, \zeta_1\}$ are not identified under fixed effects, it is still necessary to demonstrate that the time-demeaning procedure does not induce linear dependence that would inhibit identification of the remaining parameters. Recall that identification relies on both absence of collinearity among the explanatory variables and absence of collinearity between the explanatory variables and the unobserved components. The latter condition is achieved by the FE assumption $\mathsf{E}(\epsilon_{ip}|X_i, Z_i) = 0$. The former condition is the absence of linear dependence between $\{(\tilde{a}_{ip} - U), \tilde{x}_{ip}^{\mathbb{A}'}, \tilde{x}_{ip}^{\mathbb{P}'}, \tilde{z}_{ip}\}$. A proof of the absence of this linear dependence is given in Appendix 4.B.

Under fixed effects assumptions as well as technical existence and finiteness conditions, the FE estimator of $\{\xi_1, \zeta_1\}$ converges to the true parameter at rate N and is asymptotically normal. The variance of the asymptotic normal distribution will depend on the choice of variance-covariance matrix for the noise component ϵ_{ip}. As a baseline the variance-covariance matrix is assumed to be homoskedastic and diagonal, but other options exist (see for example Arellano, 1987).

Under an additional assumption, the parameters $\{\xi_2, \zeta_2\}$ of the time-invariant variables can be estimated, using a two-stage method (Hausman & Taylor, 1981). The required assumption is $\mathsf{E}(\omega_i|X_{2,i}, Z_{2,i}) = 0$. This assumption is slightly weaker than that made in the RE and POLS settings, $\mathsf{E}(\omega_i|X_i, Z_i) = 0$.

4.3.3.1 APC sub-models and fixed effects

In the fixed effects setting, the main acceleration model and standard reductions are different to those described in §4.3.1.1 and §4.3.2.1. The main model is the "FAP" model, and three sub-models are also considered.

FAP This is the model in equation (4.10). This has accelerations in both age and period but only one slope, in the age dimension. The lack of second slope is because the individual fixed effects absorb the slope in the cohort dimension.

FA This model omits the period accelerations from the FAP model: $\forall p : \Delta^2 \beta_p = 0$. This model is used in §4.4 to describe hospital in-patient stays.

FP This model omits the age accelerations from the FAP model: $\forall a : \Delta^2 \alpha_a = 0$.

Ft This model omits both age and period accelerations from the FAP model: $\forall a, p : \Delta^2 \beta_p = \Delta^2 \alpha_a = 0$. It retains a single slope in age, which combines the linear effects of age and period. This model is also used in §4.4.

As in the other two panel settings, the appropriate sub-model is determined by a series of Wald tests. First, all models are tested against the most general model (APC or FAP). Then, the smallest model not rejected by that first round of Wald tests is tested against the models which it nests. If all reductions are rejected in this second round of testing, the model is retained; otherwise the smallest model not rejected in the second round of testing goes to a third round of testing, and so on until a final model is selected.

4.3.4 Choice of panel setting for APC acceleration model

In this section, I provide guidance on the choice between the three panel settings in the context of the APC acceleration model. I show that there is little impact of the inclusion of APC explanatory variables on the correlation between explanatory variables and unobserved components. This strengthens the case for the use of the random effects setting.

Typically, explanatory variables affect the choice between the panel settings in two ways. First, the parameters associated with certain explanatory variables may not be identified in all settings. I have already shown in §4.3.3 that cohort effects are not identified in the FE setting. This reduces the attractiveness of the FE setting. Second, the three panel settings differ in their assumptions about the relationship between explanatory variables and unobserved components. In this section, I show that the deterministic nature of the APC variables renders redundant the differences between the panel settings in their assumptions about correlation between the APC explanatory variables and unobserved components, except where the dataset is unbalanced in a way that is linked to age or period. Therefore the choice between panel settings need not be affected by consideration of this correlation. Instead, it should be based on other considerations, such as

whether all parameters are identified in each setting, assumptions the different settings imply about correlation between other explanatory variables (covariates) and unobserved components, restrictions they impose on the within-individual correlation of the unobserved components, and efficiency.

In general terms, there are two major points of difference between the panel settings in terms of their assumptions about the relationship between the explanatory variables and unobserved components. The differences are as follows:

1. **Correlation between explanatory variables and ϵ_{ip}, within individuals i but across periods p:** The POLS setting requires only contemporaneous exogeneity of the unobserved components with respect to the explanatory variables, whereas both RE and FE require strict exogeneity.
2. **Correlation between explanatory variables and ω_i:** The FE setting allows some correlation between the explanatory variables and the unobserved components; any time-varying explanatory variable may be correlated with the individual-specific unobservable component ω_i. Neither the RE nor the POLS setting permits this.

In the context of the APC acceleration model, I show that these two points of difference are redundant with respect to the APC explanatory variables. This redundancy is due to the deterministic properties of the APC explanatory variables. Regarding the first point of difference, I show in §4.3.4.1 that if the unobservable components exhibit contemporaneous exogeneity with respect to the APC explanatory variables, this implies strict exogeneity; so the advantage of the POLS setting in this respect is eliminated. Regarding the second point of difference, I show in §4.3.4.2 that the ability of the FE setting to allow for correlation between time-varying APC explanatory variables, $x_{1,ip}$, and the unobserved component ω_i is useful only if the panel is unbalanced in a way that is linked to age and period. This unbalancedness is the only mechanism which could generate correlation between $x_{1,ip}$ and ω_i. In the absence of such unbalancedness, there is no risk of correlation between $x_{1,ip}$ and ω_i, so the advantage of the FE setting is eliminated.

The redundancy of these two points of difference eliminates some advantages of both the FE and POLS settings. This strengthens the case for the use of

the RE setting, which efficiently accounts for within-individual correlation of the unobserved components while permitting identification of all APC parameters.

Since the two points of difference regarding the correlation between APC explanatory variables and the unobserved components are redundant, the choice between the three settings should be made on the basis of other considerations. These considerations include the two points of difference as they relate to correlation between unobserved components and covariates z_{ip}; the differences between the settings in terms of assumptions about the within-individual correlation of unobserved components; and differences between the settings regarding efficiency and identification. Except in the special case of unbalancedness linked to age and period, which is discussed in §4.3.4.2, there is no possibility of correlation between the APC variables and unobserved components, so this need not be considered. The rest of this section explains why there is no possibility of such correlation.

4.3.4.1 Strict exogeneity

In model (4.1), strict exogeneity of the unobserved component ϵ_{ip} with respect to the APC variables, x_{ip}, is implied by the contemporaneous exogeneity of the unobserved component ϵ_{ip} with respect to x_{ip}. This means that the validity of the strict exogeneity assumption as it applies to x_{ip}, required by the RE and FE settings but not by POLS, need not be taken into consideration when choosing between the three panel settings (note that it still needs to be considered as it applies to the covariates z_{ip}). The implication is that the relative value of POLS, which unlike RE and FE does not require this assumption, is reduced.

Strict exogeneity (SE) in the model without covariates is defined as $\mathsf{E}(\epsilon_{ip}|X_i) = 0$. Contemporaneous exogeneity (CE) is defined as $\mathsf{E}(\epsilon_{ip}|x_{ip}) = 0$. I show that CE implies SE by showing that $\mathsf{E}(\epsilon_{ip}|X_i) = \mathsf{E}(\epsilon_{ip}|x_{ip})$.

1. Fix the dimensions of the age-cohort space; then U and L are fixed so x_{ip} is a function of $\{a_{ip}, p_{ip}\}$ only via the $m(r, s)$ function defined in §2.3.3.4.
2. Note the there is a bijective mapping between x_{ip} and $\{a_{ip}, p_{ip}\}$. Therefore $\mathsf{E}(\epsilon_{ip}|x_{ip}) = \mathsf{E}(\epsilon_{ip}|a_{ip}, p_{ip})$.
3. Let $\{a_{i,L+1}, p_{i,L+1}\}$ be a random variable. Then for any p, $\{a_{ip}, p_{ip}\}$ is a function of $\{a_{i,L+1}, p_{i,L+1}\} : \{a_{ip}, p_{ip}\} = \{a_{i,L+1} + p - 1, p_{i,L+1} + p - 1\}$. Therefore

there is also a $1:1$ relationship between $\{a_{ip}, p_{ip}\}$ and $\{a_{i,L+1}, p_{i,L+1}\}$ for any p. Therefore, $\mathsf{E}(\epsilon_{ip}|a_{ip}, p_{ip}) = \mathsf{E}(\epsilon_{ip}|a_{i,L+1}, p_{i,L+1})$.

4. The above logic also implies $\mathsf{E}(\epsilon_{ip}|X_i) = \mathsf{E}(\epsilon_{ip}|a_{i,L+1}, p_{i,L+1})$.
5. Therefore, $\mathsf{E}(\epsilon_{ip}|X_i) = \mathsf{E}(\epsilon_{ip}|x_{ip})$

Where covariates z_{ip} (which satisfy the random effects assumptions) appear in the model, the parameter ξ will remain identified, but the special property of equivalence between strict and contemporaneous exogeneity will be lost. That is because, in general, the value of a covariate z_{ip} is not a function of $z_{i,L+1}$. Therefore, $\mathsf{E}(\epsilon_{ip}|z_{ip}) = 0$ does not guarantee $\mathsf{E}(\epsilon_{ip}|Z_i) = 0$. However, if the researcher is satisfied to assume strict exogeneity of covariates, then the above result means that there is strict exogeneity for the full model. If there is strict exogeneity for the full model, then the relative advantage of the POLS framework is reduced.

4.3.4.2 Correlation with time-invariant unobservables

The deterministic nature of the APC variables x_{ip} means that there is only one, rare mechanism which could generate correlation between the time-varying APC explanatory variables, $x_{1,ip}$, and the time-invariant unobserved component, ω_i. This mechanism is a particular form of unbalancedness linked to age and period. The implication is that the relative advantage of the FE setting, which can accommodate correlation between explanatory variables and ω_i, is reduced for APC models. The FE setting is only useful, in the APC context, if the above-mentioned unbalancedness exists, or if there is concern about correlation between a time-varying covariate $z_{1,ip}$ and the time-invariant unobserved component ω_i.

Generally speaking, there are two mechanisms that could generate correlation between time-varying explanatory variables and ω_i. Thee first involves a causal relationship between the unobserved component ω_i and the explanatory variables, while the second involves non-random unbalancedness of the panel dataset. I will argue that the deterministic nature of the time-varying APC variables, $x_{1,ip}$, prevents the first mechanism; the second mechanism is possible for $x_{1,ip}$, but unlikely.

The first mechanism which can generate correlation between an explanatory variable and the unobserved component ω_i arises where the explanatory variable and ω_i are causally related. Typically, this would be because ω_i is a composite of some omitted variables that directly affect the explanatory variables. Let w_{ip}

be a generic vector of explanatory variables. Then this causal mechanism for generating correlation between w_{ip} and ω_i can be expressed as a data-generating process (DGP) of the form:

$$y_{ip} = \delta + w'_{ip}\psi + \omega_i + \epsilon_{ip}$$

$$w_{ip} = f_i + u_{ip} \tag{4.12}$$

$$\mathsf{cov}(\omega_i, f_i) \neq 0 \tag{4.13}$$

Here ψ is a vector of coefficients on the explanatory variables w_{ip} and f_i is a time-constant composite including some of the same variables as ω_i. In this DGP, $\mathsf{E}(\omega_i|w_{ip}) \neq 0$, i.e. there is correlation between the unobservable component ω_i and the explanatory variables in the model for y_{ip}.

This sort of causal relationship is not possible when the explanatory variables consist of only time-varying APC effects. The DGP described in the previous paragraph is impossible if $w_{ip} = x_{ip}$. We already know that x_{ip} is fully determined by a_{ip} and p_{ip}. Since a_{ip} and p_{ip} are time-varying, they could not appear in the composites f_i or ω_i. Therefore there is no possibility of correlation arising via this mechanism in a model with only time-varying APC effects.

The second mechanism that can produce correlation between an explanatory variable and ω_i arises due to a particular type of non-random unbalancedness in the data. This non-random unbalancedness can be understood as follows. Define an indicator h_{ip} which takes the value 1 if individual i appears in the dataset at time p. The dataset thus contains only observations for which $h_{ip} = 1$, and all statements about the dataset can be written with conditioning on h_{ip}. While I argued above that it is not possible to have a data-generating process under which $\mathsf{E}(\omega_i|x_{1,ip}) \neq 0$, it is possible, once unbalancedness in the data is accounted for, that $\mathsf{E}(\omega_i|x_{1,ip}, h_{ip}) \neq \mathsf{E}(\omega_i|x_{1,ip})$. Therefore, in the actual dataset it may be the case that $\mathsf{E}(\omega_i|x_{1,ip}, h_{ip}) \neq 0$, i.e. there is correlation between the unobservable component ω_i and the explanatory variable which necessitates the use of fixed effects rather than random effects or POLS.

This mechanism can be defined formally via the dataset selection process

$$\begin{aligned} y_{ip} &= x'_{1,ip}\xi_1 + \omega_i + \epsilon_{ip} \\ h_{ip} &= \mathbb{1}(h^*_{ip} > 0) \\ h^*_{ip} &= w'_{ip}\phi + g_i + e_{ip}. \end{aligned} \tag{4.14}$$

Here, a latent variable, h^*_{ip}, determines whether h_{ip} equals 1 or not. This latent variable depends on variables w_{ip}, which may include the APC variables $x_{1,ip}$; a noise term e_{ip}; and an individual-specific effect g_i. If there is correlation between ω_i and g_i, then $\mathsf{E}(\omega_i|x_{1,ip}, h_{ip}) \neq E(\omega_i|x_{1,ip})$; and so even though $\mathsf{E}(\omega_i|x_{1,ip})$, it need not be the case that $\mathsf{E}(\omega_i|x_{1,ip}, h_{ip}) = 0$.

This dataset selection process is possible when the only explanatory variables in the main model for y_{ip} are time-varying APC variables. To give a concrete example: let h_{ip} be labour force participation, and y_{ip} be productivity. The unobservable characteristics which determine labour force attachment (high g_i) may also increase productivity (high ω_i). The effect of the business cycle on labour force participation and productivity manifests as period effects in the models for h_{ip} and y_{ip} respectively. If an upswing in the business cycle increases labour force participation, then more individuals with low g_i and ω_i will have $h_{ip} = 1$ in that period, creating a relationship between ω_i and the period effects in x_{ip} that is conditional on h_{ip} i.e. $\mathsf{E}(\omega_i|x_{1,ip}, h_{ip}) \neq 0$.

If there exists correlation between ω_i and $x_{1,ip}$ due to such a dataset selection process, then under one additional condition the FE estimator will be preferred to the RE estimator or POLS estimator. In the presence of such correlation the RE and POLS estimators will always be inconsistent. Verbeek & Nijman (1992) show that, under the relatively weak condition that $\mathsf{E}(\tilde{\epsilon}_{ip}|h_{ip}) = 0$, the fixed effects estimator is consistent. However, if that condition is not satisfied then the FE estimator will also be inconsistent and is no better than the RE or POLS estimators. In that situation a full model for the dataset selection process is needed, using some of the techniques described by Verbeek & Nijman (1992). An analysis of how these techniques interact with the APC acceleration model is beyond the scope of this thesis and is left for future work.

Overall, there are few contexts in which an FE estimator will be necessary for a model with only APC effects. That is because the fixed effects estimator is

designed to account for correlation between the unobservable component ω_i and the explanatory variables, but where the explanatory variables are x_{ip} only there is only one mechanism which could generate that correlation. That mechanism is a particular form of panel unbalancedness, which will affect only some panel datasets. Where this mechanism does not operate, for example in a balanced panel, then this relative advantage of the FE setting is lost.

In a model with covariates in addition to age, period, and cohort effects, the choice between the three panel settings must consider the relationship between those covariates and the unobserved component ω_i. Where the covariates z_{ip} are correlated with ω_i, the RE or POLS estimators of $\{\xi, \zeta\}$ will be inconsistent but the FE estimator will be consistent and therefore preferred.

4.4 Application: evaluation of control variables in a study of the effect of commuting on hospital stays

I use the framework developed in this paper, permitting APC accelerations to be identified and estimated from panel data regressions, as a diagnostic tool to evaluate the treatment of controls in a health economics model. Period fixed effects and a quadratic in age are "standard" controls in such models, but in this application I show that a more careful modelling of APC effects can lead to improvements in the model. There is scope for the APC acceleration framework to be more widely used as a diagnostic tool in health economics and other applied areas to improve the standard treatment of age, period, and cohort as controls.

The health economics model I consider is that of Künn-Nelen (2016), which examines the impact of commute time on the probability of being a hospital in-patient. There is a clear policy interest in developing good models of the relationship between commute time and health outcomes; since most people commute and public investment in commuting infrastructure affects commute time, it is valuable to know whether this is a policy lever which can be used to affect public health. The particular analysis of in-patient hospital stays by Künn-Nelen is worth revisiting for three reasons. First, as Künn-Nelen points out, hospital in-patient stays

are a more objective measure of public health than many others commonly used in the health literature, but are relatively understudied. Second, Künn-Nelen's finding (that commute time does not affect the probability of being a hospital in-patient) is marginal: the coefficients on both commute time and its square were individually significant at the 10% level, but were jointly insignificant. There is therefore added value to improving the model in order to get a more clear-cut result. Third, Künn-Nelen specifies the age control as quadratic but provides little justification for this assumption. She seems to adopt it from the literature on self-reported health and well-being, outcomes which may have a different relationship with age than do hospital in-patient stays. There is therefore value to more critically examining this specification.

In my analysis I use the APC acceleration framework as a diagnostic to evaluate the treatment of controls in the analysis of Künn-Nelen (2016), and to identify improvements to the model which allow me to unambiguously conclude that commute time is not a significant determinant of the probability of being a hospital in-patient. This finding suggests that policy interventions to change commute time will not substantially change this indicator of public health, strengthening the original conclusion of Künn-Nelen. The APC acceleration framework is employed to reach this conclusion as follows. First, I estimate the APC acceleration model and test the quadratic restriction of the age accelerations. The restriction is rejected. Second, I plot the detrended sum of age accelerations and from this discover the need to account for childbirth among women. Once childbirth is accounted for, the fit of the overall model for hospital in-patient stays improves and the uncertainty regarding the impact of commuting is resolved: all three of the coefficient on commute time, the coefficient on its square, and the joint test are insignificant.

The remainder of this §4.4 is as follows. First in §4.4.1, I review the use of controls - in particular APC controls - in the literature on the relationship between commute time and health. Then in §4.4.2 I describe the data I use in my analysis, and how it differs from the data used by Künn-Nelen (2016). I give evidence that my data is a good proxy for that used by Künn-Nelen in §4.4.3, by showing that I can replicate her regression results. In §4.4.4 I show how the APC acceleration framework can be used as a diagnostic tool to improve Künn-Nelen's model.

4.4.1 Literature review

There is a substantial literature investigating the impact of commute times on health outcomes, which is of interest to policy makers. This literature relies on regressions of health outcomes on commute time, with a standard set of controls, including age and period. The existence of some contradictory results in the literature has generated interest in how seemingly minor modelling choices, such as the functional form of the commute time variable or the period range of the data, can influence the findings. The treatment of controls is another seemingly minor modelling choice which has not been carefully investigated but could be expected to influence the findings. The APC acceleration framework developed in this paper can be used as a diagnostic tool to evaluate the treatment of controls in such studies. In this paper I illustrate this with a health outcome which has been relatively under-studied in this literature: hospital in-patient stays.

In an era of growing cities, there is a clear public policy interest in the literature investigating the impact of commute time on health. There is substantial evidence to suggest a detrimental effect of commute time on subjective measures such as life and health satisfaction (Stutzer & Frey, 2008; Roberts et al., 2011; Nie & Sousa-Poza, 2018; Lorenz, 2018; Dickerson et al., 2014; Ingenfeld et al., 2019; Künn-Nelen, 2016; Clark et al., 2019), perceived health status (Künn-Nelen, 2016; Hansson et al., 2011; Oliveira et al., 2015), and quality of sleep (Hansson et al., 2011; Halonen et al., 2020). Positive relationships have been found between commute time and the number of health problems (Costa et al., 1988; Künn-Nelen, 2016); reported days off work due to illness (Costa et al., 1988; Gimenez-Nadal et al., 2018; Hansson et al., 2011; Van Ommeren & Gutiérrez-i-Puigarnau, 2011); perceived stress (Gottholmseder et al., 2009); and mortality (Sandow et al., 2014). However, some contradictory studies exist which have found no correlation with diagnosed health problems (Costa et al., 1988; Künn-Nelen, 2016); overall happiness or life satisfaction (Dickerson et al., 2014; Lorenz, 2018); or reported days off work due to illness (Künn-Nelen, 2016).

Due to conflicting results, this literature has recently paid attention to the impact of seemingly small modelling choices on findings. For instance, Dickerson et al. (2014) find no effect of commute time on happiness, whereas Roberts et al.

(2011) find a negative effect, although both use the British Household Panel Survey (BHPS) and the same measure of happiness. Dickerson et al. (2014) establish that the difference is due to the choice of BHPS waves included in the sample; while they use waves 6 through 18, Roberts et al. used waves 4 through 14. There is also evidence that modelling the relationship between health and commute time as monotonic may be restrictive and lead to inaccurate conclusions. Ingenfeld et al. (2019) find that only very long commutes reduce overall life satisfaction, whereas health satisfaction is most negatively affected by mid-length commutes. Hansson et al. (2011) find that mid-length car commutes are more detrimental to sleep quality, and produce more stress symptoms, than long ones.

A modelling choice of interest is the use of a "standard" set of controls, with little consideration to whether they are appropriate to the particular dataset and health outcome. Künn-Nelen (2016) justifies her choice of controls by stating "This set of control variables is common in the literature on health outcomes such as heath satisfaction, BMI, and sickness absence (e.g., Rietveld et al., 2014; Roberts et al., 2011; Hansson et al., 2011; Stutzer and Frey, 2008)," although she studies a wider range of health outcomes than are considered in these studies. Two of the studies she cites, Roberts et al. (2011) and Hansson et al. (2011), do little to justify their choice of controls. A third, Stutzer & Frey (2008), justify their choice of controls by reference to their own earlier paper from 2004.

The risk of relying on standard controls is that the set of controls becomes ossified, so that new important controls are omitted while old unimportant controls are retained. This is particularly likely if the set of "standard" controls comes from analyses of a different outcome variable, a different country, or a different time period. For instance, it is unclear whether the controls used by Stutzer and Frey in Germany in 2004 - prior to the development of the smartphone - remain appropriate in a study of UK commuting in 2016.

A second modelling choice is the use of "standard" functional forms for non-categorical controls, such as age; again with little consideration to their suitability for the model at hand. The need to control for age in studies of the impact of commute time on health is widely accepted; see the discussions in Dickerson et al. (2014); Ingenfeld et al. (2019); Gimenez-Nadal et al. (2018). It is typical to model age using a quadratic (Stutzer & Frey, 2008; Gimenez-Nadal et al., 2018;

Lorenz, 2018; Dickerson et al., 2014; Roberts et al., 2011; Ingenfeld et al., 2019; Künn-Nelen, 2016) or a small number of categories (Hansson et al., 2011; Clark et al., 2019) or even as linear (Gottholmseder et al., 2009; Oliveira et al., 2015; Nie & Sousa-Poza, 2018). Sometimes the functional form restriction is justified by reference to existing literature; for instance, Dickerson et al. (2014) and Roberts et al. (2011) choose a quadratic in age because Blanchflower & Oswald (2008) found that happiness is U-shaped in age. However, the choice of functional form is often neither explained nor evaluated. The APC acceleration framework provides the tools to perform such an evaluation.

In this paper I focus on modelling choices for controls in relation to a relatively under-studied outcome: hospital in-patient stays. This outcome is of interest for two reasons. First, hospital in-patient stays constitute a direct cost to the taxpayer in publicly-funded systems like the UK's National Health Service (NHS). Second, hospital in-patient stays are an objective measure of health, unlike self-reported health or sleep quality. The only study to date investigating the relationship between commute time and hospital in-patient stays was Künn-Nelen (2016). The outcome of that study was ambiguous: the coefficients on commute time and its square were individually significant at the 10% level, but jointly insignificant. In this study, Künn-Nelen relied on a set of standard controls, and functional forms for those controls, drawn from models of other health outcomes such as health satisfaction and BMI. I use the APC panel regression framework developed in this paper to examine whether the uncertainty in Künn-Nelen (2016)'s analysis of the relationship between commute time and hospital in-patient stays may be driven by the reliance on standard controls.

4.4.2 Description of the data

I did not have access to Künn-Nelen's exact dataset for my analysis, but I replicated the data based on the description in her paper. The similarity between the two datasets can be seen by comparing descriptive statistics and the results of regression models; the former in this section, the latter in the next section.

I use annual data from the British Household Panel Survey 1991-2008. The explanatory variables of interest are one-way morning commute time in minutes

and age in single-year increments. The main health outcome of interest is an indicator for whether the person was a hospital in-patient in the past year.

Following Künn-Nelen, I restrict the data to individuals aged between 18 and 65 who report being in full-time employment, who commute to work using a "passive" mode (i.e. travel by public transport, car, or motorbike, rather than walking or cycling), and who report a strictly positive commuting time. I top-code commute time at 90 minutes. Summary statistics for the data are provided in Table 4.1. These are not an exact match to the statistics in Table 1 of Künn-Nelen (2016), but they are close. The overall discrepancy is likely driven by a slightly different approach to top-coding of commuting time, and possible revisions to the data between access times. I was unable to replicate Künn-Nelen's variable recording relationship to the household head so I use an alternative, recording relationship to the household reference person.

I also replicated the descriptive Figures 2 and 3 from Künn-Nelen (2016), seen in Figures 4.1a and 4.1b of this paper, respectively. The first shows the average commute time for those who report a strictly positive commute time and for whom the place of work is not the home. The figure shows what Künn-Nelen describes as a "a clear though small increase" in average commute time with period. Although the Y-axis in my Figure 4.1a is shifted down about two minutes from her Figure 2, they are much closer to the numbers she reports in the text: an increase "from 22.8 min in 1992 to 24.6 min in 2007". My Figure 4.1b shows that commute time is right-skewed, with most commutes less than thirty minutes in duration. This is closely aligned with Künn-Nelen's Figure 3.

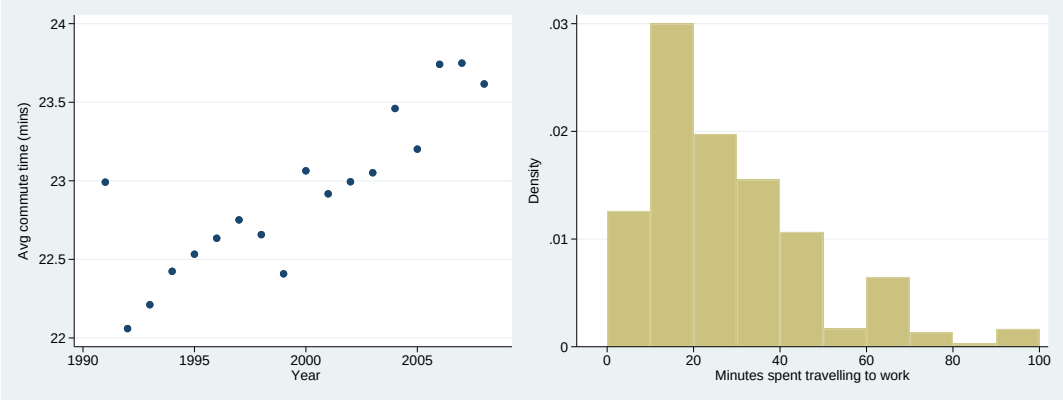
4.4.3 Replication of results from the original model

To further demonstrate that my data is comparable to that used by Künn-Nelen, I show that I get similar regression results to those in her paper when running the same model on my data. I replicate the fixed effect regression of the hospital in-patient stay indicator on commute time, seen in Table 3 of Künn-Nelen (2016). This regression is a linear probability model for hospital in-patient stays as a function of a quadratic in commute time, a quadratic in age, and other controls,

Table 4.1: BHPS Summary statistics (cf. Künn-Nelen Table 1)

Variable	mean	sd	min	max
In-patient hospital stay	0.07	0.25	0	1
Commuting time	25.26	18.47	1	90
Age	38.15	11.46	18	65
Female	0.43	0.50	0	1
Number of children	0.60	0.92	0	7
Highest qualification level	5.51	2.98	1	12
Relationship to household ref. person	2.26	3.13	1	30
Marital status	2.44	2.02	1	9
Length (days) current employment	1826.00	2223.28	0	18 570
Overtime hours	4.52	6.52	0	80
Net household income	586.36	302.32	37.78	10 010.67
Hospital stays were childbirth	0.01	0.11	0	1

Figure 4.1: BHPS summary figures (cf. Künn-Nelen Figures 2 and 3)



(a) commute time increases with year (b) most commutes are < 30 minutes

per the following equation

$$H_{ip} = \phi_1 CT_{ip} + \phi_2 CT_{ip}^2 + \lambda_1 a_{ip} + \lambda_2 a_{ip}^2 + \beta_p + z'_{ip}\zeta + \omega_i + \epsilon_{ip}. \tag{4.15}$$

Here, H_{ip} is an indicator which takes the value 1 if individual i was a hospital in-patient in year p and 0 otherwise; CT_{ip} is the commute time of i at p; a_{ip} is the age of i at p; β_p is a time fixed effect; z_{ip} is a set of controls; ω_i is an individual fixed effect; and ϵ_{ip} is a noise term.

The results of this regression are seen in the second column of my Table 4.2. The original results as reported in Künn-Nelen (2016) appear in the first column of this table, for comparison. Note that although her model did include age and its square as explanatory variables, she did not report her estimates; nor did she report the various R^2 measures I report at the bottom of the table. The third column contains results from models which use the APC panel regression framework developed earlier in this paper; these will be discussed in §4.4.4.

The coefficients I report from my regression are in line with those reported by Kunn-Nelen. The F-statistic from a test of the hypothesis that $\phi_1 = \phi_2 = 0$, i.e. that the combined effect of commute time and its square is zero, is also in line with that of Künn-Nelen. Note that the F-test is used here to align with Künn-Nelen; since no assumption of normality is imposed on the error term, this is justified on an approximate basis, see §4.2.2 of Wooldridge (2010).

Table 4.2: Regression results for in-patient stay

	In-patient hospital stay		
	Original Künn-Nelen results	Replication	APC accelerations
CT	0.0005*	0.0005*	0.0005*
	(0.0002)	(0.0002)	(0.0002)
CT^2	−5.19e-06*	−5.22e-06*	−5.17e-06*
	(2.99e-06)	(2.90e-06)	(2.90e-06)
Age		−0.0065***	
		(0.0014)	
Age^2		7.95e-05***	
		(1.57e-05)	
APC model			FA
H_0: age quadratic			0.0158
H_0: $\phi_1 = \phi_2 = 0$	0.1742	0.15	0.1543
Individuals	14,066	13,889	13,889
Observations	71,589	71,492	71,493
Full R^2		0.2902	0.2908
Within R^2		0.0041	0.0050
Adjusted within R^2		−0.2372	−0.2367

Note: *p<0.1; **p<0.05; ***p<0.01.
Coefficients reported with se below. Hypothesis tests report p-values.

These findings support my claim that my dataset is a good proxy for that of Künn-Nelen. Further support comes from the results of regression models for five of the remaining seven outcomes considered in her paper (the two other outcomes were not evaluated due to data limitations). The results of applying these regression models to my data are very similar to the results presented by Künn-Nelen (2016). These results are reported in Appendix 4.D.

As well as verifying that my data is a good proxy for Künn-Nelen's data, by running these regression models I am able to examine estimates which were not reported by Künn-Nelen, in particular the estimated age effects and R^2 values. The most striking result from the replication is the very low value of the within R^2. The within R^2 captures the proportion of within-individual variation that is explained by the model, i.e. the R^2 net of the individual fixed effects ω_i. The values of the within R^2 in Table 4.2 are low; for reference, Cameron & Trivedi (2005) consider a within R^2 of .015 as low in their §21.3. The low R^2 indicates that this model does a poor job of explaining within-individual variation in the probability of a hospital in-patient stay. This suggests that important determinants of hospitalization have been left out of the model.

The low value of the within R^2 increases the potential benefit from investigating the treatment of controls in the model. If omitted determinants which have sizeable effects are even slightly correlated with commute time, they may bias its estimated effect. Although Künn-Nelen is primarily interested in the effect of commute time on hospital in-patient stays and so low explanatory power of the model may not seem to be a problem, in fact it is a cause for concern because of the potential for omitted variable bias.

4.4.4 Results from the APC acceleration model

Having shown in §4.4.2 and §4.4.3 that my data is a good proxy for that of Künn-Nelen (2016), I proceed to use it in conjunction with the APC panel regression framework to evaluate the treatment of control variables by Künn-Nelen. All analysis in this section is performed using the software described in §5. First, I show that the quadratic restriction on the age effect is inappropriate for the data, by testing the hypothesis that all age accelerations are equal. I plot detrended sums

of age accelerations, as described in §2.3.4.2, in an effort to identify the reasons for which the quadratic restriction on the age effect is inappropriate. This suggests that an important control, childbirth, should be added to the model. Adding childbirth to the model has the following three effects: the need for non-linearities in age is eliminated, the within-R^2 is increased, and the uncertainty around the effect of commute time is resolved in favour of insignificance. This application illustrates that the APC acceleration framework developed in this paper can be used as a diagnostic to identify improvements in panel data models.

4.4.4.1 Test: original model vs acceleration model

I re-estimate the model for hospital in-patient stays using the ξ parametrization of APC effects described in §4.2.2, and find that the estimated age accelerations are inconsistent with a quadratic age effect. I use the fixed effects setting, rather than the random effects or pooled OLS settings, to align with Künn-Nelen and the replication of her results in §4.4.3. Künn-Nelen chooses the fixed effects setting to allow for correlation between the unobserved component, ω_i, and commute time, for instance if ω_i captures underlying health conditions. The most general form of the ξ parametrization of APC effects in the fixed effects setting, the FAP model, nests the quadratic age model. The quadratic functional form can therefore be tested as a restriction of the FAP model. The FAP model is

$$\begin{aligned} H_{ip} =& \delta + \phi_1 CT_{ip} + \phi_2 CT_{ip}^2 + (a_{ip} - U)v_a + x_{ip}^{\mathbb{A}\prime}\xi^{\mathbb{A}} \\ &+ x_{ip}^{\mathbb{P}\prime}\xi^{\mathbb{P}} + z_{ip}\zeta + \omega_i + \epsilon_{ip}. \end{aligned} \tag{4.16}$$

Relative to Künn-Nelen's model (4.15), the age quadratic and period fixed effects are replaced by a single slope combining the linear effects of age and period, plus accelerations in age and period. This model has the same generality in period as Künn-Nelen's model, but is more general in age.

A process of testing whether sub-models of the FAP model are sufficient to describe the data reveals that significant age non-linearities are present in the data. In Table 4.2, the row labelled "APC model" records which of the sub-models described in §4.3.3.1 is selected by the testing procedure outlined in that section. The selected "FA" model indicates that non-linearities are present in the

age dimension but not the period dimension. Note that the lack of non-linearities in the period dimension means that the period indicators included by Künn-Nelen in her original model are unnecessary.

A direct test of whether the identified age non-linearities can be modelled as quadratic is rejected. This test is implemented as an F-test for equality of the age accelerations. To see why, consider the definition of an age acceleration:

$$\Delta^2\alpha_s = \alpha_s - \alpha_{s-1} - (\alpha_{s-1} - \alpha_{s-2}) \tag{4.17}$$

If the effect of age is quadratic, then

$$\alpha_s = \lambda_1 s + \lambda_2 s^2. \tag{4.18}$$

Substituting this into the expression for $\Delta^2\alpha_s$, it can be shown that

$$\Delta^2\alpha_s = \lambda_2. \tag{4.19}$$

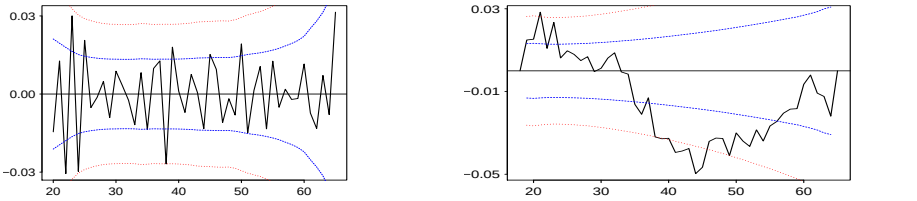
Under the null that the age effect is quadratic, all age accelerations equal the constant λ_2. The p-value associated with the F-test of this joint hypothesis is seen in the line labelled "H_0: age quadratic". The low p-value (0.016) indicates rejection of the quadratic restriction. Given the large sample size, even small deviations from a quadratic may produce a statistically significant rejection; therefore to evaluate the seriousness of the deviation, a visual representation of the relationship between age and hospital in-patient stays implied by the accelerations must be considered.

4.4.4.2 The shape of the age effect

By examining a visual representation of the shape of the relationship between age and hospital in-patient stays, I find that the relationship is not well-described by a quadratic. I use detrended sums of age accelerations to construct this representation. The representation reveals an inflection of hospital in-patient stays at midlife, which would require at least a third-order polynomial to be adequately described. A quadratic is not sufficient.

Detrended sums of age accelerations can be used to visualize the non-linear part of the shape of the relationship between age and hospital in-patient stays. Recall from §2.3 that the design matrix used to estimate APC accelerations from

Figure 4.2: Age accelerations, FA model for inpatient stays



(a) Age accelerations (b) Detrended sum of age accelerations

Blue dashed, red dotted lines = 1, 2 standard deviations from zero.

regression is constructed by summing accelerations relative to a linear plane chosen by the researcher. The summed accelerations reflect deviations of the relationship from that linear plane. Recall also from §2.3.4.2 that the linear plane used for estimation is not ideal for visualisation. An alternative linear plane and sum of accelerations, constructed as a bijective mapping of those used for estimation, is preferred for visualisation.

To visually isolate the non-linear part of the relationship, the accelerations are summed in such a way that the plots of the summed age accelerations are anchored to begin and end in zero. The linear plane which achieves this effectively detrends the sums of APC accelerations. Details of this plane and summation of accelerations, as well as how they can be constructed via a bijective mapping from the original $x'_{ip}\xi$ representation, can be found in §2.3.4.2 and Nielsen (2015).

In Figure 4.2b, the non-linear part of the relationship between age and the probability of being a hospital in-patient has an elongated Z shape. This is clearly not a quadratic relationship. Instead, there is concavity up to the late 30s and convexity thereafter. The interpretation of this depends to some extent on the slope over which this non-linear shape appears. Remember that the slope of the age effect is unidentified. If the slope were positive, this shape would correspond to a gradual fall in the amount by which the probability of being a hospital in-patient increases with age, followed by an increase from the late 30s. If the slope were negative, the amount by which the probability of being an hospital in-patient falls with age would gradually increase, and level off from the late 30s. Either way, there is a transition in the late 30s from decreasing acceleration in the probability of being an in-patient to increasing acceleration.

4.4.4.3 A new control to explain the age effect: childbirth

I find evidence that the Z-shaped relationship between age and hospital in-patient stays is due to childbirth, and that once this is accounted for there is no significant non-linearity in the relationship between age and hospital in-patient stays. Preliminary evidence comes from examining the estimated age accelerations when the data is separated by sex, as the non-linearities in age are only significant among women. I then construct an additional control variable from the BHPS data, reflecting whether hospital in-patient stays were for childbirth, and find that this control is significant and has a large effect size. Once the childbirth control is included there are no significant age non-linearities in either the model for women alone, or the model for men and women together.

The fact that the Z-shape has its inflection point around the mid 30s suggests that childbirth may be important to explaining the non-linearity in age. If childbirth is the main explanation, one would expect acceleration in the probability of being a hospital inpatient only up to the beginning of the peak childbearing years (late twenties and early thirties), with negative acceleration through the childbearing years and thereafter. Eventually, acceleration would return due to biological deterioration with age. This is consistent with the shape in Figure 4.2b.

The hypothesis that the Z-shape in the probability of being a hospital in-patient is due to childbirth is supported by the fact that the age non-linearities observed in Figure 4.2b are entirely attributable to women. This is seen from a supplementary analysis which separates the sample by sex, documented in Table 4.3 and Figure 4.3. The first two columns of Table 4.3 simply replicate the Künn-Nelen quadratic age model for men and women separately. Columns 3 and 4 are of more interest. They reveal that the selected submodel of the FAP model for men is "Ft", i.e. a model with no accelerations in period or in age. However, the selected model for women is "FAP", with accelerations in both period and age. The hypothesis that the age accelerations present among women are quadratic is rejected. Further, an examination of the shape of the detrended sum of age accelerations for women in Figure 4.3b shows that they have almost the same shape as those for the combined population of men and women, from Figure 4.2b. Since it is overwhelmingly women

who give birth, the results when split by sex support the idea that the inflection in hospital in-patient stays is due to childbirth.

Note that to estimate the model for men and women separately it was necessary to reduce the maximum age from 65 to 60, to ensure sufficient observations in each age-cohort cell. A supplementary analysis available in Appendix 4.E verifies that the reduction of the age range alone had almost no effect on the analysis.

Table 4.3: Regression results for in-patient stay, split by sex

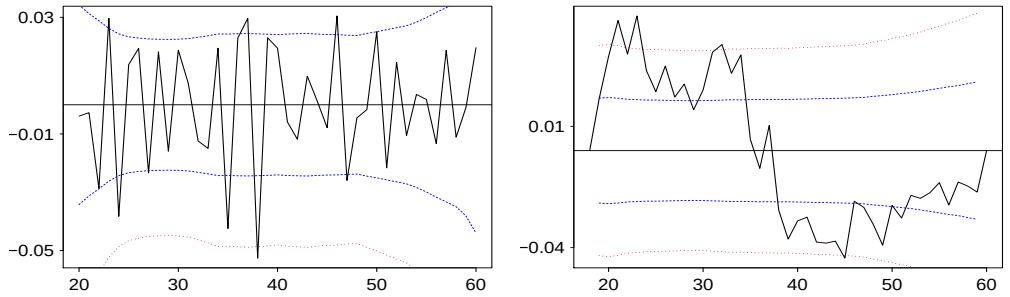
	Age quadratic		APC accelerations	
	Men 18-60	Women 18-60	Men 18-60	Women 18-60
CT	0.0004	0.001*	0.0004	0.001*
	(0.0003)	(0.0004)	(0.0003)	(0.0004)
CT^2	−0.00000	−0.00001	−0.00000	−0.00001
	(0.00000)	(0.00001)	(0.00000)	(0.00001)
Age	−0.007***	−0.006**		
	(0.002)	(0.003)		
Age^2	0.0001***	0.0001**		
	(0.00002)	(0.00003)		
APC model			Ft	FAP
H_0: age quadratic				0.0012
H_0: $\phi_1 = \phi_2 = 0$	0.2896	0.2516	0.3011	0.2503
Individuals	7,271	6,435	7,271	6,435
Observations	39,474	30,538	39,474	30,538
Full R^2	0.2771	0.3068	0.2764	0.3089
Within R^2	0.002	0.020	0.001	0.023
Adjusted within R^2	−0.225	−0.245	−0.226	−0.243

Note: *p<0.1; **p<0.05; ***p<0.01.
Coefficients reported with se below. Hypothesis tests report p-values.

The strongest evidence for the importance of childbirth in explaining hospital in-patient stays comes from Table 4.4, in which an indicator for childbirth is added to the model. In the first column of the table, we see that including an indicator for whether the hospital in-patient stays were due to childbirth increases the within R^2 of the model for women aged 18-60 from 0.023 to 0.304. It changes the selected FAP sub-model model from "FAP" to "Ft".

Although the evidence at this point indicates that men and women should be

Figure 4.3: Age accelerations, FAP model for inpatient stays, women



(a) Age accelerations (b) Detrended sum of age accelerations

Blue dashed, red dotted lines = 1, 2 standard deviations from zero.

modelled separately, in order to verify that results from the inclusion of childbirth are not due to the reduced age range or the smaller sample size, I consider the combined dataset. Column 2 of Table 4.4 shows the model for men and women together, aged 18-65, with childbirth. Childbirth is significant and has a large effect size. Comparing this model to the model for the same dataset without childbirth, in column 3 of Table 4.2, we see that the model fit has improved and all age accelerations have been eliminated. The within-R^2 has increased from 0.005 to 0.166. The FAP sub-model selected has been reduced from "FA" to "Ft".

4.4.4.4 The effect of commuting in the new model

Two improvements to the analysis were identified by careful consideration of the age accelerations estimated using the APC framework: separate analysis of men and women, and the inclusion of childbirth. These improvements resolved the ambiguity around the effect of commuting on the probability of a hospital in-patient stay in favour of its insignificance. First, once the data is separated by sex it is clear that there is no relationship between commute time and the probability of being a hospital in-patient among men. This is seen in columns 1 and 3 of Table 4.3, which report results for men alone: the coefficients on commute time and its square are not significant, and the p-value of the joint test is high at 0.30, indicating that the joint hypothesis that both coefficients are zero is not rejected. Second, considering the model for women, the coefficient on commute time alone is still significant before childbirth is accounted for (columns 2 and 4 of Table 4.3); but

Table 4.4: Regression results for in-patient stay, with childbirth indicator

	Women 18-60	Joint 18-65	Joint 18-65
CT	0.0003	0.0003	0.0003
	(0.0004)	(0.0002)	(0.0002)
CT^2	−0.00000	−0.00000	−4.20e-06
	(0.00000)	(0.00000)	(2.64e-06)
Age-period slope	−0.0003	−0.0002	−0.0002
	(0.0005)	(0.0003)	(0.0003)
Childbirth	0.923***	0.926***	0.926***
	(0.009)	(0.009)	(0.0087)
APC model	Ft	Ft	Ft
Controls:			
Nkids	Y	Y	
Education	Y	Y	Y
Marital status	Y	Y	
Government Office Region	Y	Y	
Relationship to household head	Y	Y	
H_0: $\phi_1 = \phi_2 = 0$	0.7812	0.3048	0.2806
Observations	30,538	71,492	71,492
Full R^2	0.5079	0.4054	0.4051
Within R^2	0.304	0.166	0.165
Adjusted within R^2	0.117	−0.036	−0.0362

Note: *p<0.1; **p<0.05; ***p<0.01.
Coefficients reported with se below. Hypothesis tests report p-values.

once childbirth is accounted for this becomes insignificant (column 1 of Table 4.4). Finally, considering the models for men and women together, including a childbirth indicator is sufficient to both provide the necessary sex differentiation (since most of those giving birth are women) and account for the impact of childbirth among women. Comparing column 3 of Table 4.2 and column 2 of Table 4.4, we see that adding the childbirth indicator has eliminated the significance of commute time and its square and has increased the p-value of the joint significance test from 0.15 to 0.30. Adding the childbirth indicator improves the model in a way that strengthens the finding of Künn-Nelen, i.e. that commute time and hospital in-patient stays are unrelated.

It is interesting to note that introducing the childbirth indicator, which was found to be the cause of the non-quadratic age non-linearity, improved the model fit and clarified the findings in a way that simply allowing for a more general model of the age non-linearity did not. This can be seen by comparing the models in columns 2 and 3 of Table 4.2; moving from the quadratic age effect in column 2 to the more general "FA" model in column 3 barely changes the R^2 values or the results regarding commute time. This highlights the importance of the additional step of understanding and interpreting the estimated age accelerations, in this case finding that they are driven by childbirth.

Further efforts to improve the model by omitting insignificant controls were found to have little impact. This can be seen by comparing columns 2 and 3 of Table 4.4. Column 3 contains estimates from a "final" model for hospital in-patient stays, including the crucial childbirth control and omitting the controls included by Künn-Nelen that were found to be insignificant. Künn-Nelen included controls for region, marital status, relationship to household head, number of children, and educational qualification, as well as the quadratic in age and indicators for each period. Of all of this, only a single slope in age/period and the education indicators were retained. However, the results of this model are nearly identical to the model which retains these controls, in column 2.

These results suggest that there is benefit to spending time searching for additional controls, and that there is little cost to including many insignificant controls. Both of these points strengthen the case for the use of the ξ parametrization of APC effects in the analysis of controls. First, the detrended sums of accelerations

can provide visual clues regarding additional controls which should be included, particularly since they allow the non-linear effects of age, period, and cohort to be isolated both from each other and from their linear parts. Second, the fact that including many irrelevant controls does not damage the model allays concerns about the large number of parameters involved in the ξ parametrization. The ξ parametrization has many parameters because it is flexible and non-parametric; this allows it to avoid any potential omitted variable bias arising from imposing an inaccurate functional form on any of age, period, or cohort. In a large panel dataset such as the BHPS, there is no reason to not use a flexible non-parametric model, as the comparison of columns 2 and 3 of Table 4.4 demonstrates that there is little downside to including many potentially redundant controls.

4.5 Conclusion

This paper develops a framework for separate identification of some of the effects of age, period, and cohort in panel data regressions. I rely on the APC acceleration parametrization of Kuang et al. (2008), which represents the age, period, and cohort (APC) effects as a combined linear plane and separable age, period, and cohort accelerations. In the pooled OLS and random effects panel settings, accelerations in all three of age, period, and cohort are separately identified. In the fixed effects setting only age and period accelerations are identified; the cohort effects are entirely absorbed into the individual fixed effects.

I accounted for the impact of the APC acceleration parametrization on the choice between the three different panel settings. I showed that it is only necessary to consider the possible correlation between covariates in the model and unobservable components, as well as the restrictions imposed on the within-individual correlation of unobservable components and the implications for efficiency and identification. There is no need to consider the possibility of correlation between the APC explanatory variables and unobservable components, because that is controlled due to the deterministic nature of the APC variables.

In the application, I illustrate how accelerations in age can be identified in a fixed effects panel data model and used as a diagnostic tool to evaluate the treatment of controls. The specific setting is a study of the impact of commute time

on hospital in-patient stays using UK data, where age is an important control that is assumed to be quadratic. By testing the hypothesis that all age accelerations are equal, I show that the quadratic functional form restriction is invalid. Closer inspection of the shape of the relationship between age and the probability of being a hospital in-patient allows me to identify an improvement to the model: the addition of a control for hospital stays due to childbirth. After including the childbirth control, I find that the coefficients on commute time and its square change from being significant at the 10% level to insignificant. The model R^2 is improved and the age non-linearities are eliminated.

Given the relative ease of testing such functional form restrictions with the APC acceleration framework developed in this paper, and the potential to identify improvements to the model, it is hoped that this will become a standard part of any panel data analysis where age (or period or cohort) is used as a control. Tests for any age, period, or cohort non-linearities in the explanatory variable could also easily be performed using the tools developed in this paper.

Future work should pursue two directions: applying the tools developed in this paper in contexts other than hospitalization, and expanding the tools beyond what is presented in this paper. In the first direction, further use cases that should be demonstrated include policy evaluation and forecasting. In the second direction, methodological extensions should include: testing for the absence of interactions between age, period, and cohort as has already been done for repeated cross section data (see §3.4); developing a modelling framework that can incorporate such interactions; and allowing for non-continuous outcome variables. Further work could also explore the interaction of the age, period, and cohort accelerations with other explanatory variables. It would also be of interest to consider the appropriate treatment of APC controls in dynamic panel settings. A major limitation of the present framework is the requirement for a contiguous dataset; it is likely that this is a sufficient but not necessary condition for identification of accelerations, and further theoretical work should be done to clarify this point.

4.A APC acceleration model design vector

The design vectors $x_{ip}^{\mathbb{A}}$, $x_{ip}^{\mathbb{P}}$, and $x_{ip}^{\mathbb{C}}$, are defined as follows, with $m(r,s) = max(r - s + 1, 0)$:

$$x_{ip}^{\mathbb{A}} = m(1, a_{ip}), \ldots, m(U-1, a_{ip}), m(a_{ip}, U+2), \ldots, m(a_{ip}, A), \tag{4.20}$$

$$x_{ip}^{\mathbb{P}} = \begin{cases} \mathbb{1}(p_{ip} = 2U-2), m(p_{ip}, 2U+1), \ldots, m(p_{ip}, 2U-3+P) & \text{for } L \text{ odd} \\ m(p_{ip}, 2U+1), \ldots, m(p_{ip}, 2U-2+P) & \text{for } L \text{ even} \end{cases} \tag{4.21}$$

$$x_{ip}^{\mathbb{C}} = m(1, c_{ip}), \ldots, m(U-1, c_{ip}), m(c_{ip}, U+2), \ldots, m(c_{ip}, C) \tag{4.22}$$

See also equations (2.25) through (2.27) in §2.3.3.4.

4.B Linear dependence among APC parameters under fixed effects

It is required to prove the absence of linear dependence of the APC variables in the fixed effects FAP model, $\{(\tilde{a}_{ip} - U), \tilde{x}_{ip}^{\mathbb{A}\prime}, \tilde{x}_{ip}^{\mathbb{P}\prime}\}$. Equivalently, it must be proved that that the parameters $\{v_a, \xi^{\mathbb{A}}, \xi^{\mathbb{P}}\}$ are separately identified from time-demeaned data. I first provide a proof that the parameters are identified for data with a period-cohort structure, i.e. where each cohort is observed for a fixed number of periods (e.g. individuals born 1990-1996 observed from 2000 through 2010). I then explain that the same approach works for data with either an age-cohort structure, where each cohort is observed for a fixed number of ages, or an age-period structure, where a certain age range is observed for a number of periods (the different data structures are shown in Figure 2.1 of §2.3.2). At present I omit covariates $\tilde{z}_{1,ip}$ and assume a balanced panel for ease of exposition; I will relax these assumptions at the end of the appendix.

The core idea of the proof is that the parameters are identified if a change in the parameter vector changes the linear predictors of the FAP model. Letting $\tilde{\mu}_{ip}$ represent the linear predictor of the FAP model, the requirement is to show that if $\{v_a, \xi^{\mathbb{A}}, \xi^{\mathbb{P}}\}^{\flat} \neq \{v_a, \xi^{\mathbb{A}}, \xi^{\mathbb{P}}\}^{\natural}$, then $\tilde{\mu}_{ip}^{\flat} \neq \tilde{\mu}_{ip}^{\natural}$. I do this by showing that each parameter is equal to a combination of either two or four linear predictors, so that if the parameter changes, one of the four must change. This approach is

adapted from the proof of Theorem 1 of Kuang et al. (2008). However, the within-individual demeaning in the FAP model introduces additional complexity relative to that paper, which I address.

Consider a period-cohort dataset, with a total number of cohorts C, total number of periods P, and thus total number of ages $A = P + C - 1$. Let the cells of this dataset be indexed by cohort and period, so that the cell containing observations on individuals in cohort 4 at the first period in the data is indexed by $\{4, L+1\}$, recalling that L is a data-determined offset. The parameter L can be calculated from A, P, C as outlined in §4.2.2. For now, assume L is even, so $U = (L+2)/2$. To keep things simple, assume a balanced panel, so each individual is observed in each period, and assume there are no covariates in the model.

By defining and comparing the linear predictors associated with different cells in the dataset, I show that the parameters $\{v_a, \xi^{\mathbb{A}}, \xi^{\mathbb{P}}\}$ are separately identified from this dataset. The FAP model for this dataset is

$$\begin{aligned} y_{ip} =& \mu_{ip} + \omega_i + \epsilon_{ip} && (4.23)\\ =& (a_{ip} - U)v_a + x_{ip}^{\mathbb{A}\prime}\xi^{\mathbb{A}} + x_{ip}^{\mathbb{P}\prime}\xi^{\mathbb{P}} + \omega_i + \epsilon_{ip}, && (4.24) \end{aligned}$$

and the within-individual demeaned version used for estimation is

$$\begin{aligned} \tilde{y}_{ip} =& \tilde{\mu}_{ip} + \tilde{\epsilon}_{ip} && (4.25)\\ =& (\tilde{a}_{ip} - U)v_a + \tilde{x}_{ip}^{\mathbb{A}\prime}\xi^{\mathbb{A}} + \tilde{x}_{ip}^{\mathbb{P}\prime}\xi^{\mathbb{P}} + \tilde{\epsilon}_{ip}, && (4.26) \end{aligned}$$

with $\tilde{w}_{ip} = w_{ip} - P^{-1}\sum_p w_{ip}$ for any variable w. The subtraction of $\bar{\mu}_{ip} = P^{-1}\sum_p \mu_{ip}$ in the construction of $\tilde{\mu}_{ip}$ means that all age or period elements of the parameter vector ξ that appear in any period's μ_{ip} for a given i will appear in every period's $\tilde{\mu}_{ip}$ for that i. This introduces additional complexity relative to Kuang et al. (2008), which deals with the identification of the parameter vector ξ from the original linear predictors μ_{ip}. I account for this additional complexity.

Since the panel is balanced and only age, period, and cohort appear as explanatory variables, the within-individual mean, $\bar{w}_{ip} = P^{-1}\sum_p w_{ip}$, of any explanatory variable is common to all members of the same cohort. This includes the within-individual mean of the linear predictor, $\bar{\mu}_{ip}$. Therefore the linear predictor, $\tilde{\mu}_{ip}$, is

common to all observations with the same period-cohort combination. I write $\tilde{\mu}_{c,p}$ for the linear predictor for any individual of cohort c in period p.

I first show that v_a is identified. Let the linear predictor in the cell with $c = U, p = L + 2$, i.e. cell $\{U, L + 2\}$, be given by

$$\tilde{\mu}_{U,L+2} = v_o + v_a - \bar{\mu}_U \tag{4.27}$$

Here the notation $\bar{\mu}_U$ refers to the within-individual mean of the APC linear predictor $\mu_{U,L+2}$; it is indexed by U because it is the same for all individuals within cohort U, since the dataset is a balanced panel. Similarly, the time-demeaned APC linear predictor in the cell $\{U, L + 1\}$, is given by

$$\tilde{\mu}_{U,L+1} = v_o - \bar{\mu}_U \tag{4.28}$$

The difference between these two linear predictors is $\tilde{\mu}_{U,L+2} - \tilde{\mu}_{U,L+1} = v_a$. By choosing two linear predictors from the same cohort, I am able to cancel the two $\bar{\mu}_U$ when I take the difference. This resolves the additional complexity of dealing with identification from the time-demeaned model in terms of $\tilde{\mu}$ rather than the original model in terms of μ, which I highlighted earlier. The within-individual comparisons between these two cells allows us to identify the first parameter, v_a.

I next consider identification of the age acceleration $\Delta^2\alpha_{U+2}$. The difference between the linear predictors for cells $\{U-1, L+2\}$ and $\{U-1, L+1\}$ is $\tilde{\mu}_{U-1,L+2} - \tilde{\mu}_{U-1,L+1} = v_a + \Delta^2\alpha_{U+2}$. Recalling how v_a was identified above, we can use a double-difference to identify $\Delta^2\alpha_{U+2}$ as follows:

$$(\tilde{\mu}_{U-1,L+2} - \tilde{\mu}_{U-1,L+1}) - (\tilde{\mu}_{U,L+2} - \tilde{\mu}_{U,L+1}) = \Delta^2\alpha_{U+2}. \tag{4.29}$$

Proceeding in a logical sequence, and using the same double-differencing approach, all period accelerations, and all age accelerations from $U + 2$ through $A - U + 2$, can be identified using observations on cohorts U and $U - 1$. To identify the age accelerations from $A - U + 3$ through A, observations for all cohorts preceding cohort $U - 1$ in periods $L + P$ and $L + P - 1$ are required. To identify the age accelerations from 3 through $U + 1$, observations for all cohorts succeeding U in periods $L + 2$ and $L + 1$ are required. This is seen from Table 4.5.

The proof for the case where L is odd works in exactly the same way. The main difference is in the set of cohorts that are used for identification. Observations on

Table 4.5: FAP model parameter identification, period-cohort data

Parameter	Linear predictor combination
Slope; period accelerations; age accelerations $U+2$ through $A-U+2$	
v_a	$(\tilde{\mu}_{U,L+2} - \tilde{\mu}_{U,L+1})^{\dagger}$
$\Delta^2\alpha_{U+2}$	$(\tilde{\mu}_{U-1,L+2} - \tilde{\mu}_{U-1,L+1}) - (\tilde{\mu}_{U,L+2} - \tilde{\mu}_{U,L+1})$
$\Delta^2\beta_{L+3}$	$(\tilde{\mu}_{U,L+3} - \tilde{\mu}_{U,L+2}) - (\tilde{\mu}_{U-1,L+2} - \tilde{\mu}_{U-1,L+1})$
$\Delta^2\alpha_{U+3}$	$(\tilde{\mu}_{U-1,L+3} - \tilde{\mu}_{U-1,L+2}) - (\tilde{\mu}_{U,L+3} - \tilde{\mu}_{U,L+2})$
$\Delta^2\beta_{L+4}$	$(\tilde{\mu}_{U,L+4} - \tilde{\mu}_{U,L+3}) - (\tilde{\mu}_{U-1,L+3} - \tilde{\mu}_{U-1,L+2})$
$\Delta^2\alpha_{U+4}$	$(\tilde{\mu}_{U-1,L+4} - \tilde{\mu}_{U-1,L+3}) - (\tilde{\mu}_{U,L+4} - \tilde{\mu}_{U,L+3})$
$\Delta^2\beta_{L+5}$	$(\tilde{\mu}_{U,L+5} - \tilde{\mu}_{U,L+4}) - (\tilde{\mu}_{U-1,L+4} - \tilde{\mu}_{U-1,L+3})$
etc.	$\vdots$
$\Delta^2\beta_{L+P}$	$(\tilde{\mu}_{U,L+P} - \tilde{\mu}_{U,L+P-1}) - (\tilde{\mu}_{U-1,L+P-1} - \tilde{\mu}_{U-1,L+P-2})$
$\Delta^2\alpha_{A-U+2}$	$(\tilde{\mu}_{U-1,L+P} - \tilde{\mu}_{U-1,L+P-1}) - (\tilde{\mu}_{U,L+P} - \tilde{\mu}_{U,L+P-1})$
Age accelerations $A-U+3$ through A	
$\Delta^2\alpha_{A-U+3}$	$(\tilde{\mu}_{U-2,L+P} - \tilde{\mu}_{U-2,L+P-1}) - (\tilde{\mu}_{U-1,L+P} - \tilde{\mu}_{U-1,L+P-1})$
$\Delta^2\alpha_{A-U}$	$(\tilde{\mu}_{U-3,L+P} - \tilde{\mu}_{U-3,L+P-1}) - (\tilde{\mu}_{U-2,L+P} - \tilde{\mu}_{U-2,L+P-1})$
etc.	$\vdots$
$\Delta^2\alpha_A$	$(\tilde{\mu}_{2,L+P} - \tilde{\mu}_{2,L+P-1}) - (\tilde{\mu}_{1,L+P} - \tilde{\mu}_{1,L+P-1})$
Age accelerations $U+1$ through 3	
$\Delta^2\alpha_{U+1}$	$(\tilde{\mu}_{U+1,L+2} - \tilde{\mu}_{U+1,L+1}) - (\tilde{\mu}_{U,L+2} - \tilde{\mu}_{U,L+1})$
$\Delta^2\alpha_U$	$(\tilde{\mu}_{U+2,L+2} - \tilde{\mu}_{U+2,L+1}) - (\tilde{\mu}_{U+1,L+2} - \tilde{\mu}_{U+1,L+1})$
etc.	$\vdots$
$\Delta^2\alpha_3$	$(\tilde{\mu}_{C,L+2} - \tilde{\mu}_{C,L+1}) - (\tilde{\mu}_{C-1,L+2} - \tilde{\mu}_{C-1,L+1})$

† only a single difference is required to identify v_a
Note that $\tilde{\mu}$ are indexed by first cohort, then period

cohorts $U-1$ and $U-2$ are sufficient to identify v_a, all period accelerations, and all accelerations from $U+2$ through $A-U+3$. To identify the age double-differences from $A-U+4$ through A, observations for all cohorts preceding cohort $U-2$ from periods $L+P$ and $L+P-1$ will be required. To identify the age double differences from 3 through $U+1$, observations for all cohorts succeeding $U-1$ from periods $L+2$ and $L+1$ will be required.

Similar proofs to that outlined above exist for the two other data structures: age-cohort and age-period data. For age-cohort data, observations on cohorts 1 and 2 at all ages are sufficient to identify v_a, all age accelerations, and period accelerations from 3 through $A+1$. Observations at ages A and $A-1$ for all cohorts are needed to identify the remaining period accelerations. The requisite combinations of linear predictors are given in Table 4.6. For age-period data, the same idea holds. Observations on two cohorts are sufficient to identify the slope, plus age and period accelerations within a certain range. Outside of that range, additional observations are needed from the first two periods and the last two ages in order to identify the remaining acceleration parameters. The requisite combination of linear predictors are given in Table 4.7.

The assumption of a balanced panel can be relaxed. It is sufficient for identification that for each pair of adjacent ages and periods, there is at least one individual observed in both cells of the pair. This only applies to the pairs of cells used to construct the differences in Tables 4.5 through 4.7.

Covariates can also be introduced to the model. Where there are covariates, those covariates must satisfy the condition in §3 for identification of covariates in an APC model: they must not be linear functions of x_{ip}. Additionally, the number of individuals observed in each pair of cells will need to increase in order to identify the additional parameters. Most panel datasets are large so this requirement should be easily satisfied. Further, the set of cohorts and periods on which observations are required by the tables above is a minimal sufficient set; where further cells are observed, this will increase the possible pairs from which each parameter could be identified and thus introduce more degrees of freedom for the identification of covariate coefficients.

Table 4.6: FAP model parameter identification, age-cohort data

Parameter	Linear predictor combination
Slope; age accelerations; period accelerations 3 *through* $A+1$	
v_a	$(\tilde{\mu}_{2,1} - \tilde{\mu}_{1,1})^{\dagger}$
$\Delta^2\beta_3$	$(\tilde{\mu}_{2,2} - \tilde{\mu}_{1,2}) - (\tilde{\mu}_{2,1} - \tilde{\mu}_{1,1})$
$\Delta^2\alpha_3$	$(\tilde{\mu}_{3,1} - \tilde{\mu}_{2,1}) - (\tilde{\mu}_{2,2} - \tilde{\mu}_{1,2})$
$\Delta^2\beta_4$	$(\tilde{\mu}_{3,2} - \tilde{\mu}_{2,2}) - (\tilde{\mu}_{3,1} - \tilde{\mu}_{2,1})$
$\Delta^2\alpha_3$	$(\tilde{\mu}_{4,1} - \tilde{\mu}_{3,1}) - (\tilde{\mu}_{3,2} - \tilde{\mu}_{2,2})$
$\Delta^2\beta_5$	$(\tilde{\mu}_{4,2} - \tilde{\mu}_{3,2}) - (\tilde{\mu}_{4,1} - \tilde{\mu}_{3,1})$
etc.	$\vdots$
$\Delta^2\alpha_A$	$(\tilde{\mu}_{A,1} - \tilde{\mu}_{A-1,1}) - (\tilde{\mu}_{A-1,2} - \tilde{\mu}_{A-2,2})$
$\Delta^2\beta_{A+1}$	$(\tilde{\mu}_{A,2} - \tilde{\mu}_{A-1,2}) - (\tilde{\mu}_{A,1} - \tilde{\mu}_{A-1,1})$
Period accelerations $A+2$ *through* P	
$\Delta^2\beta_{A+2}$	$(\tilde{\mu}_{A,3} - \tilde{\mu}_{A-1,3}) - (\tilde{\mu}_{A,2} - \tilde{\mu}_{A-1,2})$
$\Delta^2\beta_{A+3}$	$(\tilde{\mu}_{A,4} - \tilde{\mu}_{A-1,4}) - (\tilde{\mu}_{A,3} - \tilde{\mu}_{A-1,3})$
etc.	$\vdots$
$\Delta^2\beta_P$	$(\tilde{\mu}_{A,C} - \tilde{\mu}_{A-1,C}) - (\tilde{\mu}_{A,C-1} - \tilde{\mu}_{A-1,C-1})$

† only a single difference is required to identify v_a
Note that $\tilde{\mu}$ are indexed by first age, then cohort

Table 4.7: FAP model parameter identification, age-period data

Parameter	Linear predictor combination
Slope; age accelerations $U+2$ through A;	
period accelerations $L+3$ through $U+A-2$	
v_a	$\tilde{\mu}_{U+1,L+2} - \tilde{\mu}^{\dagger}_{U,L+1}$
$\Delta^2\alpha_{U+2}$	$(\tilde{\mu}_{U+2,L+2} - \tilde{\mu}_{U+1,L+1}) - (\tilde{\mu}_{U+1,L+2} - \tilde{\mu}_{U,L+1})$
$\Delta^2\beta_{L+3}$	$(\tilde{\mu}_{U+2,L+3} - \tilde{\mu}_{U+2,L+2}) - (\tilde{\mu}_{U+2,L+2} - \tilde{\mu}_{U+1,L+1})$
$\Delta^2\alpha_{U+3}$	$(\tilde{\mu}_{U+3,L+3} - \tilde{\mu}_{U+2,L+2}) - (\tilde{\mu}_{U+2,L+3} - \tilde{\mu}_{U+2,L+2})$
$\Delta^2\beta_{L+4}$	$(\tilde{\mu}_{U+3,L+4} - \tilde{\mu}_{U+2,L+3}) - (\tilde{\mu}_{U+3,L+3} - \tilde{\mu}_{U+2,L+2})$
$\Delta^2\alpha_{U+4}$	$(\tilde{\mu}_{U+4,L+4} - \tilde{\mu}_{U+3,L+3}) - (\tilde{\mu}_{U+3,L+4} - \tilde{\mu}_{U+2,L+3})$
$\Delta^2\beta_{L+5}$	$(\tilde{\mu}_{U+4,L+5} - \tilde{\mu}_{U+3,L+4}) - (\tilde{\mu}_{U+4,L+4} - \tilde{\mu}_{U+3,L+3})$
etc.	$\vdots$
$\Delta^2\alpha_A$	$(\tilde{\mu}_{A,U+A-3} - \tilde{\mu}_{A-1,U+A-4}) - (\tilde{\mu}_{A-1,U+A-3} - \tilde{\mu}_{A-2,U+A-4})$
$\Delta^2\beta_{U+A-2}$	$(\tilde{\mu}_{A,U+A-2} - \tilde{\mu}_{A-1,U+A-3}) - (\tilde{\mu}_{A,U+A-3} - \tilde{\mu}_{A-1,U+A-4})$
Period accelerations $U+A-1$ through $L+P$	
$\Delta^2\beta_{U+A-1}$	$(\tilde{\mu}_{A,U+A-1} - \tilde{\mu}_{A-1,U+A-2}) - (\tilde{\mu}_{A,U+A-2} - \tilde{\mu}_{A-1,U+A-3})$
$\Delta^2\beta_{U+A}$	$(\tilde{\mu}_{A,U+A} - \tilde{\mu}_{A-1,U+A-1}) - (\tilde{\mu}_{A,U+A-1} - \tilde{\mu}_{A-1,U+A-2})$
etc.	$\vdots$
$\Delta^2\beta_{L+P}$	$(\tilde{\mu}_{A,L+P} - \tilde{\mu}_{A-1,L+P-1}) - (\tilde{\mu}_{A,L+P-1} - \tilde{\mu}_{A-1,L+P-2})$
Age accelerations $U+1$ through 3	
$\Delta^2\alpha_{U+1}$	$(\tilde{\mu}_{U+1,L+2} - \tilde{\mu}_{U,L+1}) - (\tilde{\mu}_{U,L+2} - \tilde{\mu}_{U-1,L+1})$
$\Delta^2\alpha_U$	$(\tilde{\mu}_{U,L+2} - \tilde{\mu}_{U-1,L+1}) - (\tilde{\mu}_{U-1,L+2} - \tilde{\mu}_{U-2,L+1})$
$\Delta^2\alpha_{U-1}$	$(\tilde{\mu}_{U-1,L+2} - \tilde{\mu}_{U-2,L+1}) - (\tilde{\mu}_{U-2,L+2} - \tilde{\mu}_{U-3,L+1})$
etc.	$\vdots$
$\Delta^2\alpha_3$	$(\tilde{\mu}_{3,L+2} - \tilde{\mu}_{2,L+1}) - (\tilde{\mu}_{2,L+2} - \tilde{\mu}_{1,L+1})$

† only a single difference is required to identify v_a
Note that $\tilde{\mu}$ are indexed by first age, then period

4.C A linear plane in age and period

It may be thought that a second APC slope could be identified under fixed effects assumptions if a different initial linear plane were chosen in the reparametrization, with slopes in age and period rather than slopes in age and cohort. The number of identifiable age, period, and cohort parameters in the fixed effects model might therefore be increased. Here, I show that this is not the case. Under fixed effects assumptions only one APC slope can be estimated, which combines the within-individual linear variation in age and period.

In the initial definition of the ξ parametrization in §2.3, the linear plane was defined in terms of a slope in the age dimension and a slope in the cohort dimension. Due to the exact relationship $period = age + cohort - 1$, the same linear plane could be defined in terms of a slope in the age dimension and a slope in the period dimension. In §2.3.3.4, the three slope terms in equation (2.21) were combined into two slopes in age and cohort in equation (2.22). While the reduction from three to two slopes is required by the exact relationship $period = age + cohort - 1$, the two slopes need not have been in age and cohort; they could have been in cohort and period, or period and age. In the fixed effects setting, we are particularly interested in what would happen if the three slopes in equation (2.21) were reduced to two slopes in age and period. Then (2.22) would be replaced by

$$[\Delta\alpha_{U+1} - \Delta\gamma_{U+1}][a_{ip} - U] + [\Delta\gamma_{U+1} + \Delta\beta_{2U}][p_{ip} - (2U-1)]. \tag{4.30}$$

The fixed effects model (4.4) would then become

$$\begin{aligned} y_{ip} =& \mu_{ip} + z'_{ip}\zeta + \omega_i + \epsilon_{ip} \\ \mu_{ip} =& x^{\dagger\prime}_{ip}\xi^{\dagger} \\ =& v_o + (a_{ip} - U)v^{\dagger}_a + (p_{ip} - (2U-1))v^{\dagger}_p \\ &+ x^{\mathbb{A}'}_{ip}\xi^{\mathbb{A}} + x^{\mathbb{P}'}_{ip}\xi^{\mathbb{P}} + x^{\mathbb{C}'}_{ip}\xi^{\mathbb{C}}, \end{aligned} \tag{4.31}$$

with

$$v^{\dagger}_a = \Delta\alpha_{U+1} - \Delta\gamma_{U+1} \quad ; \quad v^{\dagger}_p = \Delta\gamma_{U+1} + \Delta\beta_{2U} \tag{4.32}$$

(note that $v_o, \xi^{\mathbb{A}}, \xi^{\mathbb{P}}$, and $\xi^{\mathbb{C}}$ are unchanged).

The perceived benefit to defining the linear plane in terms of age and period, as in (4.31), rather than in terms of age and cohort is that it means both slopes are time-varying. In the original model (4.4), the slope in cohort is time-invariant. The parameters associated with time-varying variables can be identified under fixed effects assumptions, whereas the parameters associated with time-invariant variables cannot. The idea of moving to model (4.31), where both slopes are time-varying, is that it will be possible to identify the parameter associated with the second APC slope.

However, this alternative parametrization does not change the parameters that are identified under fixed effects. To understand this, consider the time-demeaned version of model (4.31), which is

$$\tilde{y}_{ip} = (\tilde{a}_{ip} - U)v_a^\dagger + [\tilde{p}_{ip} - (2U-1)]v_p^\dagger + \tilde{x}_{ip}^{\mathbb{A}\prime}\xi^{\mathbb{A}} + \tilde{x}_{ip}^{\mathbb{P}\prime}\xi^{\mathbb{P}} + \tilde{z}'_{1,ip}\zeta_1 + \tilde{\epsilon}_{ip}. \tag{4.33}$$

In this model, the two slope variables, $(\tilde{a}_{ip} - U)$ and $[\tilde{p}_{ip} - (2U-1)]$ are exactly collinear. This means that only one of the two slopes can be included in estimation and it will have a combined coefficient

$$v_a^\dagger + v_p^\dagger = \Delta\alpha_{U+1} - \Delta\gamma_{U+1} + \Delta\gamma_{U+1} + \Delta\beta_{2U} = \Delta\alpha_{U+1} + \Delta\beta_{2U}. \tag{4.34}$$

This is equal to the single slope coefficient v_a from the original model (4.4). Therefore, starting from an alternative parametrization of the slopes does not alter the set of APC parameters that can be estimated under fixed effects assumptions.

4.D Evaluating the control variables for other health outcomes

I analysed five of the seven other health outcomes considered by Künn-Nelen (2016), using the APC acceleration framework developed in this chapter to evaluate, and identify improvements to, the standard set of controls she uses. I did not consider the remaining two of the seven outcomes due to data limitations. My evaluation showed that the standard set of controls, in particular the assumption of a quadratic age effect, was inappropriate for two of the five outcomes considered. I identified improvements which increased the fit for one of the two outcomes, GP

Table 4.8: Other health outcomes summary statistics

Variable	mean	sd	min	max
Health satisfaction	5.21	1.35	1	7
Health status	4.02	0.82	1	5
Health problems	0.47	0.50	0	1
Sickness absence	0.02	0.13	0	1
Nr. GP visits (Kunn-Nelen variable)	1.14	1.04	0	4
Nr. GP visits (using band minimum values)	1.81	2.46	0	11

visits, but was not able to identify substantial improvements to the model for the other health outcome, health satisfaction. In neither case did the estimated effects of commute time on the health outcome change.

The five health outcomes considered here are defined as follows, with descriptive statistics available in Table 4.8. The first two outcomes are self-reported: health satisfaction, on a 7-point Likert scale, and health status, on a 5-point Likert scale. On both scales a higher number is better. The third outcome is an indicator for whether the person has any diagnosed health conditions. The fourth is an indicator for whether the person was off sick in the last week (Künn-Nelen reports this as off sick in the last year, but the BHPS only has a variable recording sickness absence in the last week, and it corresponds to Künn-Nelen's summary statistics). The final outcome is the number of GP visits in the year, which is banded and top-coded. Künn-Nelen uses two further outcomes: an indicator for whether the person exercises regularly, and body mass index. I do not consider these as they are recorded too infrequently to be suitable for the APC acceleration model.

I first verify that the data I have for these outcomes is a good proxy for Künn-Nelen's data. The descriptive statistics in Table 4.8 align with those in her Table 1. I reproduce her regressions for each outcome, as I did for hospital in-patient stays in §4.4.3. Recall that her model has a quadratic age effect:

$$y_{ip} = \phi_1 CT_{ip} + \phi_2 CT_{ip}^2 + \lambda_1 a_{ip} + \lambda_2 a_{ip}^2 + \beta_p + z'_{ip}\zeta + \omega_i + \epsilon_{ip}. \tag{4.35}$$

Here y_{ip} is the outcome variable, CT_{ip} is commute time, a_{ip} is age, β_p is a period fixed effect, and z_{ip} is a vector of other explanatory variables. I find that the estimates obtained using this model with my data are similar to Künn-Nelen's

original results, supporting the claim that my data is a good proxy for hers. The similarity can be seen by comparing the columns titled "KN" (for Künn-Nelen) and "Rep" (for replication) in Tables 4.9 through 4.13. The only notable difference is that the joint hypothesis that neither commute time nor its square affects the probability of having a diagnosed health problem is closer to significance in my data than in Künn-Nelen's; however, the hypothesis is still not rejected at the 5% significance level. The low values of the within R^2 for all models are worth noting; recall that these were not reported by Künn-Nelen. Note that there are fewer observations for health satisfaction because it was only recorded from 1996. Further, the regressions for health satisfaction include observations for 2001 which are constructed by simple mean interpolation of observations for the same individuals in 2000 and 2002. The interpolation was needed to estimate the APC model but does not affect the results of the quadratic model.

I apply the APC acceleration framework to my data to evaluate the use of standard controls in model (4.35) for each of the five outcomes, with a particular focus on age as a control. The model I estimate is the FAP model,

$$\begin{aligned} y_{ip} =& \phi_1 CT_{ip} + \phi_2 CT_{ip}^2 + (a_{ip} - U)v_a + x_{ip}^{\mathbb{A}'}\xi^{\mathbb{A}} \\ &+ x_{ip}^{\mathbb{P}'}\xi^{\mathbb{P}} + z_{ip}\zeta + \omega_i + \epsilon_{ip}. \end{aligned} \tag{4.36}$$

This model generalizes model (4.35). Results from this analysis are seen in the columns title "APC" in Tables 4.9 through 4.13. Age non-linearities are present in all models except the model for absence due to sickness; the lack of age non-linearities in this model is seen in Table 4.12 where the selected sub-model is "Ft", indicating that age accelerations are not needed (see §4.3.3.1).

For the four outcomes where age accelerations were needed, I tested the quadratic functional form restriction by testing for equality between all age accelerations. The result of testing this restriction is reported as "H_0: age quadratic in Tables 4.9 through 4.13. The null is that the quadratic restriction is acceptable. This quadratic restriction was not rejected for two outcomes (self-reported health status and health problems indicator) but was rejected for the other two outcomes (GP visits and self-reported health satisfaction).

Table 4.9: Regression results for health satisfaction

	Health satisfaction		
	KN	Rep.	APC
CT	-0.0043^{***}	-0.0038^{***}	-0.004^{***}
	(0.0013)	(0.0012)	(0.001)
CT^2	4.49e-05***	0.00004***	0.00004***
	(1.53e-05)	(1.39e-05)	(0.00001)
Age		0.0007	
		(0.0075)	
Age^2		-0.0002^{**}	
		(8.57e-05)	
APC model			FAP
H_0: age quadratic			0.0202
H_0: $\phi_1 = \phi_2 = 0$	0.0087	0.01	0.0068
Individuals	11,693	11,610	11,610
Observations	50,503	54,094	54,094
Within R^2		0.0195	0.021
Adjusted within R^2		−0.2499	−0.249

Note: *p<0.1; **p<0.05; ***p<0.01.
Coefficients reported with se below. Hypothesis tests report p-values.
"KN"= Künn-Nelen original findings; "Rep." = replication of Künn-Nelen; "APC" = APC panel regression.

Table 4.10: Regression results for health status

	Health status		
	KN	Rep.	APC
CT	-0.0019^{***}	-0.0015^{**}	-0.0015^{**}
	(0.0007)	(0.0006)	(0.0006)
CT^2	1.9e-05**	1.47e-05*	1.47e-05*
	(8.05e-06)	(7.58e-06)	(7.58e-06)
Age		0.0172^{***}	
		(0.0035)	
Age^2		-0.00043^{***}	
		(4.10e-05)	
APC model			FAP
H_0: age quadratic			0.7474
H_0: $\phi_1 = \phi_2 = 0$	0.0329	0.04	0.0418
Individuals	13,702	13,888	13,888
Observations	66,857	71,483	71,483
Within R^2		0.0353	0.0359
Adjusted within R^2		−0.1985	−0.1986

Note: *p<0.1; **p<0.05; ***p<0.01.
Coefficients reported with se below. Hypothesis tests report p-values. "KN"= Künn-Nelen original findings; "Rep." = replication of Künn-Nelen; "APC" = APC panel regression.

Table 4.11: Regression results for health problems

	Health problems		
	KN	Rep.	APC
CT	0.0008**	0.00077**	0.00076**
	(0.0004)	(0.00036)	(0.00036)
CT^2	−9.01e-06**	−8.32e-06*	−8.13e-06*
	(4.37e-06)	(4.25e-06)	(4.25e-06)
Age		−0.0071***	
		(0.00198)	
Age^2		0.0002***	
		(2.30e-05)	
APC model			FAP
H_0: age quadratic			0.6410
H_0: $\phi_1 = \phi_2 = 0$	0.1538	0.09	0.0999
Individuals	14,065	13,889	13,889
Observations	71,559	71,492	71,492
Within R^2		0.0168	0.0175
Adjusted within R^2		−0.2214	−0.2215

Note: *p<0.1; **p<0.05; ***p<0.01.
Coefficients reported with se below. Hypothesis tests report p-values. "KN"= Künn-Nelen original findings; "Rep." = replication of Künn-Nelen; "APC" = APC panel regression.

Table 4.12: Regression results for sickness absence

	Sickness absence		
	KN	Rep.	APC
CT	0.0002	0.00013	0.00013
	(0.0001)	(0.00012)	(0.00012)
CT^2	−1.83e-06	−7.65e-07	−7.83e-07
	(1.60e-06)	(1.48e-06)	(1.48e-06)
Age		−0.0011	
		(0.00069)	
Age^2		2.09e-05***	
		(8.02e-06)	
APC model			Ft
H_0: age quadratic			
H_0: $\phi_1 = \phi_2 = 0$	0.2323	0.26	0.2660
Individuals	14,067	13,889	13,889
Observations	71,609	71,492	71,492
Within R^2		0.0010	0.0007
Adjusted within R^2		−0.2410	−0.2411

Note: *p<0.1; **p<0.05; ***p<0.01.
Coefficients reported with se below. Hypothesis tests report p-values.
"KN"= Künn-Nelen original findings; "Rep." = replication of Künn-Nelen; "APC" = APC panel regression.

Table 4.13: Regression results for GP visits

	Nr. GP visits		
	KN	Rep.	APC
CT	0.0029***	0.0027***	0.0027***
	(0.0008)	(0.0008)	(0.0008)
CT^2	−3.99e-05***	−3.73e-05***	−3.65e-05***
	(1.02e-05)	(9.89e-06)	(9.89e-06)
Age		−0.0424***	
		(0.0046)	
Age^2		0.0006***	
		(5.35e-05)	
APC model			FAP
H_0: age quadratic			0.0026
H_0: $\phi_1 = \phi_2 = 0$	0.0012	0.00	0.0009
Individuals	14,059	13,882	13,882
Observations	71,563	71,446	71,446
Within R^2		0.0052	0.0065
Adjusted within R^2		−0.2359	−0.2352

Note: *p<0.1; **p<0.05; ***p<0.01.
Coefficients reported with se below. Hypothesis tests report p-values.
"KN"= Künn-Nelen original findings; "Rep." = replication of Künn-Nelen; "APC" = APC panel regression.

4.D.1 The age effect for GP visits and health satisfaction

For the first outcome where the standard quadratic functional form restriction on the age control was rejected, my analysis of the estimated age accelerations led to two improvements to the model. The outcome variable was the number of GP visits made in the year. The improvements were to change the coding of the outcome variable and to introduce a childbirth indicator as a control. After implementing these two improvements, the within R^2 increased and the quadratic functional form restriction on age was not rejected.

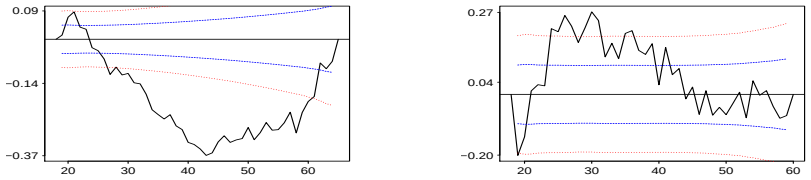
The first improvement was to change the coding of the outcome variable, the number of GP visits in the year. In the BHPS data, GP visits are coded into five categories with each category covering an increasingly large interval. Category 0 is no visits, category 1 is one or two visits, category 2 is three to five visits, category 3 is six to ten visits, and category 4 is eleven or more visits. Künn-Nelen treats this as a cardinal variable with values from 0 through 4 whereas I code it as a cardinal variable with values 0, 1, 3, 6, and 11 (since GP visits are approximately Poisson distributed I use the lowest value associated with each category). Künn-Nelen's treatment "squashes" the curve at older ages, creating a long left tail and short right tail as seen in Figure 4.4a. This squashed shape is inconsistent with a quadratic. My new coding does not have this squashing effect and so the quadratic restriction on the age effect is not rejected when that coding is used.

The second improvement was to introduce a childbirth indicator to the model. This control is significant and has a large effect size, as seen in column 1 of Table 4.14. Including childbirth increased both the within R^2 of the model and the p-value of the test statistic for the quadratic age effect, indicating that the inclusion of childbirth improves the fit of the model with the quadratic in age.

These model improvements increased the fit of the model but did not change the estimated effects of interest in Künn-Nelen's study, i.e. the coefficients associated with commute time and its square. This can be seen by comparing the estimates for "CT" and "CT^2" in column 1 of Table 4.14 and column 2 of Table 4.13. In this case, while the APC acceleration framework was effective as a diagnostic tool, the mis-specifications it diagnosed were not sufficient to affect the estimated

coefficients of interest. The effects of commute time and its square identified by Künn-Nelen are robust to the diagnosed mis-specifications.

Figure 4.4: Detrended age accelerations, GP visits and health satisfaction



(a) GP visits (b) Health satisfaction (women)

Blue dashed, red dotted lines = 1, 2 standard deviations from zero.

For the second health outcome where the standard approach to the age control was rejected, I was not able to identify improvements to the model, but was able to learn more about the relationship with age, using the APC framework. The second outcome was self-reported health satisfaction. I found that there was a clear difference between men and women in the patterns of health satisfaction by age. For men, no non-linearities in age were present; the "FP" model was sufficient to describe the data. For women, non-linearities in age were present, and the hypothesis that those non-linearities were quadratic in form was rejected.

Since the non-linear relationship between age and health satisfaction was primarily due to women, I inspected the detrended sums of age accelerations for women alone, see Figure 4.4b. There is negative acceleration in health satisfaction until the early 40s, and little acceleration or deceleration thereafter.

I tested several explanations for the non-linear relationship between age and health satisfaction among women. I tested variables capturing career and child-care pressures, which could reduce time for self-care and thus be detrimental to health satisfaction. I controlled for the number of health problems. None of these variables explained the relationship. I tested an ad hoc functional form simplification, imposing a quadratic age effect up to age 44 and no age non-linearities thereafter. This model was rejected, with a p-value of 0.0302. I conclude that the non-linearity of health satisfaction in age among women may be a "pure" age effect. This could be explained by the societal association of youth, health, and beauty among women, leading women to be dissatisfied with aging in a way that men are not. Further exploration of this is left for future research.

Although the relationship between age and health satisfaction is not quadratic, permitting the age non-linearities to have a more flexible functional form did not alter estimated coefficients on Künn-Nelen's objects of interest, commute time and its square. This can be seen by comparing the estimates for "CT" and "CT^2" in Columns 2 and 3 of Table 4.9. This indicates that Künn-Nelen's estimated relationship between commute time and health satisfaction is robust to the quadratic mis-specification of the age effect.

The final models for both GP visits and health satisfaction are seen in Table 4.14. For GP visits, this is a joint model of men and women, where the outcome is recoded following my earlier description and the indicator for childbirth is included. Since the quadratic age restriction was not rejected for this model, I impose it to achieve a parsimonious representation. I also omit controls included by Künn-Nelen but found to be insignificant, such as the relationship to the household head, again in the interests of parsimony. The within R^2 of this model has increased relative to Künn-Nelen's original specification in Table 4.13. It can also be seen that the estimated coefficients on commute time and its square, and the associated joint test, are largely unchanged relative to the original specification.

For health satisfaction, I include two final models: one for men and one for women. This reflects the fact that the shape of the relationship with age was found to be different between the two. For men, the FP model is presented. For women, the FAP model is presented. Those controls which were found to be significant are included in both models. For women this includes additional explanatory variables relating to health problems; however, despite these additional covariates, the restriction of the age effect to have a quadratic functional form is clearly rejected. This example highlights that although the APC framework acceleration is effective at diagnosing mis-specification, it may not be straightforward to resolve that mis-specification.

Table 4.14: Regression results for health satisfaction and GP visits, final

	Nr. GP visits	Health satisfaction	
		men	women
CT	0.0066^{***}	-0.003^{**}	-0.006^{***}
	(0.0019)	(0.001)	(0.002)
CT^2	-0.00009^{***}	0.00004^{**}	0.0001^{**}
	(0.00002)	(0.00002)	(0.00002)
Age	-0.0794^{***}		
	(0.0107)		
Age^2	0.0012^{***}		
	(0.0001)		
Childbirth	3.6765^{***}		-0.043
	(0.0772)		(0.051)
Health problems			-0.421^{***}
			(0.025)
Nr. new health problems			-0.052^{***}
			(0.010)
APC model		FP	FAP
Controls:			
Nkids	Y	N	N
Education	Y	N	N
Marital status	Y	N	Y
Relationship to household head	N	N	N
Government region	N	N	N
H_0: age quadratic			0.0087
H_0: $\phi_1 = \phi_2 = 0$	0.0005	0.0593	0.0027
H_0: quadratic to age 44			0.0302
Observations	71,446	29,633	21,218
Within R^2	0.0421	0.023	0.050
Adjusted within R^2	-0.1897	-0.229	-0.238

Note: *p<0.1; **p<0.05; ***p<0.01.
Coefficients reported with se below. Hypothesis tests report p-values.

4.E Results from APC acceleration models for ages 18-60

Table 4.15: APC acceleration model for men 18-60

	Health satisfaction	Health status	Health problems	Sickness absence	GP visits	In-patient
CT	-0.003^{**}	-0.002^{***}	0.001	-0.0002	0.002^{**}	0.0004
	(0.001)	(0.001)	(0.0005)	(0.0001)	(0.001)	(0.0003)
CT^2	0.00004^{**}	0.00002^{**}	-0.00001	0.00000	-0.00003^{**}	-0.00000
	(0.00002)	(0.00001)	(0.00001)	(0.00000)	(0.00001)	(0.00000)
APC model	FP	FAP	FAP	Ft	FAP	Ft
H_0: $\phi_1 = \phi_2 = 0$	2.8259^{*}	4.4518^{**}	1.0284	0.8756	2.8118^{*}	1.2005
	(0.0593)	(0.0117)	(0.3576)	(0.4166)	(0.0601)	(0.3011)
H_0: age quadratic		1.0367	0.8215		1.3647	
		(0.4066)	(0.7810)		(0.0620)	
Observations	29,633	39,470	39,474	39,474	39,451	39,474

Note: *p<0.1; **p<0.05; ***p<0.01.
Coefficients reported with se below. Hypothesis tests report p-values.

Table 4.16: APC acceleration model for women 18-60

	Health satisfaction	Health status	Health problems	Sickness absence	GP visits	In-patient
CT	-0.006^{***}	-0.001	0.001^{*}	0.0005^{**}	0.004^{***}	0.001^{*}
	(0.002)	(0.001)	(0.001)	(0.0002)	(0.001)	(0.0004)
CT^2	0.00005^{**}	0.00001	-0.00001	-0.00000	-0.0001^{***}	-0.00001
	(0.00002)	(0.00001)	(0.00001)	(0.00000)	(0.00002)	(0.00001)
APC model	FAP	FAP	FAP	Ft	FAP	FAP
H_0: $\phi_1 = \phi_2 = 0$	5.8601^{***}	0.2523	1.5484	3.0362^{**}	4.6689^{***}	1.3853
	(0.0029)	(0.7770)	(0.2126)	(0.0480)	(0.0094)	(0.2503)
H_0: age quadratic	1.9245^{***}	1.1368	1.2582		1.8115^{***}	1.8157^{***}
	(0.0004)	(0.2550)	(0.1272)		(0.0013)	(0.0012)
Observations	23,261	30,533	30,538	30,538	30,515	30,538

Note: *p<0.1; **p<0.05; ***p<0.01.
Coefficients reported with se below. Hypothesis tests report p-values.

Table 4.17: APC acceleration model for joint men and women 18-60

	Health satisfaction	Health status	Health problems	Sickness absence	GP visits	In-patient
CT	-0.004^{***}	-0.002^{**}	0.001^{**}	0.0001	0.003^{***}	0.0005^{**}
	(0.001)	(0.001)	(0.0004)	(0.0001)	(0.001)	(0.0002)
CT^2	0.00004^{***}	0.00002^{**}	-0.00001^{**}	-0.00000	-0.00004^{***}	-0.00001^{*}
	(0.00001)	(0.00001)	(0.00000)	(0.00000)	(0.00001)	(0.00000)
APC model	FAP	FAP	FAP	Ft	FAP	FA
H_0: $\phi_1 = \phi_2 = 0$	5.8921^{***}	3.4662^{**}	2.4316^{*}	1.0318	7.0060^{***}	2.0013
	(0.0028)	(0.0312)	(0.0879)	(0.3564)	(0.0009)	(0.1352)
H_0: age quadratic	1.5368^{**}	0.7166	0.9114		1.7181^{***}	1.5986^{***}
	(0.0162)	(0.9091)	(0.6306)		(0.0032)	(0.0095)
Observations	52,894	70,003	70,012	70,012	69,966	70,012

Note: *p<0.1; **p<0.05; ***p<0.01.
Coefficients reported with se below. Hypothesis tests report p-values.

Chapter 5

`apc.indiv`: An R package for acceleration-based age-period-cohort regressions with individual-level data

5.1 Introduction

I develop an R package, `apc.indiv`, to estimate age, period, and cohort accelerations from regressions using individual-level data. This builds on an existing R package, `apc`, which estimates age, period, and cohort (APC) accelerations from regressions using aggregate data (Nielsen, 2015). In `apc.indiv`, I use the same parametrization of age, period, and cohort effects that is used in `apc` to isolate and estimate APC accelerations. I allow for other explanatory variables in the regression and account for the particular features of individual-level data. Data for use with `apc.indiv` must be a table where each row is a unique observation and each column is a variable, including at least two of age, period, and cohort. Outcomes may be either continuous or dichotomous. Both repeated cross section and panel data are accommodated. For repeated cross section, the package includes a test of fit based on comparison to a more general model. A custom algorithm is required to estimate the more general model. These new tools are illustrated using two pre-existing R datasets: the repeated cross section dataset `Wage` from the `ISLR` package, and the panel dataset `PSID7682` from the `AER` package. In both cases, I

model log wages as a function of age, period, and cohort.

The individual-level data that can be used with the `apc.indiv` package is either repeated cross section or panel data. When multiple waves of large cross section surveys are combined, repeated cross section data is produced. Examples include the Health Survey for England discussed in §3, and the Current Population Survey in the US which is used later in this chapter. In repeated cross section surveys, different individuals are observed in each period. Panel data, on the other hand, uses the same individuals in every period. Examples of panel data include the British Household Panel Survey discussed in §4, and the Panel Survey for Income Dynamics in the US which is used later in this chapter. Panel data is also called longitudinal data. The `apc.indiv` package developed here is intended for use with both repeated cross section and panel data. The data must be drawn from multiple evenly-spaced waves of a study; at least three waves are needed. The data should also have detailed information on age or year of birth.

Age, period, and cohort are used as explanatory variables across the social sciences, but their study is complicated by the well-known age-period-cohort (APC) identification problem, see §2.3.2.2 and Glenn (2005). An individual's age is equal to the year in which they are observed minus their year of birth, i.e. $age = period - cohort$. Therefore it is impossible to identify separate linear effects of these three variables.

The APC identification problem can be avoided by focusing on non-linear effects, i.e. accelerations, in each of age, period, and cohort. These accelerations capture deviations from linearity associated with age, period, and cohort. It is only the linear APC effects which are rendered inseparable by the APC identification problem; deviations from linearity are not affected, see Holford (1983), Clayton & Schifflers (1987a), Carstensen (2007), and Kuang et al. (2008).

Accelerations in age, period, and cohort can be directly estimated from regressions with individual-level data. This was demonstrated for regressions with repeated cross section data in §3, and for regressions with panel data in §4. Direct estimation of the accelerations in these settings is achieved by using the parametrization of age, period, and cohort effects developed by Kuang et al. (2008), called the APC acceleration parametrization. The APC acceleration parametrization includes accelerations in each of age, period, and cohort, as well as a single linear

plane which combines the inseparable linear APC effects. This parametrization confers two benefits: first, the estimates do not rely on any untestable identifying assumptions; and second, it is easy to test model restrictions and count associated degrees of freedom.

The `apc.indiv` package provides a set of tools to estimate APC accelerations from regression models for individual-level data, implementing the techniques described in §3 and §4. First, the package contains user-friendly functions to estimate the features of the parametrization of Kuang et al. (2008): APC accelerations, and a single linear plane combining the linear efects of age, period, and cohort. Second, the package contains tools for testing restrictions on the model, such as the hypothesis that all accelerations equal zero. Third, for repeated cross section data it contains a test of fit for the APC acceleration model against a more general "time-saturated" (TS) model. The time-saturated model is high-dimensional and so standard OLS estimation and Newton-Raphson approximation cannot be used. Instead, `apc.indiv` includes purpose-built OLS and Newton-Raphson algorithms to overcome this dimensionality issue.

I demonstrate the use of `apc.indiv` with an application to the estimation of age profiles of log wages, using both repeated cross section and panel data. There is a long-standing interest in economics in estimating age profiles of wages which exclude period and cohort effects; see for example Hanoch & Honig (1985); Meghir & Whitehouse (1996); Kalwij & Alessie (2007); Lagakos et al. (2018). The datasets I use to estimate these age profiles are drawn from existing R packages: the repeated cross section dataset `Wage` from the `ISLR` package, and the panel dataset `PSID7682` from the `AER` package. Both datasets pertain to the US labour force. The repeated cross section data covers the 2000s, while the panel data covers the late 1970s and early 1980s. Both datasets show concavity in log wages with age. This may reflect diminishing marginal returns of experience to wages, and selection out of the labour force at older ages by high earners with the financial resources to retire early. I also identify a period deceleration in log wages that is consistent with the 1979 oil price shock, which may have depressed real wages via both the economic slowdown and inflation. I demonstrate the use of `apc.indiv` for binary outcomes with repeated cross section data by examining the probability that an individual's job is classified as "industrial" rather than "information".

The `apc.indiv` software is the first software which enables direct estimation of APC accelerations from individual-level data. Existing software implements alternative approaches which have been used in an attempt to resolve the APC identification problem. One example is the `apcd` package for Stata by Chauvel (2012), which tries to separate the linear and non-linear effects, but does so using a different approach to that implemented in this chapter. Other examples which do not separate the linear and non-linear effects include the `epi` package for R of Carstensen (2013), which was adapted for individual-level data by Diouf et al. (2010) and Peeters et al. (2015), and Stata packages by Schulhofer-Wohl & Yang (2006) and O'Dea (2012). Other approaches exist for which software packages have not been written; these include the hierarchical age-period-cohort model, of (Yang & Land, 2006) which was implemented by Reither et al. (2009) using SAS PROC GLIMMIX, and An & Xiang (2016) using Stata 14.1.

The approaches used in existing software impose untestable assumptions on the linear effects of age, period, and cohort in order to achieve identification. The limitation is that effects are then estimated with respect to this imposed constraint. Estimated effects are sensitive to the choice of constraint; for example, Lagakos et al. (2018) consider the approach used in `epi` and highlight the sensitivity of estimated age-wage profiles to whether the linear trend set to zero is that in period or that in cohort. The choice of constraint cannot be validated by testing because of the APC identification problem, so researchers must rely on theory to select their constraints; but it is unclear whether any theory in the social sciences is sufficiently accepted to be used as a basis for further analysis without the possibility of empirical verification. The parametrization strategy implemented in `apc` and `apc.indiv` avoids the imposition of untestable assumptions, and explicitly includes a framework for testing the modelling assumptions. This is achieved by recognizing that in the context of APC effects, one can characterize mathematically that which can be estimated and that which is unidentifiable. We focus on the identifiable effects, i.e. the accelerations, and choose to ignore the unidentifiable linear effects.

The remainder of this chapter is as follows. §5.2 summarises the theoretical framework employed in `apc.indiv`. More detail on this framework can be found in chapters §2 through §4. I provide an overview of the `apc.indiv` package in §5.3.

I demonstrate how this package can be used with an application to US log wages in §5.4. §5.5 concludes.

5.2 Overview of the APC acceleration framework

The `apc.indiv` software is designed around a general regression model which permits estimation of age, period, and cohort accelerations from individual-level data. It therefore unites the work in §3 and §4. Central to the regression model is a linear predictor, which includes age, period, and cohort (APC) effects as well as covariates of interest. The APC effects in the linear predictor are represented by the parametrization of Kuang et al. (2008), henceforth the APC acceleration parametrization. The APC acceleration parametrization includes accelerations in each of age, period, and cohort as well as a single, combined linear plane. The linear predictor is embedded in a statistical model suitable for the data at hand. This regression-based approach to estimating APC accelerations also allows for testing of restrictions on the model and testing against a more general model.

5.2.1 Linear predictor

A single linear predictor is suitable to describe the APC acceleration model for both repeated cross section and panel data. The linear predictor is given by

$$\eta_{ip} = \mu_{ip} + z'_{ip}\zeta \quad ; \quad \mu_{ip} = x'_{ip}\xi \tag{5.1}$$

The linear predictor η_{ip} is defined for each observation, where an observation is an individual-time pair $\{ip\}$. Individuals are indexed by $i = 1, \ldots, N$ and time is indexed by $p = L+1, \ldots, L+P$, for L a data-specific constant defined as in §2.3.2.1. The term μ_{ip} is the part of the linear predictor that captures the APC effects. It is composed of a design vector x_{ip} and a parameter vector ξ. These two terms constitute the APC acceleration parametrization, and are described in further detail in §2.3.3.4; a summary is given in §5.2.3 of this chapter. The vector z_{ip} captures information on any other explanatory variables, here referred to as covariates; ζ is the associated coefficient.

5.2.2 Statistical models

The effects of age, period, and cohort as well as any covariates are estimated by embedding the linear predictor in an appropriate statistical model, determined by the data. For repeated cross section data, the statistical model is a generalized linear model, estimated by maximum likelihood. This framework is preferred for repeated cross section data because it can easily accommodate binary and other non-Gaussian outcomes. For repeated cross section data where survey weights are available, a generalized linear model, estimated by pseudo-maximum likelihood, is used. For panel data, the statistical model is a linear projection estimated by generalized least squares. This framework is preferred for panel data because it is easy to allow for correlation of unobserved components between observations on the same individual.

5.2.2.1 Repeated cross section

Repeated cross section data is assumed to be IID across individuals. That is, the vectors $\{y_{ip}, x_{ip}, z_{ip}\}$ are independent draws from an identical distribution. All asymptotic analysis will therefore be at a rate determined by N, the total number of individuals in the sample.

For repeated cross section analysis, the linear predictor is embedded in a generalized linear model (GLM); see Nelder & Wedderburn (1972), McCullagh & Nelder (1989), Dobson & Barnett (1990), Dunteman & Ho (2005). This framework is described in detail in §3.3.1. The GLM is adapted to reflect the distribution of the outcome variable y_{ip}; in `apc.indiv` this may be continuous and approximately Gaussian, or binary. Where the outcome of interest y_{ip} is a continuous variable, a Gaussian distribution and identity link function are used:

$$y_{ip} = \eta_{ip} + \varepsilon_{ip} \qquad ; \qquad \varepsilon_{ip} \sim N(0, \sigma^2). \tag{5.2}$$

Here ε_{ip} is unobserved, observation-specific, random noise. Where the outcome is dichotomous, a binomial distribution and logit link function are used

$$\log \frac{\mathsf{P}(y_{ip} = 1)}{\mathsf{P}(y_{ip} = 0)} = \eta_{ip}. \tag{5.3}$$

Estimation of the GLM for repeated cross section data is by maximum likelihood. The maximum likelihood estimator can be represented analytically for model (5.2) and can be approximated by Newton-Raphson iteration for model (5.3). Both estimators are subject to regularity conditions detailed in §3.3. There are two approaches to inference in the maximum likelihood framework: finite sample inference and asymptotic inference. Finite sample inference is possible where an exact distribution of a test statistic can be derived given the sample size and specified distribution of the outcome variable. For outcomes that are modelled using a Gaussian distribution, it is possible to perform finite sample inference using an F-test. Asymptotic inference is possible for all maximum likelihood estimators. Under the regularity conditions detailed in §3.3, the maximum likelihood estimators of the parameters are consistent and have asymptotically normal distributions, and likelihood ratio tests are asymptotically χ^2.

For repeated cross section data where survey weights are available, a generalized linear model is used but estimation is by pseudo-maximum likelihood rather than maximum likelihood. Survey weights are used to account for selective survey participation. The use of GLMs is inherited from repeated cross section data without survey weights. However, pseudo-maximum likelihood is preferred to full maximum likelihood due to the complexity associated with developing a full likelihood model for survey participation (see the discussion in Lumley & Scott, 2017). The pseudo-maximum likelihood estimator is generated by multiplying each observation in the likelihood function by its sampling weight:

$$\hat{\ell}(\xi;\zeta) = \sum_{i=1}^{N} w_i \ell_i(\xi;\zeta). \tag{5.4}$$

Note that the p index is omitted here for clarity of exposition, since it is redundant where the data is repeated cross section. Here $\ell_i(\xi;\zeta)$ is the likelihood of observation i in the population given ξ and ζ. This is multiplied by the sampling weight w_i. The estimator obtained from maximising the pseudo-likelihood $\hat{\ell}$ is consistent and asymptotically normal under regularity conditions, see Lumley & Scott (2017). Inference is therefore by appeal to large-sample asymptotic theory.

5.2.2.2 Panel

Panel data is again assumed to be IID across individuals. Because each individual is observed more than once, this means that it is $\{Y_i, X_i, Z_i\}$ that are independent draws from an identical distribution, rather than the individual-time units $\{y_{ip}, x_{ip}, z_{ip}\}$. Here Y_i is a P-length vector that stacks all units y_{ip} for a given individual i, and X_i and Z_i are similarly defined. All asymptotic analysis is performed at a rate determined by N.

For panel data, the linear predictor is embedded in a generalized least squares (GLS) framework. The model is

$$y_{ip} = \eta_{ip} + \varepsilon_{ip} \qquad ; \qquad \varepsilon_{ip} = \omega_i + \epsilon_{ip}. \tag{5.5}$$

The unobserved, observation-specific term ε_{ip} is broken into two parts; an observation-specific random noise term ϵ_{ip}, and an individual-specific term ω_i. The treatment of these two components depends on which of three panel settings is being used: pooled OLS, random effects, or fixed effects. These settings are discussed in detail in §4.3. In pooled OLS ω_i is constrained to be zero for all individuals i. The vector ϵ_i, which contains all observation-specific terms ϵ_{ip} for a given individual i, is permitted to have an unconstrained variance-covariance matrix. In random and fixed effects, ϵ_{ip} is assumed $IID(0, \sigma_\epsilon^2)$ at the observation level but structure is imposed on ω_i to account for correlation in unobserved components within a given individual. In random effects, ω_i is treated as a random noise term: $\omega_i \sim IID(0, \sigma_\omega^2)$. In fixed effects, ω_i is treated as an individual-specific, freely-varying parameter.

Estimation under the generalized least squares framework is by minimizing the sum of estimated squared residuals, $\hat{\varepsilon}_{ip}^2$. In pooled OLS, this is achieved by ordinary least squares applied to the original data, $\{y_{ip}, x_{ip}, z_{ip}\}$. In random effects, ordinary least squares is applied to data which has been adjusted for the possibility of non-zero ω_i, which produces correlation in individual observations over time. This adjustment is done by a procedure called θ-differencing. OLS applied to θ-differenced data produces the same results as the feasible generalized least squares (FGLS) procedure described in §4.3.2. The θ-differencing procedure, and its equivalence to the FGLS representation of the random effects estimator, are described in Croissant & Millo (2008). In fixed effects, ordinary least squares is

applied to data which has been within-individual demeaned, see §4.3.3. This demeaning adjusts for both non-zero ω_i and the possibility of correlation between ω_i and the explanatory variables x_{ip}, z_{ip}. The practical situations in which correlation between ω_i and the APC variables x_{ip} might arise are further discussed in §4.3.4.2.

Inference in the generalized least squares framework is performed by appeal to large-sample asymptotic theory. Under regularity conditions, the least squares estimators of the parameters are consistent, with asymptotically normal distributions, and Wald test statistics are asymptotically χ^2.

If an additional assumption can be made, that ϵ_{ip} is normally distributed and homoskedastic, an F-test can be used instead of a Wald test. The F-test is the same as that described earlier in the context of GLM and repeated cross section data; the additional assumption on the distribution of ϵ_{ip} results in equivalence between the GLM and GLS frameworks for the linear model. The F statistic is a normalization of the Wald statistic and should be compared to an F distribution, see §7.2.1 of Cameron & Trivedi (2005). The F-test may be used in place of the Wald test, even without the additional assumptions on ϵ_{ip}. It is argued that as long as ϵ_{ip} is approximately normal, then the F-test is approximately valid; see §4.2.2 of Wooldridge (2010).

The use of different analytical frameworks, GLM and GLS, for repeated cross section and for panel reflects common practice in social science. This common practice is driven by historical factors. The GLM/maximum likelihood framework was adopted for repeated cross section data as a comprehensive way to handle outcome variables that could not be approximated by a Gaussian distribution. The GLS framework was adopted for panel data because it could handle the more complex error structure needed to account for correlation between observations on the same individual. While it is now possible to handle correlated errors in the GLM framework, the results of a GLM model with correlated errors are slightly different to those of a GLS model with correlated errors. These slight differences mean the GLS framework is still preferred for panel data, while the comprehensiveness of the GLM framework means that it remains preferred for repeated cross section data. I chose to conform to this common practice with `apc.indiv`, using GLM for repeated cross section analysis and GLS for panel analysis.

5.2.3 Age-period-cohort acceleration parameters

The effects of age, period, and cohort that appear in the linear predictor η_{ip} are accelerations in each of age, period, and cohort as well as a combined linear plane. These are captured by the parameter vector ξ, which is defined:

$$\xi = \{v_o, v_a, v_c, \Delta^2\alpha_3, \ldots, \Delta^2\alpha_A, \Delta^2\beta_{L+1}, \ldots, \Delta^2\beta_{L+P}, \Delta^2\gamma_3, \ldots, \Delta^2\gamma_C\}. \quad (5.6)$$

The design vector associated with the parameter vector ξ is x_{ip}.

The first three elements of the parameter vector ξ define a linear plane combining the inseparable linear effects of age, period, and cohort. These three elements, v_o, v_a, v_c, are the intercept and slopes of a linear plane in age-cohort space. Age-cohort space is a three-dimensional space with age $a = 1, \ldots A$ and cohort $c = 1, \ldots, C$ on the two horizontal axes. See Figure 2.3 in §2.3.3.3 for an illustration of age-cohort space. Period, indexed by $p = L + 1, \ldots, L + P$, is given by the 45° line in the horizontal plane defined by the age and cohort axes, since $p = a + c - 1$. The vertical axis of the three-dimensional space measures the value of the APC part of the linear predictor, μ_{ip}.

The remaining elements are the accelerations in age, period, and cohort. The acceleration at age a captures the difference between the effect of aging from $a-2$ to $a-1$ and the effect of aging from $a-1$ to a. If the effect of aging from $a-1$ to a is greater than that of aging from $a-2$ to $a-1$, the acceleration at a will be positive. The age accelerations are denoted $\Delta^2\alpha_a$; the period accelerations are denoted $\Delta^2\beta_p$; and the cohort accelerations are denoted $\Delta^2\gamma_c$. These accelerations are summed in the design vector x_{ip} to reflect the cumulative effects of age, period, and cohort over the life of the individual i up to time p. The derivation of this summation, x_{ip}, is given in §2.3.3.4.

5.2.4 Uses of the framework

The regression framework with age, period, and cohort accelerations captured via the parameter vector ξ can be used in several ways. I describe five use cases in this section: inference on individual accelerations; testing restrictions on the full model; testing generalizations of the model; using the sums of accelerations to

evaluate the shape of the relationship between an outcome and age, period, or cohort; and forecasting.

First, inference can be performed on the individual accelerations to evaluate the impact of a policy change or exogenous shock. This inference is by appeal to the asymptotic theory described in §5.2.2. An individual acceleration in age, period, or cohort may be generated by a policy or other shock. For example, McKenzie (2006) identified a negative period acceleration in aggregate consumption data associated with the Mexican peso crisis. With the regression approach used in `apc.indiv` such accelerations can be identified from repeated cross section or panel data, controlling for individual covariates.

Second, restrictions on the APC acceleration model can be tested. Restrictions may be sequentially tested in order to achieve a parsimonious model. Alternatively a particular restriction implied by economic theory may be tested; for example, the life cycle hypothesis implies no age accelerations in consumption. The restrictions are tested using Wald, F, or likelihood ratio tests. Two categories of restriction are considered: sub-models and functional form restrictions.

The first category of restrictions pertains to sub-models of the APC acceleration model. sub-models of the APC model constrain groups of elements of ξ to be zero. For example, a "period-cohort" (PC) model implements the restriction implied by the classical life cycle model, that all age accelerations are zero: $\forall a, \Delta^2\alpha_a = 0$. An "age-cohort" (AC) model constrains all period accelerations to be zero: $\forall p, \Delta^2\beta_p = 0$. A more parsimonious model for BMI of English men was obtained in §3.5.4 by showing that a restriction of the APC model to the AC sub-model was not rejected. A full list of sub-models suitable for repeated cross section data is found in §2.3.4.3, and full lists for the three panel data settings can be found in the relevant sub-sections of §4.3. The software provides for estimation of any of these sub-models, as well as for recursive testing of sub-models to achieve a parsimonious representation.

The second category of restriction pertains to functional form restrictions. For example, a quadratic restriction on the effect of age can be tested by imposing equality of the age accelerations: $\forall a, \Delta^2\alpha_a = \lambda$, for λ an unknown constant. This restriction was rejected for panel data on hospital in-patient stays among British women in §4.4. Restrictions on the covariate coefficients ζ can also be tested; for

instance in §4.4 I test the joint significance of commute time and its square in models for hospital in-patient stays. It is straightforward to perform such tests using `apc.indiv` in conjunction with existing packages for hypothesis testing.

The third use case of the APC acceleration framework is testing generalizations of the APC acceleration model. This is done by comparing the APC acceleration model to a larger model which "nests" it, meaning that the APC model acceleration can be obtained by imposing restrictions on the larger model. A particularly interesting general model is the time-saturated model; this model allows for interaction effects between age, period, and cohort, which are omitted from the parameter vector ξ. This model is high-dimensional and so requires a custom algorithm for estimation. Testing against the time-saturated model is implemented in `apc.indiv` for repeated cross section data, as described in §5.3.3.1 through §5.3.3.3, but is not yet implemented for panel data. Generalizations of the covariate vector z_{ip} can also be tested by adding covariates, as seen in §4.4 where the addition of a childbirth indicator is tested.

The fourth use case of the APC acceleration framework is that sums of accelerations can be used to examine the shape of the relationship between an outcome of interest and age, period, and cohort. For example, in §3.5 I used sums of cohort accelerations to identify a concave relationship between obesity and cohort among English. In §4.4, an investigation of sums of age accelerations was pivotal in identifying the importance of accounting for childbirth in the model of hospital in-patient stays. Since accelerations capture deviations from linearity, the sums of accelerations capture cumulative deviation from linearity, i.e. the non-linear part of the shape of the relationship between the outcome and age, period, and cohort. To visually isolate the non-linear part of the shape, the summation is performed in such a way that the sums of age, period, and cohort accelerations are anchored to start and end in zero. This is achieved as a bijective mapping of the summation implied by the design vector x_{ip}. As the mapping is bijective, the degrees of freedom of the two summations are the same. Further details of the mapping are provided in §2.3.4.2. All visual representations of the sums of accelerations in the software use the procedure with anchoring at zero by default, although there is an option to use the procedure implied by the original design vector x_{ip}.

The fifth use case for the APC acceleration framework is forecasting. This has not yet been implemented for individual-level data and is not provided for in this software. This is due to the additional complexity associated with forecasting covariates in an individual data setting. An example of forecasting with aggregate data can be found in Martínez Miranda et al. (2015).

5.2.5 Advantages of the framework

The approach to age-period-cohort analysis used here, of estimating accelerations via a regression model, improves on other available approaches to age-period-cohort analysis for individual-level data. The existing approaches to age-period-cohort analysis fall into two categories. In the first category are approaches where the age, period, and cohort parameters are only identified by untestable constraints; the current framework is preferred to this because the identification of accelerations does not require untestable constraints. In the second category are approaches where the age, period, and cohort parameters to be estimated are also accelerations, but the estimation procedure is cumbersome. The approach in this paper is unique because it combines identification that does not rely on untestable constraints with simple, regression-based estimation.

Most other approaches to age-period-cohort analysis for individual-level data involve regression parameters that are only identified by untestable constraints. For example, a common approach is to constrain the linear effect in one of age, period, and cohort to be zero (Deaton & Paxson, 1994; Carstensen, 2007). This permits estimation of level effects for each age, period, and cohort in the data, but those estimates are sensitive to the constraints. An illustration of this sensitivity can be found in Lagakos et al. (2018), where the ordering of countries in terms of the slope of the age-wage profile depends on whether the slope in period or the slope in cohort is constrained to zero. Other examples of constraints-based approaches to individual-level APC identification include the HAPC model of Yang & Land (2006) and the procedure developed in Schulhofer-Wohl (2018).

In contrast to these approaches, the identification of APC accelerations here does not rely on untestable constraints. The invariance of the accelerations to constraints is outlined in §2.3.3.1. Formal inference and testing in `apc.indiv` is

based on these accelerations. The sums of accelerations do rely on a choice of linear plane, but this differs fundamentally from the constraints methods because the plane is recognised as being a composite of all linear effects rather than attributed to two, and separation is maintained between the linear and non-linear effects.

There is some work with APC identification and individual data which focuses on accelerations as the parameters of interest, but the estimation procedure used is cumbersome. Van Landeghem (2012) investigates the claim that well-being is U-shaped in age by examining accelerations of well-being in age. A first-stage regression in differences of well-being with age is estimated, and from the parameters of this regression the accelerations are constructed. This two-stage procedure is laborious to implement. In contrast, the approach developed in this paper embeds the accelerations as parameters to be estimated directly from a regression. This makes it easy not only to estimate accelerations, but also to test restrictions and generalizations for both the APC parameters and any covariates.

5.3 Overview of the R package `apc.indiv`

The development of `apc.indiv` was guided by two principles: ease of use and ease of adaptation. In keeping with the principle of ease of use, `apc.indiv` contains three main functions which reflect the three core steps of econometric analysis: model estimation, selection, and interpretation. In this section, I introduce each of these functions. In keeping with the principle of ease of adaptation, at each stage the output has been designed to be compatible with other common R packages for econometric analysis, allowing the user to move from the `apc.indiv` environment to alternative environments where necessary. For example, the output of the estimation command is compatible with the hypothesis testing infrastructure in the packages `car` and `lmtest`.

The `apc.indiv` package depends on several existing R packages. For construction of the design vector x_{ip} in the estimation function and plotting of cumulated accelerations for interpretation, I drew on the existing `apc` package (Nielsen, 2015). For estimation of the APC models I drew on existing estimation packages: `stats` for repeated cross section data (R Core Team, 2019), `survey` for repeated cross

section data with survey weights (Lumley, 2019), and `plm` for panel data (Croissant & Millo, 2008). For model selection I rely on the testing packages `car` and `lmtest` (Fox & Weisberg, 2019; Zeileis & Hothorn, 2002).

The topics covered in this section are as follows. In §5.3.1 I discuss the existing software for analysis of APC effects using individual-level data. In §5.3.2 I describe the data format required by `apc.indiv`. Estimation is discussed in §5.3.3, including algorithms for estimation of the more general time-saturated model which nests the APC acceleration model. Model selection is discussed in §5.3.4, and interpretation is discussed in §5.3.5.

5.3.1 Existing software for age-period-cohort analysis

There is no existing software which can provide direct estimates of APC accelerations from individual-level data. The `apc` package in R provides direct estimates of APC accelerations, but is only suitable for aggregate data (Nielsen, 2015). The available packages for individual-level data rely on methods involving the imposition of untestable constraints to identify level effects, and so do not produce estimates of accelerations.

The `apc` package in R is closely related to the software developed in this paper. The `apc` package is the first to use the parametrization developed by Kuang et al. (2008), and described in §5.2, to estimate APC accelerations in a regression framework. The `apc` package is designed for aggregate data tabulated in a Lexis structure, for example with a row for each age and a column for each cohort. The commands are not adaptable to individual-level data, which has a different structure. The software developed in this paper borrows some internal computations from the `apc` package, but is optimized for individual-level data.

A number of packages are available for APC analysis of individual-level data, but they do not produce estimates of the APC accelerations. Instead of accelerations, they produce estimates of level effects of each age, period, and cohort, which are identified by the imposition of untestable constraints. There is a risk with such approaches that the estimated level effects may be misinterpreted if their sensitivity to the identifying constraints is not appreciated.

The existing package which carries the least risk of misinterpretation is Chauvel's `apcd` package for Stata (Chauvel, 2012). The approach in this package relies on just-identifying constraints which "de-trend" the estimated age, period, and cohort level effects. In this respect it bears some resemblance to the sums of accelerations available in `apc.indiv`; the parameters estimated by `apcd` are sums of accelerations at each age, period, and cohort, associated with a linear plane chosen in order to ensure those sums have no overall trend. Estimates from `apcd` can therefore be used in a similar way to the detrended sums of accelerations from `apc.indiv`, to visually evaluate the non-linear part of the shape of the relationship between age, period, and cohort. However, the lack of estimates of accelerations limits its utility for other applications, such as evaluating the impact of shocks.

More caution is required when working with other methods, such as those in the Stata packages `apc` from O'Dea (2012) and `apc` from Schulhofer-Wohl & Yang (2006), and in the adaptation of `epi` in R for individual-level data by Peeters et al. (2015). Like `apcd`, the methods in these packages use just-identifying constraints to get estimates of level effects of age, period, and cohort. However, more caution is required when working with estimates from these packages, as the imposed constraints assign linear effects to age, period, and cohort by assumption.

It is preferable to directly estimate the accelerations, which are invariant to assumptions about the linear plane, and use these for statistical testing, as is done in `apc.indiv`. Sums of accelerations which rely a linear plane that isolates the non-linear shape may then be used to examine shape of the relationship, as is done in `apc.indiv` and Chauvel's `apcd`, for R and Stata respectively.

There are also packages in both Stata and R which are only suitable for aggregate data and also do not produce direct estimates of the accelerations. Instead, they rely on constraints to identify estimates of the age, period, and cohort effects. In some cases, it is easy to understand how the constraints affect the estimates; this is true of the `epi` package for R of Carstensen (2013) and the `apcfit` package for Stata of Rutherford et al. (2010). Both of these rely on a simple assumption of no slope in one of the three of age, period, and cohort. In others, the relationship between the constraints and the estimates is less clear. This is true of `st0245` for Stata by Sasieni (2012) and `BAMP` by Schmid & Held (2007).

5.3.2 Data format

The functions in `apc.indiv` can be used with either repeated cross section or panel data. The data must have a time dimension, i.e. several "waves" must be included. The waves must be evenly spaced; `apc.indiv` currently cannot allow for missing waves. Repeated cross section data should be an $N \times d$ table, where N is the number of individuals and d is the number of variables. Thus there is a row for each individual and a column for each variable. Panel data should be in "long" format. This is similar to the repeated cross section data in that there is a column for each variable and a row for each observation. The unit is the record for a single individual at a particular time period. Each individual is observed at more than one time period, so there will be multiple units for each individual. These appear as sequential rows within the dataset. If the panel is balanced (i.e. all individuals observed in all periods), there are then $P \times N$ rows and d columns.

The data must have one of three possible age-period-cohort structures. The three structures are: age-period data, period-cohort data, and age-cohort data. In data with an age-period structure, individuals within a particular age range are sampled over several years. An example of this format would be an annual survey of working-age adults. In data with a period-cohort structure, individuals of certain birth-years (cohorts) are sampled over several years. The National Longitudinal Survey of Youth (USA) is an example of period-cohort data. Data with an age-cohort structure, where cohorts are sampled at successive ages within a given range, is rarer but may be encountered in the context of insurance reserving.

The set of variables d must include time indices for age, period, and cohort. Any two can be used to construct the third by the identity $age = period - cohort$. The time indices are required to construct the design vector x_{ip}, describing the APC effects. The construction of x_{ip} is discussed further in §2.3.3.4.

In addition to the time indices, the dataset must include at least one additional variable, the outcome variable y_{ip}. For panel data there must also be an individual identifier. Survey weights may be present for repeated cross section data. Any number of covariates z_{ip} can be included.

5.3.3 Model estimation

The first stage of econometric analysis is to develop a framework for estimation of models. The command for estimation is `apc.indiv.est.model()`. This command is run on data described in §5.3.2. In addition to the data, the name of the outcome variable and the names of any covariates are required inputs. The choice of model to be estimated must also be specified; it may be the full APC model, one of the sub-models, or the more general TS (time-saturated) model. For repeated cross section data, the assumed distribution of the outcome variable must be stated. If using survey weights, the weight variable must be specified. If using panel data, the individual identifying variable must be specified as well as the choice of panel data setting (i.e. pooled OLS, random effects, or fixed effects).

The output of the estimation command `apc.indiv.est.model()` is as follows. Of primary interest are two matrices containing the estimates and standard errors of the coefficients ξ and ζ respectively. The output object `fit` can be used in conjunction with functions from the `car` and `lmtest` packages to perform hypothesis tests. There are also outputs that contain information about the data structure required by subsequent interpretative functions, discussed in §5.3.5.

The `apc.indiv.est.model()` command is composed of three sub-functions, which can be accessed independently. These three sub-functions are `apc.indiv.design.collinear()`, `apc.indiv.design.model()`, and `apc.indiv.fit.model()`. The first sub-function is used to produce a general version of the APC design matrix X, which stacks x_{ip} over i an p. The function `apc.indiv.design.collinear()` appends X to the main dataset. For a researcher confident of their ability to correctly use the elements of X in models which this `apc.indiv` package does not provide for, the dataset produced by this command would be of use. The second sub-function, `apc.indiv.design.model()` takes provided information about the chosen model (APC, a sub-model, or TS), as well as the selected outcome and covariates, and produces a model formula of the format commonly used in R. If it is desired to make use of the APC acceleration parametrization in conjunction with a statistical model not currently implemented in this package, the output of this command may be of use. The command `apc.indiv.fit.model()` performs the estimation using one of `glm()`,

`svyglm()`, or `plm()`. It also structures the output of these commands to be compatible with the plotting commands described in §5.3.5.

This division into three sub-functions can be exploited where it is needed to estimate multiple models. The first sub-command, producing a general APC design matrix, is computationally intensive, but need only be run once for a given dataset. A researcher could run this once and then use `apc.indiv.design.model()` and `apc.indiv.fit.model()` to fit build and estimate the various models. This would reduce the overall time required to estimate all models. This is the strategy employed in the model selection command described in §5.3.4.

5.3.3.1 The time-saturated model

One of the models that can be estimated using `apc.indiv.est.model()` is the TS (time-saturated) model, introduced in §3.4. The TS model is a more general model that is used to test whether the APC specification, which does not allow for interactions between age, period, and cohort, is overly parsimonious. At present the TS model is only available for repeated cross section data. It replaces the APC element of the linear predictor in equation (5.1), $\mu_{ip} = x'_{ip}\xi$, with a new element, $t'_{ip}\kappa$. Here t_{ip} is a vector of indicators for whether an observation $\{ip\}$ belongs to a particular age-cohort combination, and κ is the associated n-length parameter vector for n the number of age-cohort combination observed in the data. Covariates are specified exactly as in the APC model, so the TS linear predictor

$$\eta_{ip} = t'_{ip}\kappa + z'_{ip}\zeta \tag{5.7}$$

nests the APC linear predictor

$$\eta_{ip} = x'_{ip}\xi + z'_{ip}\zeta. \tag{5.8}$$

By stacking the vectors t'_{ip} over all units $\{ip\}$, we obtain T, a $N \times n$ matrix of indicators; and by stacking the vectors z'_{ip} over all units we obtain Z, a $N \times d_z$ matrix where d_z is the number of covariates present in z_{ip}.

The design matrix $M = \{T, Z\}$ of the TS model has dimension $N \times (n + d_z)$. Estimation of the TS model requires that this matrix be squared and inverted, which introduces storage problems for large n. An algorithm to overcome this

problem was proposed in §3.4. The idea of this algorithm is to separate $\{T, Z\}$ for estimation by partitioned inversion, noting that the particular fixed effects structure of T means that $T'T$ is diagonal and thus simple to invert. The `apc.indiv` package includes two custom functions which implement this algorithm for the Gaussian and logit cases respectively.

5.3.3.2 Algorithm for the normal time-saturated model

I estimate the time-saturated (TS) normal model

$$y_{ip} = \eta_{ip} + \varepsilon_{ip} \quad ; \quad \eta_{ip} = z'_{ip}\zeta + t'_{ip}\kappa \quad ; \quad \varepsilon_{ip} \sim \mathsf{N}(0, \sigma^2) \tag{5.9}$$

by maximum likelihood. In a single function, I implement a partial regression where the order in which variables are partialled out is carefully specified to exploit the fixed effects structure of T. This function is `apc.indiv.estimate.TS()` and can be called as an option from `apc.indiv.est.model()`.

Inside `apc.indiv.estimate.TS()`, estimation proceeds in four steps. First, each covariate (i.e. each column of Z) is regressed on T. That is, I construct

$$Z = T(T'T)^{-1}T'Z + \upsilon = \breve{Z} + \upsilon. \tag{5.10}$$

Due to the fixed effects structure of T, the matrix $T'T$ is diagonal and thus can be inverted by taking the reciprocal of all elements on the main diagonal. Then $\breve{Z}$ can be constructed by simple matrix multiplication.

In the second step, the outcome Y is regressed on $\upsilon = [I - T(T'T)^{-1}T']Z$. This gives an estimator of ζ in model (5.9).

In the third step, a reparametrization of (5.9) is constructed, which has orthogonal regressors and so is more amenable to numerical estimation. This is

$$Y = T\rho + [I - T(T'T)^{-1}T']Z\zeta + \varepsilon. \tag{5.11}$$

The orthogonality means that an estimate of ρ can be obtained by regression of Y on T, while an estimate of ζ can be obtained by regression of Y on $[I - T(T'T)^{-1}T']Z$. The latter is straightforward to perform because $[I - T(T'T)^{-1}T']Z$ is of low dimension. For the former, we can again make use of the fact that $T'T$

is diagonal to simplify estimation of ρ. The fit from this specification in terms of ρ and ζ is identical to that from the model in (5.9) in terms of κ and ζ.

The fourth step is the construction of κ from the estimates $\hat{\rho}$ and $\hat{\zeta}$:

$$\hat{\kappa} = \hat{\rho} - (T'T)^{-1}T'Z\hat{\zeta}. \tag{5.12}$$

5.3.3.3 Algorithm for the logit time-saturated model

I estimate the time-saturated (TS) logit model

$$\log\left\{\frac{\mathsf{P}(y_{ip}=1)}{\mathsf{P}(y_{ip}=0)}\right\} = \eta_{ip} = z'_{ip}\zeta + t'_{ip}\kappa \tag{5.13}$$

by Newton-Raphson approximation to the maximum likelihood estimator. The function for this is `apc.indiv.logit.TS()`. This function is called internally by `apc.indiv.est.model()`. Aspects of the Newton-Raphson procedure e.g. starting values, loop limits, and convergence criteria can be fine-tuned as needed.

The algorithm follows a five-step iterative procedure. Before entering this iteration, starting values are chosen for the parameters $\{\xi, \zeta\}$. The default is to use the OLS estimators from a linear probability model, calculated using the procedure outlined in §5.3.3.2.

The first step in the iterative procedure is to construct the score. To begin, the linear predictor $\eta_{ip} = t'_{ip}\kappa + z'_{ip}\zeta$ is calculated, using the assigned starting values for $\{\kappa, \zeta\}$. Logistic probabilities for each observation are calculated as

$$\pi_{ip} = \frac{\exp(\eta_{ip})}{1+\exp(\eta_{ip})}. \tag{5.14}$$

These are stacked into an N-length vector Π and the score is constructed:

$$\dot{\ell} = \begin{pmatrix} T' \\ Z' \end{pmatrix}(Y - \Pi). \tag{5.15}$$

The second step of the iterative procedure is to construct the second derivative of the log-likelihood. This is

$$J = -\ddot{\ell} = \begin{pmatrix} T' \\ Z' \end{pmatrix} W \begin{pmatrix} T & Z \end{pmatrix} = \begin{pmatrix} J_{TT} & J_{TZ} \\ J_{ZT} & J_{ZZ} \end{pmatrix}. \tag{5.16}$$

Here W is a $N \times N$ diagonal matrix with weights on the main diagonal given by $\pi_{ip}(1-\pi_{ip})$. Each of the four block matrices, J_{TT}, J_{TZ}, J_{ZT}, J_{ZZ}, is constructed

separately within the code, in preparation for the subsequent partitioned inversion of the second derivative J. These operations are performed using a vector of the diagonal elements of W.

In the third step, the inverse of the second derivative is constructed. This is done in stages by partitioned inversion so that

$$J^{-1} = \begin{pmatrix} J_{TT}^{-1} + J_{TT}^{-1} J_{TZ} J_{ZZ \cdot T}^{-1} J_{ZT} J_{TT}^{-1} & -J_{TT}^{-1} J_{TZ} J_{ZZ \cdot T}^{-1} \\ -J_{ZZ \cdot T}^{-1} J_{ZT} J_{TT}^{-1} & J_{ZZ \cdot T}^{-1} \end{pmatrix}, \tag{5.17}$$

where $J_{ZZ \cdot T} = J_{ZZ} - J_{ZT} J_{TT}^{-1} J_{TZ}$. Crucially, this has been structured in such a way that only two inversions are needed: $J_{ZZ \cdot T}^{-1}$ and J_{TT}^{-1}. Because $J_{ZZ \cdot T}$ is of dimension equal to the number of covariates, it is small and easy to invert. Since J_{TT} has inherited the diagonal structure of $T'T$, it can be inverted by simply storing the diagonal elements as a vector and taking the reciprocal of each element.

The fourth step is to combine the initial parameter values, the score, and the inverted second derivative in the Newton-Raphson updating procedure. The updated estimator is

$$(\zeta, \kappa)_{(j)} = (\zeta, \kappa)_{(j-1)} - \lambda \ddot{\ell}_{(j-1)}^{-1} \dot{\ell}_{(j-1)} \tag{5.18}$$

where $(\zeta, \kappa)_{(j)}$ is the j-th iteration of the procedure. The starting values are considered the first iteration. The parameter λ is a line search parameter, the setting of which is discussed further below.

The fifth and final step is to compare the log-likelihood associated with the parameter estimates at iteration j and the log-likelihood associated with the parameter estimates at iteration $j-1$. The logit log-likelihood is

$$\ell(\zeta, \kappa) = \sum_{i=1}^{N} \eta_{ip} y_{ip} - \sum_{i=1}^{N} \ln (1 + \exp \eta_{ip}). \tag{5.19}$$

Note that it is not necessary to sum over p here, since it is a redundant index in repeated cross section data where each individual i is observed only once. If the log-likelihood has increased from iteration $j-1$ to iteration j, then the updated parameters are adopted and the entire procedure is repeated to update from them. At the iteration where the change in the log-likelihood between $j-1$ and j is less than the pre-set tolerance value, then subject to a condition on the first derivative convergence is declared and the routine stops at those parameter estimates.

If the log-likelihood has decreased from iteration $j-1$ to iteration j, the procedure is said to have over-stepped. The line search parameter λ, which controls the "length of the step" taken in the updating algorithm, is then reduced. It is repeatedly halved until the difference in the log-likelihoods is positive, the limit on line search iterations is reached, or convergence occurs.

5.3.4 Model selection

The frameworks to estimate APC and TS models described in the preceding §5.3.3 are combined to perform model selection in the function `apc.indiv.model.table()`. This function estimates the APC acceleration model and all sub-models, as well as the TS model where that is available, and provides test statistics allowing the user to compare between them. The user can choose between likelihood ratio tests or Wald tests; where appropriate, the Akaike Information Criterion is also provided. Note that F-tests are available as a special case of Wald tests. The objective is to choose the most parsimonious model supported by the data.

The command `apc.indiv.model.table()` is run directly from the data described in §5.3.2. The user must list the outcome variable and any covariates. It is also possible to choose which models and test statistics appear in the final table, including whether or not to estimate the TS model, and whether tests should be Wald or likelihood ratio. Likelihood ratio tests will be compared to a χ^2 distribution. If Wald tests are selected, the user must specify either the χ^2 or F distribution. If an F distribution is chosen then the F-test is performed, which is a special case of the Wald test suitable for finite sample inference, as described in §5.2.2. For repeated cross section data, the assumed distribution of the outcome variable must be specified as either Gaussian or binomial. Where survey weights are used, that variable must be specified. For panel data, the distribution should be specified as Gaussian; this reflects the fact that the outcome is continuous. The individual identifier must also be specified.

An additional, supplementary command is available where it is desired to test between two models; this command is `apc.indiv.compare.direct()`. This works in a similar way to `apc.indiv.model.table()`, but the user must additionally specify the names of the two models to be compared.

5.3.5 Model interpretation

To facilitate interpretation of the estimates produced by `apc.indiv.est.model()`, the command `var.apc.plot.fit()` can be used. This produces a nine-panel graphic of the form seen in Figure 5.2 of §5.4.1.2. The top row of plots displays the sequences of estimated APC accelerations. These accelerations can be used to evaluate the impact of policy changes or shocks.

The bottom row of plots displays detrended sums of the APC accelerations. They are referred to as detrended sums because they are anchored at zero at both the first and last age (or period or cohort). This enables the user to clearly visualise the departure from linearity of the relationship between age, period, and cohort and the outcome of interest.

The central row of plots displays the linear plane that corresponds to these detrended sums of APC accelerations. This linear plane can be used in conjunction with the APC accelerations for forecasting.

5.4 Example: analysis of log wages with `apc.indiv`

I illustrate the use of the `apc.indiv` software to estimate age, period, and cohort accelerations in log wages. This is a topic of interest to economists, for which datasets already exist in the R environment. I use one repeated cross section dataset from the `ISLR` package (James et al., 2017), and one panel dataset from the `AER` package (Kleiber & Zeileis, 2008). I also illustrate the use of the `apc.indiv` software for binary outcomes by estimating an APC acceleration model for the probability that a person has an industrial (rather than an information) job, using the `ISLR` dataset.

There is a long-standing interest among economists in estimating age profiles of wages which exclude period and cohort effects. An early example is Hanoch & Honig (1985), who used a constraints-based procedure to isolate the non-linear part of the relationship between age and log earnings. Their approach was similar to that in the Stata package of Chauvel (2012), discussed in §5.3.1, but used aggregated data. Meghir & Whitehouse (1996) use a primarily graphical analysis to explore the age, period, and cohort effects on UK wages. Kalwij & Alessie (2007)

use the Deaton-Paxson constraints approach in an attempt to identify age and period effects on British log wages, and Low et al. (2010) use the same approach to estimate an age profile of earnings against which to evaluate their structural model. Lagakos et al. (2018) show the sensitivity of estimates of age-wage profiles to the constraints in the Deaton-Paxson approach, and instead carefully select constraints based on economic theory.

I examine a repeated cross section dataset from the 2000s and a panel dataset from the early 1980s. Both datasets show concavity in log wages with age. I test whether this concavity can be modelled using a quadratic, as is often done in applications, and find that the quadratic model is supported. This approximately quadratic concavity may reflect the diminishing marginal return of experience to wages, and potentially selection out of the labour force at older ages by high earners, who are more likely to have the financial resources to retire early. I also identify a period deceleration in log wages that is consistent with the 1979 oil price shock, which may have depressed real wages.

The purpose of my analysis is to illustrate the use of the `apc.indiv` software, using an example that can be easily understood and replicated. As was mentioned above, both datasets are already publicly available in the R environment and have been extensively studied. I therefore do not attempt to present novel findings, instead focusing on demonstrating how widely-accepted stylized facts (that the age profile of wages is concave, and that wages are depressed by an economic downturn) manifest in the estimated age, period, and cohort accelerations. It is hoped that these examples will guide researchers in the use of the accelerations to identify novel stylized facts in other, less well-studied settings.

5.4.1 Repeated cross section

To illustrate the use of the code for repeated cross section data, I use the `Wage` data from the `ISLR` package (James et al., 2017). This data records information about 3000 male workers in the Mid-Atlantic region of the US, and was manually assembled from the March 2011 supplement to the American Current Population Survey. I examine the age, period, and cohort accelerations lined to the log wage of these workers (a continuous outcome), and to the probability that they hold a

job classified as "industrial" rather than "information" (a binary outcome). Plots of detrended sums of age accelerations reveal a concave non-linear relationship between age and the log wage, but there is no clear shape to the non-linear relationship between period or cohort and log wage. There is a large acceleration in the probability of holding an industrial job in 2008, followed by a compensating deceleration; this may indicate a temporary surge in layoffs in response to the financial crisis that are job class-specific. This data does not contain weights and there is no evidence that the wage information has been corrected for inflation.

5.4.1.1 Data assessment and cleaning

I begin by visually examining the age-period-cohort structure of the data. There are three possible structures, as outlined in §5.3.2: age-period, period-cohort, and age-cohort. I suspect that this data will have an age-period structure, containing data on those of working age for a number of years, and so I produce an age-period array of the counts of observations. This tabulation of the data permits me to verify that age, period, and cohort are recorded at regular intervals, and that the data is contiguous. The array is produced using the following code:

```
library("plyr")
library("reshape")
library("ISLR")
data("Wage")
summary(Wage)
AP_count <- count(Wage, c("age", "year"))
AP_show <- cast(AP_count, age ~ year)
View(AP_show)
```

The object `AP_show` is a table with a column for each of the seven periods in the data and a row for each of the 61 ages, i.e. an age-period array. Each cell shows the number of observations in the data for that age-period (equivalently, age-cohort) combination. There are some cells with zero observations, but the lack of pattern to the location of these cells within the array confirms that this data has an age-period structure. Since the `apc.indiv` functions require a contiguous dataset, I

restrict the data to eliminate the cells with zero observations. For this ISLR Wage data these cells are all among the youngest and oldest ages, which are sparsely observed, and so I censor the sample to those aged between 25 and 55.

```
Wage2 <- Wage[Wage$age >= 25 & Wage$age <= 55, ]

names(Wage2)[names(Wage2) %in% c("year", "age")] <-
 c("period", "age")

cohort <- Wage2$period - Wage2$age
indust_job <- ifelse(Wage2$jobclass=="1. Industrial", 1, 0)
hasdegree <- ifelse(Wage2$education %in% c("4. College Grad",
 "5. Advanced Degree"), 1, 0)
married <- ifelse(Wage2$maritl == "2. Married", 1, 0)
Wage3 <- cbind(Wage2, cohort, indust_job, hasdegree, married)
```

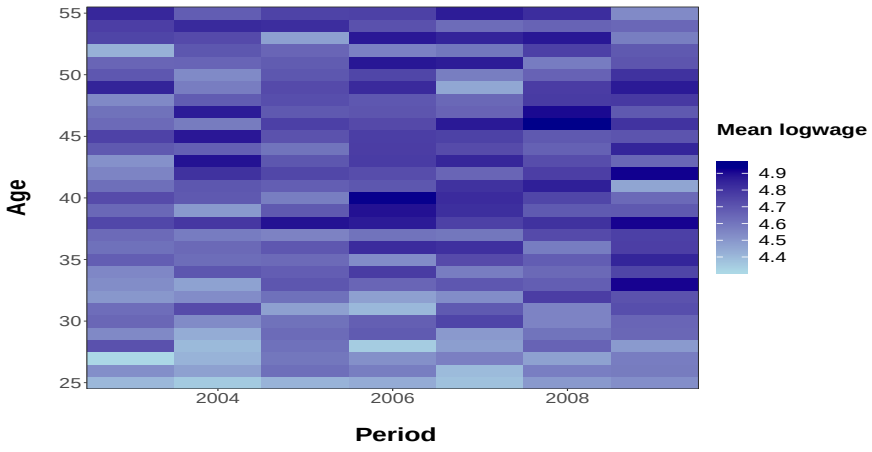
I make a number of other changes to clean the data. First, I rename the variable year to period; the apc.indiv functions require that at least two of the variables age, period, and cohort are present in the data. I also construct the variable cohort from age and period. I create indicator variables for whether a job is classified as industrial (rather than information), whether the worker has a college degree, and whether the worker is married.

I will be interested in how the wage of the worker and the classification of their job (industrial or information) is related to their age, cohort, and period of observation. Before performing a formal analysis of these relationships using the apc.indiv functions, I use a visualisation to conduct a preliminary search for patterns in log wage and the job classification along age, period, or cohort. The visualisation is constructed using ggplot2 (Wickham, 2016), as follows.

```
library("ggplot2")

mean_logwage <- ddply(Wage3, .variables=c("period", "age"),
function(dfr, colnm){mean(dfr[, colnm])}, "logwage")
names(mean_logwage)[3] <- "Mean_logwage"
```

```
plot_mean_logwage <- ggplot(mean_logwage, aes(period, age)) +
theme_bw() +
xlab('\n Period') +
ylab('Age\n') +
geom_tile(aes(fill = Mean_logwage)) +
scale_fill_gradientn(colours=c("lightblue", "darkblue"),
space = 'Lab', name="Mean logwage \n") +
scale_x_continuous(expand=c(0,0)) +
scale_y_continuous(expand=c(0,0)) +
theme(axis.text=element_text(size=18),
axis.title=element_text(size=24, face="bold"),
legend.title=element_text(size=20, face="bold"),
legend.key.size = unit(1, "cm"),
legend.text=element_text(size=18))

plot_mean_logwage
```

Figure 5.1: Mean of log wage by age and period, `ISLR Wage` data



The visualization produced by the above code is seen in Figure 5.1. Period is on the X-axis, and age is on the Y-axis, so each cell corresponds to a unique age-period combination in the data. The colour of the cell reflects the mean log wage

of the individuals in that cell. The light blue colour corresponds to a lower mean log wage and the dark blue to a higher mean log wage. There is a concentration of light blue at the bottom-left of the graph. This could be a combination of age and period effects: young people have lower wages, and in later years people have higher nominal wages (recall that the data may not be adjusted for inflation). It is unlikely to be due to cohort effects, since cohorts are counted from the top-left to the bottom-right of the graph; the prevalence of light blue cuts across cohorts.

Similar code can be used to produce an analogous graph, showing the mean value of the indicator `indust_job` in each age-period cell. That mean indicates the proportion of people in that cell whose job is classified as industrial rather than information. The graph for `indust_job` is shown in Appendix 5.A; it has a similar pattern to that for log wage, with a higher probability of being in an industrial job concentrated among the young in the early 2000s.

5.4.1.2 A model for log wages

I now use the functions from `apc.indiv` to investigate accelerations in age, period, and cohort that are identified from this data. There are three stages to the analysis. First, the appropriate model must be selected. Then, that model must be estimated. Finally, the estimates of the accelerations and other coefficients from that model must be interpreted.

Prior to the analysis, a number of packages on which `apc.indiv` depends must be loaded if they are not already in the environment.

```
library("apc")
library("plyr")
library("lmtest")
library("car")
library("plm")
library("survey")
```

In the first stage of the analysis I use `apc.indiv.model.table()`, described in §5.3.4, to select the appropriate model for this data. The relevant code is below: I specify the dataset, dependent variable, and covariates (I include an indicator for whether a person has a degree, which is expected to increase their wage).

Since the data is repeated cross section, the estimation procedure is maximum likelihood. The assumed distribution of the data must therefore be specified, which is Gaussian. It is also necessary to specify whether a Wald or Likelihood Ratio test should be used to compare models. Here, I use an F-test, which is a transformation of the Wald test that is suitable for Gaussian data with finite sample inference. This is accessed by specifying the `test` option to be `"Wald"` and the `dist` option (for distribution) to be `"F"`. I include the time-saturated (TS) model in the table by setting `TS = TRUE`. The TS model, described in §5.3.3.1, is a more general model which nests the APC acceleration model.

```
logwage_tab <- apc.indiv.model.table(Wage3, dep.var="logwage",
covariates="hasdegree",
model.family="gaussian",
test="Wald", dist="F", TS=TRUE)

View(logwage_tab$table)
```

The output of the above code is seen in Table 5.1. The left-most column indicates the name of the model under consideration. The models considered are the general TS model, the APC model with accelerations in age, period, and cohort, and various sub-models of the APC acceleration model. Details of all sub-models can be found in §4.3.2.1. Columns two through four contain the results of F-tests of these models against the TS model, while columns five through seven contain the results of F-tests of these models against the APC model. The eighth column contains the Akaike Information Criterion and the final column contains the log-likelihood.

To select a model, I first consider the Akaike Information Criterion (AIC) in the eighth column of Table 5.1. The AIC is a likelihood-based statistic which incorporates a penalty for the size of the model. It is defined as follows:

$$AIC = -2 \times \ell + 2 \times d, \tag{5.20}$$

where ℓ is the log-likelihood and d is the number of parameters to be estimated. The model with the lowest value of the AIC achieves the best balance between fit

Table 5.1: Model selection table: log wage

Model	Wald(F) vs TS	DF (*,2197)	p	Wald(F) vs APC	DF (*,2342)	p	AIC	lik
TS							1178.74	-370.37
APC	1.05	145	0.32				1050.92	-451.46
AP	1.03	180	0.38	0.94	35	0.57	1014.56	-468.28
AC	1.06	150	0.29	1.36	5	0.24	1047.90	-454.95
PC	1.14	174	0.10	1.60	29	0.02	1040.46	-475.23
Ad	1.04	185	0.35	0.98	40	0.51	1010.99	-471.50
Pd	1.24	209	0.02	1.65	64	0.00	1029.15	-504.58
Cd	1.15	179	0.10	1.55	34	0.02	1036.55	-478.27
A	1.14	186	0.11	1.43	41	0.04	1028.70	-481.35
P	1.55	210	0.00	2.64	65	0.00	1091.99	-537.00
C	1.37	180	0.00	2.67	35	0.00	1075.50	-498.75
t	1.23	214	0.01	1.61	69	0.00	1024.75	-507.38
tA	1.30	215	0.00	1.81	70	0.00	1038.18	-515.09
tP	1.54	215	0.00	2.54	70	0.00	1087.52	-539.76
tC	1.38	215	0.00	2.07	70	0.00	1055.78	-523.89
1	1.61	216	0.00	2.75	71	0.00	1102.23	-548.12

DF = degrees of freedom; p= p-value; AIC = Akaike Information Criterion; lik = log likelihood

and parsimony. For this log wage data, the AIC is minimized by the Ad model. The Ad, or age-drift, model includes accelerations in age only; accelerations in period and cohort are omitted. In addition to having the lowest AIC value, there is also support for the Ad model from the F-tests; the p-values of the F-tests of this model against the more general TS and APC models are quite large, indicating that restricting those models to the Ad model is not rejected by the data.

The second stage of the analysis is to estimate the selected Ad model, using `apc.indiv.est.model()`. Estimation requires specification of the dataset, dependent variable, covariates, and distribution of the data for maximum likelihood. It is also necessary to specify that the model to be estimated is the Ad model.

```
logwage_ad <- apc.indiv.est.model(Wage3, dep.var = "logwage",
covariates="hasdegree",
model.family="gaussian",
model.design="Ad")
```

```
View(logwage_ad$coefficients.covariates)
var.apc.plot.fit(logwage_ad, main.outer="")
```

One of the outputs of `apc.indiv.est.model()` is a matrix of the estimated covariate coefficients, called `coefficients.covariates`. Since the number of covariates in this model is small, it is easy to examine this matrix directly. As expected, the coefficient on having a degree is positive (0.285) and highly significant (p-value $\ll 0.00$).

The command `apc.indiv.est.model()` also produces a matrix of the estimated accelerations, called `coefficients.canonical`. However, the number of accelerations is typically large. In this `ISLR Wage` dataset, there are 28 age accelerations. It is easiest to interpret these accelerations by plotting them.

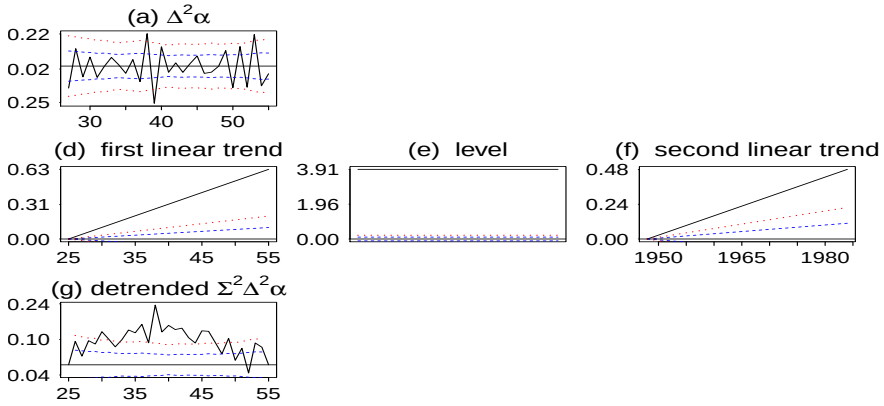
The third stage of the analysis therefore is to create a visualization which can be used to interpret the estimated accelerations. The command which produces this visualization is `var.apc.plot.fit()`. The input to this command is simply the output of `apc.indiv.est.model()`. The default output of `var.apc.plot.fit()` is seen in Figure 5.2. As outlined in §5.3.5, the top row of Figure 5.2 contains plots of the estimated accelerations, the bottom row contains plots of detrended sums of accelerations, and the middle row contains plots of the combined linear plane.

Consider first the middle row of the visualization in Figure 5.2. These plots (d) through (f) show the estimated linear plane, which combines the unidentified linear effects of age, period, and cohort. This is the "drift" part of the model. The first linear trend is plotted in the age dimension and combines the linear effects of age and period; the second linear trend is plotted in the cohort dimension and combines the linear effect of cohort and period. The plane is selected to ensure that the sums of accelerations are anchored to start and end in zero and therefore it has no natural interpretation, but it can be used in conjunction with the sums of accelerations for forecasting.

The top and bottom rows of the visualization reflect the estimated accelerations from the model. In this case there is only one plot in each row, because the model is Ad and therefore only contains age accelerations. There are empty spaces where plots of period and cohort accelerations would otherwise be. Figure 5.2(a) in the top row shows the estimated accelerations in age, of which there are 28. Figure

5.2(g) in the bottom row shows the detrended sum of accelerations, which describes the non-linear part of the relationship between age and log-wage. We see that this relationship is concave, up to some noise.

Figure 5.2: Ad model of log wage



Blue dashed, red dotted lines = 1, 2 standard deviations from zero.

The concavity of the detrended sums of age accelerations in Figure 5.2(g) may be compared with the predictions of economic theory. It is, for example, consistent with the theory that there are diminishing marginal returns to experience. It is also possible to test whether this concave relationship is approximately quadratic. A quadratic relationship between age and log wage corresponds to a constant age acceleration, so

$$\forall a \Delta^2 \alpha_a = \lambda, \tag{5.21}$$

as mentioned in §5.2.4 and explained more thoroughly in §4.4.4.1. A test for constant age accelerations can be performed as follows, using the `linearHypothesis` function from the package `car` (Fox & Weisberg, 2019).

```
allageDD <- rownames(logwage_ad$coefficients.canonical)[grep(
"DD_age", rownames(logwage_ad$coefficients.canonical))]

ageDD1 <- allageDD[-1]
ageDD2 <- allageDD[-length(allageDD)]
quadratic_hyp <- paste(ageDD2, ageDD1, sep = " = ")
```

```
rm(list=ls(pattern="ageDD"))
```

```
linearHypothesis(logwage_ad$fit, quadratic_hyp, test="F")
```

The test is performed using a F-statistic, which is compared to an F distribution. The resulting test statistic of 1.297, when compared to an $F(28, 2382)$ distribution, has a p-value of 0.14. This indicates that the hypothesis of a quadratic relationship between the age of the worker and their log wage cannot be rejected.

5.4.1.3 A model for a binary outcome variable

The `apc.indiv` functions can be used to investigate the relationship between a binary variable and age, period, and cohort. I illustrate this with a model for whether or not the worker has an industrial job. As was the case in §5.4.1.2, the first stage in the analysis is model selection. I use `apc.indiv.model.table()` to produce a table comparing the TS model, the APC acceleration model, and sub-models of the APC acceleration model. For binary outcomes, the binomial distribution is used for the maximum likelihood procedure. Likelihood ratio tests, evaluated against a χ^2 distribution, are used to compare models.

```
indust_job_tab <- apc.indiv.model.table(Wage3, dep.var="indust_job",
covariates="hasdegree",
model.family="binomial",
test="LR", dist="Chisq", TS=TRUE)
```

```
View(indust_job_tab$table)
```

There are two parts to the output of `apc.indiv.model.table()` in this analysis. The first part is the table, which I discuss further below. The second is a report on the behaviour of the custom Newton-Raphson algorithm used to estimate the TS model, which was described in §5.3.3.3. The report is a list, called `NR.report`. The most important element of this list is called `result`, which indicates whether the TS model has converged or not. If convergence is not reported, the parameters of the Newton-Raphson algorithm should be modified using the

Table 5.2: Model selection table: having an industrial job

Model	LR-test vs TS	DF	p	LR-test vs APC	DF	p	AIC	lik
TS							3292.00	-1428.00
APC	169.95	145	0.08				3171.96	-1512.98
AP	235.34	180	0.00	65.39	35	0.00	3167.35	-1545.67
AC	179.29	150	0.05	9.33	5	0.10	3171.29	-1517.64
PC	206.75	174	0.04	36.80	29	0.15	3150.76	-1531.38
Ad	245.44	185	0.00	75.49	40	0.00	3167.44	-1550.72
Pd	271.02	209	0.00	101.06	64	0.00	3145.02	-1563.51
Cd	216.14	179	0.03	46.19	34	0.08	3150.14	-1536.07
A	245.46	186	0.00	75.51	41	0.00	3165.47	-1550.73
P	275.27	210	0.00	105.32	65	0.00	3147.28	-1565.64
C	216.34	180	0.03	46.39	35	0.09	3148.35	-1536.17
t	280.95	214	0.00	111.00	69	0.00	3144.95	-1568.48
tA	281.18	215	0.00	111.23	70	0.00	3143.19	-1568.59
tP	285.61	215	0.00	115.65	70	0.00	3147.61	-1570.81
tC	280.95	215	0.00	111.00	70	0.00	3142.96	-1568.48
1	285.78	216	0.00	115.83	71	0.00	3145.79	-1570.89

DF = degrees of freedom; p= p-value; AIC = Akaike Information Criterion; lik = log likelihood

option `NR.controls` in `apc.indiv.model.table()` until convergence is achieved - for example, by increasing the number of iterations.

The output of `apc.indiv.model.table()` is shown in Table 5.2. It has a similar structure to Table 5.1: the first column contains the model name, the next three columns report tests of that model against the TS model, the next three report tests of that model against the APC model, and finally there is the AIC and the log-likelihood. Again I begin by considering the AIC in the eighth column. The model which minimizes AIC is the tC model.

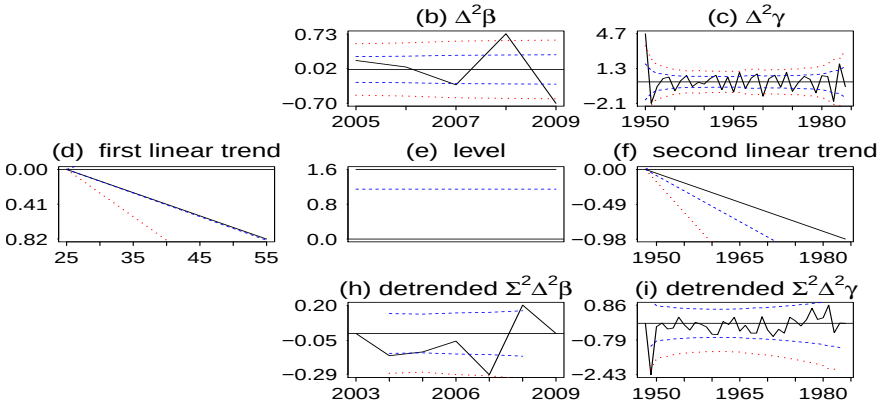
It is difficult to determine from Table 5.2 which model best fits the data. Although the AIC is minimized by the tC model, the likelihood ratio tests comparing the tC model to the TS and APC models reject the restriction. There is no model on which the AIC and likelihood ratio tests agree. In fact, the low p-value of the likelihood ratio test comparing the APC model to the TS model ($p = 0.08$) suggests that the APC model and its sub-models are close to rejection as restrictions on the TS model. Therefore this class of models may not be well-suited to the

data; alternative restrictions on the TS model should be considered.

In order to illustrate the estimation and interpretation of a model with a binomial outcome, I continue the analysis with the PC model. I choose this because it has one of the lower AIC values, is the most supported sub-model against the APC model, and is almost supported against the TS model. The model is estimated using `apc.indiv.est.model()`.

```
indust_job_pc <- apc.indiv.est.model(Wage3, dep.var="indust_job",
covariates="hasdegree",
model.family="binomial",
model.design="PC")

var.apc.plot.fit(indust_job_pc)
View(indust_job_pc$coefficients.covariates)
```

Figure 5.3: PC model of industrial job



Blue dashed, red dotted lines = 1, 2 standard deviations from zero.

I create a visualization of the period and cohort accelerations of this PC model using `var.apc.plot.fit()`. The visualization is seen in Figure 5.3. This has the same structure as Figure 5.2; the top panels show the accelerations, the bottom panels show detrended sums of APC accelerations, and the middle panels show the linear plane. Considering Figure 5.3(b), it can be seen that there is a large acceleration in the probability of having an industrial job in 2008. This may reflect

a streamlining of operations during the financial crisis, eliminating administrative positions classified as "information". There is no clear pattern in the cohort accelerations. The effect of having a degree on the probability of having an industrial job is, unsurprisingly, significant and negative.

5.4.1.4 Extensions

The analysis of log wage could be extended to account for survey weights, although none are provided in the `ISLR Wage` dataset. Survey weights are incorporated by specifying the name of the weight variable using the option `wt.var` in `apc.indiv` commands. Models incorporating survey weights are not estimated by maximum likelihood, so the likelihood column is omitted from the model selection table, and Wald tests rather than likelihood ratio tests are used. A psuedo-AIC is reported; see Lumley & Scott (2017) for details. The TS model is not currently available for data with survey weights.

A second extension relates to instability that can arise in the model due to infrequently observed ages, periods, or cohorts. In age-period data such as the `ISLR Wage` data, the earliest and latest cohorts are observed in only one age-period cell. This can lead to very large estimates of accelerations for these cohorts as they pick up features unique to those age-period cells. An example of this can be seen in the PC model for having an industrial job, where the estimated acceleration for the earliest cohort is very large compared to the other cohort accelerations. This problem can be addressed by "censoring" those early and late cohorts out of the data, by first dropping them from the data using standard R techniques and then specifying the options `n.coh.excl.start` and `n.coh.excl.end` in the `apc.indiv` functions. For age-period data this problem arises in cohort, for period-cohort data it arises in age, and for age-cohort data it arises in period. Similar censoring options are available for these formats.

5.4.2 Panel

To illustrate the application of `apc.indiv` to panel data, I build a model for log wages using the `PSID7682` data from the `AER` package (Kleiber & Zeileis, 2008). `PSID7682` is an excerpt from the Panel Survey of Income Dynamics that covers 595

individuals over a seven-year period from 1976-1982. The dataset was originally constructed by Cornwell & Rupert (1988), who used it to estimate the effect of education on log wages. That analysis has been replicated in economics textbooks such as Baltagi (2005) (Example 7.5) and Greene (2008) (Example 11.5). The data is not adjusted for inflation. This is known because Cornwell & Rupert (1988) control for price level effects in their analysis by including year indicators. Note that in this data the variables affected by the identification problem are not age, period, and cohort, but rather years of work experience, period, and year of entering the workforce. The existence of an APC-style identification problem among these three variables is discussed in Heckman & Robb (1985).

My models for log wages are comparable to those in Baltagi (2005) and Greene (2008). I use the same covariates, but replace the time effect variables (experience, the square of experience, and period indicators) with the APC structure $\mu_{ip} = x'_{ip}\xi$. As mentioned above, experience replaces age in this model. I find that the covariate estimates are robust to the change in specification of the time effects. This is not surprising, as a test of the quadratic restriction on the effect of experience is not rejected. The failure to reject this test indicates that the original specification of the time effects used by Baltagi and by Greene is valid.

5.4.2.1 Data assessment and cleaning

I begin by visually exploring the data. I create an array to show the number of observations in each age-cohort cell present in the data. I initially displayed the data in an age-period array, as was done for the repeated cross section data in §5.4.1.1. However the pattern of cells with zero observation counts in the corners of the table indicated that this data had a period-cohort rather than an age-period structure. I therefore created a period-cohort array instead, i.e. an array of observation counts by period and cohort.

```
library("AER")
data("PSID7682")
summary(PSID7682)

AP_count <- count(PSID7682, c("experience", "year"))
```

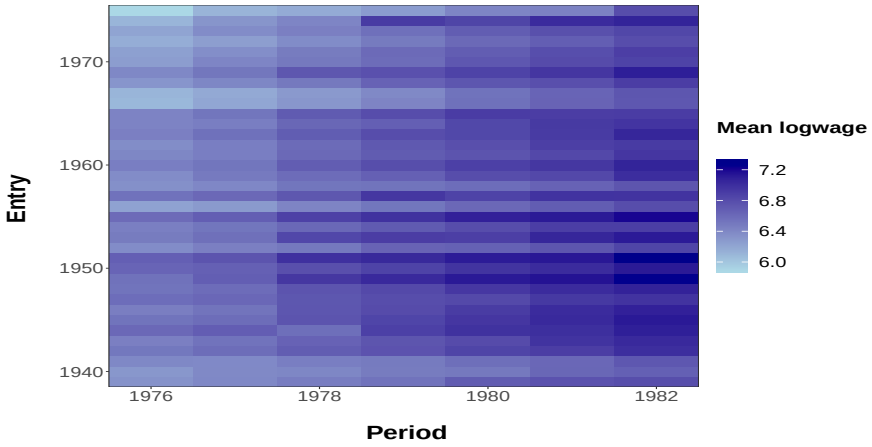
```
AP_show <- cast(AP_count, experience~year)
View(AP_show)
# the missing corners of the data show that this is actually
# period-cohort data.

period <- as.numeric(PSID7682$year) + 1975
entry <- period - PSID7682$experience
psid <- cbind(PSID7682, period, entry)

CP_count <- count(psid, c("entry", "year"))
CP_show <- cast(CP_count, entry~year)
View(CP_show)
```

It is easily seen from the array of observation counts by cell, `CP_show`, that this is a balanced panel; the number of observations in a given cohort does not change over period. I use `CP_show` to determine how the data should be restricted to eliminate cells with zero observations. In this case, I omit the oldest cohorts. I construct a set of indicator variables for use as covariates, and rename the variables corresponding to `age`, `period`, and `cohort`.

```
psid2 <- psid[psid$entry >= 1939, ]

# construct outcome and covariates
logwage <- log(psid2$wage)
inunion <- ifelse(psid2$union == "yes", 1, 0)
insouth <- ifelse(psid2$south == "yes", 1, 0)
bluecollar <- ifelse(psid2$occupation == "blue", 1, 0)
ismarried <- ifelse(psid2$married == "yes", 1, 0)
incity <- ifelse(psid2$smsa == "yes", 1, 0)
black <- ifelse(psid2$ethnicity == "afam", 1, 0)
isfem <- ifelse(psid2$gender == "female", 1, 0)
ismanuf <- ifelse(psid2$industry == "yes",1 , 0)

psid3 <- cbind(psid2, logwage, inunion, insouth, bluecollar,
```

```
ismarried, incity, black, isfem, ismanuf)

names(psid3)[names(psid3) %in% c("experience", "entry")] <-
c("age", "cohort")
```

Figure 5.4: Mean of log wage by experience and period, `AER PSID7682` data



I use a visualization to check for age, period, and cohort patterns in the outcome of interest, log wage. This is produced using `ggplot2`, and is similar to Figure 5.1 in §5.4.1.1. Since this is period-cohort data, I plot cohort (year of entry) instead of age (experience) on the Y-axis. The code is in Appendix 5.A.2. The visualization is seen in Figure 5.4. There is a clear period effect; the colour becomes more dark blue towards the right of the graph, indicating higher wages in later years. There is also evidence of cohort (year of entry) effects, appearing as horizontal bands of colour. Those starting work around 1968, for instance, appear to have lower wages throughout their lives. Finally, the predominance of light blue in the top-left corner of the graph suggests an age (experience) effect; unsurprisingly, lack of experience corresponds to low wages.

5.4.2.2 The random effects setting without covariates

The first step when building a panel data model is to determine which panel setting is appropriate: pooled OLS, random effects, or fixed effects. It seemed necessary in this dataset to account for correlation in the error terms of a given individual

over time, due to the importance of unmeasurable factors such as intelligence or diligence to log wages. The random effects setting accounts for this kind of correlation more efficiently than does pooled OLS. The additional requirement of random effects relative to pooled OLS, that the error terms be strictly exogenous with respect to the APC explanatory variables, was shown to be no stronger than the pooled OLS contemporaneous exogeneity requirement where only age, period, and cohort variables are included in the model (see §4.3.4.1). Therefore I select a random effects setting with no covariates as a reasonable starting point.

Having selected the random effects setting, I proceed to model selection using `apc.indiv.model.table()`. The inputs to this command are largely the same as they were for repeated cross section analysis. There are two additional inputs: the panel setting must be specified, using the option `plmmodel`, and the individual identifier must be specified using `id.var`.

Note that I use Wald tests here with asymptotic inference (comparison to a χ^2 distribution), since finite sample inference is not possible under the assumptions of the GLS framework used to estimate panel models.

```
panel_tab <- apc.indiv.model.table(psid3, dep.var="logwage",
model.family = "gaussian",
test="Wald", dist="Chisq",
plmmodel="random", id.var="id")

View(panel_tab$table)
```

The output of `apc.indiv.model.table()` is seen in Table 5.3. The first column is the list of models that are considered. Note that the TS model is not currently implemented for panel data. Further, some of the sub-models seen in previous tables do not appear here. Those are: the C, tC, and 1 models. This is because random effects estimation requires at least one explanatory variable which changes over time within an individual, and these models do not satisfy this requirement if no time-varying covariates are included in the model; see §4.3.2. The second through fourth columns contain the results of Wald tests against the APC acceleration model. Since the panel data models are estimated by least squares

Table 5.3: Model selection table: log wage, random effects

Model	Wald (Chisq) vs APC	DF	p
AP	69.10	35	0.00
AC	31.50	5	0.00
PC	106.31	41	0.00
Ad	100.42	40	0.00
Pd	181.51	76	0.00
Cd	150.80	46	0.00
A	1972.36	41	0.00
P	208.01	77	0.00
t	225.99	81	0.00
tA	2438.06	82	0.00
tP	252.49	82	0.00

DF = degrees of freedom; p= p-value

rather than maximum likelihood, the AIC and log-likelihood are not produced. Therefore model selection is by Wald test only.

It is clear from Table 5.3 that all of the restrictions of the APC model are rejected by the Wald tests. The model selected by this analysis is therefore the APC model. I proceed to estimate that model using `apc.indiv.est.model()`.

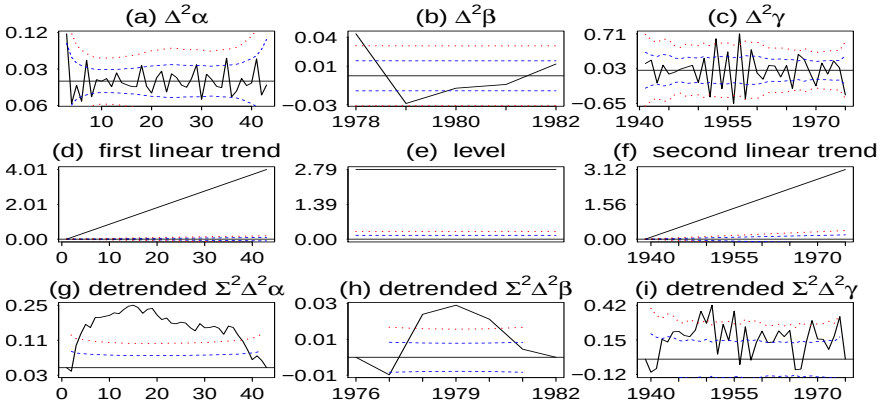
```
panel_apc <- apc.indiv.est.model(psid3, dep.var="logwage",
model.family="gaussian",
plmmodel="random", id.var="id")
```

```
var.apc.plot.fit(panel_apc)
```

As there are no covariates in the model, the only parameters to be interpreted are those relating to age, period, and cohort. The `var.apc.plot.fit()` command is used to facilitate this interpretation. The output of this command applied to the APC random effects model for log wages is seen in Figure 5.5. The detrended sums of accelerations in the bottom row reveal concavity in both age and period. The age concavity could be explained by diminishing marginal returns to experience. The period concavity may indicate an increasingly gradual return of log wages to

their previous trajectory following the economic crisis of 1979. There is no clear pattern to the non-linearity in cohort.

Figure 5.5: APC random effects model for log wage, no covariates



Blue dashed, red dotted lines = 1, 2 standard deviations from zero.

5.4.2.3 The random effects setting with covariates

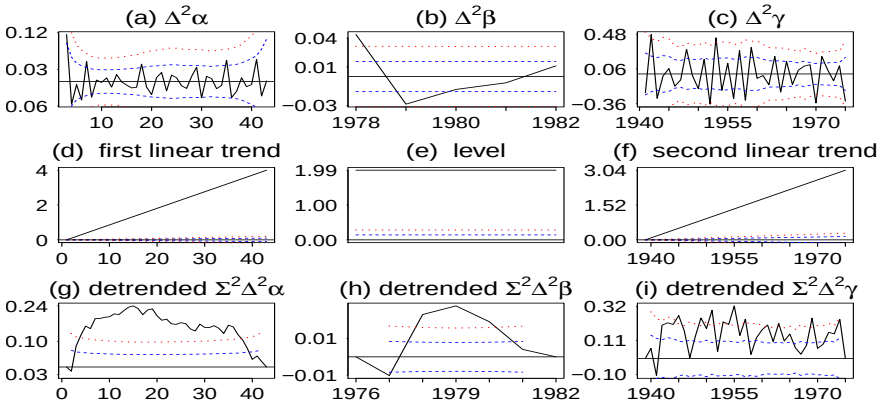
I now introduce covariates to the model. The covariates are: a continuous variable recording the number of weeks worked in the year, an indicator for holding a blue-collar job, an indicator for working in a manufacturing industry, an indicator for being located in the southern USA, an indicator for being in a city (i.e. a metropolitan statistical area), an indicator for being married, an indicator for union membership, an indicator for sex, a continuous variable recording the number of years of education, and an indicator for being Black. These are the same set covariates used by Cornwell & Rupert (1988), Baltagi (2005), and Greene (2008). They are included by specifying the `covariates` option inside `apc.indiv.model.table()` and `apc.indiv.est.model()`.

```
all_covs <- c("weeks", "bluecollar", "ismanuf", "insouth",
"incity", "ismarried", "inunion", "isfem", "education", "black")

panel_tab_cov <- apc.indiv.model.table(psid3, dep.var="logwage",
covariates = all_covs,
```

```
model.family = "gaussian",
test="Wald", dist="Chisq",
plmmodel="random", id.var="id")

View(panel_tab_cov$table)

panel_apc_cov <- apc.indiv.est.model(psid3, dep.var="logwage",
covariates = all_covs,
model.family = "gaussian",
plmmodel="random", id.var="id")

var.apc.plot.fit(panel_apc_cov)

View(panel_apc_cov$coefficients.covariates)
```

The table comparing the APC models and sub-models is very similar to Table 5.3 and so is not shown. It also results in the same conclusion; the full APC model is needed to describe the temporal variation in the data. That model is estimated as `panel_apc_cov` in the above code. The estimated APC effects are seen in Figure 5.6. These are very similar to those in Figure 5.5, indicating that the addition of covariates had little impact on the APC effects.

Figure 5.6: APC random effects model for log wage, with covariates



Blue dashed, red dotted lines = 1, 2 standard deviations from zero.

The estimated effects of the covariates are seen in Table 5.4. These can be compared to the estimated effects under the header "GLS" in Table 7.4 of Baltagi (2005). My estimates are very similar to those of Baltagi, indicating that the covariate coefficient estimates are robust to the change in specification of the time effects. This is unsurprising since the introduction of covariates had little effect on the estimated time effects; it appears that the two are largely orthogonal in terms of their influence on log wages.

Table 5.4: Covariate coefficients for random effects model of log wage

Variable	Estimate	Std. Error	z-value	Pr(>|z|)
weeks	0.001	0.001	1.621	0.105
education	0.067	0.005	14.481	0.000
inunion	0.037	0.013	2.735	0.006
insouth	-0.055	0.021	-2.645	0.008
bluecollar	-0.044	0.013	-3.394	0.001
ismarried	-0.015	0.018	-0.833	0.405
incity	0.043	0.016	2.788	0.005
black	-0.126	0.046	-2.748	0.006
isfem	-0.382	0.041	-9.224	0.000
ismanuf	0.033	0.014	2.467	0.014

5.4.2.4 The fixed effects setting with covariates

The fixed effects setting also accounts for correlation of the error terms between observations on the same individual. This setting is less restrictive than the random effects setting in how it models that correlation; see §4 for further details. The cost of this less restrictive approach is the loss of identification of coefficients on time-invariant variables, including cohort accelerations. Since there was no clear pattern in the cohort accelerations when using the random effects setting, it seems worthwhile to consider the fixed effects setting.

The first step is model selection, performed using `apc.indiv.model.table()`. The fact that the coefficients of time-invariant variables are no longer identified in the fixed effects setting changes the set of models available for the age, period,

and cohort effects. The cohort accelerations are no longer identified, and nor is the slope of the linear plane that lies in the cohort dimension of age-cohort space. The set of available models under fixed effects are as follows: FAP, FA, FP, Ft. These stand for "fixed effects with age and period accelerations", "fixed effects with age accelerations", "fixed effects with period accelerations", and "fixed effects with trend". Note that FAP, FA, and FP all also contain the single linear trend that can be identified in these models, which is represented in the age dimension and combines the linear effects of age and period.

```
panel_tab_fe <- apc.indiv.model.table(psid3, dep.var="logwage",
covariates = all_covs,
model.family = "gaussian",
test="Wald", dist="Chisq",
plmmodel="within", id.var="id")

View(panel_tab_fe$table)

panel_fap <- apc.indiv.est.model(psid3, dep.var="logwage",
covariates = c("weeks", "bluecollar",
"ismanuf", "insouth", "incity",
"ismarried", "inunion", "isfem",
"education", "black"),
model.family = "gaussian",
plmmodel="within", id.var="id",
model.design="FAP")

var.apc.plot.fit(panel_fap)

View(panel_fap$coefficients.covariates)
```

The output of `apc.indiv.model.table()` has the same structure as the output of that command in all examples thus far. The first column contains the list of models that are estimated. As was noted in the discussion on random effects, the TS model is not available for panel data. Columns two through four contain the

results of Wald tests against the most general FAP model. Since the models are estimated by least squares, no AIC or log-likelihood is available.

The model selection table for the log wage data under fixed effects is shown in Table 5.5. All standard sub-models of the FAP model are rejected.

Table 5.5: Model selection table: log wage, fixed effects

Model	Wald (Chisq) vs FAP	DF	p
FA	29.68	5	0
FP	105.00	41	0
Ft	146.80	46	0

DF = degrees of freedom; p= p-value

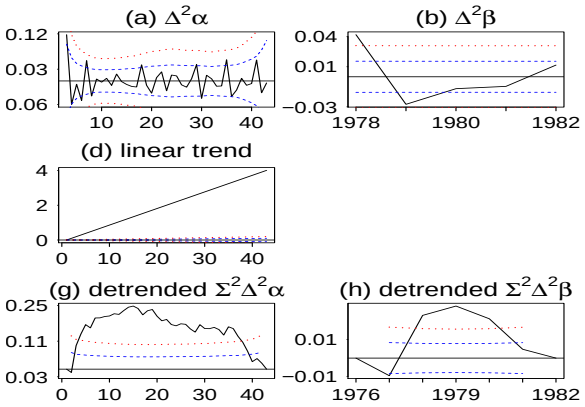
The next step of the analysis is estimation of the selected FAP model. This is done using `apc.indiv.est.model()`. The estimates of covariate coefficients from this model appear in Table 5.6. They are comparable to the estimates under the heading "Within" in Table 7.4 of Baltagi (2005) and to the estimates under the heading "Time and Ind. Effects" in Table 11.5 of Greene (2008). The effect of being in a city is significant and negative, while union membership is close to attaining significance and is positive. The remaining covariates are insignificant. These findings are consistent with those of Baltagi and Greene.

Interpretation of the estimated age and period effects relies on a visualization produced by `var.apc.plot.fit()`. This visualization appears in Figure 5.7. The age and period accelerations are largely unchanged from the random effects setting.

Table 5.6: Covariate coefficients for fixed effects model of log wage

Model	Estimate	Std. Error	t-value	Pr(>|t|)
weeks	0.001	0.001	1.217	0.224
inunion	0.028	0.015	1.838	0.066
insouth	0.008	0.034	0.245	0.806
bluecollar	-0.020	0.014	-1.397	0.162
ismarried	-0.031	0.019	-1.613	0.107
incity	-0.046	0.019	-2.378	0.017
ismanuf	0.026	0.016	1.634	0.102

Figure 5.7: FAP fixed effects model for log wage, with covariates



Blue dashed, red dotted lines = 1, 2 standard deviations from zero.

5.4.2.5 Extensions

In addition to the analysis under random and fixed effects settings illustrated here, it is also possible to estimate panel data using the pooled OLS setting. To access the pooled OLS setting, the option `plmmodel` should be specified as `pooling`, and the individual identifier must be specified. To conduct inference in the pooled OLS setting it is necessary to allow for the general variance-covariance matrix that was mentioned in §5.2.2. This is done as follows:

```
pool_apc <- apc.indiv.est.model(psid3, dep.var="logwage",
model.family="gaussian",
plmmodel="pooling", id.var="id")

# Inference with appropriate standard errors
coeftest(pool_apc$fit, vcov=vcovHC(pool_apc$fit, method="arellano",
type="HC0"))
```

The command `coeftest()`, from package `lmtest`, produces a table containing the estimate, standard error, t-statistic and p-value for each coefficient in the model. The standard errors displayed in `var.apc.plot.fit()` do not account for this general variance-covariance matrix and so should not be relied upon for inference.

As seen in the above example, the output from `apc.indiv.est.model()` is compatible with post-estimation functions from other packages, such as `lmtest`. This is true for all three panel settings. There are a number of useful functions for panel data in the `plm` package with which the output is also compatible. For example, the command `phtest()` can be used to perform a Hausman test. The Hausman test is used to compare the fixed effects and random effects estimators; if the test is rejected, one of the two estimators is inconsistent. Usually this is taken to indicate that there is correlation between some explanatory variable and the unobserved component ω_i, so the random effects estimator is inconsistent and the fixed effects estimator should be used. An example of the use of the Hausman test is given below:

```
phtest(panel_apc_cov$fit, panel_fap$fit, test="Chisq")
```

The output of the above is

```
Hausman Test


data:  model.formula
chisq = 112.07, df = 54, p-value = 6.002e-06
alternative hypothesis: one model is inconsistent
```

The results of this test favour the fixed effects model in §5.4.2.4 over the random effects model in §5.4.2.3. Again, this finding is consistent with the previous literature using this dataset.

The censoring of cohorts, ages, or periods to improve the stability of estimates, described in §5.4.1.4, is available for panel data.

5.5 Conclusion

The `apc.indiv` package for R expands the set of tools available to estimate age, period, and cohort (APC) effects from individual-level data. To my knowledge, it is the only existing software which provides estimates of APC accelerations from individual-level data. The accelerations are estimated as parameters from regression, so it is easy to conduct hypothesis tests and account for covariates.

The `apc.indiv` package focuses on APC accelerations because, unlike linear APC effects, they can be estimated without imposing untestable assumptions. Other existing software for APC analysis imposes untestable assumptions in order to estimate linear APC effects, but this creates a risk of misinterpretation if the sensitivity of the estimated APC effects to those assumptions is not recognised. The focus on accelerations in `apc.indiv` avoids this risk of misinterpretation. The accelerations can be used to evaluate the impact of shocks or policy changes, to test functional form restrictions implied by economic theory, and to explore the shape of the relationship between age, period, and cohort and an outcome of interest.

There are three core functions in `apc.indiv` which reflect the three stages of econometric analysis: model estimation, selection, and interpretation. The first function, `apc.indiv.est.model()`, provides for estimation of an APC acceleration model. It also permits estimation of sub-models, which restrict some of the accelerations to be zero, and a more general model against which the APC model can be tested. The second function, `apc.indiv.model.table()`, creates a table of test statistics which can be used to select between these models. Once a model has been selected and estimated, the third function `var.apc.plot.fit()` can be used to visualize and evaluate the APC accelerations.

I demonstrate the use of the `apc.indiv` functions with an example from labour economics. I use datasets from the R packages `ISLR` and `AER` to examine log wages and the probability of having an industrial job. I find a concave shape of log wages over the life cycle, as well as evidence of the detrimental effects of the 1979 and 2008 economic crises on employment conditions. These are not novel findings, but serve to show how stylized facts can be detected using the APC acceleration framework and `apc.indiv` software. In future it is hoped that this analysis may serve as a template for analysis of other, less well-studied datasets.

There is scope for substantial future development of the package. A natural first step would be to extend the broader capabilities that exist for repeated cross section data to panel data, allowing for binary outcomes and implementing estimation of the more general time-saturated model. It should be possible to extend the theoretical framework to allow for a wider variety of outcome variables, such as censored or count outcomes. Further options to incorporate testing of the covariates and interactions between covariates and the APC effects could be added.

5.A Additional data visualisations

5.A.1 Repeated cross section

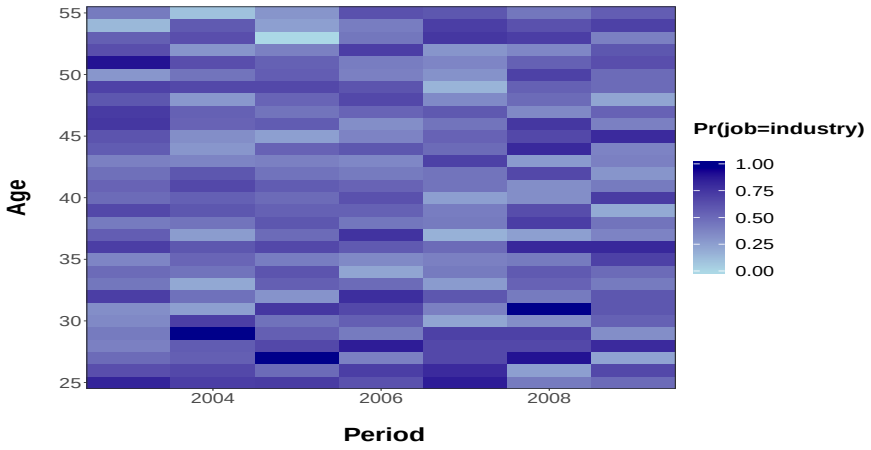
The code below produces Figure 5.8, showing the mean value of the indicator `indust_job` in each age-period cell of the repeated cross section data. That mean indicates the proportion of people in that cell whose job is classified as industrial rather than information. The graph has a similar pattern to that for log wage seen in Figure 5.1, with a higher probability of being in an industrial job concentrated among the young in the early 2000s.

```
mean_indust_job <- ddply(Wage3, .variables=c("period", "age"),
function(dfr, colnm){mean(dfr[, colnm])}, "indust_job")
names(mean_indust_job)[3] <- "Mean_indust_job"

plot_mean_indust_job <- ggplot(mean_indust_job, aes(period, age)) +
theme_bw() +
xlab('\n Period') +
ylab('Age\n') +
geom_tile(aes(fill = Mean_indust_job)) +
scale_fill_gradientn(colours=c("lightblue", "darkblue"),
space = 'Lab', name="Pr(job=industry) \n") +
scale_x_continuous(expand=c(0,0)) +
scale_y_continuous(expand=c(0,0)) +
theme(axis.text=element_text(size=18),
axis.title=element_text(size=24, face="bold"),
legend.title=element_text(size=20, face="bold"),
legend.key.size = unit(1, "cm"),
legend.text=element_text(size=18))
plot_mean_indust_job
```

5.A.2 Panel

The code below produces the visualisation in Figure 5.4.

Figure 5.8: Mean of industrial job by age and period, `ISLR Wage` data




```
# visualise the data
mean_logwage <- ddply(psid3, .variables=c("period", "cohort"),
function(dfr, colnm){mean(dfr[, colnm])}, "logwage")
names(mean_logwage)[3] <- "Mean_logwage"

plot_mean_logwage <- ggplot(mean_logwage, aes(period, cohort)) +
theme_bw() +
xlab('\n Period') +
ylab('Entry \n') +
geom_tile(aes(fill = Mean_logwage)) +
scale_fill_gradientn(colours=c("lightblue", "darkblue"),
space = 'Lab', name="Mean logwage \n") +
scale_x_continuous(expand=c(0,0)) +
scale_y_continuous(expand=c(0,0)) +
theme(axis.text=element_text(size=18),
axis.title=element_text(size=24, face="bold"),
legend.title=element_text(size=20, face="bold"),
legend.key.size = unit(1, "cm"),
legend.text=element_text(size=18))
plot_mean_logwage
```

Chapter 6

Conclusion

In this thesis, I developed a regression framework to estimate age, period, and cohort accelerations from individual-level data. This framework relies on accelerations because, unlike the levels and slopes associated with age, period, and cohort, they are identified without the imposition of untestable constraints. The framework is suitable for both repeated cross section and panel data.

In §2, I provided the theoretical background to the acceleration-based regression framework. I outlined the well-known age-period-cohort identification problem, and showed that accelerations are known to be unaffected by this problem. I gave a detailed account of the acceleration-based reparametrization of the classical age-period-cohort model developed by Kuang et al. (2008), which forms the basis of the regression framework developed in this thesis.

The regression framework for repeated cross section data was developed in §3. I embedded the acceleration-based parametrization of Kuang et al. (2008) in a generalized linear modelling framework, and accounted for the inclusion of covariates. I developed a test of this model against a more general time-saturated model. I used these tools to detect a previously unreported concavity in the relationship between cohort and obesity among English men.

The regression framework for panel data was developed in §4. I explored the implications of using the acceleration-based parametrization of age, period, and cohort effects in three commonly-used panel settings: pooled OLS, random effects, and fixed effects. I showed that the choice between the three panel settings need not be affected by concerns about correlation between APC variables and time-

invariant unobservables, since this correlation is ruled out by the deterministic nature of the APC variables. Careful consideration of the relationship between hospital in-patient stays and age using this regression framework revealed the importance of accounting for childbirth in any model of in-patient stays.

In §5 I developed an R package to implement the repeated cross section and panel data analytical framework described in the preceding chapters. This R package was designed to be easy to use: it has three core functions for model estimation, selection, and interpretation. It is compatible with existing R packages including those for hypothesis testing. I showed that the package can be used in conjunction with existing data in the R environment to recover familiar results regarding the life-cycle profile of log wages and the impact of economic downturns on the labour market.

It is hoped that the framework developed in this thesis to estimate APC accelerations from regressions using individual-level data will prove useful for economists and other social scientists dealing with age, period, and cohort effects.

Bibliography

Agresti, A. (2013). *Categorical Data Analysis* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Agyemang, C., Kunst, A., Bhopal, R., Zaninotto, P., Nazroo, J., M., N., Unwin, N., van Valkengoed, I., Redekop, K., & Stronks, K. (2015). Dutch versus English advantage in the epidemic of central and generalised obesity is not shared by ethnic minority groups: comparative secondary analysis of cross-sectional data. *International Journal of Obesity*, *35*, 1334–1346.

Akbartabartoori, M., Lean, M. E. J., & Hankey, C. R. (2005). Relationships between cigarette smoking and body shape. *International Journal of Obesity*, *29*, 236–243.

Allman-Farinelli, M. A., Chey, T., Bauman, A. E., Gill, T., & James, W. P. T. (2008). Age, period and birth cohort effects on prevalence of overweight and obesity in Australian adults from 1990 to 2000. *European Journal of Clinical Nutrition*, *62*, 898–907.

Almond, D. (2006). Is the 1918 influenza pandemic over? Long-term effects of *in utero* influenza exposure in the post-1940 U.S. population. *Journal of Political Economy*, *114*, 672–712.

An, R. & Xiang, X. (2016). Age-period-cohort analyses of obesity prevalence in US adults. *Public Health*, *141*, 163–169.

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, *49*, 431–434.

Attanasio, O. P. (1998). Cohort analysis of saving behavior by US households. *Journal of Human Resources*, *33*, 575–609.

Baltagi, B. H. (2005). *Econometric Analysis of Panel Data* (3rd ed.). Chichester, England: John Wiley & Sons.

Bardazzi, R. & Pazienza, M. G. (2018). Ageing and private transport fuel expenditure: Do generations matter? *Energy Policy*, *117*, 396–405.

Baum II, C. L. & Ruhm, C. J. (2009). Age, socioeconomic status and obesity growth. *Journal of health economics*, *28*, 635–648.

Bell, A. & Jones, K. (2014). Don't birth cohorts matter? A commentary and simulation exercise on Reither, Hauser, and Yang's (2009) age-period-cohort study of obesity. *Social Science & Medicine*, *101*, 176–180.

Bíró, A. (2017). Effect of ageing on the ownership of durable goods. *Scottish Journal of Political Economy*, *64*, 501–529.

Blanchflower, D. G. & Oswald, A. J. (2008). Is well-being U-shaped over the life cycle? *Social Science and Medicine*, *66*, 1733–1749.

Browning, M., Crossley, T. F., & Lührmann, M. (2016). Durable purchases over the later life cycle. *Oxford Bulletin of Economics and Statistics*, *78*, 145–169.

Cameron, A. C. & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Carstensen, B. (2007). Age-period-cohort models for the Lexis diagram. *Statistics in Medicine*, *26*, 3018–3045.

Carstensen, B. (2013). BMI trends in Australia – population surveys. Technical report, Steno Diabetes Center, Gentofte, Denmark & Department of biostatistics, University of Copenhangen.

Chauvel, L. (2012). *apcd: Stata module for estimating age-period-cohort effects with detrended coefficients*. Statistical Software Components.

Clark, B., Chatterjee, K., Martin, A., & Davis, A. (2019). How commuting affects subjective wellbeing. *Transportation*, 1–29.

Clayton, D. & Schifflers, E. (1987a). Models for temporal variation in cancer rates. II: age–period–cohort models. *Statistics in Medicine*, *6*, 469–481.

Clayton, D. & Schifflers, E. (1987b). Models for tempral variation in cancer rates. II Age-period-cohort models. *Statistics in Medicine*, *6*, 469–481.

Cornwell, C. & Rupert, P. (1988). Efficient estimation with panel data: An empirical comparison of instrumental variables estimators. *Journal of Applied Econometrics*, *3*, 149–155.

Costa, G., Lickup, L., & Di Martino, V. (1988). Commuting - a further stress factor for working people: Evidence from the European Community II. An empirical study. *International Archives of Occupational and Environmental Health*, *60*, 377–385.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Croissant, Y. & Millo, G. (2008). Panel data econometrics in R: The `plm` package. *Journal of Statistical Software*, *27*.

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., & Swinton, J. (2018). *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-3.

Davidson, R. & MacKinnon, J. (1993). *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.

Deaton, A. S. & Paxson, C. (1994). Saving, growth, and aging in Taiwan. In *Studies in the Economics of Aging* (pp. 331–362). University of Chicago Press.

Department of Health (2011). Healthy lives, healthy people: A call to action on obesity in England. Technical Report 16166, HM Government.

Devaux, M. & Sassi, F. (2011). Social inequalities in obesity and overweight in 11 OECD countries. *European Journal of Public Health*, *23*, 464–469.

Dickerson, A., Hole, A. R., & Munford, L. A. (2014). The relationship between well-being and commuting revisited: Does the choice of methodology matter? *Regional Science and Urban Economics*, *49*, 321–329.

Diouf, I., Charles, M. A., Ducimetière, P., Basdevant, A., Eschwege, E., & Heude, B. (2010). The evolution of obesity prevalence in France: an age-period-cohort analysis. *Epidemiology*, *21*, 360–365.

Dobson, A. J. & Barnett, A. (1990). *An introduction to generalized linear models*. CRC Press.

Dunteman, G. H. & Ho, M. R. (2005). *An introduction to generalized linear models*, volume 145. Sage Publications.

Ejrnæs, M. & Hochguertel, S. (2013). Is business failure due to lack of effort? Empirical evidence from a large administrative sample. *Economic Journal*, *123*, 791–830.

Fahrmeir, L. & Kaufmann, H. (1986). Asymptotic inference in discrete response models. *Statistical Papers*, *27*, 179–205.

Fannon, Z. & Nielsen, B. (2019). Age-period-cohort models. In *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press.

Fitzenberger, B., Schnabel, R., & Wunderlich, G. (2004). The gender gap in labor market participation and employment: A cohort analysis for West Germany. *Journal of Population Economics*, *17*, 83–116.

Fosse, E. & Winship, C. (2019). Analyzing age-period-cohort data: A review and critique. *Annual Review of Sociology*, *45*, 467–92.

Fox, J. & Weisberg, S. (2019). *An R Companion to Applied Regression* (Third ed.). Thousand Oaks CA: Sage.

Fu, W. (2016). Constrained estimators and consistency of a regression model on a Lexis diagram. *Journal of the American Statistical Association*, *111*, 180–199.

Fukuda, K. (2013). Decomposition of new venture growth into firm age, survey period and vintage effects. *Applied Economics*, *45*, 85–97.

Gimenez-Nadal, J. I., Molina, J. A., & Velilla, J. (2018). Commuting time and sick-day absence of US workers. Discussion Paper 11700, Institute of Labor Economics.

Glenn, N. D. (2005). *Cohort Analysis* (2nd ed.)., volume 5 of *Quantitative Applications in the Social Sciences*. SAGE Publications, Inc.

Gottholmseder, G., Nowotny, K., Pruckner, G., & Theurl, E. (2009). Stress perception and commuting. *Health Economics*, *18*, 559–576.

Greene, W. H. (2008). *Econometric Analysis*. Upper Saddle River, NJ: Pearson Prentice Hall.

Halonen, J. I., Pulakka, A., Vahtera, J., Pentti, J., Laström, H., Stenholm, S., & Hanson, L. M. (2020). Commuting time to work and behaviour-related health: a fixed-effect analysis. *Occupational and Environmental Medicine*, *77*, 77–83.

Hanoch, G. & Honig, M. (1985). "True" age profiles of earnings: Adjusting for censoring and for period and cohort effects. *Review of Economics and Statistics*, *67*, 383–394.

Hansson, E., Mattisson, K., Björk, J., Ostergren, P., & Jakobsson, K. (2011). Relationship between commuting and health outcomes in a cross-sectional population survey in southern Sweden. *BMC Public Health*, *11*, 834–847.

Harnau, J. & Nielsen, B. (2018). Over-dispersed age-period-cohort models. *Journal of the American Statistical Association*, *113*, 1722–1732.

Hausman, J. A. & Taylor, W. E. (1981). Panel data and unobservable individual effects. *Econometrica*, *49*, 1377–1398.

Heckman, J. & Robb, R. (1985). *Using longitudinal data to estimate age, period and cohort effects in earnings equations*, (pp. 137–150). New York, NY: Springer New York.

Holford, T. R. (1983). The estimation of age, period and cohort effects for vital rates. *Biometrics*, *39*, 311–324.

Howel, D. (2011). Trends in the prevalence of obesity and overweight in English adults by age and birth cohort, 1991-2006. *Public Health Nutrition*, *14*, 27–33.

Hruby, A., Manson, J. E., Qi, L., Malik, V. S., Rimm, E. B., Sun, Q., Willet, W. C., & Hu, F. B. (2016). Determinants and consequences of obesity. *American Journal of Public Health Special Section: Nurses' Health Study Contributions*, *106*, 1656–1662.

Ingenfeld, J., Wolbring, T., & Bless, H. (2019). Commuting and life satisfaction revisited: evidence on a non-linear relationship. *Journal of Happiness Studies*, *20*, 2677–2709.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *ISLR: Data for an Introduction to Statistical Learning with Applications in R.* R package version 1.2.

Jensen, J. B., McGuckin, R. H., & Stiroh, K. J. (2001). The impact of vintage and survival on productivity: Evidence from cohorts of US manufacturing plants. *Review of Economics and Statistics*, *83*, 323–332.

Ji, W., Xie, N., He, D., Wang, W., Li, H., & Wang, K. (2019). Age-period-cohort analysis on the time trend of Hepatitis B incidence in four prefectures of southern Xinjiang, China from 2005 to 2017. *International Journal of Environmental Research and Public Health*, *16*.

Kalwij, A. S. & Alessie, R. (2007). Permanent and transitory wages of british men, 1975-2001: Year, age, and cohort effects. *Journal of Applied Econometrics*, *22*, 1063–1093.

Kapteyn, A., Alessie, R., & Lusardi, A. (2005). Explaining the wealth holdings of different cohorts: productivity growth and social security. *European Economic Review*, *49*, 1361–1391.

Kleiber, C. & Zeileis, A. (2008). *Applied Econometrics with R*. New York: Springer-Verlag. ISBN 978-0-387-77316-2.

Krueger, A. B. & Pischke, J. (1992). The effect of social security on labor supply: a cohort analysis of the notch generation. *Journal of Labor Economics*, *102*, 412–437.

Kuang, D., Nielsen, B., & Nielsen, J. P. (2008). Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika*, *95*, 979–986.

Künn-Nelen, A. (2016). Does commuting affect health? *Health Economics*, *25*, 984–1004.

Lagakos, D., Moll, B., Porzio, T., Qian, N., & Schoellman, T. (2018). Life cycle wage growth across countries. *Journal of Political Economy*, *126*, 797–849.

Lean, M. E. J., Katsarou, C., McLoone, P., & Morrison, D. S. (2013). Changes in BMI and waist circumference in Scottish adults: use of repeated cross-sectional surveys to explore mulitple age groups and birth-cohorts. *International Journal of Obesity*, *37*, 800–808.

Lee, R. D. & Carter, L. R. (1992). Modelling and forecasting U.S. mortality. *Journal of the American Statistical Association*, *87*, 659–671.

Lorenz, O. (2018). Does commuting matter to subjective well-being? *Journal of Transport Geography*, *66*, 180–199.

Low, H., Meghir, C., & Pistaferri, L. (2010). Wage risk and employment risk over the life cycle. *American Economic Review*, *100*, 1432–1467.

Lumley, T. (2019). `survey`: Analysis of complex survey samples. R package version 3.35-1.

Lumley, T. & Scott, A. (2017). Fitting regression models to survey data. *Statistical Science*, *32*, 265–278.

Martínez Miranda, M. D., Nielsen, B., & Nielsen, J. P. (2015). Inference and forecasting in the age-period-cohort model with unknown exposure with an application to mesothelioma mortality. *Journal of the Royal Statistical Society, Series A*, *178*, 29–55.

Mason, K. O., Mason, W. M., Winsborough, H. H., & Poole, K. (1973). Some methodological issues in cohort analysis of archival data. *American Sociological Review*, *38*, 242–258.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. Springer.

McKenzie, D. J. (2006). Disentangling age, cohort, and time effects in the additive model. *Oxford Bulletin of Economics and Statistics*, *68*, 473–495.

McPherson, K., Marsh, T., & Brown, M. (2007). Tackling obesities: Future choices - modelling future trends in obesity and the impact on health. Technical report, Government Office for Science.

Meghir, C. & Whitehouse, E. (1996). The evolution of wages in the United Kingdom: Evidence from micro data. *Journal of Labor Economics*, *14*, 1–25.

Méndez, F. & Sepúlveda, F. (2012). The cyclicality of skill acquisition: Evidence from panel data. *American Economic Journal: Macroeconomics*, *4*, 128–52.

Moody, A. (2016). Health Survey for England 2015 adult overweight and obesity. Technical report, Health and Social Care Information Centre.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A (General)*, *135*, 370–384.

Nie, P. & Sousa-Poza, A. (2018). Commute time and subjective well-being in urban China. *China Economic Review*, *48*, 188–204.

Nielsen, B. (2015). apc: An r package for age-period-cohort analysis. *The R Journal*, *7*, 52–64.

Nielsen, B. & Nielsen, J. P. (2014). Identification and forecasting in mortality models. *The Scientific World Journal*, *2014*, Article ID 347043, 24 pages.

O'Brien, R. M. (2011). Constrained estimators and age-period-cohort models (with discussion). *Sociological Methods & Research*, *40*, 419–470.

O'Brien, R. M. (2015). Age–period–cohort models and the perpendicular solution. *Epidemiologic Methods*, *4*, 87–99.

Oddone, E., Bollon, J., Nava, C. R., Bugani, M., Consonni, D., Marinaccio, A., Magnani, C., & Barone-Adesi, F. (2020). Predictions of mortality from pleural mesothelioma in Italy after the ban of asbestos use. *International Journal of Environmental Research and Public Health*, *17*.

O'Dea, C. (2012). *APC: Stata module to estimate age, period and cohort effects*. Institute for Fiscal Studies.

O'Donovan, G., Stamatakis, E., & Hamer, M. (2018). Associations between alcohol and obesity in more than 100 000 adults in England and Scotland. *British Journal of Nutrition*, *119*, 222–227.

Ogden, C. L., Carroll, M. D., Fryar, C. D., & Flegal, K. M. (2015). Prevalence of obesity among adults and youth: United States, 2011-2014. Technical Report 219, National Center for Health Statistics, Hyattsville, MD.

Oh, C. & Holford, T. R. (2015). Age-period-cohort approaches to back-calculation of cancer incidence rate. *Statistics in Medicine*, *34*, 1953–1964.

Oliveira, R., Klebson, M., Viana, J., Tigre, R., & Sampaio, B. (2015). Commute duration and health: Empirical evidence from Brazil. *Transportation Research Part A: Policy and Practice*, *80*, 62–75.

Peeters, A., Gearon, E., Backholer, K., & Carstensen, B. (2015). Trends in the skewness of the body mass index distribution among urban Australian adults, 1980 to 2007. *Annals of Epidemiology*, *25*, 26–33.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsey, J. B. (1969). Test for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society, Series B (Methodological)*, *31*, 350–371.

Reither, E. N., Hauser, R. M., & Yang, Y. (2009). Do birth cohorts matter? Age-period-cohort analyses of the obesity epidemic in the United States. *Social Science & Medicine*, *69*, 1439–1448.

Roberts, J., Hodgson, R., & Dolan, P. (2011). "It's driving her mad": Gender differences in the effects of commuting on psychological health. *Journal of Health Economics*, *30*, 1064–1076.

Rosenthal, S. S. (2014). Are private markets and filtering a viable source of low-income housing? Estimates from a "repeat income" model. *American Economic Review*, *104*, 687–706.

Rutherford, M. J., Lambert, P. C., & Thompson, J. R. (2010). Age-period-cohort modeling. *The Stata Journal*, *10*, 606–627.

Sandow, E., Westerlund, O., & Lindgren, U. (2014). Is your commute killing you? On the mortality risks of long-distance commuting. *Environment and Planning*, *46*, 1496 – 1516.

Sasieni, P. D. (2012). Age-period-cohort models in stata. *The Stata Journal*, *12*, 44–60.

Scarborough, P., Bhatnagar, P., Wickramasinghe, K. K., Allender, S., Foster, C., & Rayner, M. (2011). The economic burden of ill health due to diet, physical inactivity, smoking, alcohol and obesity in the UK: an update to 2006-07 NHS costs. *Journal of Public Health*, *33*, 527–535.

Schmid, V. J. & Held, L. (2007). Bayesian age-period-cohort modeling and prediction - BAMP. *Journal of Statistical Software*, *21*.

Schulhofer-Wohl, S. (2018). The age-time-cohort problem and the identification of structural parameters in life-cycle models. *Quantitative Economics*, *9*, 643–658.

Schulhofer-Wohl, S. & Yang, Y. (2006). *APC: Stata module for estimating age-period-cohort effects*. Statistical Software Components.

Sproston, K. & Mindell, J. (2006). *Health survey for England 2004. Volume 1. The Health of minority ethnic groups*. London: The Information Centre.

Stutzer, A. & Frey, B. S. (2008). Stress that doesn't pay: The commuting paradox. *The Scandinavian Journal of Economics*, *110*, 339–366.

Van Landeghem, B. (2012). A test for the convexity of human well-being over the life cycle: Longitudinal evidence from a 20-year panel. *Journal of Economic Behavior and Organisation*, *81*, 571–582.

Van Ommeren, J. N. & Gutiérrez-i-Puigarnau, E. (2011). Are workers with a long commute less productive? An empirical analysis of absenteeism. *Regional Science and Urban Economics*, *41*, 1–8.

Verbeek, M. J. C. M. & Nijman, T. E. (1992). Incomplete panels and selection bias. Discussion Paper 9207, Tilberg Center for Economic Research.

Voas, D. & Chaves, M. (2016). Is the United States a counterexample to the secularization thesis? *American Journal of Sociology*, *121*, 1517–1556.

Wang, Y. C., McPherson, K., Marsh, T., Gortmaker, S. L., & Brown, M. (2011). Health and economic burden of the projected obesity trends in the USA and the UK. *Lancet*, *378*, 815–825.

Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, *63*, 27–32.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*, 817–838.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12).

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1–29.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: Massachusetts Institute of Technology.

Yang, Y. (2008). Social inequalities in happiness in the united states, 1972 to 2004: An age-period-cohort analysis. *American Sociological Review*, *73*, 204–226.

Yang, Y., Fu, W. J., & Land, K. C. (2004). A methodological comparison of age-period-cohort models: The intrinsic estimator and conventional generalized linear models. *Sociological Methodology*, *34*, 75–110.

Yang, Y. & Land, K. C. (2006). A mixed models approach to the age-period-cohort analysis of repeated cross-section surveys, with an application to data on trends in verbal test scores. *Sociological Methodology*, *36*, 75–97.

Yang, Y. & Land, K. C. (2013). *Age-period-cohort analysis: New models, methods, and applications*. Boca Raton, FL: CRC Press.

Zaninotto, P., Head, J., Stamatakis, E., Wardle, H., & Mindell, J. (2009). Trends in obesity among adults in England from 1993 to 2004 by age and social class and projections of prevalence to 2012. *Journal of Epidemiology and Community Health*, *63*, 140–146.

Zeileis, A. & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, *2*(3), 7–10.

Zhao, S., Dong, H., Qin, J., Liu, H., Li, Y., Chen, Y., Molassiotis, A., He, D., Lin, G., & Yang, L. (2019). Breast cancer mortality in Chinese women: does migrant status play a role? *Annals of Epidemiology*, *40*, 28–34.