An Investigation of Age-Period-Cohort Models for Individual-Level Data

Zoë Fannon

Trinity Term 2016

Submitted in partial fulfilment of the requirements
for the degree of the Master of Philosophy in Economics

WORDCOUNT:
28,644 (372 words on p.2, 75 pages)

Abstract

When dealing with temporal effects, there is a well-known identification problem that arises as a result of the linear relationship between age, period, and cohort. Kuang, Nielsen, and Nielsen (Kuang *et al.*, 2008b) approach this problem in the context of aggregate data by examining estimable functions of the parameters of the standard model of temporal effects. These functions are both uniquely identified from data and invariant to transformations of the parameters of the original model. This thesis extends their approach to study the standard model applied to individual-level data. A new test of fit for models of inter-temporal variation in individual-level data is developed, along with an algorithm for the test's implementation. These extensions are demonstrated in an application to BMI data from the Health Survey for England from 1995-2013, which yields suggestive evidence that age and period effects are of primary importance in this data.

Acknowledgement is due to:

My supervisor, Bent Nielsen, who has given invaluable guidance and support throughout this project. It would not have been possible without his efforts, which went far beyond his obligations, and I am very grateful to him.

Christiaan Monden, who kindly provided the data used in the example application and whose insights were of great help in interpreting the results.

My parents, my brother, and my sister, for their emotional support and encouragement over the course of my studies at Oxford.

This work was supported by the Economic and Social Research Council [grant number 1499758].

Contents

1 Introduction **1**

2 The APC Identification Problem **4**

2.1 The Problem in Data 4

2.1.1 Introduction to the Problem in Age-Cohort Space 4

2.1.2 Discussion of Age-Cohort Space and Indexing 6

2.2 The Problem in Modelling 7

2.2.1 Linear Model 8

2.2.2 Dummy Variable Model 9

2.3 Review of APC Literature 11

2.3.1 Approaches Involving Constraints on θ 12

2.3.2 The Estimable Functions Approach 14

2.3.3 Applications to Individual Data 19

3 Estimable Functions with Individual Data **22**

3.1 Age, Period, and Cohort Effects 22

3.2 Introducing Covariates 24

3.3 Submodels of the APC model 27

4 Assessing Fit of Individual-Level Models **29**

4.1 Test of Fit for Aggregate Data 29

4.2 Test of Fit for Individual Data 30

4.2.1 Problems with the Deviance Statistic 30

4.2.2 Moving Beyond Discrete Outcomes without Covariates 32

4.3 Computational Challenges for the Test of Fit 34

5 Implementation in R **38**

5.1 The `apc` Package 38

5.2 Commands to Fit APC Models to Individual Data 39

5.3 Commands to Assess Fit of Models for Individual Data 42

6 Example Application: Health Survey for England **44**

6.1 Motivation 44

6.2 Description of Data 45

6.3 Exploration of Data 49

6.3.1 BMI as a Function of APC Only 49

6.3.2 BMI with Covariates 57

6.3.3 Stability of Estimated APC Structure: Split Sample 61

6.3.4 Stability of Estimated APC Structure: Binary Dependent Variable . . . 63

6.4 Summary 65

7 Conclusion **66**

A Excerpt from HSE Data **68**

References **69**

1 Introduction

There is long-standing interest in finding methods to investigate how the passage of time, and the position of an individual in time, affect the actions and characteristics of that individual. The influence of time is multi-faceted: we expect an individual's life-goals and limitations to change as they age, but we also expect that these will be influenced by features of the era in which that person lives as well as the years through which they and their cohort have lived. These three distinct ways in which an individual relates to time - through their age, the period in which they are observed, and the cohort to which they belong - are each of substantive interest to researchers in fields ranging from economics, political science, and sociology to demography and epidemiology. And yet, due to the simple fact that a person's age can be determined by subtracting their year of birth from the current year, it is impossible to mechanically disentangle the three linear effects of age, period, and cohort.

At least one author is of the opinion that in light of this linear relationship, any attempt to find generally applicable statistical tools that would enable the separation of the three effects is a "futile quest"(Glenn, 1976). Glenn instead argues that theoretical models which rely on assumptions about age, period, and cohort effects must stand or fall on their logical consistency or contextual plausibility alone, since they cannot be validated by statistical methods. However, the questions which motivate interest in the effects are of sufficient importance that this proposal of abandoning the attempt to discriminate between them is fairly unpalatable. In the economics literature, efforts to evaluate the life-cycle savings hypothesis have repeatedly run up against this age-period-cohort (APC) identification problem (Deaton & Paxson, 1993; Attanasio, 1993). Studies of age profiles of earnings (Hanoch & Honig, 1985; Meghir & Whitehouse, 1996; Kalwij & Alessie, 2007) and lifetime labour force participation decisions (Heckman & Robb, 1985; Krueger & Pischke, 1991; Fitzenberger *et al.*, 2004; Browning *et al.*, n.d.) are similarly afflicted. Sociologists and political scientists studying phenomena such as religious practice (Schwadel, 2010; Voas, 2013) and volatility in voting behaviour (Dassonneville, 2013) have grappled with the APC problem, as have epidemiologists looking to understand changes in incidence and recovery rates of diseases ranging from cancer to obesity (Clayton & Schifflers,

1987; Robertson & Boyle, 1998). It is clear that there is potential for substantial advances across a range of disciplines if these three temporal effects can be even somewhat more clearly understood.

One potentially promising approach, which is pursued by Kuang, Nielsen, and Nielsen (2008b), involves breaking down the trend information into constituent estimable functions: a linear plane, which is not attributed to any of the three temporal effects, and deviations from that plane, which can be so attributed. Careful interpretation of these estimated functions can shed some light on the distinct roles played by age, period, and cohort in determining the outcome variable of interest. The object of this thesis is to extend this approach from the aggregate data context in which it was originally developed to individual-level data.

Individual-level data contains additional information that can enable researchers to develop a more detailed understanding of a phenomenon. It is possible, for example, to gain valuable insights by introducing covariates, looking at the differences between inter- and intra-temporal variation, and following particular individuals through time using panel studies. While this thesis does not address all of the potential advantages from working with data at the individual level, it constitutes the beginning of an effort to comprehensively examine this setting. In particular it addresses two issues. First, modifications to existing theory and code required to introduce the estimable functions approach to the individual-level context are considered. Second, a test of model fit appropriate to individual-level data and an algorithm to implement that test are developed. These are demonstrated with an example application to data on BMI (body mass index) from the Health Survey for England (HSE) from 1995-2013.

The structure of the thesis is as follows: in section 2, the age-period-cohort (APC) identification problem is explained and efforts to overcome it are discussed. In particular a summary of the estimable functions approach mentioned above is provided, along with an overview of some existing methods for addressing the APC problem in the context of individual-level data. Following this introduction to the APC problem and contextual grounding of this thesis, in section 3 the steps in extending the estimable-parts approach to individual data and introducing covariates to the analysis are outlined. In section 4 I depart substantially from existing

work in the estimable functions literature, introducing a new method of assessing the fit of the APC model based on what I term the temporally-saturated model. I also describe an algorithm developed to overcome the computational difficulties arising from this new method of assessing fit in the context of continuous data. Tools for implementation in R are described in section 5, and the application to BMI data is presented in section 6. Section 7 concludes.

2 The APC Identification Problem

In this section an overview of the APC problem is provided. First, the problem is described as it appears in data, before formal models are introduced. At this point the coordinate system and indexing used throughout the thesis are also explained. Two formal models are then presented, along with an explanation of the identification issue in each. The first model consists of a linear plane only, and is provided to give the reader an initial understanding of the nature of the identification problem. The second model, involving three sets of dummy variables, is the literature standard and forms the basis of subsequent analysis.

2.1 The Problem in Data

2.1.1 Introduction to the Problem in Age-Cohort Space

Here I introduce the APC problem for the case of data with an age-cohort structure. An example of such data is seen in figure 2.1 (note that this is an excerpt from the BMI data that will be analysed in section 6). Age appears on the horizontal axis, and cohort on the vertical axis; points in the coordinate space are referred to as age-cohort cells. At each cell the graph shows mean BMI for that age-cohort group (men and women are pooled in this excerpt). For example, when the cohort of individuals born in 1980 are 25 years old, their average BMI is 23.06, but by the time they have reached age 29 they have an average BMI of 26.33. If we were to add a period axis to this coordinate space it would extend along the 45° line from the top-left corner: the cohort born in 1980 are aged 25 in 2005; in 2006 they are aged 26 and the cohort born in 1981 are aged 25; and so on.

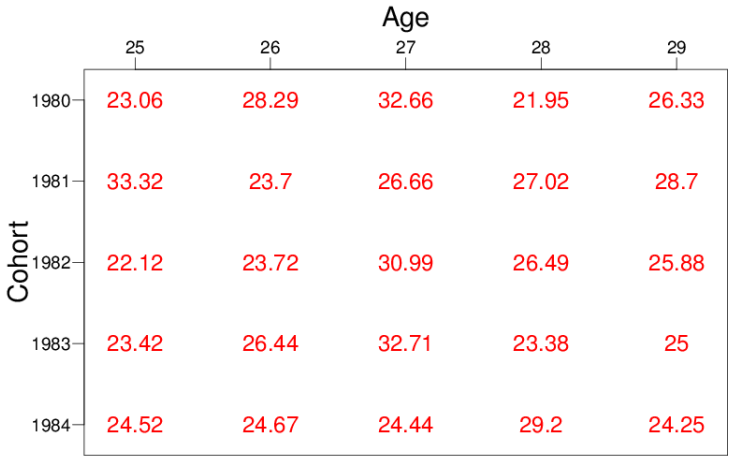


Figure 2.1: Excerpt from BMI data showing mean BMI per age-cohort cell

The critical point to realise here is that introducing this third concept of period does not alter the coordinate space described by the plot above; it simply provides us with a different way of thinking about that space. This makes sense: if it is always possible to determine the period by using the existing information about the coordinate space and the identity $period = age + cohort$ (as was done above when calculating $1980 + 25 = 2005$) then it must be true that the introduction of the concept of period does not extend the coordinate space. Descriptions of this data that seek to incorporate all three of age, period, and cohort as though they were fully independent are thus over-parameterized. To put this less formally, it is difficult if not impossible to determine which of the three factors should be held responsible for any observed patterns. For example, looking at the bottom corner of the data set we see that the elements of the second-to-last diagonal are both greater than that in the final diagonal. Should we interpret this to mean that BMI rose for all groups between 2012 and 2013? Or is it that the cohort of 1984 generally has a lower mean BMI than that of 1983 at the same ages, and that there is an increase in BMI as we age from 28 to 29?

The problem of interdependence seen in the aggregate data persists when we move to the disaggregated version of the same data. Since this disaggregated data involves many individual observations at each point in age-cohort space, it is no longer possible to visually represent it in a two-dimensional grid of numbers as in 2.1. Instead a three-dimensional representation

such as that in figure 2.2 is used, where each individual observation is represented as a point (note that the period axis in this figure runs diagonally from the front-left to the back-right of the space). It can be seen that any effort to detect patterns will encounter the same problem we met in the aggregate data: it is very difficult to make the claim that the pattern is related to one of the three dimensions rather than some combination of the other two. Essentially we are dealing with three conceptually distinct but fundamentally interdependent ideas, and so attempts to treat them as entirely separable in data are bound to go awry.

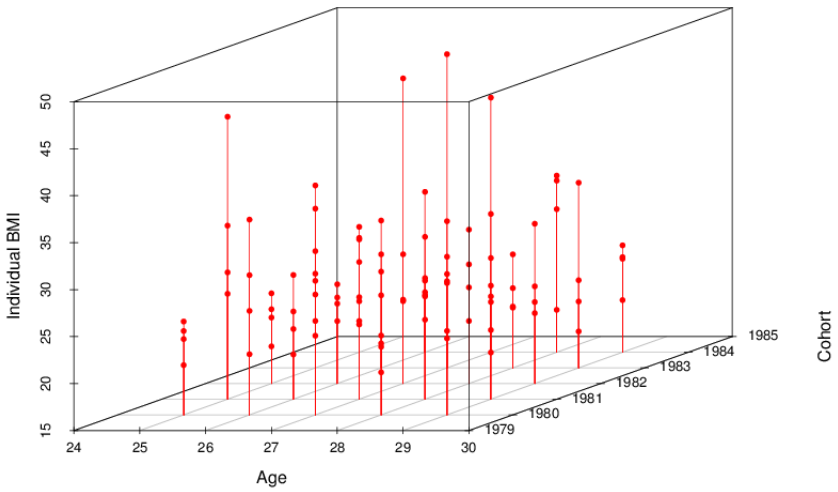


Figure 2.2: Disaggregated BMI data from figure 2.1; each point is an individual observation.

2.1.2 Discussion of Age-Cohort Space and Indexing

The discussion above focused exclusively on data that appears in age-cohort (AC) format. Other common data structures include repeated observations on particular age ranges (age-period, or AP, format), and data that follows certain cohorts over a number of years (period-cohort, or PC, format). While one might expect that in these cases a coordinate space that reflects the format in which the data was collected should be used, it is in fact preferable to continue to use the age-cohort coordinate space. The data then appears as a parallelogram in this space, as seen in the examples in figure 2.3.

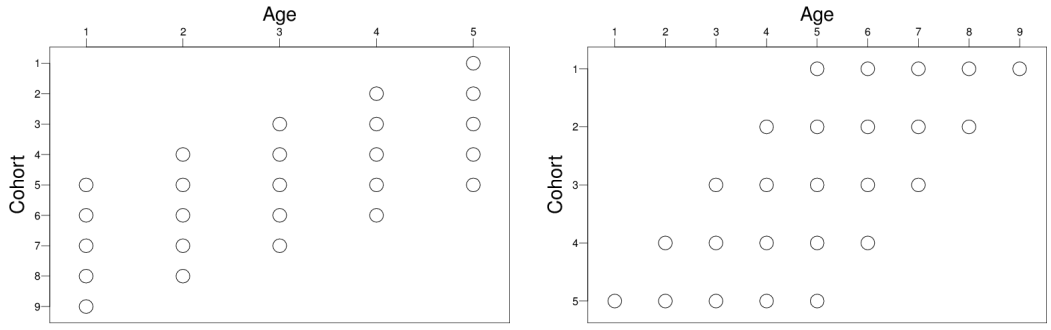


Figure 2.3: Format of age-period (left) and period-cohort (right) data in age-cohort space

The age-cohort space is preferred because when it is fully stretched out, all indices take the value 0 at the origin. This is not true of a representation in age-period space or period-cohort space, where the age index and cohort index respectively do not begin at 0. This relates to a deeper problem with these spaces: they cannot represent the fact that the oldest cohorts seen in the data were in existence prior to the data collection period. Working in age-cohort space pushes us to recognize that the oldest individuals in the data will have been subject to the relevant age, period, and cohort effects during their youths, even though these are not observed in the data.

A final point to be raised in this section relates to the indexing of the data. Looking more closely at figure 2.3, we can see that at the first period, $age = 1$ and $cohort = 1$. In this data the three concepts are related not by the identity $age + cohort = period$, but by $age + cohort - 1 = period$. This difference arises because of the convention that a person is described as being aged 1 when they are in fact in their second year of life. To avoid any potential confusion arising from this convention, it is assumed that all data is indexed in the manner shown in figure 2.3; and indeed one of the first steps in the R implementation discussed in section 5 is to relabel all indices accordingly.

2.2 The Problem in Modelling

Thus far I have only discussed the APC identification problem as it appears when looking at plots of data. In this section I consider the APC problem as it arises when thinking about

mathematical models that could be used to describe such data, using the examples of a simple linear model and an indicator model. For these I adopt the convention that age is indexed by $i = 1, ...I$; cohort by $k = 1, ...K$; and period by $j = 1, ...J$. Thus, given the discussion about indexing in section 2.1.2, the relationship that is at the centre of the APC problem is now expressed by the identity $i + k - 1 = j$.

2.2.1 Linear Model

To motivate the discussion in this section, consider the three-dimensional shape represented in figure 2.4. Letting μ_{ik} represent the height of the figure at cell ik in age-cohort coordinate space (note that the j subscript can be dropped since it provides no additional information), the fact that the shape is uniformly increasing along the period axis would suggest that it was generated by a model of period effects only, $\mu_{ik} = 2 + 2j$. However, it was actually generated by the model $\mu_{ik} = 2i+2k$, which includes only age and cohort effects; and is indistinguishable from the shape generated by $\mu_{ijk} = 1 + i + j + k$, including all three effects. The equivalence of these expressions can be verified by simple algebra.

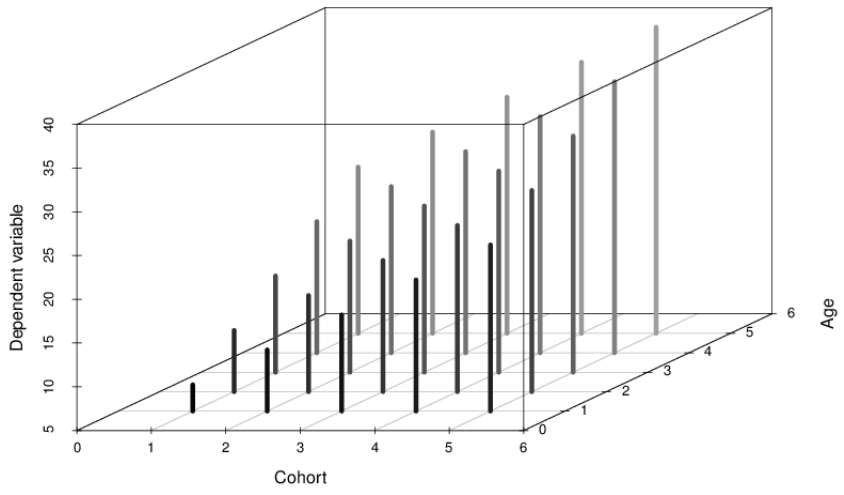


Figure 2.4: Shape generated by a linear model of age, period, and cohort effects, in age-cohort space

In fact, there are uncountable expressions of the general linear form

$$\mu_{ik} = v_o + v_a i + v_p j + v_c k \tag{2.1}$$

that could have generated this shape, where v_o is a general intercept and v_a, v_p, and v_c are slope coefficients in the age, period, and cohort dimensions. More formally, the problem can be stated as follows. Let $\mathcal{V}$ be the space on which the parameter vector $v = \{v_o, v_a, v_p, v_c\}$ is defined. Let μ be the set of observed μ_{ik} (of which there will be as many as there are ik cells in the age-cohort space on which the model is defined); then μ lies in a space $\mathcal{M}$. The problem is that the model given in (2.1) does not constitute a one-to-one mapping from $\mathcal{V}$ to $\mathcal{M}$; we may have $\mu(v) = \mu(v')$ for $v \neq v'$. While it is clearly not the case that every element of $\mathcal{V}$ will be consistent with a particular μ, so long as at least two elements of $\mathcal{V}$ are consistent with μ then it is not possible to determine from μ which of these elements is most appropriate to describe μ.

Of course, when we consider real data it will not have an appealing regular shape such as that in figure 2.4. It is important to recognise that the irregularity in this data will not help to determine which parameter vector is most appropriate, once it has been decided to use the linear model in (2.1). The data determines which shape of the form described by the model is most appropriate, i.e. it determines μ; but it cannot help us to choose the parameter vector v we use to describe this shape.

2.2.2 Dummy Variable Model

Analysing the simple linear model described in section 2.2.1 is a helpful step in understanding how the APC identification problem affects the models we might wish to use to describe data. This model is, however, quite restrictive and therefore is unlikely to be used directly in applied work. The more flexible model that is the standard in literature on the APC problem allows a dummy variable for each distinct value of each of the three indices. This model is generally represented by

$$\mu_{ik} = \alpha_i + \beta_j + \gamma_k + \delta \tag{2.2}$$

where μ_{ik} is the height of the modelled shape at cell ik as before, α_i is the effect of being at age i, β_j is the effect of being in period j, γ_k is the effect of being a member of the kth cohort, and δ is a grand mean. Of course, the representation

$$\mu_{ik} = \sum_{q=1}^{I} 1\left(i = q\right)\alpha_q + \sum_{r=1}^{J} 1(j = r)\beta_r + \sum_{s=1}^{K} 1\left(k = s\right)\gamma_s + \delta$$

is strictly speaking more accurate, but representation (2.2) is preferred as it is much cleaner. The full parameter vector for this model is

$$\theta = \{\alpha_1, \ldots\alpha_I, \beta_1, \ldots\beta_J, \gamma_1, \ldots\gamma_K, \delta\} \tag{2.3}$$

Once again, the model in (2.2) does not constitute a one-to-one mapping between the parameter space in which the parameter vector θ lies, which we will call $\mathcal{G}$, and $\mathcal{M}$. There exists a group of linear transformations of the elements of θ that will be mapped to the same μ by the model in (2.2), defined by

$$g : \begin{pmatrix} \alpha_i \\ \beta_j \\ \gamma_k \\ \delta \end{pmatrix} \longmapsto \begin{pmatrix} \alpha_i + a + (i-1)d \\ \beta_j + b - (j-1)d \\ \gamma_k + c + (k-1)d \\ \delta - a - b - c \end{pmatrix} \tag{2.4}$$

where $\{a, b, c, d\}$ are arbitrary constants. That is, $\mu(\theta) = \mu\{g(\theta)\}$. This group of transformations is discussed further in Carstensen (2007) and Kuang *et al.* (2008a).

The implication is that a wide range of parameter vectors that are elements of $\mathcal{G}$ will be consistent with a particular μ. Once again, while data can help us to determine μ (which constitutes a description of the chosen shape, and is thus the set of linear predictors resulting from the data) it cannot provide guidance as to which element of the group of θ that are consistent with that μ is most appropriate. The parameter vector θ can then be said to

be under-identified. The question at the heart of the APC literature might therefore be understood as: given that θ is under-identified, but we are interested in understanding more about age, period, and cohort effects in a range of applied contexts (such as those mentioned in the introduction), how should we proceed? I have already noted the position taken by Glenn that we must simply accept the limitations of our knowledge; in section 2.3 I provide an overview of alternative suggestions that have been made.

Before proceeding to consider these suggestions, it is worth noting as an aside that although the model in (2.2) is substantially more flexible than the linear model in section (2.2.1), it does impose some restrictions. In particular it does not allow for interaction effects between the three temporal dimensions, which may not be a credible restriction in certain situations. As an example, consider a hypothetical study looking at the probability that a given individual would choose to live abroad in a given year. We would expect there to be age effects (young people with fewer commitments may be more likely to spend a year abroad), period effects (if migration is driven by macroeconomic shocks), and cohort effects (those who moved abroad in response to macroeconomic shocks when young may have settled there). However there would also be age-period interactions as distinct from cohort shocks, for example if young people are more likely to move abroad in response to macroeconomic shocks than older adults, but do not then settle abroad. In such a study, model (2.2) would not be appropriate. This highlights that there is scope for future work to consider the implications of the APC identity for a more general model allowing for these interaction effects.

2.3 Review of APC Literature

Having outlined the APC problem as it appears both in data and in the models used to describe data, I now give a brief overview of some of the key pieces of literature around this problem. Two broad classes of approach to the problem exist are seen in the literature, both of which have been developed primarily in the context of aggregate data. The first class of approach involves imposing constraints on the vector of parameters, θ, with the aim of reducing the set of possible θ that are consistent with both data and constraints to a single element.

This approach mainly developed within the social sciences literature on cohort analysis, and has also been adopted by some economists (Krueger & Pischke, 1991; Deaton & Paxson, 1993; Kalwij & Alessie, 2007). The second class is the estimable functions approach mentioned in the Introduction. This approach moves away from consideration of θ to focus on structures which can be uniquely identified from the data, in particular the non-linear effects of age, period, and cohort. The approach has its origins in epidemiological literature, although a variation on the approach is employed by a second group of economists (Attanasio, 1993; Fitzenberger *et al.*, 2004)). The constraints-based and estimable functions approaches are described in sections 2.3.1 and 2.3.2 respectively, while a discussion of some of the existing efforts to extend analysis to individual-level data is provided in section 2.3.3.

2.3.1 Approaches Involving Constraints on θ

To understand the constraints-based approach to the APC problem, it is helpful to recognise that the under-identification of θ in the standard model (2.2) implies that if an attempt were made to estimate the parameters of that model by standard linear regression, the attempt would fail because of rank deficiency of the design matrix. There are two sources of this rank deficiency. First, within the set of dummies for each temporal dimension (i.e. age, period, and cohort) the well-known dummy variable trap problem must be contended with. This can be addressed by the standard approach to the dummy variable trap, of eliminating one dummy from each of age, period, and cohort. Yet this does not fully resolve the issue, because the identity $j = i + k - 1$ constitutes a second source of deficiency.

In one of the earliest papers in this branch of the literature, Mason *et al.* (1973) show that the deficiency can be fully resolved by dropping one more dummy variable. They suggest choosing for the fourth constraint one of the dummies adjacent to one of the three already dropped to address the dummy variable trap problem. For example, one might drop the dummies for the first and the second age group, in addition to one each from the sets of period and cohort dummies. Mason et al. argue that it is reasonable to assume that two adjacent age groups are similar, and consequently imposing the constraint that they are identical should

not substantially alter the estimation of the remaining unconstrained parameters. It would then be possible to interpret these estimated parameters as differences from a reference case that now consists of two age groups.

Subsequent approaches in this category of approach incorporate more sophisticated statistical techniques, but continue to rely fundamentally on the imposition of constraints on the parameter vector θ. For example, two widely-used methods, the Bayesian approach employed by Sasaki & Suzuki (1987) and the Intrinsic Estimator introduced by Yang *et al.* (2008), incorporate the assumption that all of the dummy variables in each dimension sum to zero, i.e. $\sum_i \alpha_i = \sum_j \beta_j = \sum_k \gamma_k = 0$.

The problem with any such approach is that the resulting estimated θ is, unsurprisingly, sensitive to the assumptions made to reach it. The particular assumption made by the Bayesian and Intrinsic Estimators has the unappealing consequence that altering the dimensions of the data set completely alters the set of parameter vectors that are consistent with the sum-to-zero assumption. This means that the parameter vector found to be consistent with both the sum-to-zero assumption and μ in the original data may be entirely unrelated to that chosen when the number of, for instance, periods included in the sample is extended or reduced. In addition to this, it is unclear how one should incorporate the sum-to-zero assumption in the interpretation of estimated coefficients. In the Mason et al. approach, interpretation at least seems more straightforward, as one simply extends the reference case logic from the standard dummy variable trap context. However, that approach only works in the event that the assumption that the two adjacent age categories are equivalent is valid. If the assumption is invalid, it is unclear how the estimated values of other coefficients are influenced by the imposed constraint that two non-identical categories are equivalent. Since the assumption of equivalence is untestable, to use the Mason et al. approach a researcher must strongly believe that it is a valid assumption (which Mason and her co-authors do stress in the paper). In practice it is difficult to imagine a context in which such a belief could be convincingly justified. Even if such a justification may be available in some contexts, any constraints-based approach is certainly not useful for exploratory data analysis.

It seems that any approach whereby constraints are imposed on some elements of θ, and the remaining unconstrained elements are then estimated, is liable to lead to confusion and misinterpretation of the resulting output because it is generally difficult to understand exactly how the constraints are driving the results. Recognising the limitations of this approach, we turn to the second strand of the literature, which took a very different approach to the problem.

2.3.2 The Estimable Functions Approach

The literature on the APC problem in epidemiology slightly post-dates that in the social sciences, with the earliest papers appearing in the 1980s (Holford, 1983, 1985; Clayton & Schifflers, 1987). Recognising that the problem of confounding is confined to the linear parts of the temporal effects, these researchers focus instead on the non-linear components of the temporal effects. The clearest exposition is provided by Robertson and Boyle (1998), who explain that the second differences of the parameter estimates in the classical model can be identified from data. Recalling that α_i is the coefficient on the indicator variable for age i , these second differences take the form

$$\Delta^2\alpha_3 = \Delta\alpha_3 - \Delta\alpha_2 = (\alpha_3 - \alpha_2) - (\alpha_2 - \alpha_1) = \alpha_1 - 2\alpha_2 + \alpha_3$$

where α can be replaced by β or γ. It can be verified that these are invariant to the group of transformations g from equation (2.4) by substituting the transformation $\alpha_i + a + (i-1)\,d$ in place of α_i in the expression above.

Kuang *et al.* (2008b), henceforth KNN, prove that by cumulating these second, or double, differences and combining them with a linear plane, it is possible to arrive at an exact reparameterization of the classical model. The paper begins with a proof that such a reparameterization exists for data in an age-cohort format. Starting from the standard model (2.2)

$$\mu_{ik} = \alpha_i + \beta_j + \gamma_k + \delta$$

with under-identified parameter vector

$$\theta = \{\alpha_1, ...\alpha_I, \beta_1, ...\beta_J, \gamma_1, ...\gamma_K, \delta\}$$

they show that by introducing telescopic sums of the form $\alpha_i = \alpha_1 + \sum_{t=2}^{i} \Delta\alpha_t$ and$\Delta\alpha_t = \Delta\alpha_2 + \sum_{s=3}^{t} \Delta^2\alpha_s$ for the age, period, and cohort parameters one can arrive at the representation

$$\begin{aligned}\mu_{ik} &= \alpha_1 + (i-1)\Delta\alpha_2 + \sum_{t=2}^{i}\sum_{s=3}^{t}\Delta^2\alpha_s + \beta_1 + (j-1)\Delta\beta_2 \\ &\quad + \sum_{t=2}^{j}\sum_{s=3}^{t}\Delta^2\beta_s + \gamma_1 + (k-1)\Delta\gamma_2 + \sum_{t=2}^{k}\sum_{s=3}^{t}\Delta^2\gamma_s + \delta\end{aligned}$$

which can be shown to be equivalent to the following:

$$\begin{aligned}\mu_{ik} &= \mu_{11} + (i-1)\mu_{21} + (k-1)\mu_{12} \\ &\quad + \sum_{t=2}^{i}\sum_{s=3}^{t}\Delta^2\alpha_s + \sum_{t=2}^{j}\sum_{s=3}^{t}\Delta^2\beta_s + \sum_{t=2}^{k}\sum_{s=3}^{t}\Delta^2\gamma_s \qquad (2.5)\end{aligned}$$

This is a linear plane in age-cohort space plus a cumulation of double differences in each temporal dimension. The new parameter vector,

$$\xi = \left\{\mu_{11}, \mu_{21}, \mu_{12}, \Delta^2\alpha_3, ...\Delta^2\alpha_I, \Delta^2\beta_3, ...\Delta^2\beta_J, \Delta^2\gamma_3, ...\Delta^2\gamma_K\right\} \qquad (2.6)$$

has three appealing properties. First, it is uniquely identified from μ, in the sense that $\mu(\xi) \neq \mu(\xi')$ for $\xi \neq \xi'$. Second, it is invariant to the group of transformations g as defined in equation (2.4); that is, $\xi(\theta) = \xi\{g(\theta)\}$. Finally, the interpretation of the parameters is reasonably straightforward: the first three are the coefficients that describe a linear plane in age-cohort space, while the double differences represent accelerations.

KNN also generalize their reparameterization result in two ways. First, they show through simple matrix algebra that one can replace the initial points $\{\mu_{11}, \mu_{21}, \mu_{12}\}$ with almost any set of points $\{\mu_{i_1k_1}, \mu_{i_2k_2}, \mu_{i_3k_3}\}$ that lie within the age-cohort coordinate space on which the data

is defined, and the identification result will hold[1] . The only change is to the sums of double differences in equation (2.5), as if the linear plane is changed the extent to which a particular observation deviates from the plane must change accordingly (it should be emphasized that it is the sums, not the double differences themselves, that change).

The key point to realise from this is that the particular linear plane and sums of double differences chosen to describe a given shape are arbitrary, and also that the linear plane and sums of double differences are interdependent and must be interpreted jointly. For example, to describe a dependent variable that follows a convex path in age-cohort space, we might choose a representation in which the initial points selected are close to one another, such that the linear plane would lie mostly below the observations and the sums of double differences would be greatest at the extreme ends of the index set. Alternatively if the chosen initial points were distant from one another the plane would lie mostly above the observations and the sums of double differences would be greatest at the centre of the index set. This example demonstrates that it is essential to evaluate the plane and sums of double differences together.

The ability to relocate the initial points also means that the relevance of this model is not restricted to data sets that contain the origin point of the age-cohort space. This implication is the basis of KNN's second generalization, which allows the scope of analysis to be widened from data that has an age-cohort structure to so-called "generalized trapezoids", including for example the age-period or period-cohort index sets seen earlier in figure 2.3. Formally, the index set $\mathcal{I}$ is a generalized trapezoid if, for some I, K, $J \in N$ and $L \in N_0$and $L+J \leq I+K-1$, then

$$\mathcal{I} = \{(i,\, k) \,:\, i = 1,\, ...,\, I,\, k = 1,\, ...K,\, j = i + k - 1 = L + 1,\, ...L + J\}$$

The implication of this definition is that any ik cell such that $i+k-1 < L+1$ or $i+k-1 > L+J$ is not included in the index set. This means that by choosing the appropriate $L > 0$, an index can be defined set that describes the subset of age-cohort space occupied by the trapezoid shapes of age-period and period-cohort data seen above in figure 2.3 and below in figure 2.5. L is determined by calculating the number of periods that would need to be added to the

[1]The only restriction is that the points cannot all lie in a straight line along one of the dimensions.

data such that the earliest period of interest in the lifetime of the oldest group observed in the data would be included. For example, if the data is in age-period format with $I = 5$, then the group with age $i = 5$ at the first period in the data set would have been of age $i = 1$ four periods prior to the start of the data set, so $L = 4$.

To prove the extension of the identification result to such generalized trapezoids, KNN show that for any $\{\mu_{i_1k_1}, \mu_{i_2k_2}, \mu_{i_3k_3}\} \in \mathcal{I}$ the parameter vector

$$\xi = \{\mu_{i_1k_1}, \mu_{i_2k_2}, \mu_{i_3k_3}, \Delta^2\alpha_3, ...\Delta^2\alpha_I, \Delta^2\beta_{L+3}, ...\Delta^2\beta_{L+J}, \Delta^2\gamma_3, ...\Delta^2\gamma_K\} \tag{2.7}$$

uniquely identifies the set μ of aggregate linear predictors; that is, $\{\Delta^2\beta_3, ...\Delta^2\beta_{L+2}\}$are not required. Intuitively this is because the age-period-cohort combinations which would incorporate these parameters lie outside the index set $\mathcal{I}$ which defines the trapezoid space occupied by the observed data (they lie within the isosceles triangle excluded at the origin). This extension constitutes the fully general statement of their result.

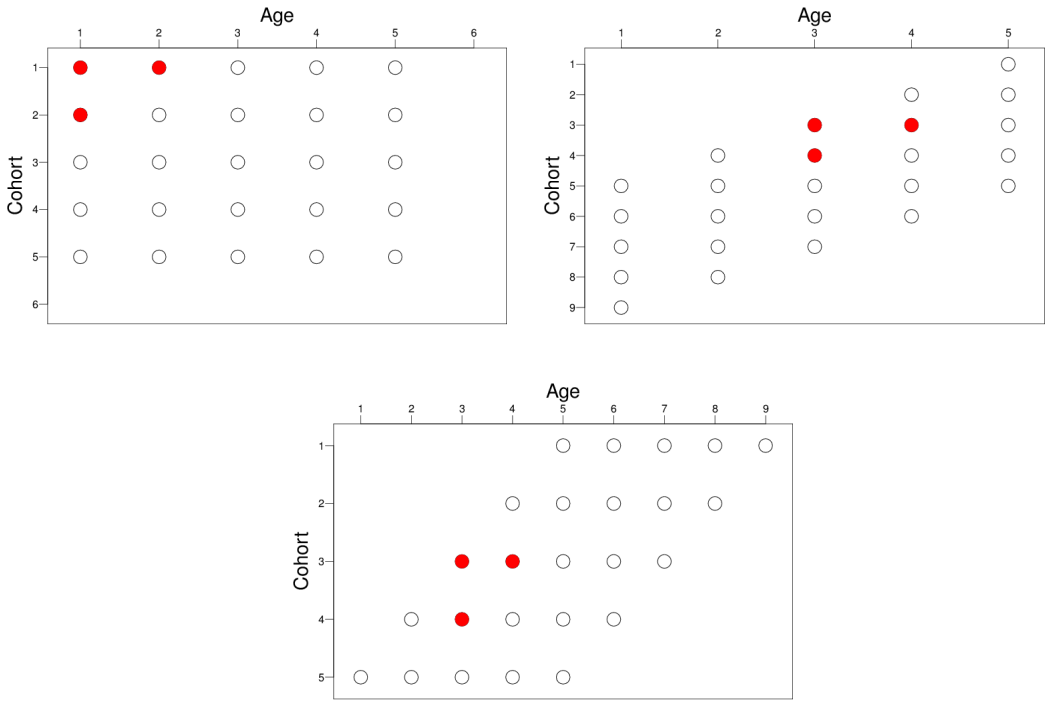


Figure 2.5: Red dots highlight the standardized initial points per Nielsen (2014). Data structures are: top-left, AC; top-right, AP; bottom, PC.

In a follow-up paper, Nielsen (2014) advocates the standardization of the choice of initial points, to facilitate interpretation of the double-sums of double differences in applied analysis. He suggests the points $\{\mu_{UU}, \mu_{U,U+1}, \mu_{U+1,U}\}$, where $U = \lfloor\left(\frac{L+3}{2}\right)\rfloor$. These points are shown for the three different data structures in figure 2.5. Looking at the figure gives some insight into the reasoning behind this choice of points: they are the first set of points within the index set that lie on the 45° line. They therefore minimize backward cumulation of double differences, and allow age and cohort double differences to be treated symmetrically. In the paper, Nielsen provides a single formula (formula (3.10), p. 7) that links the parameters to the set of linear predictors for all possible data structures. It is reproduced below for completeness as it forms the basis of subsequent analysis.

$$\mu_{ik} = \mu_{UU} + (i-U)\mu_{U+1,U} + (k-U)\mu_{U,U+1} + A_i + B_j + C_k \tag{2.8}$$

$$A_i = 1_{(i<U)} \sum_{t=i+2}^{U+1} \sum_{s=t}^{U+1} \Delta^2\alpha_s + 1_{(i>U+1)} \sum_{t=U+2}^{i} \sum_{s=U+2}^{t} \Delta^2\alpha_s$$

$$B_j = 1_{(L\,odd\,\&\,j=2U-2)}\Delta^2\beta_{2U} + 1_{(j>2U)} \sum_{t=2U+1}^{j} \sum_{s=2U+1}^{t} \Delta^2\beta_s$$

$$C_k = 1_{(k<U)} \sum_{t=k+2}^{U+1} \sum_{s=t}^{U+1} \Delta^2\alpha_s + 1_{(k>U+1)} \sum_{t=U+2}^{k} \sum_{s=U+2}^{t} \Delta^2\alpha_s$$

This can be written in matrix notation as

$$\mu_{ik} = X_{APC,ik}\xi$$

where $X_{APC,ik}$ is an ik-cell specific, $1 \times \dim(\xi)$ row vector of the form $\{1, (i-U), (k-U), ...\}$ (the remaining elements indicate how many times each double-difference is counted in μ_{ik}), and ξ is taken to be a column vector. The complete set of linear predictors for the index set

under consideration can then be written as

$$\mu = X_{APC}\xi$$

where X_{APC} is a matrix that consists of stacked $X_{APC,ik}$, and μ is a column vector of stacked μ_{ik}.

This structure provides a comprehensive and extremely flexible framework for thinking about APC effects, without needing to make substantive, untestable assumptions as was the case with constraints-based approaches. The linear predictors can be incorporated into a generalized linear model to study many forms of outcome variable. An example of such an application in the context of aggregate data can be found in Martínez Miranda *et al.* (2015).

2.3.3 Applications to Individual Data

Researchers in both the constraints-based and estimable functions traditions have worked primarily with aggregate data in their efforts to understand the APC problem. Yet it is clear that additional insights may be gained from expanding the scope of application of APC methods to individual data, as outlined in the introduction. While this has not previously been done within the estimable functions literature, some recent papers have employed other methods to study age, period, and cohort effects in individual-level data. In this section I briefly draw attention to two of these: the Hierarchical Age Period Cohort Model and the maximum entropy method.

The Hierarchical APC model (HAPC) introduced by Yang & Land (2006) employs a two-level modelling approach which is best conceptualized in period-cohort space. At the first, "within-cell", level, individual outcomes are modelled in terms of a period-cohort cell mean, a quadratic function in age, any covariates, and an error term. At the second, "between-cell", level, the cell mean at the first level is modelled as a function of period and cohort as well as a second error term. This reduces to a pooled OLS model of the form

$$Y_{jkn} = \delta + \beta_j + \gamma_k + \upsilon_{jk} + \lambda_1 i_{jkn} + \lambda_2 i_{jkn}^2 + \lambda_3 C_{jkn} + \varepsilon_{jkn}$$

where Y_{jkn} is the outcome for individual n at period-cohort cell jk; υ_{jk} is the second-level error term, ε_{jkn} is the first-level error term, and C_{jkn} is a vector of covariates. Considering the subset of this equation $\delta + \beta_j + \gamma_k + \lambda_1 i_{jkn}$, we can see that it is in fact a version of model (2.2) where the constraint $\alpha_i = \lambda_1 i$ has been imposed. This constraint is not itself sufficient to restrict the set of possible θ that are consistent with both data and model to a single element. To achieve that reduction, two further assumptions are made: that a single cohort consists of a five-year rather than single-year group (which is in practice equivalent to setting $\gamma_1 = \gamma_2 = ...\gamma_5$; $\gamma_6 = \gamma_7 = ...\gamma_{10}$, and so on), and that the cohort and period effects are random effects (eliminating the possibility of trends in these dimensions). Although a later paper (Yang & Land, 2008) shows that the second assumption is not strictly necessary, it is still the case that the model depends on assumptions that would be exceedingly difficult to justify a priori, and that it is unclear how the output of this model should be interpreted when we think those assumptions may be suspect.

A second approach, introduced by Browning *et al.* (n.d.), applies maximum entropy methods to estimate the parameters of the classical APC model. At present this method can only be applied for outcome variables with closed support that can be described by a linear model, although the conclusion of the paper mentions that further research may enable the extension of the method to additional forms of outcome variable (Browning *et al.*, n.d., p. 17). In the paper it is shown that if the outcome variable has closed support, then so must the parameter vector θ. Constraints relating to consistency with μ can then be imposed on this closed parameter space, and estimate a maximum entropy distribution over that constrained space. The maximum entropy APC estimator (MEE) is chosen as the mean of this maximum entropy distribution. In a way this can be said to be the most statistically plausible, or statistically safe, estimate of the parameter values (the authors provide an interesting game-theoretic justification for the estimator); however, a further step is needed to explain whether the statistically safest estimate might be expected to provide a good approximation to the true effects in the APC context. In other contexts, a statistically safe estimate such as MEE or the less-sophisticated OLS does provide some rough guidance as to what the data

looks like; it will be imperfect, but not misleading. However, the fundamental interdependence between concepts in the APC context can cause the MEE to be quite misleading, as Voas (2013) illustrates with regard to data on religious observance. This suggests that further consideration of the relationship between what the estimator approximates (the mean of the maximum entropy distribution) and the object of interest is required before the MEE could be used with confidence in exploratory data analysis.

Given the limited valid scope of application of the HAPC model and the need for further analysis of the MEE approach, it is clear that neither of these presently offers a generally applicable method for thinking about APC effects in individual-level data. Since the estimable functions approach has proved promising in the context of aggregate data, it seems worthwhile to seek to extend it to individual-level data. The remainder of the thesis constitutes a first attempt at such an extension.

3 Estimable Functions with Individual Data

In this section I describe the theoretical steps in the extension of the estimable functions approach to APC estimation from aggregate to individual-level data. The exposition begins with the simple case in which there are no covariates, only the outcome variable and the APC structure. While there are no substantial differences between the aggregate and individual-level data models in this case, it is included as an important step in the progression of thought and as a way to introduce the reader to the model in the individual context. Subsequently I introduce covariates to the model. Towards the end of the section I outline the possible submodels of the APC model, which may be used in place of the full APC model to give a more parsimonious description of data. The assessment of the fit of these models is discussed in section 4, and implementation of the methods discussed in both sections is described in section 5.

3.1 Age, Period, and Cohort Effects

Recall the standard dummy-variable APC model for aggregate data given by equation (2.2) in section 2.2.2:

$$\mu_{ik} = \alpha_i + \beta_j + \gamma_k + \delta$$

where μ_{ik} is the linear predictor of the value of the dependent variable for cell ik on the age-cohort grid, α_i is the effect of being at age i, β_j is the effect of being in period j, and γ_k is the effect of being a member of cohort k, while δ is a grand mean. Again, the identity $j = i + k - 1$ means that is unnecessary to add a j subscript to the linear predictor. In the aggregate-data setting there was a single data point in each ik cell, which could be for example a count variable, a rate, or a mean value of some dependent variable for observations in that cell.

Moving from aggregate to individual-level data, where there was previously a single data point in each cell there are now many. These data points might be binary outcomes, each taking a value 0 or 1, or continuous outcomes, each data point taking a unique value, or some

other form of outcome variable. In the case with no covariates, the only information that can be used to model these outcomes is the information about their position on the age-cohort grid; thus all data points within the same cell must have the same modelled outcome, i.e. the same linear predictor. Then letting $n = 1, ...N$ index all of the individuals in the sample (note that there are N individuals in total rather than N in each cell; note also that we are agnostic as to how these N individuals are distributed across the cells) we can write

$$\mu_{ikn} = \mu_{ik} = \alpha_i + \beta_j + \gamma_k + \delta$$

where μ_{ik} is the linear predictor for any n that is a member of cell ik. Consequently, the reparameterization of Kuang, Nielsen, and Nielsen described in section 2.3.2 above will proceed exactly as before, resulting in the following expression for the linear predictor of an observation in a rectangular age-cohort data set:

$$\begin{aligned} \mu_{ikn} \quad = \mu_{ik} = \quad & \mu_{11} + (i-1)\mu_{21} + (k-1)\mu_{12} \\ & + \sum_{t=2}^{i}\sum_{s=3}^{t} \Delta^2\alpha_s + \sum_{t=2}^{j}\sum_{s=3}^{t} \Delta^2\beta_s + \sum_{t=2}^{k}\sum_{s=3}^{t} \Delta^2\gamma_s \end{aligned} \tag{3.1}$$

Since the generalisations discussed in section 2.3.2 (changing the initial points that determine the linear plane, and expanding from rectangular rectangular age-cohort structures to generalized trapezoids) are both achieved by algebraic manipulation of the expression for the predictor, and since this expression is unchanged in model (3.1), both will continue to be valid for individual-level data. We therefore arrive with little difficulty at the standardized model from Nielsen (2014) that is valid for any data structure

$$\mu_{ikn} = \mu_{UU} + (i-U)\mu_{U+1,U} + (k-U)\mu_{U,U+1} + A_i + B_j + C_k$$

with A_i, B_i, and C_i defined as in equation (2.8).

The only difference between the cases of aggregate data and individual-level data without

covariates will arise when we come to estimate the parameter

$$\xi = \{\mu_{UU},\, \mu_{U,U+1},\, \mu_{U+1,U},\, \Delta^2\alpha_3,\, ...\Delta^2\alpha_I,\, \Delta^2\beta_{L+1},\, ...\Delta^2\beta_{L+J},\, \Delta^2\gamma_3,\, ...\Delta^2\gamma_K\}$$

in that the estimation procedure must be equipped to deal with a substantially increased number of data points. This is discussed further in section 5.2.

3.2 Introducing Covariates

In the preceding section the KNN reparameterization was shown valid for individual-level data without covariates. I now add covariates to the individual-level form of the standard APC model, which becomes

$$\mu_{ikn} = \alpha_i + \beta_j + \gamma_k + \delta + C_{ikn}\zeta \tag{3.2}$$

Here C_{ikn} is a row vector of covariates of length d_C (the number of covariate variables included in the model) and has a value that is both individually and temporally specific, while the parameter on C_{ikn}, ζ, is constant across observations. Because C_{ikn} introduces an individual-specific component to the right-hand-side of the expression for the linear predictor, it is no longer the case that all individuals within the same ik cell will have a common linear predictor. The size of the set of unique linear predictors will depend on the nature of the included covariates. If only categorical variables are used then some individuals will share a linear predictor, whereas if even a single continuous variable is included it is highly unlikely that any two individuals will share a predictor.

The steps taken in section 2.3.2 to reach the KNN reparameterization can mostly be repeated here. The first substantial difference arises at the line

$$\begin{aligned}\mu_{ikn} \;=\; & \alpha_1 + (i-1)\Delta\alpha_2 + \sum_{t=2}^{i}\sum_{s=3}^{t}\Delta^2\alpha_s + \beta_1 + (j-1)\Delta\beta_2 \\ & + \sum_{t=2}^{j}\sum_{s=3}^{t}\Delta^2\beta_s + \gamma_1 + (k-1)\Delta\gamma_2 + \sum_{t=2}^{k}\sum_{s=3}^{t}\Delta^2\gamma_s + \delta + C_{ikn}\zeta\end{aligned}$$

In the aggregate data case it was possible to substitute μ_{11}, μ_{12}, and μ_{21} into this equation.

These common linear predictors longer exist in the individual data model, and so instead we come to

$$\begin{aligned}\mu_{ikn} &= v_o + (i-1)v_a + (k-1)v_c \\ &\quad + \sum_{t=2}^{i}\sum_{s=3}^{t} \Delta^2\alpha_s + \sum_{t=2}^{j}\sum_{s=3}^{t} \Delta^2\beta_s + \sum_{t=2}^{k}\sum_{s=3}^{t} \Delta^2\gamma_s + C_{ikn}\zeta\end{aligned} \tag{3.3}$$

where $v_o = \alpha_1 + \beta_1 + \gamma_1 + \delta$; $v_a = \Delta\alpha_2 + \Delta\beta_2$; and $v_c = \Delta\gamma_2 + \Delta\beta_2$. It should be noted that although v_o, v_a, and v_c no longer have straightforward interpretation as elements of the set of linear predictors, their relationship to the other components of the reparameterized model is the same as that of the linear predictors μ_{11}, μ_{12},and μ_{21} to the other components in the reparameterized aggregate model.

The full parameter vector for this model is then

$$\{\xi, \zeta\} = \left\{v_o, v_a, v_c, \Delta^2\alpha_3, ...\Delta^2\alpha_I, \Delta^2\beta_3, ...\Delta^2\beta_J, \Delta^2\gamma_3, ...\Delta^2\gamma_K, \zeta\right\} \tag{3.4}$$

Under standard assumptions on C, the $N \times d_C$ covariate matrix created by stacking all N observations on C_{ikn}, the vector $\{\xi, \zeta\}$ is uniquely identified from the set of linear predictors. These standard assumptions are as follows:

1. no perfect multicollinearity between C and the APC structure, and

2. no perfect multicollinearity between the elements of C

Under these assumptions, the model in equation (3.3) constitutes a one-to-one mapping from $\mathcal{G}+\mathcal{Z}$, the space on which the parameter vector $\{\xi, \zeta\}$ is defined, and $\mathcal{M}$, the space on which the set of linear predictors is defined, so $\mu(\xi, \zeta) \neq \mu(\xi', \zeta')$. Further, the parameter vector $\{\xi, \zeta\}$ is invariant to the set of transformations of θ described by g in equation (2.4). These are the same properties that were earlier attributed to the model without covariates. In that model, these properties are fundamentally a result of the absence of exact linear relationships between elements of that model. Since the assumptions above preclude the addition of covariates which would introduce exact linear relationships to the model, the model with covariates inherits

these two appealing properties from the model without covariates.

The two extensions to the original KNN result (that an alternative linear plane may be chosen, and that the framework applies not only to age-cohort data structures but to generalized trapezoids) also remain valid here, since the first depends only on an algebraic manipulation of the APC structure and the parameter vector and the second arises by virtue of a simple observation about the summation of the double differences. Because these extensions remain valid, the same standardization of initial points proposed in Nielsen (2014) and shown above in figure 2.5 can be employed. The standardized parameter vector will be exactly as in equation (3.4), but with v_o, v_a, and v_c redefined so that $v_o = \alpha_U + \beta_U + \gamma_U + \delta$; $v_a = \Delta\alpha_{U+1} + \Delta\beta_{U+1}$; and $v_c = \Delta\gamma_{U+1} + \Delta\beta_{U+1}$. This standardized model can be written compactly as

$$\mu = X_{APC}\xi + C\zeta \tag{3.5}$$

It is worth briefly discussing the covariates which are, and are not, permissible under the first assumption on C. Since this APC structure is a reparamaterisation (i.e. linear combination) of the original APC structure in equation (3.2), any C that can be constructed a linear combination of elements of the dummy structure will result in a failure of identification. For example, a dummy variable for recession years or war years would be inadmissible. A researcher interested in the effect of recession on some outcome variable would instead need to examine the estimated period double differences for evidence of an effect. However, a dummy for some particular age-period combination would be acceptable. For example, if studying unemployment trends over time one could include a dummy for being 18-25 during a recession (note the difference between "a dummy for being 18-25 during recession" and "a dummy for those who were 18-25 during recession"; the latter will be perfectly collinear with a set of cohort dummies). Particular caution would also be required if including dummies for single combinations lying near the beginning or end of the sample; for example, if the data has $I = K = 3$ then a dummy for $i, k = 3, 3$ would be perfectly collinear with the dummy for the fifth period. Even if this caution is exercised it may be inadvisable to include dummies of this nature in C, as they complicate interpretation of the coefficients; furthermore, they will

create multicollinearity problems later in the analysis, if it is wished to test the validity of the restriction of temporal effects to the APC structure (see section 4.2.2).

It is also worth noting that the covariate has been introduced here in linear additive form. This sort of model is most appropriate for factors that are anticipated to affect the outcome variable independently of the three temporal effects, but might also be related to them; that is, variables that are candidates for omitted variable bias in estimation of the APC structure. Alternatively covariates could be introduced by interacting them with the elements of the APC structure. This would be appropriate for variables that are expected to mediate the effect of the APC structure on the dependent variable. Introducing covariates in this alternative way should not introduce additional complications regarding identification, as these interaction terms can simply be counted as elements of C. It is, however, suggested that such interactions should be considered in groups. For example, it would be unwise to interact a covariate with only one of the two slopes or only some of the double differences in age, as it would make interpretation of the results extremely difficult.

3.3 Submodels of the APC model

The reparameterized APC model described above consists of a linear plane in age-cohort co-ordinate space and three series of sums of double differences: one for each temporal dimension. This model is not fully general, as it inherits the restriction of the classical APC model from which it is derived of no interaction effects between the three temporal factors, mentioned in section 2.2.2. That said, it still permits a non-linear non-parametric effect in all three temporal dimensions as well as a linear plane. In applications, we will be interested in investigating whether a more parsimonious model can accurately describe the temporal variation. In particular we will want to consider submodels of the APC model, which exclude one or more of its constituent components.

The possible submodels are described in table 1, copied from Nielsen (2014, p. 14). They are ordered vertically from most to least general. submodels $\{AP, AC, PC\}$ each exclude one set of sums of double differences, while the submodels $\{Ad, Pd, Cd\}$ (termed "drift" models

	level	slopes		double differences			df
		age	cohort	$\Delta^2\alpha$	$\Delta^2\beta$	$\Delta^2\gamma$	
APC	*	*	*	*	*	*	$I+J+K-3$
AP	*	*	*	*	*	0	$I+J-1$
AC	*	*	*	*	0	*	$I+K-1$
PC	*	*	*	0	*	*	$J+K-1$
Ad	*	*	*	*	0	0	$I+1$
Pd	*	*	*	0	*	0	$J+1$
Cd	*	*	*	0	0	*	$K+1$
A	*	*	0	*	0	0	I
P	*	=	=	0	*	0	J
C	*	*	0	0	0	*	K
t	*	*	*	0	0	0	3
tA	*	*	0	0	0	0	2
tP	*	=	=	0	0	0	2
tC	*	0	*	0	0	0	2
1	*	0	0	0	0	0	1

Table 1: Overview of submodels. "0": factor set to zero, "*": factor unrestricted, "=": two factors restricted to have same coefficients.

by Clayton & Schifflers (1987)) exclude two; that is, deviation from the linear plane is allowed in the dimensions included in the name of the submodel only. Note that for all six models, the linear plane is still plotted in age-cohort space. Restrictions on the linear plane are first introduced in submodels $\{A, P, C\}$. These submodels consider whether a good description of the variation in the dependent variable can be achieved using one temporal dimension. In addition to restricting the double differences the linear plane is allowed a slope in one direction only, with that direction determined by the choice of submodel. While it would be possible to think about and plot these models in two-dimensional space (rather than the three-dimensional age-cohort-dependent variable space used up to this point), for comparability it is better to study all models in the same coordinate space. This is why the P model is described by restricting the age and cohort slopes to be equal to one another, rather than in terms of a new period slope. Even more restrictive submodels are possible: t allows for a two-dimensional linear plane only (with no double differences); $\{tA, tP, tC\}$ further restrict us to a one-dimensional slope; and finally 1 includes an intercept only. In practice these last few submodels are very unlikely to be accepted by any test of fit on real data.

4 Assessing Fit of Individual-Level Models

In this section, I describe a test of fit which can be used to determine which of the submodels described in the previous section best describes the inter-temporal variation in a particular individual-level data set. In general, in order to assess the validity of a proposed model one may look at t-statistics for individual elements of the model or at aggregate measures of fit. For APC models, the complexity of the proposed structure means that there are too many elements, which must in any case be interpreted in light of one another, to take a t-statistic based approach. Therefore the focus throughout this discussion will be on aggregate measures of fit.

It is also important to establish that when we consider the fit of the APC model we are thinking specifically about its adequacy as a description of inter-temporal variation in the data. We are not concerned with the extent to which any included covariates adequately model the intra-temporal variation in the data. This is because the temporal patterns are assumed to be the primary object of interest in APC analysis, with covariates introduced primarily to avoid omitted variable bias as discussed in section 3.2 (note that if we were primarily interested in the covariates and wished only to control for temporal effects this could be achieved without need for an APC model; see for example Ejrnæs & Hochguertel (2013)).

I begin this section with a description of the way in which the fit of the reparameterized model is assessed in existing work involving discrete, aggregate data. I then explain why this approach is unsuitable for individual-level data and introduce an alternative method of assessing fit in this context. Finally I highlight the computational challenges encountered in implementing this alternative method, and introduce a method of circumventing these challenges based on an understanding of the mechanics of multivariate regression.

4.1 Test of Fit for Aggregate Data

In existing applications of the estimable functions approach to aggregate, discrete data, the APC model and its submodels are evaluated using likelihood ratios (Nielsen, 2014; Martínez Miranda *et al.*, 2015). Likelihood ratios comparing the submodels to the full APC model are

considered, as are deviance statistics for the full APC model and each of the submodels. For discrete outcome variables, the deviance statistic is equivalent to a likelihood ratio statistic comparing the model of interest to a fully-saturated model, i.e. a model which allows one parameter per observation (so in the context of aggregate APC data, the fully-saturated model includes one parameter per ik cell in the age-cohort coordinate space). The fully-saturated model then has a likelihood of 1, and thus log-likelihood of 0. To conduct inference, the deviance is compared to a χ^2 distribution with appropriate degrees of freedom, determined by the number of parameters in the two models under comparison. That is

$$D_{APC} = LR_{APC,TS} = -2\left[\ell_{APC} - \ell_{sat}\right] \sim \chi^2\left(p_{sat} - p_{APC}\right)$$

where D_q is the deviance, ℓ_q is the log-likelihood, and p_q is the number of parameters for model q. The subscripts APC and sat represent the APC and fully-saturated models respectively, with $LR_{APC,TS}$ indicating the likelihood ratio comparing the two models. Large p-values for this statistic indicate that there is no significant difference between the two models' effectiveness in describing the variation in the data. Given that we are interested in finding a parsimonious model, large p-values are generally taken as licence to focus on the smaller of the two models compared, which in this case would be the APC model or relevant submodel.

4.2 Test of Fit for Individual Data

4.2.1 Problems with the Deviance Statistic

When working with individual-level data, it is not appropriate to compare the APC model to a fully-saturated model that allows a parameter for each observation. Because each observation is now an individual and there are therefore many observations per ik cell, this fully-saturated model perfectly describes all of the variation in the data, both inter-temporal and intra-temporal (variation between individuals in the same ik cell). However, as mentioned at the beginning of this section, the objective of the APC model is to describe the inter-temporal variation only. A test of the APC model against a fully-saturated model at the individual level

is therefore overly demanding; it evaluates the effectiveness of the APC model in describing variation which it is not the intent of the APC model to describe.

Instead of the fully-saturated model, we would like to compare the APC model to one that is maximally agnostic as regards the temporal patterns in the data but does not attempt to describe intra-temporal variation. It is proposed to use for this purpose a model which introduces a dummy variable for each ik cell. Such a model allows the effect of being at any given position in temporal coordinate space to be entirely unrelated to that of being at any other position in temporal coordinate space, and can therefore be considered to be maximally agnostic regarding time patterns. It is represented by

$$\mu_{ikn} = \kappa_{11}1(ik = 11) + \kappa_{12}1(ik = 12) + ...\kappa_{IK}1(ik = IK) \tag{4.1}$$

where κ_{ik} is the estimated effect of being in cell ik and μ_{ikn} as usual represents a linear predictor for observation ikn (where in this case it is the linear predictor based on this new model rather than the APC model). Note that since there are no covariates in (4.1), all individuals in the same ik cell will share a common linear predictor: that is, $\mu_{ikn} = \mu_{ik}$ as in section 3.1. This means that there will be only IK unique linear predictors. Of course, this model is identical to the fully-saturated model for aggregate data; however, to avoid possible confusion with the fully-saturated model for individual data, it is referred to as the temporally-saturated (TS) model.

A more concise representation of the TS model in equation(4.1) is given by:

$$\mu_{ikn} = X_{TS,ikn}\kappa \tag{4.2}$$

where $\kappa = \{\kappa_{11}, \kappa_{12}, ...\kappa_{1K}, \kappa_{21}, ..., \kappa_{IK}\}$ is a column vector of all the dummy variable coefficients, and $X_{TS,ikn}$ is a $1 \times IK$ row vector containing all zeros except for the element that corresponds to the κ_{ik} appropriate to the ik cell in which observation n is found (that is, the $(i \times (k-1) + k)$th element), which takes a 1. The complete set of linear predictors can be

represented as

$$\mu = X_{TS}\kappa \tag{4.3}$$

where X_{TS} is a matrix constructed by stacking the vectors $X_{TS,ikn}$ and μ is the column vector containing all N linear predictors.

For individual-level data on discrete outcome variables without covariates, the fit of the APC model or any of its submodels can be assessed by using a likelihood ratio test to compare it with this TS model. This test is conducted for the same reason that the deviance test was used in the context of aggregate data: to get a sense of how suitable the restrictions inherent in the APC structure are for the data in question.

4.2.2 Moving Beyond Discrete Outcomes without Covariates

A likelihood ratio test against the TS model described above is an appealing alternative to the deviance statistic as a tool for assessing the fit of the APC model to individual-level data on discrete outcome variables without covariates. However, this is not the only context in which APC models are used. I next discuss the extension of the TS model to incorporate covariates when these are included in the APC model, and consider whether a direct likelihood ratio test is the best approach when dealing with continuous dependent variables.

When introducing covariates that appear in the APC model to the TS model, it is important to recall the role played by covariates in the original APC model. Since the purpose of APC models is to understand temporal variation in data, the primary function of covariates is to act as controls for omitted variable bias. Insofar as the same omitted variable bias could affect the TS model, it seems reasonable that the alternative time-saturated model should incorporate the covariates using the same structure was employed in the original APC model. It is not necessary for the TS model to allow for all of the possible forms that the relationship between covariate and outcome variable could take (higher-order polynomials, logarithmic functions, etc.) where these are not considered in the APC model. The linear predictor for the time-saturated model with covariates should then be the same as that for the APC model with covariates in equation (3.5), except that the APC structure is replaced by the TS structure.

That is, it should be of the form

$$\mu = X_{TS}\kappa + C\lambda \tag{4.4}$$

where λ is the coefficient on C in the TS model, which is expected to differ from ζ, the coefficient in the APC model. The TS model can be extended in this way for almost any covariate matrix C; however, it should be noted that covariates that constitute interaction terms between temporal dimensions, such as the young person-recession interaction described in section 3.2, will be perfectly collinear with elements of the TS structure and would have to be dropped for this fit assessment to proceed. In light of this, it may be wisest to simply not include such interactions in the first place.

Thus far I have focused on discrete outcome variables, modelled with and without covariates. In these settings we evaluate the APC against the TS model by constructing a likelihood ratio and comparing that ratio a χ^2 distribution, which has asymptotic justification. However, when we work with continuous dependent variables that have an approximately normal distribution, and can therefore employ a standard OLS model, we can use an F-test in place of the likelihood ratio test. The F-statistic is a transformation of the likelihood ratio that has two advantages in this context: first, it has a known finite-sample distribution, and second, it can be constructed based on the residual sum of squares (RSS) according to the formula

$$F = \frac{(RSS_{APC} - RSS_{TS})/(p_{TS} - p_{APC})}{RSS_{TS}/(N - p_{TS})} \tag{4.5}$$

where p_q is the number of parameters in model q and N is the number of observations in the data. This statistic has an $F(p_{TS} - p_{APC}, p - p_{TS})$ distribution, which can be used to conduct inference. Again, a large p-value indicates that there is no significant difference between the two models, so that with an eye to parsimony we can proceed on the basis that the APC model or relevant submodel provides an adequate description of inter-temporal variation in the data.

4.3 Computational Challenges for the Test of Fit

In the preceding section it was suggested that the fit of the APC model as a description of the inter-temporal variation in data may be assessed by comparison with the TS model, using either an LR test or F-test as appropriate. In order to make this comparison, it is necessary to explicitly estimate the TS model in addition to the APC model. While direct estimation of the TS model using standard methods is feasible for small data sets, a computational problem was encountered when trying to estimate the TS model for the larger BMI data set. This data covers 19 years and 85 cohorts, so the TS model contained over 1000 explanatory variables which proved beyond the capacity of the usual regression command of a standard statistical package. An alternative approach was required.

In the case of a binary dependent variable modelled on the APC structure alone (i.e. without covariates), one simply estimate the deviance statistic from the aggregate data model, as it can be shown that in this case the aggregate deviance statistic is equal to the likelihood ratio between the APC and TS models at the individual level. However, many of the models we want to examine for individual-level data will involve a continuous dependent variable, covariates, or both, and so a more general solution is needed. Efforts to find a method to estimate the TS model in the context of discrete data with covariates have not yet proved fruitful. However, a method for continuous dependent variables, with and without covariates, has been developed and is outlined here.

First, a special feature of the TS model for continuous data without covariates was noticed. Since the matrix of explanatory variables (i.e. the TS dummies) consists of stacked unit vectors, the matrix to be inverted when calculating the OLS estimator is diagonal. This diagonality should substantially reduce the computational burden in estimation. To see this, let the relationship between the outcome variable Y and the TS structure X_{TS} be given by

$$Y = \mu_{TS} + \varepsilon = X_{TS}\kappa + \varepsilon \tag{4.6}$$

where κ, and X_{TS} are as defined in equation(4.3), μ_{TS} is the vector of linear predictors from

the TS model, and ε is an error term. We can then write

$$\hat{\kappa} = \left(X'_{TS}X_{TS}\right)^{-1} X'_{TS}Y$$

Now, $(X'_{TS}X_{TS})$ is an $IK \times IK$ diagonal matrix where each element of the main diagonal is the number of observations in the associated ik cell (that is, the $(iK + k)$th element is the number of observations in the grid cell ik). Thus even for a very large age-cohort coordinate space, the inverse is straightforward to construct by simply inverting each element of the main diagonal. Looking to the second part of the estimator, we see that $X'_{TS}Y$ is an $IK \times 1$ vector where each element is the sum of the observed values of y_{ik} for a given ik cell. Combining these, $\hat{\kappa}$ can be seen to have the following structure:

$$\hat{\kappa} = \left(X'_{TS}X_{TS}\right)^{-1} X'_{TS}Y = \begin{bmatrix} \frac{1}{n_{11}} & 0 & \cdots \\ 0 & \frac{1}{n_{12}} & \\ \vdots & & \ddots \end{bmatrix} \times \begin{bmatrix} \sum_{ik=11} Y_{ik} \\ \sum_{ik=12} Y_{ik} \\ \vdots \end{bmatrix} = \begin{bmatrix} \frac{1}{n_{11}} \sum_{ik=11} Y_{ik} \\ \frac{1}{n_{12}} \sum_{ik=12} Y_{ik} \\ \vdots \end{bmatrix} \tag{4.7}$$

That is, $\hat{\kappa}$ is an $IK \times 1$ vector that simply contains the mean values of the dependent variable among the observations where the associated dummy variable takes value 1, i.e. the mean among the observations which fall into the relevant ik cell. This means that it is not in fact necessary to run a regression, as the TS coefficients can be estimated by simply calculating the ik cell means. These estimates are sufficient to construct the vector of predicted values $\hat{Y}$, which we can use to calculate the residual sum-of-squares and thereby construct an F-statistic.

Extending this insight to the case of continuous dependent variables with covariates requires exploitation of our understanding of the internal workings of multivariate regression. Allow the model to be estimated to be represented as:

$$Y = X_{TS}\kappa + C\lambda + \varepsilon \tag{4.8}$$

where Y and X_{TS} are defined as above, C is the covariate matrix, and λ and κ are coefficient vectors to be estimated. Let d_C be the number of covariates, so that $\dim(C) = N \times d_C$ and

$\dim(\lambda) = d_C \times 1$. Given this structure, a standard statistical software package will combine the two sets of regressors into a single matrix and try to compute the OLS estimator for the column vector $\left\{\hat{\lambda},\, \hat{\kappa}\right\}$, which has $\dim\left((d_C + IK) \times 1\right)$, directly. This involves inverting a non-diagonal matrix of dimension $(d_C + IK) \times (d_C + IK)$, which is computationally expensive for large IK.

A better approach is to recognise that the Frisch-Waugh-Lovell Theorem can be used to take advantage of the diagonality of $(X'_{TS}X_{TS})$. Rather than estimating the full set of parameters simultaneously, we proceed by "partialling out" X_{TS} from C. We can then separately regress Y on X_{TS} and on the part of C that is orthogonal to X_{TS} (which we call $\tilde{C}$). The resulting coefficient estimates can be used to calculate the multivariate regression estimators for κ and λ. More completely, the algorithm follows the procedure outlined below.

First, note that C can be modelled in terms of X_{TS} according to

$$C = X_{TS}\phi + \omega$$

where $\dim(C) = N \times d_C$, $\dim(X_{TS}) = N \times IK$, $\dim(\phi) = IK \times d_C$, and $\dim(\omega) = N \times d_C$. This model has the same general form as (4.6), so the construction of the estimator $\hat{\phi}$ is straightforward due to the diagonality of $(X'_{TS}X_{TS})$. For each column C_d of C, where each column represents a single covariate, we construct the $IK \times 1$ estimator $\hat{\phi}_d$ according to the logic in equation (4.7), and then combine these into the $IK \times d_C$ matrix $\hat{\phi}$. Using the estimated $\hat{\phi}$ we can then construct $\hat{C} = C - X_{TS}\hat{\phi}$, where $\tilde{C}$ approximates ω.

Now consider equation (4.8). $X_{TS}\hat{\phi}\lambda$ can be added and subtracted to this equation as follows

$$\begin{aligned} Y &= X_{TS}\kappa + C\lambda - X_{TS}\hat{\phi}\lambda + X_{TS}\hat{\phi}\lambda + \varepsilon \\ &= X_{TS}\left(\kappa + \hat{\phi}\lambda\right) + \left(C - X_{TS}\hat{\phi}\right)\lambda + \varepsilon \\ Y &= X_{TS}\rho + \tilde{C}\lambda + \varepsilon \end{aligned} \tag{4.9}$$

where we have defined $\rho = \kappa + \hat{\phi}\lambda$. Now, since X_{TS} and $\tilde{C}$ are orthogonal by virtue of the

construction of $\tilde{C}$, the OLS estimators of ρ and λ will be given by

$$\hat{\rho} = \left(X'_{TS}X_{TS}\right)^{-1} X'_{TS}Y \quad \textit{and} \quad \hat{\lambda} = \left(\tilde{C}'\tilde{C}\right)^{-1} \tilde{C}'Y$$

Thus $\hat{\rho}$ can be calculated from a model of Y on X_{TS} only (which is straightforward because of the diagonality of $(X'_{TS}X_{TS})$), and $\hat{\lambda}$ from a regression of Y on $\tilde{C}$ only. Since the number of covariates included in the model for Y is not expected to be unduly large, the inversion of $\left(\tilde{C}'\tilde{C}\right)$ should be feasible, so standard OLS regression methods can be used to construct $\hat{\lambda}$.

The estimated residuals from the model (4.9) will be identical to those from the model (4.8), and so it is not necessary to construct $\hat{\kappa}$ to generate the RSS we need to construct the F-statistic. However, if we do want estimates of κ, they can be constructed by simply combining the existing estimators according to $\hat{\kappa} = \hat{\rho} - \hat{\phi}\hat{\lambda}$.

This procedure makes it possible to estimate the TS model for continuous data even for large data sets with in excess of 1000 distinct ik cells, and so to examine the general fit of the APC model and its submodels to inter-temporal variation in data. Unfortunately the procedure does not work for discrete outcome variables, as the coefficients for these models are not estimated by OLS and so the Frisch-Waugh-Lovell Theorem does not apply. Further research is therefore needed to develop a method to enable estimation of the TS model for discrete data with samples involving covariates and many ik cells.

5 Implementation in R

A complete R package for estimation of the reparameterized APC models for aggregate data, called `apc`, is available online (Nielsen, 2015). Because the commands in the package contain substantial embedded structure pertaining to aggregate data and are designed to feed into one another, it was mostly not possible to use them directly to conduct estimation on individual-level data. A separate sequence of commands was therefore created for this estimation, albeit with much reliance on the internal structure of the existing package. In this section a brief overview of the structure of the existing package is given, followed by an account of the major modifications required for use with individual-level data.

5.1 The `apc` Package

This overview of the `apc` package follows closely that in the package reference manual (Nielsen, 2015), and is provided here for the reader's convenience. The first step in the use of this package is to introduce the data. This is done using the command `apc.data.list`. The user must provide the data in matrix format and indicate the structure of the data (age-cohort, age-period, etc.). It is assumed that the aggregate data is provided in the grid format of the relevant structure; so for example AC data is provided in a table with ages as columns and cohorts as rows, and so on (see for example the mean BMI data in figure2.1). The output of this command is a list containing the data and key information about the data, which is required by later commands.

Once the data has been introduced, it can be visualised using `apc.plot.data`. This produces a series of two-dimensional plots which can be used to get a general idea of the sort of patterns that might be plausible for a given data set. A good example of how similar plots might be used to infer patterns in the data is given by Voas (2013).

Following visualisation one can proceed with model fitting. Two fitting commands are provided. The first, `apc.fit.table`, internally fits the full APC model and all submodels to the data, and produces a table containing the measures of fit described in section 4.1 (the deviance statistic, likelihood ratio for submodels compared to the full APC model, and Akaike

Information Criterion). The purpose of this command is to indicate to the user which of the submodels should be considered for more detailed analysis. This command takes two inputs only: the output from `apc.data.list`, and the choice of distribution used in the model estimation (e.g. Poisson, Binomial). The second fitting command, `apc.fit.model`, fits only one of the APC submodels, which the user must indicate as an input (in addition to the two inputs already required for `apc.fit.table`).

The output of `apc.fit.model`, including for example a table of coefficients, can be examined directly to understand a specific model. Alternatively the output can be used as an input to `apc.plot.fit`, which produces a grid of nine plots: the double differences, the slopes and level of the linear plane, and the sums of double differences. All nine are provided together because, as was discussed in section 2.3.2, the sums of double differences must be interpreted jointly with the linear plane. Examples of these plots can be seen in section 6.

Other commands are available in the `apc` package, but because they do not currently have counterparts in the sequence of commands for the individual case they are not discussed here.

5.2 Commands to Fit APC Models to Individual Data

Data at the individual level cannot be provided in the grid format that the `apc` package requires of aggregate data, as a third dimension would be required to record the individual information. Instead it is assumed that individual-level data is provided in a standard long-table format, with a column for each variable and rows representing individuals. It is further assumed that the data contains columns labelled `age`, `period`, and `cohort` (if only two of these are present, the third must be constructed before using the commands below; further, if one of the columns is misnamed, such as `year` instead of `period`, this must be changed before using the commands). For an example of the assumed structure see the excerpt from the BMI data in Appendix A.

The fitting of the model proceeds in three steps. First, a rank-deficient design matrix is constructed which includes the full APC structure and all additional variables; then, we a separate command is used to select desired columns from the rank-deficient matrix to form

the design matrix for the model of interest. This procedure was split into two steps to allow multiple models to be tested on the same data by simply changing the column selection, rather than needing to build a matrix from scratch each time. In the final step, the model of interest is estimated. The output of this command is compatible with `apc.plot.fit` from the original `apc` package.

The command to construct the rank-deficient design matrix builds on a function internal to the fitting commands in the `apc` package, called `apc.get.design.collinear`. This function first converts the indices from the "true" ages, periods, and cohorts to standardized indices beginning with the number 1 (although for data structures other than age-cohort it should be noted that the first observed period in the data set will not be period 1 but will instead be period $L + 1$, where L is defined as in section 2.3.2). It then uses the general equation for the linear predictor in equation (3.5), incorporating the standardized choice of initial points advocated by Nielsen (2014), to construct a row of the APC design matrix for each different ik value. The design matrix, given by X_{APC} in equation (3.5), indicates the position of the ik cell on the linear plane (through $(i - U)$ and $(k - U)$) as well as the number of times each double-difference should be counted for that ik cell. The only slight difference from equation (3.5) is that the slope in the period dimension relative to the initial points is included in the structure matrix, along with the usual slopes in the age and cohort dimension. This period slope is used in estimation of submodels $\{P, tP\}$, and is what renders this design matrix collinear.

A section of an example collinear design matrix is provided in table 2. This particular section shows only the first eight columns (notice that only age double differences are shown here; the period and cohort double differences appear in later columns) for four individuals of the same age but in different cohorts. The level appears in all linear predictors, so a 1 appears in the first column in every row. These four individuals are at an age 33 years before that in the reference ik cell UU, so -33 appears in the column for the age slope. The period slope column indicates how many periods after that of UU the individual is observed, and the cohort column indicates how many cohorts lie between them and the reference cohort in UU.

level	age slope	period slope	cohort slope	$\Delta^2\alpha_{22}$	$\Delta^2\alpha_{23}$	$\Delta^2\alpha_{24}$	$\Delta^2\alpha_{25}$
1	-33	0	33	1	2	3	4
1	-33	1	34	1	2	3	4
1	-33	2	35	1	2	3	4
1	-33	3	36	1	2	3	4
1	-33	4	37	1	2	3	4

Table 2: Section of example collinear APC design matrix

The remaining columns show how many times each double-difference in age is included in the linear predictor for that ik cell.

In the individual case, this collinear version of the X_{APC} design matrix must include as many rows for each ik cell as there are individual observations in that ik cell, and will have the matrix C (containing the values of the covariates for each individual) appended to it. It would, however, be exceedingly inefficient to construct all of the rows of this much larger matrix individually, when the X_{APC} elements will be common to all individuals in the same ik cell. Therefore in the command for individual data the first task is to extract a list of the unique ik values appearing in the data set. The construction described above is then performed on the list of unique values, and the resulting structure matrix is merged with the original data matrix (containing the ik values and covariates) based on the ik values. Since the merge is a computationally non-trivial matching exercise, requiring a matrix of several thousand rows to be matched to one of potentially several hundred values, this was one of the main reasons for the decision to split the design matrix construction into two steps. Some time was also invested in determining which of the several available matching commands was most efficient, and the `join` command from the package `plyr` was selected (Wickham, 2011).

The resulting collinear matrix is then fed to a second command along with an indication of which APC submodel is desired, the name of the dependent variable, and a list of names of covariates of interest as they appear in the data set. The use of this information to generate the particular model design matrix from the collinear design matrix follows the logic in the `apc` package's internal function `apc.get.design`. The elements of the APC structure excluded by the various submodels are exactly as indicated in table 1, with the exception that submodels

$\{P, tP\}$ are fit using the period slope directly rather than by constraining the age and cohort slopes to be equal (in practice, these two procedures are equivalent, but the former is easier to implement).

The last step is estimation of the model. A third and final command takes the new design matrix and dependent variable along with an an indication of the distribution to be used when fitting the model (at present this must be either Gaussian or binomial), and feeds this information through to `glm.fit` from the basic `stats` package which actually performs the estimation(R Core R Core Team, 2015). The command then reorganises the resulting information so that it is compatible with the `apc.plot.fit` command.

5.3 Commands to Assess Fit of Models for Individual Data

A few minor modifications to `apc.fit.table` were required to fit the original deviance-type tables to individual data; see for example table 11 in section 6.3.4. The bulk of the work to implement tests of fit, however, lay in implementing the algorithm for estimation of the temporally-saturated model described in section 4.

The procedure for estimating the TS model from section 4.3 is implemented in R in a four-step function. In the first step, the command `ddply` from the package `plyr` is used to generate the mean values of each covariate in C within each ik cell. As discussed in section 4.3, this $IK \times d_C$ matrix of mean values is equivalent to the OLS estimator of the parameter ϕ in the model $C = X_{TS}\phi + \varepsilon$. Further, since all rows of X_{TS} are unit vectors, the elements of $\hat{C} = X_{TS}\hat{\phi}$ based on this model will all simply be elements of $\hat{\phi}$, with the unit vectors that make up X_{TS} indicating which row of $\hat{\phi}$ should be assigned to each row of $\hat{C}$. This association is made using the `join` command in `plyr`, with the new matrix $\hat{C}$ appended to the original data matrix. It is then straightforward to construct a further matrix, $\tilde{C} = C - \hat{C}$. Note that it is not at this stage necessary to consider the possible relationships between the different elements of C, as this will be accounted for in the next step; at this point we simply wish to partial out the X_{TS} structure from C.

In the next step, `glm.fit` is used to regress Y on $\tilde{C}$. As outlined in section 4.3, it is assumed

that the number of covariates is small enough that this regression is not computationally demanding. The resulting estimated coefficient vector is $\hat{\lambda}$. Any possible relationships between elements of C are accounted for here as all elements of λ are estimated jointly.

In the third step, ρ is estimated following the same procedure outlined above for estimation of ϕ. The explanatory variable matrix, X_{TS}, is the same, but the dependent variable is Y rather than C. Finally, we construct the residuals for the TS model according to $Y - X_{TS}\hat{\rho} - \tilde{C}\hat{\lambda}$. These are also appended to the data matrix. In addition to the expanded data matrix incorporating these newly-constructed elements, the function outputs the residual sum of squares (RSS) which will be needed to calculate the F-statistic. Note that the function also calculates $\hat{\kappa} = \hat{\rho} - \hat{\phi}\hat{\lambda}$, although as discussed in section 4.3 it is not required to calculate the RSS.

The calculation of the F-statistic and its comparison to an appropriate F-distribution are reasonably straightforward in R. To facilitate comparison between models and submodels, a final function was written to fit the TS model and the full set of APC submodels, and produce a table of F-tests comparing these. This table parallels the likelihood-based output of `apc.fit.table` in the aggregate data package. Examples of these tables can be seen in the following section, where a full demonstration of the sequence of individual-data estimation commands is provided for an example application to the study of temporal trends in BMI.

6 Example Application: Health Survey for England

In this section the methods and tools described earlier in the thesis are employed to study the trends in BMI among English adults over a 19-year period. Section 6.1 highlights the economic interest of the topic and mentions some existing research which attempts to model BMI trends. A description of the HSE data that was made available is then provided. Finally a series of models are examined which both demonstrate some of the possibilities of individual-level APC analysis and provide interesting insight into the patterns in adult BMI in recent years.

6.1 Motivation

Rising obesity rates are a major public health concern worldwide, with associated healthcare costs alone estimated to approach £2bn in the UK by 2030 (Wang *et al.*, 2011). Obesity is also one of the public health concerns that is more amenable to economic analysis, since in most cases it is a consequence of individual decisions. Many plausible theoretical explanations have been offered in an effort to understand individuals' diet and exercise choices, using both neoclassical (Finkelstein & Strombotne, 2010; Cawley, 2011) and behavioural (Downs & Loewenstein, 2011) frameworks. Working from the premise that the high levels of obesity seen today are socially non-optimal, there is a clear interest in determining which if any of the proposed causes of rising obesity rates has the greatest explanatory power for the phenomenon so that interventions to reduce the rate of obesity can be targeted accordingly.

One way to narrow down the set of viable candidate explanations is to analyse the time trends in obesity rates and BMI data more generally. Certain time trends will be consistent with some of the proposed explanations but not others. For example, universal environmental factors such as the range and relative prices of available foods and portion sizes, or the shift towards sedentary leisure activities in the era of computers and the internet, should result in period effects. However, if the childhood formation of lifelong eating and exercise preferences are more salient, cohort effects will dominate. If this is the case, resources should be targeted towards altering environmental factors which most influence children, such as exercise opportunities and meal options in schools. A further possibility, given the ageing population

of OECD countries, is that rising obesity rates over time are at least in part due to the combination of this demographic shift and an age effect, whereby older adults are more likely to be overweight or obese and indeed have been for some time. Policy-makers may then want to consider more long-standing lifestyle factors, such as the tighter time constraints facing working-age adults which increase the opportunity costs of exercise and healthy eating, or the limited range of exercise options for older bodies.

It seems apparent that distinguishing between age, period, and cohort effects may be of use in discriminating between competing explanations of rising obesity rates. However, many existing studies of time trends in BMI data do not appear to have adequately accounted for the three different dimensions of temporal change. In the statistical methods section of their paper, Wang et al. note that they included age and calendar year as covariates, but make no mention of birth cohorts (Wang *et al.*, 2011, p. 819). A recent meta-study (NCD Risk Factor Collaboration 2016) which received widespread media coverage[2] relies on a modelling approach that also seems not to make provision for cohort effects (Finucane *et al.*, 2014). Not only does this mean that these studies may have lost valuable information which could help to distinguish between the candidate explanations for rising obesity, but it also implies that the estimated age and period effects are subject to possible confounding and may be unreliable. Some sociological studies have noted this problem and tried to account for all three of age, period, and cohort effects when analysing BMI trends, but they tend to make use of methods such as those discussed in section 2.3, which rely on strong and untestable assumptions (see for example Reither *et al.* (2009); Allman-Farinelli *et al.* (2008)). Thus there is clear scope for an analysis of BMI trends using the estimable functions approach.

6.2 Description of Data

The data set used for this project is an excerpt from the Health Survey for England (HSE) from 1995-2013. The HSE is an annual study funded by the Department of Health, which covers the population living in private households only. Households are selected for inclusion

[2]See for example http://www.bbc.co.uk/news/health-35933691 and
http://www.theguardian.com/society/2016/mar/31/one-fifth-of-worlds-adults-will-be-obese-by-2025-study-predicts (last accessed 07/05/2016)

in the survey based on "a multi-stage stratified probability sampling design" (Methods and Documentation 2015, p. 13), which is intended to ensure that the survey is representative of the UK population, and does not appear to have changed significantly over the lifetime of the survey. The response information is collected by a combination of face-to-face interview and direct measurement. In particular it is noted that height and weight are directly measured and so are not expected to suffer from self-reporting bias[3].

Our data set contained information on the height, weight, smoking history, and sex of the respondent as well as the age at which the respondent left full-time education and the highest educational qualification attained, for a purported 194201 observations. The first ten lines of this data set are reported in Appendix A to give the reader a sense of the data available. BMI was constructed using the appropriate formula for height reported in cm and weight reported in kg:

$$BMI = \frac{weight}{(height/100)^2}$$

To give an example of the calculation of BMI: the basketball player Stephen Curry, who stands at 191cm and weighs 86kg, has a BMI of 23.6. The World Health Organization defines BMI ranges as follows: $<$18.5 underweight; 18-25 normal weight; 25-30 overweight; $>$30 obese[4].

Examining the descriptive statistics for this raw data highlighted a series of potential concerns that it was necessary to address before beginning our analysis. The elimination of NAs in the anthropometric data resulted in the loss of 2371 observations. A subsequent query for minimum and maximum values revealed that implausible values remained in the anthropometric data, as seen in table 3. Further investigation suggested coding errors: for a number of observations, weight had been recorded as either 888 or 1179 and height had been recorded as 513 or 675. However, with a small number of implausibly low BMI and weight values also found in the data, it was determined that a series of cut-off rules would most appropriately address these problems. Height was top-censored at 250cm and weight at 560kg, while a minimum threshold of 12 was imposed for BMI. This reduced the number of

[3]Respondents do have the option to refuse height and weight measurement, which could be an alternative source of bias.

[4]see http://apps.who.int/bmi/index.jsp?introPage=intro_3.html (last accessed 08/05/2016)

	height	weight	BMI
min	73.20	10.00	0.46
mean	204.80	181.46	37.57
max	675.00	1179.00	2200.35

Table 3: Mean, minimum, and maximum of BMI, height, and weight in raw HSE data

observations to 172159, with most eliminated by the height and weight restrictions.

The non-numeric variables were then considered. There were no missing responses for the variable `sex`, and only 79 for the variable `smoke` (which records whether an individual currently smokes, smoked in the past, or has never smoked). The 79 observations with missing information were dropped. More substantial difficulties arose with the education variables. The possible responses "No answer/refused", "Don't Know", and "Item not Applicable" were determined to be unusable; these observations were dropped. 29165 of the remaining observations had given no response to `topqual2` (recording the highest educational qualification attained). As most of these had provided useful responses to `educend` (which records the age at which an individual left full-time education), it was decided to focus on the `educend` variable as the measure of education for the analysis. A further problem related to the treatment of those reporting that they were still in education. Since the `educend` variable is top-coded at "19 or over", the problem was resolved by eliminating all those aged 18 or 19 from the data set; thus, anyone still in full-time education must have been in education until the age of 19 at least, and these two groups could be merged.

age / period	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	104
2005	29	14	9	7	8	3	3	1	2	1						
2006	56	27	19	16	12	14	12	7	6	3	1	2				
2007	22	19	7	6	11	3	1	3	5	2	1					
2008	53	52	35	17	15	9	10	11	10	3	3		1			
2009	17	12	8	8	3	4	3	1		1	2					
2010	19	23	15	14	11	9	5	2	1	1	1					
2011	24	19	18	17	9	15	6	1	2	1	4					
2012	29	11	19	10	8	12	9	5	2		2	1	1			
2013	24	20	21	17	8	10	8	8	1	2	1		2		1	1

Table 4: Count of observations in BMI data at oldest ages, 2005-2013

Examination of the age, period, and cohort ranges for the remaining observations highlighted a final concern around sparsity among the earliest cohorts and oldest age groups. At this point the maximum age observed in the data was 104, and it was expected that there would be very few individuals of this age in the data. This is confirmed by the excerpt from the age-period array given in table 4 showing the count of observations per ij cell for a section of the data. To address this problem, it was decided to censor age such that there would be a minimum of ten observations per ij cell. This resulted in an age cut-off of 86. Descriptive statistics for the resulting data are provided in table 5.

Observation Count						
Total	166080	*Men*	76128	*Women*	89952	
Age, Period, and Cohort Ranges						
Age	[20, 86]	*Period*	[1995, 2013]	*Cohort*	[1909, 1993]	
BMI Quantiles						
0	*0.2*	*0.4*	*0.5 (mean)*	*0.6*	*0.8*	*1*
13.20	22.97	25.31	27.08	27.57	30.67	82.27
Age Leaving Full-time Education						
0	*14 or under*	*15*	*16*	*17*	*18*	*19 or over*
609	18393	34236	44908	13900	15493	38541
Smoking Behaviour						
Current smoker	40806	*Past smoker*	46127	*Never smoked*	79147	

Table 5: Descriptive statistics for cleaned HSE data used in analysis

6.3 Exploration of Data

The investigation of BMI trends in the cleaned data proceeded in three stages. First, models of BMI as a function of the APC structure only were considered. Subsequently covariates for sex, education, and smoking history were introduced. Finally, a selection of models intended to test the robustness of the identified relationship between BMI and the APC structure were examined.

6.3.1 BMI as a Function of APC Only

The analysis began with an investigation of the relationship between BMI as a continuous variable and the APC structure. The theoretical model here is of the form

$$\mathbb{E}\left(\texttt{BMI}_{ikn}\right) = X_{APC,ikn}\xi$$

where as before i and k index age and cohort, while n indexes the individual. Since we do not consider covariates at this point, this model is akin to that in equation (3.1) in section 3.1 and the predicted value of `BMI` for a given individual based on this model will be equal to the predicted value of `BMI` for the ik cell to which they belong.

The first step in consideration of this model is to inspect the table of fit statistics, seen here in table 6, to determine which of the APC submodels best describes the temporal variation in the data. The first column of this table identifies the submodel under consideration in a given row. The second and third columns report the F-statistic and appropriate degrees of freedom for the comparison of this model to the temporally-saturated (TS) model, and the fourth column computes the p-value found when the F-statistic is compared to the F-distribution with that many degrees of freedom. A p-value in excess of 0.05 suggests that there is no significant difference between the submodel under consideration and the TS model. If this is the case, the APC model or submodel is favoured on the basis of parsimony. In this example, it can be seen that there is support for reductions from the TS model to the APC, AP, or AC models.

	F-test vs TS	DF (*,164807)	p-value	F-test vs APC	DF (*,165912)	p-value	AIC
APC	0.99	1105	0.62				998518.09
AP	1.04	1188	0.16	1.76	83	0.00	998497.94
AC	1.07	1122	0.05	6.62	17	0.00	998596.66
PC	1.63	1170	0.00	12.55	65	0.00	999202.41
Ad	1.12	1205	0.00	2.60	100	0.00	998578.65
Pd	3.21	1253	0.00	19.85	148	0.00	1001137.14
Cd	1.68	1187	0.00	11.02	82	0.00	999255.99
A	2.05	1206	0.00	13.66	101	0.00	999691.60
P	6.27	1254	0.00	45.48	149	0.00	1004868.03
C	4.00	1188	0.00	44.20	83	0.00	1001984.72
t	3.28	1270	0.00	18.61	165	0.00	1001234.23
tA	4.19	1271	0.00	25.52	166	0.00	1002373.41
tP	6.29	1271	0.00	41.62	166	0.00	1004962.25
tC	5.46	1271	0.00	35.24	166	0.00	1003940.44
1	7.44	1272	0.00	50.12	167	0.00	1006358.60

Table 6: F-table comparing APC model and submodels: `BMI` as dependent variable, no covariates

The fifth, sixth, and seventh columns are similar to the second through fourth, differing only in that the TS model is replaced by the full APC model as the standard against which the submodels are compared. It can be seen that in this case, no significant difference is found between the APC model and either the AP or AC models. The final column reports the Akaike Information Criterion (AIC), which in this example is minimised by the AP model. The APC and AC models come close, although it is interesting to note that the Ad model actually has a lower AIC value than does the AC model. This collection of results suggests that the key deviations from linearity in this data will be in the age dimension.

On the basis of these results, one could reasonably choose to examine either the APC or the AP model (or both) more closely. Since this was a first examination of the data it was decided to look initially at the larger APC model, to get a fuller picture of the behaviour of the data in all three dimensions. Figure 6.1 gives a visual representation of this model. Sub-plots (a), (b), and (c) show the double differences as estimated at each age, period, and cohort. The black line represents the point estimates, while the blue and red lines are given at one and two standard deviations around zero respectively. Notice that in the plot of cohort double differences, the

estimated confidence bands become substantially wider and the estimates themselves are more variable at the extremes of the cohort plot; we will address this observation in a moment.

The linear plane is presented in sub-plots (d), (e), and (f). Although the indices on the x-axes of plots (d) and (f) are those of age and cohort, this is only because the linear plane is estimated in age-cohort space. It does not mean that these slopes can be interpreted as being driven by age or cohort effects; as discussed in section 2.3.2, the linear plane cannot be attributed to any of the three dimensions on the basis of data alone. Notice that the confidence intervals in these two sub-plots are narrowest at $age = 53$ and $cohort = 1942$; as one would expect, these are the values at the initial point μ_{UU} (where $U = 34$, referring to the standardized indexation used internal to the commands). Both trends are of low magnitude and are not statistically significant. The level in plot (e) represents the mean BMI for those in ik cell UU; and indeed the estimate of a BMI of 27.3 for those born in the early 1940s when aged about 50 seems appropriate.

Finally, subplots (g), (h), and (i) show the sums of double differences in age, period, and cohort respectively, i.e. the deviation from the linear plane. Again these are anchored at the initial points μ_{UU}, $\mu_{U,U+1}$, and $\mu_{U+1,U}$. From these subplots, it is clear that substantial linearity remains in the estimated deviation from trend, which will make interpretation challenging. To overcome this problem, an alternative representation is used which detrends the estimated sums of double differences by constraining the first and last of each to be zero, and redistributes the resulting trends to the estimated linear plane. An example of this is provided in figure 6.2. Since this form of representation is far easier to interpret, it is used for all subsequent models.

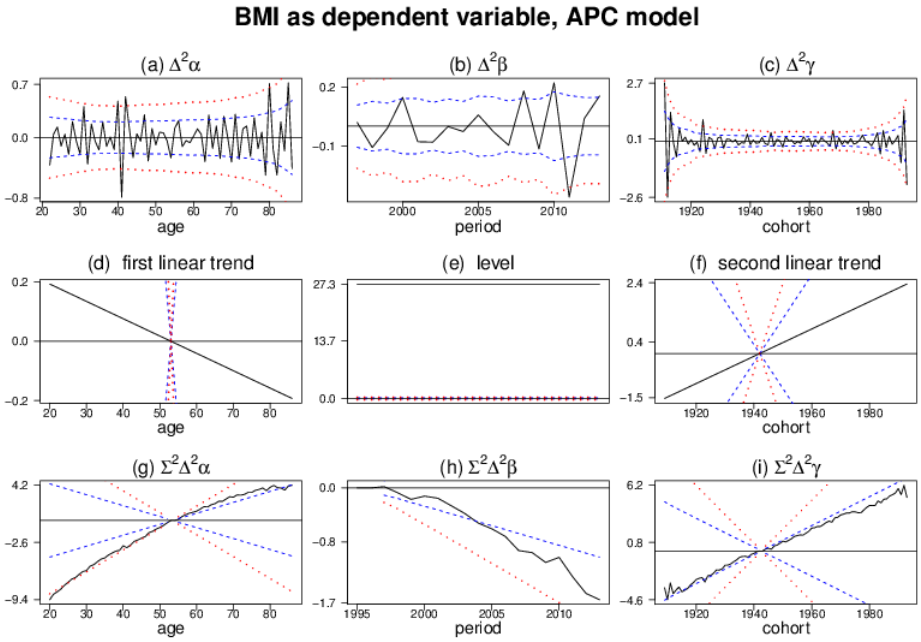


Figure 6.1: Plots of estimated double differences, linear plane, and sums of double differences: APC model with `BMI` as dependent variable, no covariates

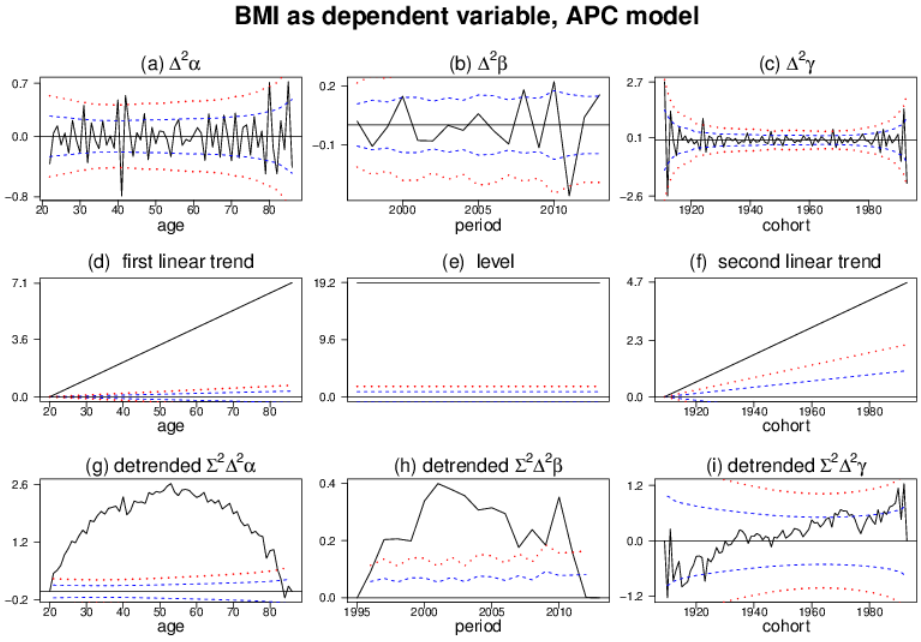


Figure 6.2: Plots of estimated double differences, linear plane, and sums of double differences: APC model with `BMI` as dependent variable, no covariates. Detrended representation.

In figure 6.2, subplots (a), (b), and (c) are as expected unchanged from figure 6.1. The representation of the trends in (d) and (f) has been changed so that instead of being anchored at zero at μ_{UU} and allowed to pass below zero, they are anchored at the point of origin of the age-cohort space, in this case at $age = 20$, $cohort = 1914$. These trends are the composite of the linear trends and the trends inherent to the double-sums of double differences in figure 6.1. In (e), the level is lower than that in figure 6.1; this is consistent with the fact that the level is now taken at the origin point rather than the point μ_{UU}, given that the two trends slope upwards (it also makes intuitive sense that 20-year-olds born in 1914 would have lower BMI than 52-year-olds born in 1942). The final row shows the detrended double-sums of double differences, and close inspection of these subplots can reveal the correspondence between them and those in the the the final row of figure 6.1.

It should be noted at this point that the confidence bands seen in the plots (g), (h), and (i) are based on the standard errors of the estimated double differences. It is not strictly speaking appropriate to use these to construct a confidence interval for even a single sum of double differences, let alone to construct confidence bands for the evolution of the sums of double differences along the index in question. This is clear from a simple example: if $\alpha \in [4, 5]$ and $\beta \in [1, 2]$ with 95% confidence, then whenever α and β lie within these ranges we have $\alpha + \beta \in [5, 7]$. However, even if α and β both lie outside these ranges we may see $\alpha + \beta \in [5, 7]$; for example if $\alpha = \beta = 3$. Thus we should have more than 95% confidence that $\alpha + \beta \in [5, 7]$. Constructing more appropriate confidence bands is a subject for future study.

The detrended representation in figure 6.2 allows for substantially clearer interpretation than did the original representation in figure 6.1. The slopes in both the age and cohort dimensions are significantly positive, which is consistent with rising BMI in the UK over the time period covered in the data set (recall that positive slopes in the age and cohort dimensions can alternatively be represented as a positive slope in the period dimension, see figure 2.4 in section 2.2.1). The shape of the deviation from the linear plane in age, as given by the sums of double differences, is quite interesting in that it is both fairly regular and apparently significantly different from zero (subject to the caveat above regarding confidence

bands). This suggests that there is support for some form of age effect beyond the ambiguous linear plane, with BMI increasing rapidly from young adulthood to middle age, and levelling off from the mid-50s onwards. There is also see a significant deviation from the plane in the period dimension. While this does not have a regular shape, and is more difficult to make general statements about given that we only observe 20 periods, it appears that there is a peak in the early 2000s and a general decline thereafter. This suggests that the rate of increase of BMI has slowed since 2000, a claim also made by the NCD Risk Factor Collaboration (2016, p. 1389).

The behaviour of the deviations in the cohort dimension are quite unusual, following an S-shaped pattern around the linear plane which defies easy interpretation. This may, however, be an artefact of the uncertainty evident in plot (c) around the estimated effects of belonging to one of the early or late cohorts in the data set. Examination of the underlying data did not reveal any outliers in the early and late cohorts that could have been the source of the problem, nor were the mean BMIs of these cohorts dramatically different to those of nearby cohorts. The absence of any such anomalies in the data suggests that the uncertainty is likely attributable to the small number of observations available in an AP data set on the cohorts at the extremities of the data. First, these cohorts are observed on fewer occasions than the more central cohorts; at the extremes, we have only one cell worth of data on the earliest and latest cohorts. Furthermore, given that the earliest cohorts are also quite old when they are observed it is likely that there will be fairly few observations per cell, compounding the problem. One way to address this problem is to exclude the extreme cohorts from the sample entirely, creating a data set that has a hexagonal structure in age-cohort space. An example of such a structure is seen in figure 6.3; we can see that the two first and two last cohorts have been eliminated (represented by crossed-out data points). Note that this censoring will not affect the identification results from sections 2.3.2 and 3: the parameter vector required to describe the censored data will be a subset of ξ (some of the cohort double differences will be excluded) and if ξ is uniquely identified from the data and invariant to θ then any subset of ξ must share these properties.

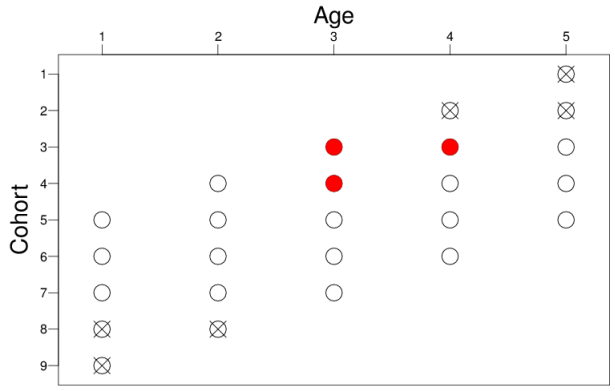


Figure 6.3: Example of cohort-censored AP data in age-cohort coordinate space

For the HSE data, it was decided that the bulk of the volatility at the extreme cohorts could be eliminated by cutting five cohorts from the beginning and the end of the sample. The first few lines of the F-table for this censored data set are shown in table 7 (the rows after the PC submodel were dropped as there was no support for the remaining submodels). It is interesting to note that there is very little change in the p-values or the AIC following the elimination of the extreme cohorts.

The detrended visual representation of the APC model for this data is given in figure 6.4. We see that the estimated age and period double differences are largely unchanged by the restriction on cohorts, as we would expect. The main changes therefore appear in the sub-plots in the right-most column of the plot. It is evident from the change to the scale of the Y-axis in subplots (c) and (i) that the extreme cohort effects have been eliminated as intended. The sums of double differences in cohort no longer exhibit any consistent deviation from the linear

	F-test vs TS	DF (*,163061)	p-value	F-test vs APC	DF (*,164146)	p-value	AIC
APC	0.98	1085	0.65				987601.29
AP	1.03	1158	0.20	1.81	73	0.00	987587.81
AC	1.07	1102	0.06	6.53	17	0.00	987678.44
PC	1.64	1150	0.00	12.57	65	0.00	988286.91

Table 7: F-table comparing APC model and submodels: cohort-censored data, `BMI` as dependent variable, no covariates

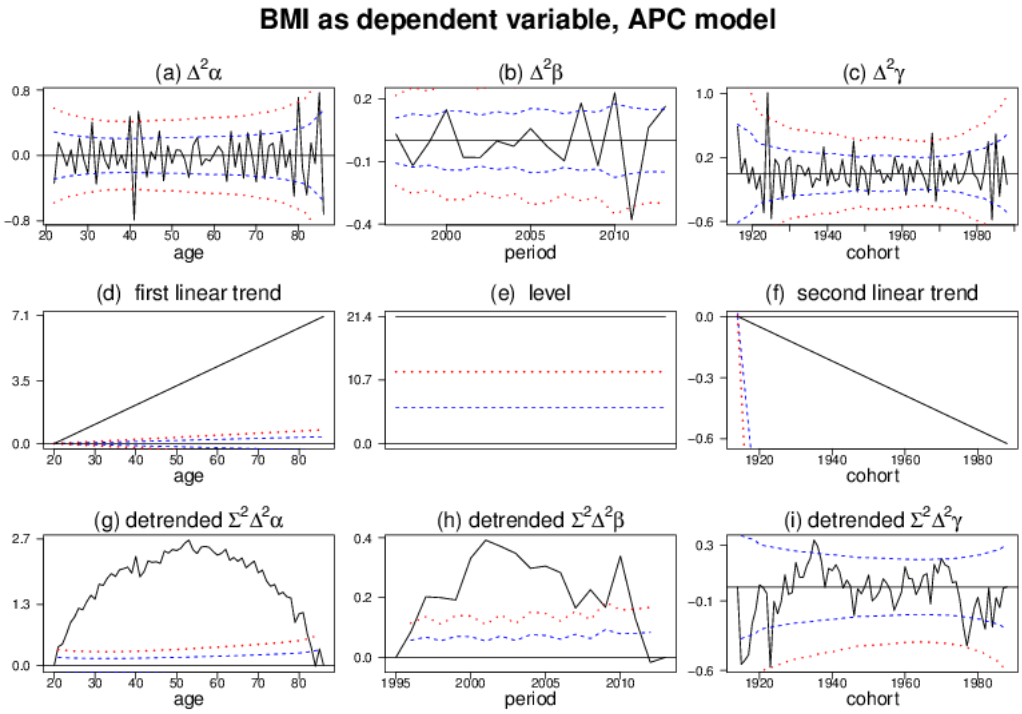


Figure 6.4: Plots of estimated double differences, linear plane, and sums of double differences: APC model for cohort-censored data with `BMI` as dependent variable, no covariates. Detrended representation.

plane; the variation appears random, with some persistence from cohort to cohort (as is to be expected given that cohort deviations from the plane likely reflect cultural and environmental variation). Interestingly, the elimination of the cohorts also affects the linear plane: the level has increased, and the second linear trend has become essentially insignificant. An investigation of the non-detrended representation (not provided here) showed that the original estimated linear plane was mostly unchanged, but the linear trend inherent to the sums of double differences was substantially reduced. This suggests that the changes between figure 6.2 and figure 6.4 are almost entirely driven by the elimination of the extreme double differences in cohort.

6.3.2 BMI with Covariates

Continuing with the cohort-censored data set, the introduction of covariates to the model for mean BMI is now considered. Recall that covariates which are candidates for omitted variable bias are of primary interest; that is, covariates which are potentially correlated with both BMI and the APC structure. Given the information available in the HSE data, the first obvious covariates to consider are sex and education. Since women have longer life expectancy than men it is probable that women are over-represented at higher ages. Thus if it were the case that women had lower BMI on average, we might expect there to be a sharper deceleration in BMI in the age dimension in a model that does not account for sex, as the proportion of women in the population increases with age.

Education, meanwhile, is expected to be related to cohort, with later cohorts staying in full-time education for longer (especially in light of reforms that raised the legal school-leaving age in 1947 and again in 1973, which would affect cohorts in this data). Education may also be related to BMI: those with fewer years of education relative to their peers will likely have lower lifetime incomes, which can be related to a range of factors influencing BMI such as the opportunity costs of exercise and healthy cooking and the affordability of healthy food or healthcare more generally. Since this education-BMI relationship operates primarily within generational groupings, the potential for omitted variable bias from this will be greatest in instances of sudden and dramatic shifts in educational attainment, such as the changes to the school-leaving age mentioned above. It is also possible that there is a direct effect of education on BMI if education increases the ability to process information about health and healthy behaviours, which would act as an omitted variable across generations.

Sex is introduced as a covariate by way of the constructed indicator variable `ismale`, that takes 1 for men and 0 for women. To introduce education to the model, the variable `educend` is used to construct a variable `age.fin.ed`, which indicates the age at which an individual exited full-time education. The values assigned to `age.fin.ed` for certain values of `educend` are presented in table 8, along with a count of the number of observations at each value. Recall that since the minimum age in the data is 20, anyone responding "Not yet finished" cannot

educend	age.fin.ed	freq ismale=1	freq ismale=0
14 or under	14	8076	9853
15	15	15625	18526
16	16	20746	23887
17	17	5628	8152
18	18	6149	9132
19 or over	19	17112	17057
Not yet finished	19	1803	1956
Never went to school	0	243	359

Table 8: Frequency of observations for `age.fin.ed`

have left school before age 19.

It is clear that one could raise several concerns about the variable `age.fin.ed`. First, the censoring at 14 and 19 mean that all of our analysis is subject to the assumption that the group of individuals with some education up to the age of 14, and the group of individuals who remained in education beyond the age of 18, are essentially homogeneous. If it is in fact the case that the relationship between education and BMI is linear, as the manner of introduction of `age.fin.ed` to the model implies, then the estimated slope based on this variable will be less steep than it ought to be as we are clustering individuals who in reality have more or fewer years of education at 19 and 14. Further, there may be concerns about the validity of the estimation methods when treating a discrete variable as though it were continuous. If the object here were to provide a rigorous analysis of the relationship between BMI, education, and the APC structure, then it would be necessary to think seriously about methods to mitigate some of these problems. However, since the main object here is simply to provide an illustration of the APC method, we only note that any results relating to analyses incorporating the `age.fin.ed` variable must be interpreted in light of these limitations.

The F-table for this model, seen here in table 9, is similar to those in tables 6 and 7. There is continued support for reductions to the APC and AP models, which has slightly increased following the introduction of covariates. There is also slight increase in support for the Ad model. Given the lack of support for reduction from APC to AP in the second column, it was decided to again focus on the full APC model. The plots of the APC structural elements are

	F-test vs TS	DF (*,163059)	p-value	F-test vs APC	DF (*,164144)	p-value	AIC
APC	0.97	1085	0.72				986258.20
AP	1.02	1158	0.30	1.73	73	0.00	986238.83
AC	1.06	1102	0.09	6.48	17	0.00	986334.40
PC	1.64	1150	0.00	12.85	65	0.00	986962.42
Ad	1.10	1175	0.01	2.66	90	0.00	986317.63
Pd	3.17	1223	0.00	20.48	138	0.00	988787.45
Cd	1.70	1167	0.00	11.26	82	0.00	987015.50

Table 9: F-table comparing APC model and submodels: cohort-censored data, `BMI` as dependent variable, covariates `ismale` and `age.fin.ed`

presented in figure 6.5. These plots show very little change from the earlier models, suggesting that there is less omitted variable bias than was initially suspected; although there are minor shifts in plot (i) and the level in plot (e) has increased (suggesting that a woman reporting no formal education at the origin point of the age-cohort space has a higher BMI than the average individual at that point). Coefficient estimates for the covariates `age.fin.ed` and `ismale` in both this model and a simple regression of BMI on `ismale` and `age.fin.ed` with an intercept only (i.e. without the APC structure) are seen in table 10. These estimates support the claim that the relationships of these variables to BMI are orthogonal to that of the APC structure to BMI, since the estimates are similar across the two models. As expected based on the reasoning above, BMI decreases significantly with education, and men have higher BMI than women on average.

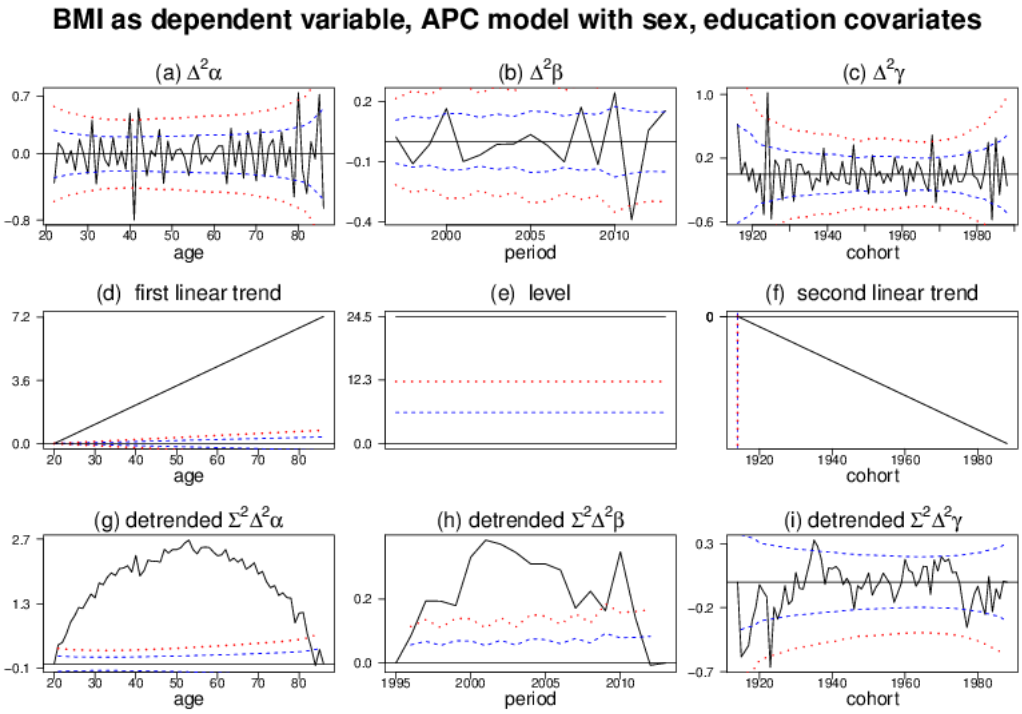


Figure 6.5: Plots of estimated double differences, linear plane, and sums of double differences: APC model for cohort-censored data with `BMI` as dependent variable, covariates `ismale` and `age.fin.ed`. Detrended representation.

	Estimate	Std. Error	t value	Pr(>|t|)
Model: APC structure and covariates				
ismale	0.34	0.02	14.19	0.00
age.fin.ed	-0.23	0.01	-34.19	0.00
Model: intercept and covariates				
intercept	31.89	0.10	308.12	0.00
ismale	0.35	0.02	14.42	0.00
age.fin.ed	-0.30	0.01	-48.36	0.00

Table 10: Estimated coefficients on `ismale` and `age.fin.ed` for models with `BMI` as dependent variable, with and without APC structure.

Further models were considered which introduced additional covariates: one incorporated a variable `smokenow` indicating whether or not an individual smoked at the time of the survey, and another replaced the `age.fin.ed` variable with a binary variable which took the value 1 if the person had remained in school until at least the age of 16. Both variables were significant, with those who smoked having a BMI about 1 point lower than non-smokers on average and those who had stayed in school until 16 or older also having significantly lower BMI than those who left early. However neither appeared to be confounded with the APC structure, which remained fairly stable for all models. As the APC structure is the primary object of interest, these models are not presented here.

6.3.3 Stability of Estimated APC Structure: Split Sample

Since the estimated APC structure appeared robust to the introduction of plausible confounding covariates, it was decided to investigate its robustness to changes to the sample analysed and to the form of model used. Changes to the sample analysed were considered first. The survey sample was split in half by period, with 2004 taken as the end point of the first sample and also the starting point of the second, so that there would be ten years in each. Both samples were cohort-censored, with 5 cohorts removed from the beginning and end of each.

The F-table from the analysis of the first half of the sample supported a reduction to the AC submodel, the visualisation of which is presented in figure 6.6. The curvature in the sums of double differences in age is preserved, as is the first linear trend, which becomes somewhat steeper. The second linear trend has also become steeper but is still insignificant. In the second half of the sample, the F-table suggested that an Ad submodel would be sufficient to capture the temporal variation in the data. This model is presented in figure 6.7. The curvature of age deviations is again preserved, although it is interesting to note that the single estimated slope is far gentler than it was in the first half of the sample. This gentle slope from the early-2000s onward may be related to the apparent deceleration in the increase in BMI seen in the period double-sums of double differences in the earlier models.

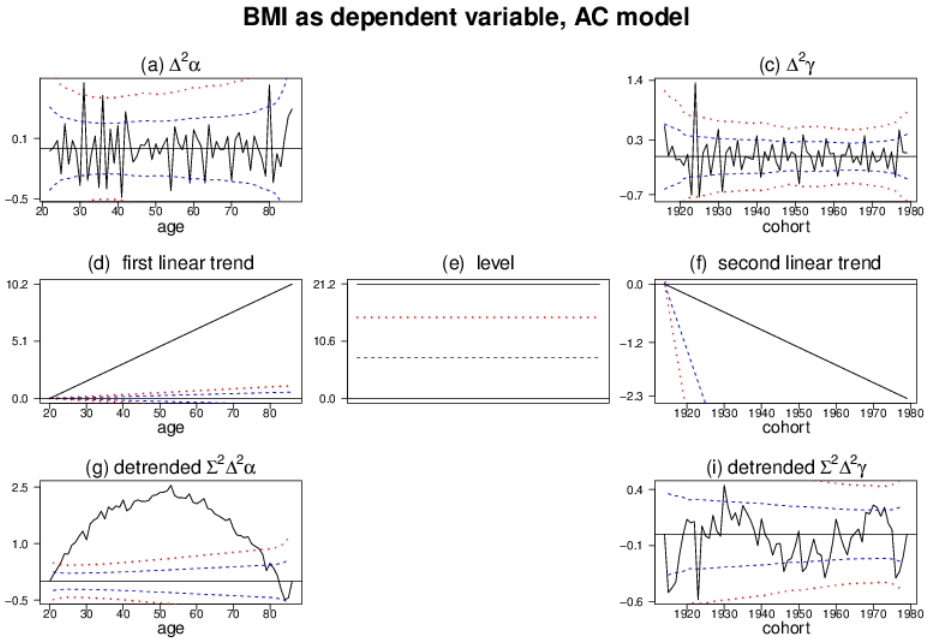


Figure 6.6: Plots of estimated double differences, linear plane, and sums of double differences: AC model for 1995-2004 cohort-censored data with `BMI` as dependent variable, no covariates. Detrended representation.

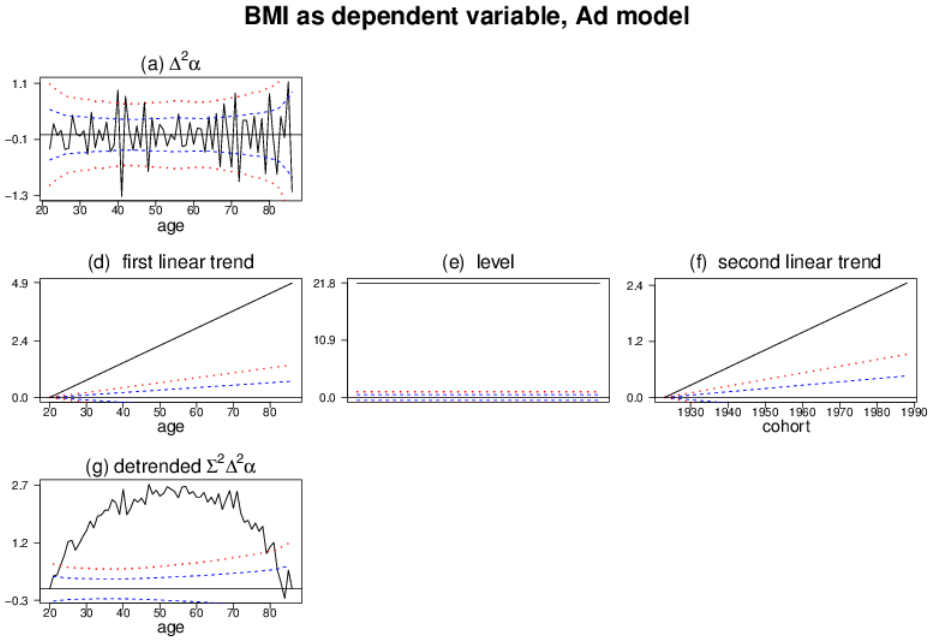


Figure 6.7: Plots of estimated double differences, linear plane, and sums of double differences: Ad model for 2004-2013 cohort-censored data with `BMI` as dependent variable, no covariates. Detrended representation.

	LR.vs.APC	df.vs.APC	prob(>chi_sq)	aic
APC				174410.34
AP	107.94	73	0.00	174372.28
AC	112.35	17	0.00	174488.69
PC	441.94	65	0.00	174722.28
Ad	221.24	90	0.00	174451.58
Pd	1340.00	138	0.00	175474.33
Cd	536.03	82	0.00	174782.37
A	1261.82	91	0.00	175490.16
P	1665.81	139	0.00	175798.14
C	1938.56	83	0.00	176182.90
t	1442.67	155	0.00	175543.01
tA	2494.45	156	0.00	176592.78
tP	1766.22	156	0.00	175864.56
tC	2910.20	156	0.00	177008.54
1	3008.72	157	0.00	177105.06

Table 11: LR-table comparing APC model and submodels: cohort-censored data, `obese` as dependent variable, covariates `ismale`, `age.fin.ed`, and `smokenow`

6.3.4 Stability of Estimated APC Structure: Binary Dependent Variable

Finally, a complete change of modelling approach was considered. An indicator variable `obese`, taking the value 1 if an individual had a BMI in excess of 30 and could therefore be considered clinically obese, was constructed and used as the dependent variable in a logistic model incorporating the APC structure in addition to the covariates `ismale`, `age.fin.ed`, and `smokenow`. As discussed in section 4.3, it is not currently possible to compare the APC model to the TS model in a discrete model of this form, because we lack the capacity to estimate the TS model. However, since the APC model and its submodels can all be estimated, comparisons between them using likelihood ratios are feasible. These likelihood ratios are seen in table 11. This table follows a similar logic to the tables of F-statistics earlier in this section: the first column indicates the submodel under consideration, the second column contains the likelihood ratio statistic, and the third column gives the degrees of freedom of the χ^2 distribution against which this statistic should be compared. The p-value obtained from this comparison is found in the fourth column, while the final column reports the Akaike Information Criterion for the model.

In this case the AIC is minimised by the AP model; however, the difference between the AIC for this model and the APC model is small, and the p-value indicates that the LR test does not support the reduction from APC to AP. Given that there is limited support for a reduction and that we are interested in a general comparison between this discrete-variable model and the continuous-variable models used above, it was decided to again focus on the full APC model. The plots for this model are seen in figure 6.8. It can be seen that the general structure found in the models of BMI as a continuous variable is preserved, with one significant linear trend, one insignificant linear trend (although in this case slightly positive), a clear concave shape in the deviations from the linear plane in the age dimension and an apparent peak in the period deviations in the early 2000s (with the last of these more pronounced in this model than it was in the models of mean BMI). Again there is no clear pattern in the cohort double differences.

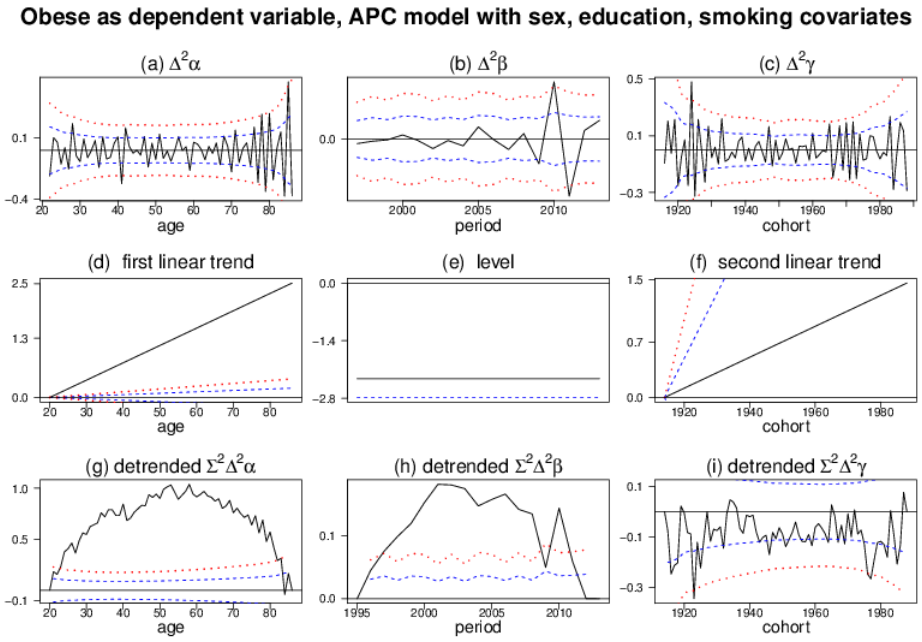


Figure 6.8: Plots of estimated double differences, linear plane, and sums of double differences: APC model for cohort-censored data with `obese` as dependent variable, covariates `ismale`, `age.fin.ed`, and `smokenow`. Detrended representation.

6.4 Summary

Based on the models above, a few substantive conclusions can be drawn. First, the primary feature of the estimated linear plane is a positive slope, which is reassuring for the modelling approach in that it is consistent with the generally accepted result that average BMI rose over the time horizon in which this data was collected. It is not possible to determine based on the data whether this estimated linear plane in the age dimension is attributable to age, two offsetting positive and negative cohort effects, or some combination of all three; however, further contextual information may help to judge the relative plausibility of these. With regard to the double differences, it does not seem unreasonable to argue that this analysis supports the following structures: a roughly convex rate of change of BMI in age, with BMI increasingly rapidly up to the mid-50s and then levelling off; a peak in period effects around 2000, after which the rate of increase of BMI among adults became slower; and little meaningful variation in cohort effects. The period pattern is likely due to an increase in public awareness of the risks of obesity from 2000 onwards (NCD Risk Factor Collaboration and others, 2016). Meanwhile, a plausible interpretation of the age pattern would be that increasingly tight time constraints from the age of 20 to the mid-50s combined with demographic shifts played a role in the rising obesity rates. That is, the pressures of full-time employment with increasing workplace responsibility over the course of a career, in addition to family responsibilities for many, make it more difficult and costly to maintain healthy behaviours (note that both of these pressures can be expected to ease off in later adulthood, as retirement approaches and children have reached adulthood, which could explain the concavity of the age effects). The lack of significant cohort effects suggests that the analyses of Wang et al. and the NCD may not have been greatly affected by their omission of cohort. Finally, it can be concluded that sex, education level, and smoking behaviour do all seem to be related to BMI, but the relationships are not confounded with time effects. Of course, this should be considered very much a preliminary analysis, and it would be unwise to draw any policy conclusions about obesity reduction based on these findings without substantial further research.

Methodologically, the main point of interest revealed by this analysis that was not discussed

in sections 2 through 5 relates to the effect of censoring extreme cohorts from the data set. If it is true, as seems to be the case, that this is an artefact of the small number of cells in which these cohorts appear, then it may be necessary to consider such censoring in all applications of APC analysis. Note that the dimension on which censoring should occur would depend on the data structure, so that while cohort-censoring is required when data has an AP structure, periods should be censored when working with an age-cohort structure and ages when the structure is period-cohort.

7 Conclusion

Since several active research areas in economics involve studying the relationship between individual decisions and temporal factors, it is important for economists to be aware of the three different ways in which time influences individual outcomes. The temporal context in which an individual is observed, the amount of time that has passed in the life of that individual, and the specific years that the individual has lived through - that is, their period, age, and cohort - are distinct concepts which must be recognised as such and accounted for in research. The estimable functions approach provides a method to describe a model of temporal variation in data in a way that accounts for all three concepts, without making substantive, untestable assumptions about the relationships between these concepts and the outcome variable. Furthermore, since it relies on a reasonably simple statistical modelling framework, it has the potential for generalization to a range of different contexts.

This project sought to execute one such generalization, by extending the applicability of the estimable functions approach to APC identification from the aggregate data context in which it originated to questions involving individual-level data. In the first stage of this project, it was verified that that the identification results for aggregate data are also supported in disaggregated data and data with covariates, and the set of R tools in the `apc` package were expanded upon to accommodate this data. A conceptual framework was developed for thinking about the assessment of fit of models of inter-temporal variation when working with individual-level data, which relied on comparison of these models to a temporally-saturated

model. While this framework was presented here as a way of assessing the literature-standard APC model and submodels described above, it can be used more generally to assess any model of age, period, and cohort effects (e.g. models assuming a particular functional form in one of the temporal dimensions). Once the conceptual framework had been developed, difficulties in estimation were encountered which led to the development of the algorithm described in section 4.3, which enables estimation of the temporally-saturated model by exploiting its particular dummy variable structure. This algorithm was also implemented in R.

It is hoped that the R tools for fitting APC models to individual data, and the conceptual framework and R tools for assessing their fit, will facilitate research seeking to explain how individual characteristics and decisions are related to time. It is expected that such research would further benefit from the resolution of a number of outstanding issues mentioned in the text, including the development of a method to fit a temporally-saturated model to discrete data with covariates, better construction of the standard errors in the plots of the linear plane and double differences seen in section 6, and consideration of models with interaction effects as was briefly mentioned in section 2.2.2. Other subjects for future research which were not mentioned in this thesis include multinomial discrete choice models, forecasting, and the potential differences arising when we move from repeated cross-section to panel data in the context of age, period, and cohort effects.

A Excerpt from HSE Data

age	period	cohort	sexfinal	htfinal (cm)	wtfinal (kg)	bmivalid	smoke	smkquant	educend	topqual2
20	2013	1993	3	173.30	58.20	19.38	never regular		19 or over	NVQ3/ GCE A Level equiv
22	2013	1991	2	157.50	48.20	19.43	never regular		19 or over	NVQ4/ NVQ5/ Degree or equiv
53	2013	1960	3	166.40	82.50	29.80	never regular		16	No qualification
44	2013	1969	2	153.80	59.00	24.94	never regular		14 or under	No qualification
46	2013	1967	3	166.60	66.10	23.82	never regular		17	NVQ1/ CSE other grade equiv
55	2013	1958	3	173.30	101.10	33.66	current	moderate 10-19	16	NVQ2/ GCE O Level equiv
38	2013	1975	2	159.50	59.80	23.51	never regular		18	NVQ3/ GCE A Level equiv
39	2013	1974	3	175.50	77.00	25.00	never regular		18	NVQ3/ GCE A Level equiv
40	2013	1973	3	166.90	72.50	26.03	ex-regular		19 or over	NVQ4/ NVQ5/ Degree or equiv
37	2013	1976	2	158.90	51.50	20.40	never regular		19 or over	NVQ4/ NVQ5/ Degree or equiv

Table 12: Excerpt from HSE data. Note: sexfinal=2 to indicate women, sexfinal=3 to indicate men

References

Allman-Farinelli, MA, Chey, T, Bauman, AE, Gill, T, & James, WPT. 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**(7), 898–907.

Attanasio, Orazio P. 1993. A cohort analysis of saving behavior by US households. *NBER Working Paper Series*.

Bell, Andrew, & Jones, Kelvyn. 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.

Browning, Martin, Crawford, Ian, & Knoef, Marike. An Entropy-Based Approach to the APC Problem, with Applications to Female Mortality and Labour-Force Participation. *Download https://www.nuffield.ox.ac.uk/People/sites/Ian.Crawford/Documents/apc.pdf.*

Carstensen, B. 2007. Age-period-cohort models for the Lexis diagram. *Statistics in medicine*, **26**(15), 3018–45.

Cawley, John. 2011. The Economics of Obesity. *In:* Cawley, John (ed), *The Oxford Handbook of the Social Science of Obesity*. Oxford University Press.

Clayton, D, & Schifflers, E. 1987. Models for temporal variation in cancer rates. I: Age-period and age-cohort models. *Statistics in medicine*, **6**(4), 449–67.

Dassonneville, Ruth. 2013. Questioning generational replacement. An age, period and cohort analysis of electoral volatility in the Netherlands, 1971–2010. *Electoral Studies*, **32**(1), 37–47.

Deaton, Angus, & Paxson, Christina H. 1993. Saving, Growth, and Aging in Taiwan. *NBER Working Paper Series*.

Downs, Julie S., & Loewenstein, George. 2011. Behavioural Economics and Obesity. *In:* Cawley, John (ed), *The Oxford Handbook of the Social Science of Obesity*. Oxford University Press.

Ejrnæs, Mette, & Hochguertel, Stefan. 2013. Is business failure due to lack of effort? Empirical evidence from a large administrative sample. *The Economic Journal*, **123**(571), 791–830.

Finkelstein, Eric A, & Strombotne, Kiersten L. 2010. The economics of obesity. *The American Journal of Clinical Nutrition*, **91**(5), 1520S–1524S.

Finucane, Mariel M, Paciorek, Christopher J, Danaei, Goodarz, Ezzati, Majid, *et al.* 2014. Bayesian estimation of population-level trends in measures of health status. *Statistical Science*, **29**(1), 18–25.

Fitzenberger, Bernd, Schnabel, Reinhold, & Wunderlich, Gaby. 2004. The gender gap in labor market participation and employment: A cohort analysis for West Germany. *Journal of Population Economics*, **17**(1), 83–116.

Glenn, Norval D. 1976. Cohort analysts' futile quest: Statistical attempts to separate age, period and cohort effects. *American Sociological Review*, 900–904.

Glenn, Norval D. 2005. *Cohort Analysis*. 2nd edn. Quantitative Applications in the Social Sciences, vol. 5.

Hanoch, Giora, & Honig, Marjorie. 1985. "True" Age Profiles of Earnings: Adjusting for Censoring and for Period and Cohort Effects. *Review of Economics & Statistics*, **67**, 383–94.

Heckman, James, & Robb, Richard. 1985. Using Longitudinal Data to Estimate Age, Period and Cohort Effects in Earnings Equations. *Pages 137–150 of:* Mason, William M., & Fienberg, Stephen E. (eds), *Cohort Analysis in Social Research: Beyond the Identification Problem*. New York, NY: Springer New York.

Holford, T R. 1985. An alternative approach to statistical age-period-cohort analysis. *Journal of Chronic Diseases*, **38**(10), 831–40.

Holford, Theodore R. 1983. The Estimation of Age, Period and Cohort Effects for Vital Rates. *Biometrics*, **39**(2), 311–324.

Kalwij, Adriaan S, & Alessie, Rob. 2007. Permanent and transitory wages of British men, 1975–2001: Year, age and cohort effects. *Journal of Applied Econometrics*, **22**(6), 1063–1093.

Krueger, Alan B, & Pischke, Jorn-Steffen. 1991. The effect of social security on labor supply: A cohort analysis of the notch generation. *NBER Working Paper Series*.

Kuang, D., Nielsen, B., & Nielsen, J. P. 2008a. Forecasting with the age-period-cohort model and the extended chain-ladder model. *Biometrika*, **95**(4), 987–991.

Kuang, D., Nielsen, B., & Nielsen, J. P. 2008b. Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika*, **95**, 979–986.

Ligges, Uwe, & Mächler, Martin. 2003. Scatterplot3d - an R Package for Visualizing Multivariate Data. *Journal of Statistical Software*, **8**(11), 1–20.

Martínez Miranda, María Dolores, Nielsen, Bent, & Nielsen, Jens Perch. 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 (Statistics in Society)*, **178**(1), 29–55.

Mason, Karen Oppenheim, Mason, William M, Winsborough, Halliman H, & Poole, W Kenneth. 1973. Some methodological issues in cohort analysis of archival data. *American sociological review*, 242–258.

Meghir, Costas, & Whitehouse, Edward. 1996. The Evolution of Wages in the United Kingdom: Evidence from Micro Data. *Journal of Labor Economics*, **14**(1), 1–25.

NCD Risk Factor Collaboration and others. 2016. Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19·2 million participants. *The Lancet*, **387**(10026), 1377–1396.

Nielsen, Bent. 2014. Deviance analysis of age-period-cohort models. *Download: http://www. nuffield. ox. ac. uk/economics/papers/2014/apc_ deviance. pdf.*

Nielsen, Bent. 2015. apc: A Package for Age-Period-Cohort Analysis. R package version 1.1.

Nielsen, Bent, & Nielsen, Jens P. 2014. Identification and forecasting in mortality models. *The Scientific World Journal*, **2014**.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reither, Eric N, Hauser, Robert M, & Yang, Yang. 2009. Do birth cohorts matter? Age-period-cohort analyses of the obesity epidemic in the United States. *Social science & medicine*, **69**(10), 1439–1448.

Robertson, C, & Boyle, P. 1998. Age-period-cohort analysis of chronic disease rates. I: Modelling approach. *Statistics in medicine*, **17**(12), 1305–23.

Sasaki, Masamichi, & Suzuki, Tatsuzo. 1987. Changes in religious commitment in the United States, Holland, and Japan. *American Journal of Sociology*, **92**(5), 1055–1076.

Schwadel, Philip. 2010. Age, period, and cohort effects on US religious service attendance: The declining impact of sex, southern residence, and Catholic affiliation. *Sociology of Religion*, **71**(1), 2–24.

The Health and Social Care Information Centre. 2015. *Health Survey for England 2014 Volume 2: Methods and Documentation*. The Health and Social Care Information Centre.

Voas, David. 2013. Identifying age, period, and cohort effects: Are the new methods really better? *In: Department of Sociology Podcasts*. University of Oxford Podcasts.

Wang, Claire Y, McPherson, Klim, Marsh, Tim, Gortmaker, Steven L, & Brown, Martin. 2011. Health and economic burden of the projected obesity trends in the USA and the UK. *The Lancet*, **378**(9793), 815–25.

Wickham, Hadley. 2007. Reshaping data with the reshape package. *Journal of Statistical Software*, **21**(12), 1–20.

Wickham, Hadley. 2011. The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, **40**(1), 1–29.

Yang, Yang, & Land, Kenneth 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**(1), 75–97.

Yang, Yang, & Land, Kenneth C. 2008. Age-period-cohort analysis of repeated cross-section surveys: fixed or random effects? *Sociological methods & research*, **36**(3), 297–326.

Yang, Yang, Schulhofer-Wohl, Sam, Fu, Wenjiang J, & Land, Kenneth C. 2008. The Intrinsic Estimator for Age-Period-Cohort Analysis: What It Is and How to Use It. *American Journal of Sociology*, **113**(6), 1697–1736.