Age-period-cohort modelling and covariates, with an application to obesity in England 2001-2014

Zoë Fannon* Christiaan Monden† Bent Nielsen‡

November 10, 2019
latest version here

Abstract

We develop a framework to study the non-linear effects of age, period, and cohort on an outcome variable, for repeated cross-sectional data with covariates. These non-linear effects of age, period, and cohort are relevant to outcomes like consumption, wages, and health. The framework we develop is suitable for any continuous or binary outcome variable. The age, period, and cohort effects in the model are represented by a recently suggested parametrization, which isolates the identifiable non-linear effects from the unidentifiable linear effects. Our framework introduces this parametrization to the repeated cross-sectional setting, and includes a new test of the parametrization against a more general "time-saturated" model. A custom algorithm is developed to estimate the time-saturated model. The framework is applied to an analysis of the obesity epidemic in England using survey data. We find that the non-linear effects present in English obesity data are concavity in age among women and concavity in cohort among men. The existence of concavity in cohort is a novel finding in this literature.

Keywords: age-period-cohort, cohort, identification, collinearity, obesity, BMI

JEL Codes: C23, C52, C81, J11, I18

*Department of Economics, University of Oxford
†Department of Sociology, University of Oxford
‡Department of Economics, University of Oxford

1 Introduction

We use repeated cross-sectional data to examine the socio-demographic determinants of obesity rates in England, using both continuous and binary measures of obesity. The explanatory variables of interest are an individual's age, birth cohort, and period of observation, as well as other individual characteristics such as sex, race, education, and socio-economic status. There is a well-known identification problem when working with age, period, and cohort as explanatory variables. Since we are interested in the contribution of these to obesity, we use a recently suggested parametrization of the age, period, and cohort (APC) effects that is both freely varying and invariant to the identification problem (Kuang et al., 2008). We develop a test of this parametrization against a more general model in the context of repeated cross-sections. This resembles a deviance test and can be used for both continuous and binary outcomes. Applying the methods to English obesity data, we find that the data can be parsimoniously described using either an age-cohort or cohort-drift model for men, and an age-drift model for women.

Adult obesity is a major public health concern. Rates of obesity in the UK almost tripled between 1980 and 2011, with over a quarter of adults estimated to be obese by 2016 (Department of Health, 2011; Moody, 2016). An individual is considered obese if their body mass index (BMI, defined in equation (1)) exceeds 30. Obesity is linked to immediate and long term health risks, the most well-known being type II diabetes. The direct healthcare costs of obesity in 2006-07 were estimated to be 5.1 billion, 6% of the NHS budget (Scarborough et al., 2011). Reducing obesity has thus been a policy goal for many years, with specific government directives issued in 2007, 2011, and 2016.

In this paper we investigate the socio-demographic determinants of obesity. We use data from the 2001 through 2014 waves of the Health Survey for England. Our dependent variable is either the continuous measure, log BMI, or a binary obesity indicator. The explanatory variables include age and the period of observation, from which we construct cohort through *cohort* = *period* - *age*, as well as other socio-demographic variables including education and smoking behaviour.

We employ a generalized linear model with APC effects (age, period, and cohort) and other explanatory variables. The generalized linear model allows us to use a common framework to analyse continuous and binary outcomes. For the continuous outcome, log BMI, we use a normal model; and for the binary outcome, obesity, we use a logit model. The three distinct APC effects may reflect different underlying factors that are difficult to measure directly. Age effects may reflect life-cycle determinants of health, such as metabolic changes. Period effects may reflect environmental conditions which affect health contemporaneously. Cohort effects may capture habits around diet and exercise formed by generation-specific experiences. We can enhance our understanding of the contributions of these factors by determining stylized facts about the effects of age, period, and cohort, and comparing these stylized facts to hypotheses about difficult-to-measure underlying factors.

The APC effects in the model are not fully identified, which is a well-known problem, see Holford (1983), Clayton & Schifflers (1987), Glenn (2005), and Carstensen (2007). We reparametrize the APC effects in the model in terms of freely varying parameters as suggested by Kuang et al. (2008), henceforth KNN. The freely varying parameters consist of deviations from linearity attributed to each of age, period, and cohort; and a linear plane, which

combines the inseparable linear effects of age, period, and cohort. In our analysis we use a normal model for the continuous outcome, log BMI, and a logit model for the binary outcome, obesity, so these freely varying parameters are canonical, in the generalized linear model (GLM) sense.

Since the parameters are canonical, statistical analysis is simplified in the following two ways. First, it is easy to impose restrictions on the APC effects and count the associated degrees freedom. Second, it is simple to incorporate extensions beyond the time horizon of the sample. For instance, we may want to conduct recursive analysis where the number of waves change, or forecast beyond the last sample period. The canonical parametrization is invariant to such changes.

In order to conduct inference on our model, we must be clear about the repetitive structure from which we generate asymptotic analysis. Our asymptotic analysis can be understood by thinking of the data as a two-way array in age and cohort with individual information accumulating in each cell of the array. In the data we have 53 age groups, 14 period groups, and 56 cohorts. We keep the dimension of the age-cohort array fixed and exploit the individual level information for inference. In light of the canonical parametrization the statistical analysis is then fairly simple. This asymptotic approach resembles earlier work for aggregate data by Martínez Miranda et al. (2015), who developed a Poisson model for counts of cancer data in which the count in each cell increases corresponding to an increase in the number of individuals in the cell.

We develop a test of the reparametrized age-period-cohort model against a more general model. The age-period-cohort model implies assumptions about the time effects: in particular, that interactions between any two of age, period, and cohort are absent. Such interaction effects would capture, for instance, the effect on obesity of a time- and age-limited government programme promoting healthy eating in schools. The assumption that such effects are absent must be tested. Our focus on the age-cohort array yields a natural specification of the more general model where each age-cohort cell has its own parameter. We call this a time saturated (TS) model. The proposed test, comparing the reparametrized APC model to the TS model, resembles a deviance test. It works for both continuous and binary dependent variables. Inference is standard, but there are computational challenges which we address.

When applying our methods to the data from the Health Survey for England, 2001-2014, we find that an age-drift model fits the data on women while an age-cohort model fits the data on men. The age-drift model for women includes two parts. The first is an increasing linear plane, or "drift", that combines the inseparable linear effects of age, period, and cohort. The second is a concave deviation from linearity in the age dimension which has a kink at age 50. This concavity is consistent with previous research (Lean et al., 2013; Wang et al., 2011; Howel, 2011). For men, the age-cohort model includes a combined linear plane; a concave deviation from linearity in age, similar to that for women; and a concave deviation from linearity in cohort, which is particularly pronounced from the 1960s cohort onwards. The deviation from linearity in cohort was not detected by previous studies, which used different age and period ranges and often imposed constraints on the cohort effects. The effects of covariates are broadly consistent with existing literature.

Other papers have used the KNN parametrization to study deviations from linearity in the effects of age, period, and cohort on a dependent variable. Recently, Harnau & Nielsen (2017) presented a model for over-dispersed Poisson data similar to the model of Martínez Miranda

et al. (2015). Other papers working with aggregate data include Fu (2016) in which the author studies a class of constrained estimators where the dimension of the array increases. That approach would be inappropriate for our data given its small period range. There is also a Bayesian approach to aggregate data presented by Smith & Wakefield (2016).

Alternative, Bayesian approaches have been used to study APC effects in models with individual data. A prominent model is the hierarchical age-period-cohort model by Yang & Land (2006), which is generalized in the cross-classified random effects model by Yang (2008). The latter has been used to study obesity by Reither et al. (2009) and An & Xiang (2016). These models impose a quadratic age structure, which is a testable restriction in our model. More importantly, the models do not fully address the identification problem since priors are imposed on both identified and non-identified parameters. It is well-known that the likelihood cannot update all of these parameters. In particular, the conditional prior for the non-identified parameters given the identified parameters is not updated, see Poirier (1998) and Nielsen & Nielsen (2014).

The paper is outlined as follows: §2 introduces the elements needed to understand the approach, including the data used in the application, the notation employed, and a more detailed summary of both the APC identification problem and the ideas in KNN. §3 and §4 contain the main theoretical contributions of this paper. In §3, we discuss the conditions for standard inference on continuous and binary outcomes using the KNN parametrization of APC effects and repeated cross-sectional data. In §4 a new test is proposed which compares the APC model to a more general model, and an algorithm for this test is developed. In §5 the situation in which the covariates are the direct object of interest, and a researcher only needs to control for APC effects, is considered. §6 contains the application of the methods to analyse obesity dynamics in England, while §7 concludes.

2 Preliminaries: the obesity data and the age-period-cohort model

In this section we introduce preliminary concepts, building on our data and the existing literature on age-period-cohort effects, which are needed to understand the theoretical contributions of sections 3 and 4. We describe the data, which are repeated cross-section. We explain how this data can be viewed as a structure in age-period or age-cohort space. The statistical model is defined. The use of the KNN parametrization to handle the age, period, and cohort effects is described. The idea of testing for the adequacy of more parsimonious sub-models of the APC model is introduced. The need for a parallel test of the adequacy of the APC model against a more general model is highlighted.

2.1 Obesity data

The data used is the core sample from the Health Survey for England (HSE) and analysed using the R-package apc, version 1.3.3; see Nielsen (2015). The data is a repeated cross-section of a representative sample of the English population. We use waves from 2001-2014. We exclude waves prior to 2001 as they do not include the National Statistics Socio-Economic

Classification (NSSEC), one of our explanatory variables. We end the sample at 2014 because from 2015 onwards age is only recorded in five-year bands.

We observe 43,077 women and 38,316 men, who we analyse separately. We index the individual observations by $h = 1, \ldots, H$, reserving the letter i for the age index. For each individual we have information on weight and height, measured by a registered nurse. From these we compute body mass index as

$$BMI = (\text{weight in kg})/(\text{height in metres})^2. \tag{1}$$

A small number of observations with BMI outside the range 12 to 60 were presumed to be subject to measurement error and excluded. In addition to BMI, age, and period of observation, we have data on the following covariates: ethnicity, level of education, NSSEC, smoking history, and alcohol consumption. Descriptive statistics are reported in tables 12 and 13 in Appendix C.

We consider two choices of dependent variable: either log BMI or an indicator for obesity defined as $BMI \geq 30$. For each individual h then Y_h is the dependent variable, i_h is the individual's age, j_h indicates the period in which the individual is observed, and k_h is the cohort of the individual which is constructed from i_h and j_h.

In this dataset age and period vary in a rectangular age-period array, where age is between 28 and 80 and period is between 2001 and 2014 (see Figure 1). We therefore have $I = 53$ age groups and $J = 14$ period groups. Cohort then varies between 1921 and 1986. However, we exclude the first and last five cohorts as they are sparsely observed. This leaves $K = 56$ cohort groups. The final data, as an age-period array, is shown in Figure 1. The shading in that figure reflects the variation in survey size across waves.

The data is an example of a generalized trapezoid in the sense of KNN. Although the data is best visualized in an age-period array, for indexing we switch to an age-cohort coordinate system, because of the age-cohort symmetry in the relation $age + cohort = period$. Thus, throughout the paper we consider an age-cohort array, where $i = 1, \ldots, I$ is the age index and $k = 1, \ldots, K$ is the cohort index. We define the period index through $j = i + k - 1$ and get an index set of the form

$$1 \leq i \leq I, \qquad 1 \leq k \leq K, \qquad L + 1 \leq j \leq L + J. \tag{2}$$

Here L is the necessary offset in the period index due to beginning the age and cohort indices at 1. With the present data, we then have that $I = 53$, $J = 14$, $K = 56$, $L = 48$ so that $age = 28$, $per = 2001$, $coh = 1921$ correspond to $i = 1$, $j = L + 1$, $k = 1$.

2.2 Generalized linear model with age, period, and cohort

We use a generalized linear model for the dependent variable, Y_h, where the linear predictor η_h is a function of the covariates and the age-period-cohort structure as described below. For continuous dependent variables a normal model is employed of the form

$$Y_h = \eta_h + \varepsilon_h \qquad \text{for } h = 1, \ldots, H. \tag{3}$$

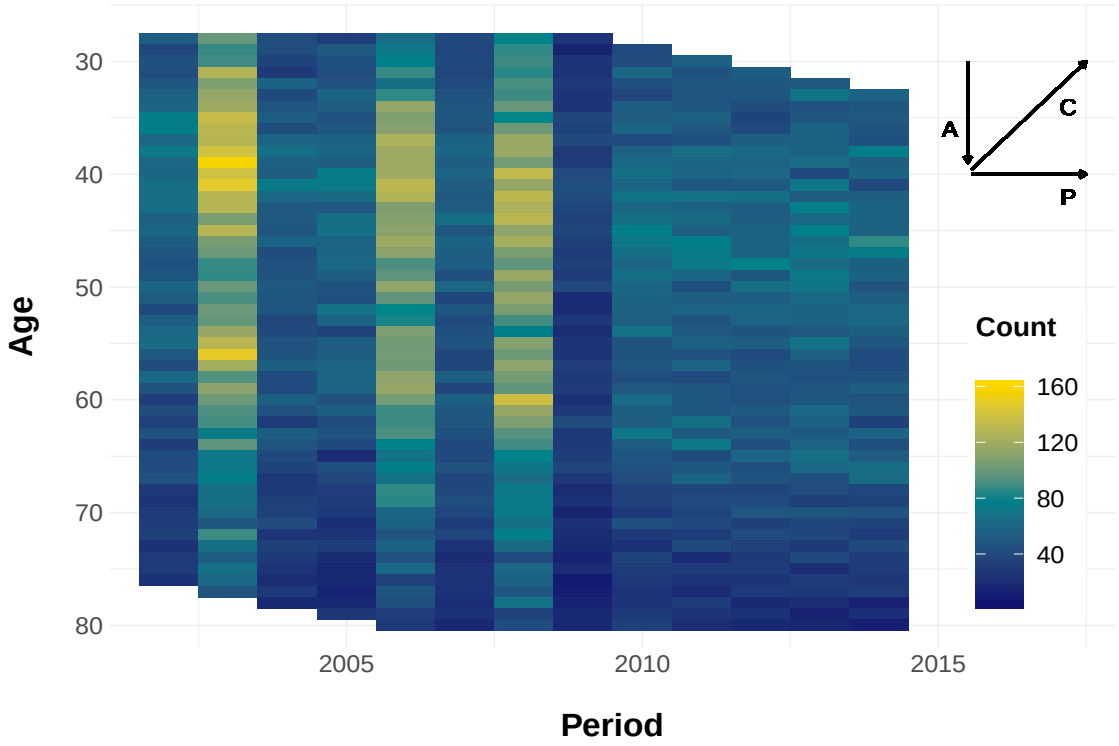


Figure 1: Within-cell observation counts women

The errors ε_h are independent over individuals and normally distributed conditional on the linear predictor: $\varepsilon_h \sim \mathsf{N}(0, \sigma^2)$. For binary dependent variables, a logistic model is employed with

$$\log \frac{\mathsf{P}(Y_h = 1)}{\mathsf{P}(Y_h = 0)} = \eta_h \qquad \text{for } h = 1, \ldots, H. \tag{4}$$

The linear predictor η_h is individual-specific and has the form

$$\eta_h = Z_h'\zeta + \mu_{i_h k_h}. \tag{5}$$

Here, ζ is a d_z-length vector of parameters and Z_h is a d_z-length vector of covariates. The term $\mu_{i_h k_h}$ describes the age-period-cohort structure as follows: an individual h with age i_h and cohort k_h observed in period $j_h = i_h + k_h - 1$ will have linear predictor $\mu_{i_h k_h}$ where

$$\mu_{ik} = \alpha_i + \beta_j + \gamma_k + \delta. \tag{6}$$

In the above α_i, β_j, and γ_k are fixed effects for age i, period j, and cohort k respectively; we refer to these as APC fixed effects. The full set of APC fixed effects is of dimension q, where $q = I + J + K + 1$. Collecting the APC fixed effects as

$$\theta = (\alpha_1, ..., \alpha_I, \beta_{L+1}, ..., \beta_{L+J}, \gamma_1, ..., \gamma_K, \delta)' \tag{7}$$

we can write $\mu_{i_h k_h} = D_h'\theta$ where D_h is a q-dimensional design vector of age, period, and cohort indicators and an intercept.

2.3 Identifying non-linearities in the age-period-cohort model

It is not possible to identify the vector of APC fixed effects θ from the likelihood, since for any constants $a, b, c, d \in \mathbb{R}$ the predictor μ_{ik} in (6) satisfies

$$\begin{aligned} \mu_{ik} = \; & \{\alpha_i + a + (i-1)d\} + \{\beta_j + b - (j-1)d\} \\ & + \{\gamma_k + c + (k-1)d\} + \{\delta - a - b - c\}, \end{aligned} \tag{8}$$

see for instance Carstensen (2007). The fact that this holds for any set of arbitrary constants makes it impossible to identify the linear parts of the APC fixed effects.

There are two ways to address the identification problem. First, we could work with a just-identified version of θ, by imposing four constraints. This would require us to keep track of the consequences for interpretation, count of degrees of freedom, plotting, and forecasting (see Nielsen & Nielsen (2014) and §5 of this paper). Second, we can reparametrize the model in terms of a freely varying parameter, which is invariant to the class of transformations in (8), and which is of a lower dimension, $p = q - 4$, than the original parametrization in terms of θ. We follow the second approach.

The reparametrization is expressed in terms of a p-dimensional parameter vector ξ and a design vector X_h following KNN. The parameter vector is

$$\xi = (\upsilon_o, \upsilon_a, \upsilon_c, \Delta^2\alpha_3, \ldots, \Delta^2\alpha_I, \Delta^2\beta_{L+3}, \ldots, \Delta^2\beta_{L+J}, \Delta^2\gamma_3, \ldots, \Delta^2\gamma_K)'. \tag{9}$$

Here, $\upsilon_o, \upsilon_a, \upsilon_c$ are the origin and two slopes of a linear plane. The double differences, e.g. $\Delta^2\alpha_i = (\alpha_i - \alpha_{i-1}) - (\alpha_{i-1} - \alpha_{i-2})$, measure deviation from that plane in age, period, and cohort. A p-dimensional design vector X_h is constructed, which combines the available information on an individual's age and cohort such that $\mu_{i_h k_h} = X_h'\xi$. The reparametrization can be expressed in terms of a $p \times q$ transformation matrix A' which satisfies $D = XA'$ and $A'\theta = \xi$. Further details are given in Appendix A.

The KNN parametrization confers the following four advantages. First, the vector ξ is invariant to the transformations in (8). The unidentified age effects $\alpha_i + a + (i-1)d$, $\alpha_{i-1} + a + (i-2)d$, etc. yield double difference $\Delta^2\alpha_i$ regardless of the values of a, d. The linear plane parameters $\upsilon_o, \upsilon_a, \upsilon_c$ are chosen to be invariant to the transformation in (8), see Appendix A. Second, the design matrix X, formed from stacking X_h, has full column rank, whereas the design matrix D, formed from D_h, has reduced column rank. Third, there is a unique ξ that can satisfy the relation between X_h and the linear predictor for all $h = 1, ..., H$, so that $\forall\ \xi^\dagger \neq \xi$ it holds that $\mu(\xi^\dagger) \neq \mu(\xi)$. Finally, for linear exponential family models such as the normal and logit models used in this paper, ξ is the canonical parameter.

2.4 Sub-models of the age-period-cohort model

We can test whether a more parsimonious sub-model of the KNN parametrization of the APC structure is sufficient to describe the data. This is done in the same way as it was for aggregate data (Nielsen, 2014). Four categories of restriction are of interest.

First, we can test for the absence of deviations from linearity in one of the age, period, and cohort. For instance, we test period non-linearities by imposing $\Delta^2\beta_{L+3} = \cdots = \Delta^2\beta_{L+J} = 0$. This gives an age-cohort (AC) model. In terms of the unidentified original parametrization

θ, this is written as $\beta_{L+1} = \ldots \beta_{L+J} = 0$. The two formulations of the hypothesis are in fact equivalent, see Nielsen & Nielsen (2014) for a formal analysis. The latter formulation gives the misleading impression that we simultaneously test for the absence of both non-linear and linear period effects, which is not the case.

Second, we can test for the absence of non-linearities in two components. The approach is similar. To test the period and cohort non-linearities we impose $\Delta^2\beta_{L+3} = \cdots = \Delta^2\beta_{L+J} = 0$ and $\Delta^2\gamma_3 = \cdots = \Delta^2\gamma_K = 0$, while leaving the linear plane unrestricted. Clayton & Schifflers (1987) refer to this as the age-drift (Ad) model.

Third,we can test a model with only an age effect, i.e. no non-linearities or linearities in period and cohort. We do this by imposing $\Delta^2\beta_{L+3} = \cdots = \Delta^2\beta_{L+J} = 0$ and $\Delta^2\gamma_3 = \cdots = \Delta^2\gamma_K = 0$, as well as $\upsilon_c = 0$. This leaves a model of the type $\mu_{ik} = \alpha_i$, which is called an Age (A) model.

Finally, we will be interested in a linear plane model, where there are no non-linearities at all. Then $\Delta^2\alpha_3 = \cdots = \Delta^2\alpha_I = 0$, and $\Delta^2\beta_{L+3} = \cdots = \Delta^2\beta_{L+J} = 0$ and $\Delta^2\gamma_3 = \cdots = \Delta^2\gamma_K = 0$. This model corresponds to making the APC fixed effects linear functions so that $\alpha_i = \alpha_0 + \alpha_1 i$ and $\beta_j = \beta_0 + \beta_1 j$ and $\gamma_k = \gamma_0 + \gamma_1 k$. We call this a (t) model.

2.5 A more general model than the age-period-cohort model

It is also of interest to test whether the APC model is too parsimonious, meaning that a more general model is needed to capture the effects of time in the data. A more general model would include interaction effects between age, period, and cohort, which are ruled out by the APC model. The fact that the APC model restricts these interaction effects is recognised by Kupper et al. (1985). In Martínez Miranda et al. (2015), where aggregate data is used, the restrictions on interaction effects implied by the APC model are tested using a deviance test. However, the deviance test cannot perform this function in the repeated cross-sectional setting, because it simultaneously tests the modelling of the covariates as well as age, period, and cohort. Therefore a test that can evaluate the modelling of the effects of age, period, and cohort alone is required.

3 The age-period-cohort model for repeated cross-sectional data

We have described the data and the model we want to use, and have shown how identification is achieved through reparametrization. We proceed to discuss estimation and inference for the KNN parametrization of the APC model with repeated cross-sectional data and covariates. We consider first a normal model for continuous outcome variables, as in equation (3), and second a logistic model for binary outcome variables, as in equation (4). Conditions for estimation and inference are mostly standard.

3.1 The normal model for continuous outcome variables

We discuss analysis of a normal model of form (3), with η specified as in (5) and using the KNN parametrization (9) to capture the effects of age, period, and cohort.

3.1.1 Estimation

For the purposes of discussing estimation, it is convenient to stack observations. We define Y as the $H \times 1$ vector of individual observations on the dependent variable, $Y = [Y_1, ..., Y_H]'$. Correspondingly let η, μ, X, and Z represent stacked individual information. Recall that X_h is the design vector associated with the KNN parametrization (9) while Z_h is the vector of covariates. We then have

$$\eta = Z\zeta + \mu, \qquad \mu = X\xi. \tag{10}$$

We embed this in the normal model (3) which can be estimated by least squares regression of Y on (X, Z), noting that ζ and ξ are freely varying parameters.

3.1.2 Inference

Following estimation, we conduct inference on the estimated model. We investigate whether some elements of the APC structure are unnecessary, as described in §2.4. Achieving a more parsimonious representation is desirable for forecasting purposes and for ease of interpretation.

Exact inference on the OLS parameter estimates $(\hat{\zeta}, \hat{\xi})$ can be performed by appealing to the classical results for analysis of variance in the linear model. Since it is desirable to relax the strict assumption of conditional normality of ε, we also consider asymptotic inference. For this we must be clear about the repetitive structure. We treat the dimensions of the generalized trapezoid in the age-cohort array as fixed: the numbers of ages, periods, and cohorts, I, J, and K, do not change. It is the number of individual observations H that is assumed to increase.

To justify asymptotic inference we make the following assumptions:

1. The triplets Y_h, X_h, Z_h are independent and identically distributed across individuals h.

2. The regressors X_h, Z_h jointly have a positive definite covariance matrix.

3. The errors ε_h have zero conditional mean and finite variance: $\mathsf{E}(\varepsilon_h|X_h, Z_h) = 0$ and $\mathsf{Var}(\varepsilon_h|X_h, Z_h) = \sigma^2$.

It is a consequence of these assumptions that as we increase the sample size H the relative frequency of individuals at all age-cohort combinations should remain constant. This could be a concern since the size of the HSE survey varies from year to year in a way that does not reflect changes to the underlying UK population. However, from Wooldridge (2010, §19.4) we learn that the estimates $(\hat{\zeta}, \hat{\xi})$ should remain consistent and asymptotically normal if the distribution of log BMI among those selected for the HSE is the same, conditional on X and Z, as the distribution of log BMI among those not selected. Since the variation in sample size across years is due to financial constraints of the surveying body, and thus is not related to the distribution of log BMI, it seems reasonable to assume this holds. In a similar vein, these assumptions imply that selection into the HSE is independent of the covariates Z_h. This is plausible as the HSE is a representative sample. The distribution of Z_h may vary between age-cohort cells.

Under these assumptions, inference can be performed in the usual way. Under exact normality, t- and F-tests can be used. Under asymptotic inference, likelihood ratio tests are asymptotically χ^2. Such tests can be used to investigate the sub-models from §2.4. We defer discussion of a test against a more general model, advocated in §2.5, to §4.

3.2 The logit model for binary outcome variables

We discuss analysis of a logit model of form (4), with η specified as in (5) and using the KNN parametrization (9) to capture the effects of age, period, and cohort.

3.2.1 Estimation

The logit log-likelihood is

$$\ell(\zeta, \xi) = \sum_{h=1}^{H} \eta_h Y_h - \sum_{h=1}^{H} \ln(1 + \exp \eta_h). \tag{11}$$

There is no closed-form expression for the ζ, ξ that maximise this log-likelihood. However, the log-likelihood is strictly concave when the design matrix has full rank so the maximum likelihood estimator is unique (Wedderburn, 1976). It is finite in the absence of separation or quasi-separation (Agresti, 2013, §6.5). Under these conditions the maximum likelihood estimator can be found by Newton iteration.

3.2.2 Inference

The asymptotic theory of the logit estimator is outlined by Fahrmeir & Kaufmann (1986). Their Theorem 2 shows consistency and asymptotic normality under the following assumptions:

1. the triplets $\{Y_h, X_h, Z_h\}$ are independent, identically distributed;
2. the regressors X_h, Z_h have a positive definite covariance matrix.

The asymptotic variance-covariance matrix of this estimator is given by $J = -\ddot{\ell}$, for $\ddot{\ell}$ the second derivative of the log-likelihood. Theorem 3 of Fahrmeir & Kaufmann (1986) shows that likelihood ratio test statistics are asymptotically χ^2. They can be used to test the necessity of different parts of the APC structure as outlined in §2.4.

4 Testing the age-period-cohort model against a more general model

We describe a new test of the APC model that addresses the concerns raised in §2.5. We explain how computational issues encountered in developing this test were overcome.

4.1 Motivation for the test

The APC model that we use implies assumptions about the structure of the time effects that must be tested, as discussed in §2.5. In particular, the APC model in equation (6) assumes that there are no interaction effects between age, period, and cohort, a point that is recognised in the literature (Kupper et al., 1985). Such interaction effects would capture, for instance, the effect on obesity of a time- and age-limited government programme promoting healthy eating in schools. To test the assumptions imposed by the APC model, we compare the APC model to a more general model that includes the interaction effects described above. In the more general model the part of the linear predictor describing the time effects, μ_{ik}, has the form:

$$\mu_{ik} = \alpha_i + \beta_j + \gamma_k + \delta + \phi_{ik} + \psi_{ij} + \omega_{jk} \tag{12}$$

In this model, ϕ_{ik} captures the specific effect of being both a member of cohort k and aged i; ψ_{ij} captures the effect of being age i in period j; and ω_{jk} captures the effect of being a member of cohort k in period j. These are the interaction effects.

The model in (12) is over-parametrized and cannot be estimated in its present form. To understand this, recall that the μ_{ik} part of the linear predictor describes the position of an individual h in an age-cohort space, as discussed in §2.1. Therefore, the scope for using the position in age-cohort space to explain an outcome variable is limited by the dimensions of that age-cohort space. There are only as many unique positions as there are cells in the space, i.e. cells within the generalized trapezoid. This means that a linear predictor can have at maximum the same number of parameters as there are cells. However, each set of interaction effects ϕ_{ik} , ψ_{ij} , and ω_{jk} on its own contains as many parameters as there are cells.

Since any one set of the interaction effects will include a parameter for each cell in the age-cohort space, and such a parameter set completely describes the possible variation due to position in age-cohort space, the fit of a model that includes one parameter per cell will be equal to the fit of model (12). Similarly, we know that the fit of the reparametrized APC model with linear predictor

$$\eta_h = Z_h\zeta + \mu_{i_h k_h}, \qquad \mu_{i_h k_h} = X_h\xi \tag{13}$$

is equal to the fit of the oroginal APC model with $\mu_{i_h k_h}$ as in (6). We can therefore compare the fit of (6) to the fit of the more general model (12), which clearly nests (6), by comparing the fit of the reparametrized APC model to the fit of a model with a parameter for each age-cohort cell. This provides a test of the assumption of no interaction effects.

4.2 The more general model: the time-saturated model

The model with a parameter for each age-cohort cell, described in the preceding section, is referred to as the time-saturated model. This name highlights the similarity between our testing approach and the use of deviance tests in other contexts. A deviance test can be thought of as a comparison between the model of interest and a "fully-saturated" model. The fully-saturated model has degrees of freedom exactly equal to the number of observations, so all variation in the data is described by the model. In our analysis, the time-saturated

model effectively saturates the age-cohort array.

To construct the time-saturated model we replace the design vector X_h, from the reparametrized APC model, with a vector T_h of dimension n, where n is the number of cells in the age-cohort array. T_h is a unit vector, indicating the age-cohort cell to which the individual h belongs.

All of the earlier statistical analysis in §3 carries through, so it is possible to compare a model of the form where $\eta_h = Z_h\zeta + X_h\xi$ with the more general model where

$$\eta_h = Z_h\zeta + T_h\kappa. \tag{14}$$

Under classical normality assumptions, F-tests are used to compare the models. Under asymptotic assumptions, likelihood ratio tests are used, which have χ^2 distributions. This is true for both normal models (continuous outcomes) and logit models (binary outcomes). However, the dimension of the general time-saturated model is sufficiently large that it poses numerical issues, which we now discuss.

4.3 Estimation of the normal time-saturated model

The overall dimension of the combined design matrix $M = (Z, T)$ in the normal model is $H \times (d_z + n)$. In the obesity example, $d_z = 15$ with $n = 684$. Consequently it is challenging to evaluate and to invert $M'M$ using a computer due to memory allocation. We can address this problem by orthogonalizing the regressors and exploiting the unique structure of the design matrix T. Instead of estimating a regresion of Y on M directly, we evaluate the partitioned regression

$$Y = \left[Z - T(T'T)^{-1}T'Z\right]\zeta + T\rho + \varepsilon. \tag{15}$$

Here $[Z - T(T'T)^{-1}T'Z] = \tilde{Z}$ is the residual of a first-stage regression of Z on T.

Since $T'T$ is diagonal by virtue of the unit vector structure of the rows of T, we do not need to store the entire matrix $T'T$; we need only store the vector of elements of the main diagonal. We can thus avoid the memory allocation problem associated with $M'M$. Because $T'T$ is diagonal the inverse is found by taking the reciprocal of the diagonal elements. It is therefore easy to calculate $\tilde{Z}$. Since $\tilde{Z}$ and T are orthogonal by construction, ζ and ρ can easily be estimated by regression of Y on $\tilde{Z}$ and T, respectively. This poses no computational challenge since $\tilde{Z}$ is of dimension $H \times d_z$ and d_z is small.

We can retrieve κ from $\hat{\kappa} = \hat{\rho} - (T'T)^{-1}T'Z\hat{\zeta}$. Note that (15) and the equation in terms of M, ζ, and κ give equivalent models with the same fit and the same residual variance. As a consequence we are normally not interested in the value of $\hat{\kappa}$.

We can test the APC model against the saturated model using

$$F = \frac{(RSS_X - RSS_T)/(d_T - d_X)}{RSS_T/(H - d_T)}, \tag{16}$$

where d is the number of parameters in a model and H is the number of individuals. RSS is the residual sum of squares from a model. Subscripts indicate the model in question: X refers to the APC model, because it has design matrix X for the time component, while T refers to the TS model. RSS_T is equal to the residual sum of squares from the model (15). This F-statistic is asymptotically χ^2, or F-distributed under exact inference.

4.4 Estimation of the logit time-saturated model

We face similar computational problems in seeking to estimate the logit TS model as we did in the normal case. We address these by exploiting the structure of the derivatives of the log-likelihood in conjunction with the unit vector structure of T. The time-saturated model here is a logit model with mean $\eta_h = Z_h\zeta + T_h\kappa$ as in (14). Thus, the score is

$$\dot{\ell} = \begin{pmatrix} Z' \\ T' \end{pmatrix} (Y - \Pi), \tag{17}$$

where Π is a H-length vector of logistic probabilities π_h that depend on individual values of Z_h and X_h through (14). The matrix J as defined in §3.2.2 is

$$J = -\ddot{\ell} = \begin{pmatrix} Z' \\ T' \end{pmatrix} W \begin{pmatrix} Z & T \end{pmatrix} = \begin{pmatrix} J_{ZZ} & J_{ZT} \\ J_{TZ} & J_{TT} \end{pmatrix}, \tag{18}$$

where W is a diagonal matrix of Bernoulli variances, $\pi_h(1-\pi_h)$. Using partitioned inversion we find, with $J_{ZZ\cdot T} = J_{ZZ} - J_{ZT}J_{TT}^{-1}J_{TZ}$, that

$$J^{-1} = \begin{pmatrix} J_{ZZ\cdot T}^{-1} & -J_{ZZ\cdot T}^{-1}J_{ZT}J_{TT}^{-1} \\ -J_{TT}^{-1}J_{TZ}J_{ZZ\cdot T}^{-1} & J_{TT}^{-1} + J_{TT}^{-1}J_{TZ}J_{ZZ\cdot T}^{-1}J_{ZT}J_{TT}^{-1} \end{pmatrix}. \tag{19}$$

Here we make use of the unit vector structure of T. In the normal model, we used the fact that $T'T$ is diagonal. Here, we exploit the fact that the large-dimensional matrix $J_{TT} = T'WT$ is diagonal, and so we need only deal with the main diagonal as a vector. This greatly reduces the computational cost of calculating J^{-1}.

To estimate the parameters of the time-saturated logit model we again use a Newton iterative procedure. We initialize the parameters at zero, and update using the score and inverse second derivative of the log-likelihood calculated by the above formulas, noting that the diagonal structure is preserved in each step.

Given the estimated values of the parameters attained by this procedure, we can calculate the log-likelihood of the time-saturated model This can then be compared to the log-likelihood of the APC model by a likelihood ratio test. Under the assumptions outlined in §3.2.2 the likelihood ratio test statistic will be asymptotically χ^2.

5 Ad hoc identification of age, period, and cohort effects

When a researcher is directly interested in the effects of age, period, and cohort the choice of identication is important and the canonical parametrization of equation (9) is preferred for the reasons outlined in §2.3. This is true of our obesity analysis. However, in other situations researchers are primarily interested in the effect of a covariate such as education, but need to control for age, period, and cohort effects. An example is found in Ejrnæs & Hochguertel (2013) where the authors are interested in the effect of insurance status on unemployment but must isolate this from the effects of age and cohort. In that situation it does not matter

how the time effects are identified.

Recall the model for the linear predictor as outlined in equation (10):

$$\eta = Z\zeta + \mu, \qquad \mu = X\xi.$$

This can be viewed as the linear predictor in the normal model, the logit model or any other generalized linear model. The maximum likelihood estimators are denoted $\hat{\xi}, \hat{\zeta}$. In the preceding sections $\hat{\xi}$ was of interest; we now consider a situation where only the estimator $\hat{\zeta}$ is of interest. Now, suppose that we identify the age-period-cohort structure differently so that

$$\eta = Z\zeta + \mu, \qquad \mu = XQ\phi,$$

where Q is a known, invertible $q \times q$-matrix and $\phi = Q^{-1}\xi$. We suppose that the researcher has arrived at this formulation not via the canonical parametrization $X\xi$ but from some other identification strategy. Maximum likelihood gives the estimators $\hat{\phi}, \hat{\zeta}_\phi$, say. The two sets of parameters are linked through

$$\begin{pmatrix} \xi \\ \zeta \end{pmatrix} = \begin{pmatrix} Q & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \phi \\ \zeta \end{pmatrix}.$$

The mapping is one-one since Q is invertible. Due to the equivariance of maximum likelihood estimators (Davidson & MacKinnon, 1993, §8.3) we have in the same way that $(\hat{\xi}, \hat{\zeta}) = (Q\hat{\phi}, \hat{\zeta}_\phi)$ and in particular $\hat{\zeta} = \hat{\zeta}_\phi$. Thus, the estimator for ζ is invariant to the choice of Q.

We note that the parametrization

$$\xi = Q\phi \tag{20}$$

covers a range of ad hoc identification schemes appearing in the age-period-cohort literature. By ad hoc we mean that the identification is not invariant to the group of transformations described in (8).

An example of ad hoc identification is the constraint, from Mason et al. (1973),

$$\alpha_1 = \alpha_2 = \beta_J = \gamma_K = 0. \tag{21}$$

To conduct estimation given this constraint, the columns of the design D corresponding to $\alpha_1, \alpha_2, \beta_J, \gamma_K$ are dropped. This gives a design matrix D_λ of dimension $n \times p$ which has full column rank. Thus, this approach replaces $D\theta = X\xi$ with $D_\lambda\psi$, for a p-vector ψ, which makes regression computationally feasible. By combining ψ with the four constraints in equation (21) a q-vector is formed. In Appendix B we show that the model implied by (21) is indeed of the form $\xi = Q\phi$ by finding Q and ϕ. Note that this ad hoc identication is not invariant to (8), since replacing α_i by $\alpha_i + a$ for some arbitrary constant a does not respect the constraint $\alpha_1 = 0$. Thus the estimated age, period, and cohort effects are with reference to these constraints.

The analysis of Ejrnæs & Hochguertel (2013) actually imposes a quadratic constraint in addition to ad hoc identification and so cannot be represented in the form $\xi = Q\phi$. Specifically they assume

$$\mu_{ik} = \alpha_{(1)}i + \alpha_{(2)}i^2 + \beta_j + \gamma_{(1)}k + \gamma_{(2)}k^2 + \delta \qquad \text{with} \qquad \beta_1 = \beta_2 = 0.$$

Here, the level and the slope of the period effect β_j are not identified, hence the need for the ad hoc identification by $\beta_1 = \beta_2 = 0$. The constraint imposed by this model can be expressed in terms of a testable linear restriction on the canonical parameter. Since the age and cohort effects are quadratic their double differences are constant, see Nielsen & Nielsen (2014, §5.4.5). This linear restriction can be tested against the unrestricted age-period-cohort model as well as the time saturated model using the tests outlined above. Subject to imposing this linear restriction on the canonical parameter the estimate for ζ will be the same whether one uses this restricted canonical parametrization or the ad hoc identication.

6 Empirical application to obesity in England

We apply the methods outlined in the preceding sections to examine the dynamics of obesity in England using the data described in §2.1. We assess whether the APC model or any of its sub-models is sufficient to describe the time effects in the data. Using the continuous measure, log BMI, as the dependent variable, we find that an age-drift model is sufficient to describe the data on women and an age-cohort model is sufficient to describe the data on men. Using the binary measure, an indicator for obesity, as the dependent variable, we find an age-drift model for women and a cohort-drift model for men. Previous studies had detected non-linearities in age for both populations but had not detected the non-linearity in cohort among men.

6.1 Preliminary data analysis

We begin by visually inspecting the data. The heatmap in Figure 2 shows the mean values of BMI in each age-cohort cell for women. The heatmap for men is similar, and therefore is not shown. We can see that there is a pattern to the variation in mean values in the data, with more yellow (light) shading concentrated towards the right and centre of the graph and more blue (dark) shading concentrated towards the top and left. These patterns suggest that there are relationships between log BMI and one or more of age, period, and cohort. Our statistical analysis will not be able to identify any linear components of such relationships, but it will be able to identify the non-linear components. For example, the area of the array around ages 30-45 and periods 2005-2010 seems more blue (dark) in colour than the central area below it (ages 45-60, periods 2005-2010). However, the central area is similar in colour to the area below it (ages 60-80, periods 2005-2010). This suggests that there may be curvature, i.e. non-linearity, in the relationship between age and log BMI.

6.2 Covariates

Recall that our covariates are ethnicity, level of education, NSSEC, smoking history, and alcohol consumption. These were adapted to facilitate regression analysis, as described below. Descriptive statistics are reported in tables 12 and 13 of Appendix C. The reference ethnicity was taken to be white, with indicators for self-identification as black, Asian, of mixed ethnicity, or of "other" ethnicity (including e.g. Arab). For education, those who

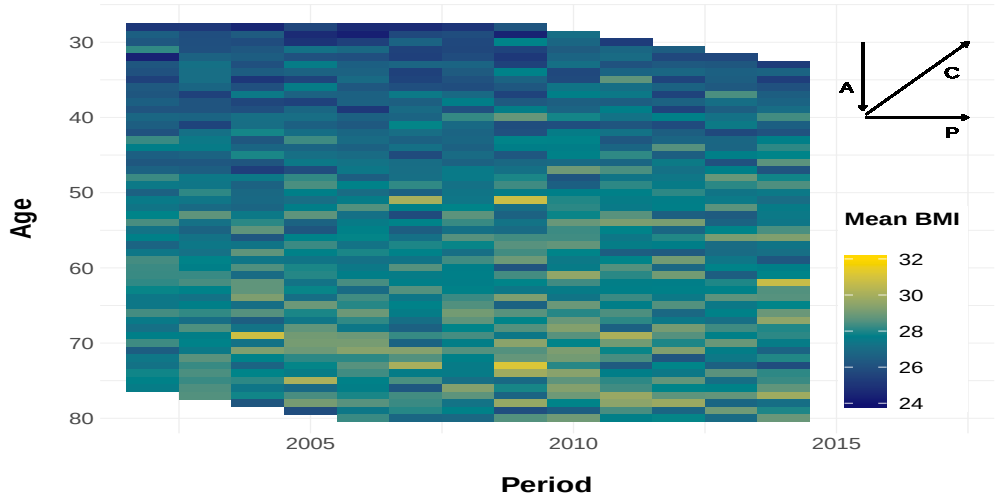


Figure 2: Within-cell BMI means women

left school after attaining a GCSE[1] or equivalent qualification were taken as the reference group, and we included three indicators: education below GCSE level, holders of a university degree, and those with education beyond GCSE but below degree level. More detail on the harmonization of different qualifications is available in the HSE documentation.

The three-class version of the National Statistics Socio-economic Classification (NSSEC) is used. The reference category is "Routine and Manual" occupations. Indicators are included for "Intermediate", "Managerial and Professional", and "Other" occupation groups. The "Other" group includes students, those permanently outside the labour force, the long-term unemployed, and anyone whose employment could not be satisfactorily classified.

Smoking behaviour is captured by two indicators. One records whether an individual currently smokes, while the other captures former regular smokers. For alcohol consumption, the casual drinking population (those drinking one to four times a week) was taken to be the reference and indicators were introduced classifying individuals as not drinking at all, drinking rarely (less than once a week), and drinking frequently (five or more times a week).

6.3 Model for the continuous outcome variable, log BMI

The techniques described in §3.1 are used to estimate and conduct inference on the non-linear effects of age, period, and cohort on log BMI. The explanatory variables are the KNN parametrization of the APC structure and the covariates described in §6.2. The analysis is performed separately for men and women. The analysis begins with fitting the general TS model, the APC model, and all sub-models described in §2.4, and comparing them using likelihood ratio tests and the Akaike Information Criterion (AIC). A preferred model is selected, and results for that model are presented and discussed. For women, the preferred model is age-drift, i.e. non-linearities in age only; for men, the preferred model is age-cohort, i.e. non-linearities in age and cohort but not in period.

[1]General Certificate of Secondary Education. This is the minimum school-leaving qualification, obtained around age 16. In 2019 most students remain in school after completing GCSEs to obtain A-levels, but in older cohorts many left school post-GCSE.

6.3.1 Women

Table 1: **Model comparisons, log BMI, women**

	Against TS			Against APC				
	F	df	p	F	df	p	AIC	ℓ
TS							-22747.33	12101.67
APC	1.02	592	0.36				-23321.91	11796.95
AP	0.99	646	0.53	0.73	54	0.93	-23390.28	11777.14
AC	1.03	604	0.32	1.38	12	0.17	-23329.33	11788.67
PC	1.11	643	0.03	2.16	51	0.00	-23313.70	11741.85
Ad	1.00	658	0.47	0.85	66	0.80	-23397.46	11768.73
Pd	1.22	697	0.00	2.35	105	0.00	-23284.82	11673.41
Cd	1.11	655	0.02	2.00	63	0.00	-23321.56	11733.78
A	1.05	659	0.20	1.29	67	0.06	-23369.27	11753.64
P	1.51	698	0.00	4.25	106	0.00	-23084.91	11572.46
C	1.28	656	0.00	3.74	64	0.00	-23210.61	11677.31
t	1.22	709	0.00	2.25	117	0.00	-23292.97	11665.49

Table 2: **Model comparisons, log BMI, women**

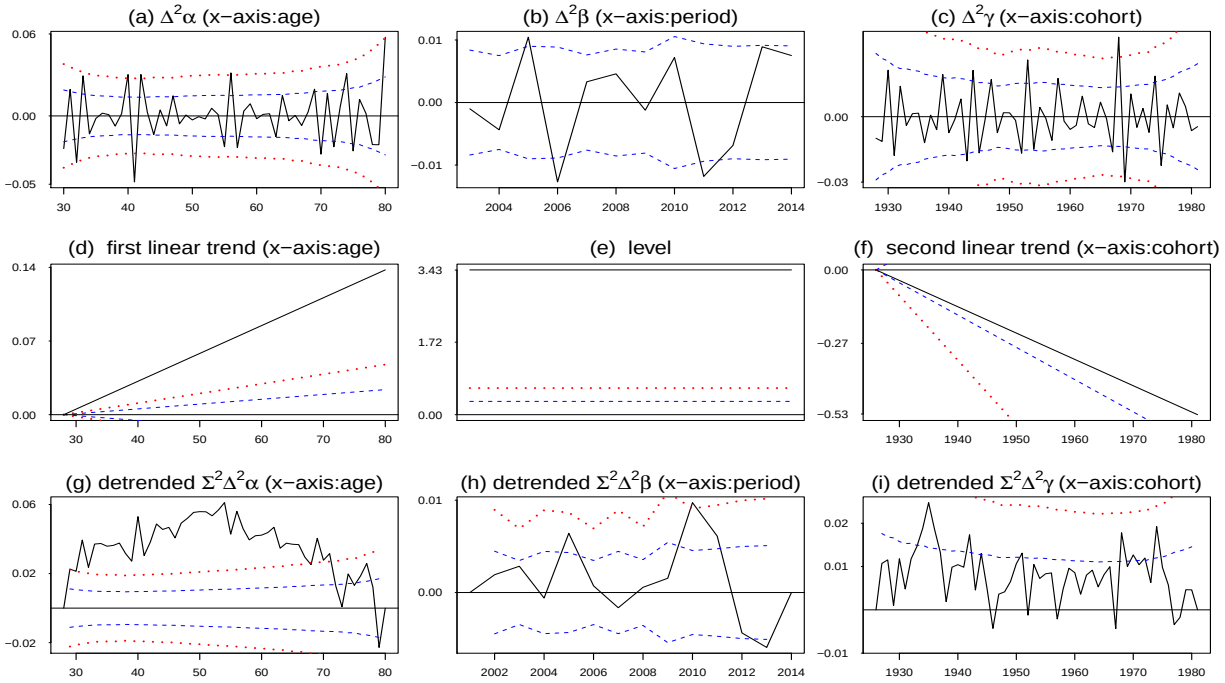
Models compared	Ad vs AC	Ad vs AP	A vs Ad
p	0.926	0.158	0.000

We begin by analysing the women. Table 1 displays statistics that facilitate comparison between models. Each row represents a different model. The differences between these models are described in detail in §2.4. The results from F-tests against the TS and APC models are shown, along with the AIC and log-likelihood. We conduct model selection by first ranking models in terms of the AIC; the most preferred model is that with the lowest AIC. We then check the results from F-tests of these models, and select the model with the lowest AIC value that is also not rejected by the F-tests comparing it to larger models.

In Table 1, the Ad model has the smallest AIC, followed by AP and A. The F-tests comparing the AP and Ad models to the APC model are not rejected. In Table 2 we conduct direct F-tests between the AP, Ad, and A models. The reduction to Ad from the larger AP model is supported, so the Ad model is selected according to our criteria above. The smallest A model is not supported as a reduction either from the APC or the Ad model.

We should now display the estimates from the Ad model; however, for illustrative purposes we instead display those from the full APC model. The KNN parametrization of the APC effects contains many parameters, and so these are best displayed graphically, as in figure 3. The middle row of panels contains the combined linear plane (KNN parameters υ_o, υ_a, and υ_c). The slopes are always estimated along the age and cohort dimensions, but they combine the linear parts of all three time effects. The top row of panels contains the series of

Figure 3: **Time effects, APC model of log BMI, women**



solid line = estimate; blue (red) dotted line = 1 (2) standard deviation

estimated double differences in age, period, and cohort (KNN parameters of the form $\Delta^2\alpha_i$, $\Delta^2\beta_j$, and $\Delta^2\gamma_k$).

Each double-difference $\Delta^2\alpha_i$ captures the extent to which the relationship between log BMI at adjacent ages deviates from a linear relationship. To calculate the overall non-linear relationship between age and log BMI, the double differences must be cumulated through the design matrix X_h, as mentioned in §2.3 and described in detail in Appendix A. The same holds for period and cohort.

The overall non-linear effect of age, period, and cohort on log BMI, calculated by cumulating double-differences, is shown in the bottom row of panels. In the panels shown here, these cumulative effects have been "de-trended" post-estimation so that they begin and end at zero; the removed trends have been added on to the linear plane in the middle panels. The de-trended cumulated time effects can be interpreted individually due to the two zero constraints. They show the non-linear development in the time effects over and above the unidentified linear trends.

Panel (g) of figure 3 shows significant concavity in age: the cumulated non-linearities exceed the red dotted line that marks two standard deviations. The concave shape in age is often found in epidemiological studies, see for instance Nielsen (2015). It is also consistent with the discussion in §6.1. The fact that neither the period non-linearities (panel h) nor the cohort non-linearities (panel i) is significant is consistent with the earlier conclusion that the Ad model would be sufficient to describe the data.

Figure 3 could be repeated for the Ad model. That model excludes the double differences

in period and cohort, so the plots in panels (b,c,h,i) fall away. The corresponding figure for the Ad model has nearly the exact same panels (a,g). This is unsurprising given that the Ad model is not rejected against the APC model.

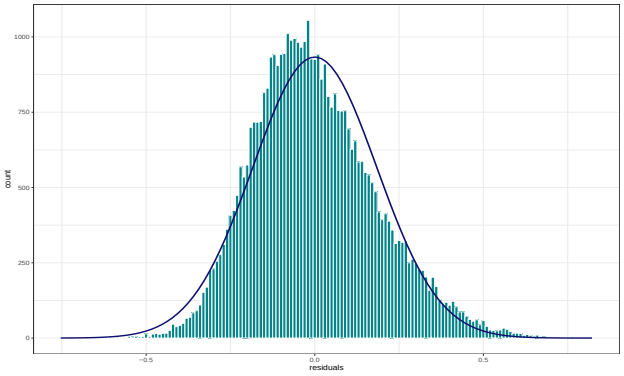
The coefficients on the covariates of the Ad model are seen in table 6. Interpretation of these is deferred to §6.3.3, where they are discussed in conjunction with the estimated effects for men.

Formal mis-specification tests for the Ad model are reported in table 3 for situations with and without log transformation of the dependent variable. The tests include a cumulant based test for normality of residuals and tests for functional form mis-specification and heteroskedasticity (Ramsey, 1969; White, 1980). The log transformation clearly improves the specification. Yet, given the large sample size, $H = 43,077$, it is difficult to avoid very small p-values. The histogram of the residuals in figure 4 suggests that the non-normality is not too severe. Nonetheless, as a precaution we conducted various robustness checks reported in §6.3.3. Those checks indicate that the mis-specification is not detrimental for inference.

Table 3: **Ad model specification tests, women**

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

Figure 4: **Residuals from Ad model of log BMI, women**



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

6.3.2 Men

For the men, a similar approach is followed. First, the table comparing all candidate models is constructed, see table 4. We see that the AIC is minimized by the Cd model. However,

the F-test comparing the Cd model to the APC model rejects, suggesting that important information is lost in moving from the APC to the Cd model. Looking at the remaining models, a case could be made for either the AC model (based on the F-test) or the PC model (based on the AIC). To aid selection we conducted direct tests comparing the AC, PC, and Cd models, seen in table 5. These tests suggest that age non-linearities are important, but period non-linearities are not. Thus an age-cohort model was deemed optimal.

The estimated reparametrized time effects from the APC model are seen in figure 5. There is some curvature in each of age and cohort, while the period non-linearity is driven by the anomalous spike in 2010. This lent support to our decision to exclude the PC model and focus on the AC model.

The double differences, linear plane, and detrended time effects for the AC model were very similar to those from the APC model apart from the omission of panels (b) and (h). The plot of the AC model is therefore omitted. The estimated coefficients on the covariates are seen in table 6. Mis-specification tests on the residuals are similar to those for women reported in table 3 and therefore are not shown.

Table 4: **Model comparisons, log BMI, men**

	Against TS			Against APC				
	F	df	p	F	df	p	AIC	ℓ
TS							-36449.64	18952.82
APC	0.98	592	0.59				-37043.86	18657.93
AP	1.06	646	0.15	1.85	54	0.00	-37051.87	18607.93
AC	0.99	604	0.56	1.22	12	0.26	-37053.16	18650.58
PC	1.03	643	0.29	1.57	51	0.01	-37065.67	18617.83
Ad	1.06	658	0.14	1.73	66	0.00	-37061.31	18600.65
Pd	1.56	697	0.00	4.80	105	0.00	-36750.82	18406.41
Cd	1.03	655	0.27	1.50	63	0.01	-37075.39	18610.70
A	1.15	659	0.00	2.63	67	0.00	-37001.23	18569.62
P	1.60	698	0.00	5.05	106	0.00	-36722.71	18391.35
C	1.21	656	0.00	3.33	64	0.00	-36958.31	18551.16
t	1.55	709	0.00	4.42	117	0.00	-36762.34	18400.17

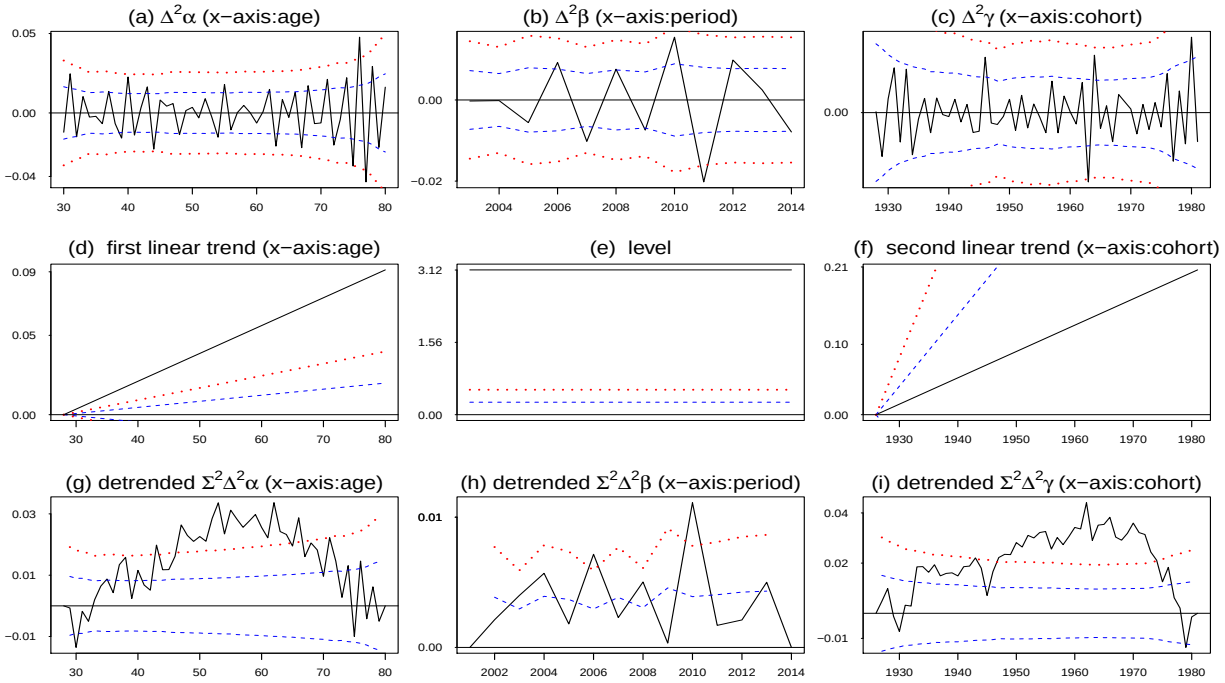
Table 5: **Model comparisons, log BMI, men**

Models compared	Cd vs AC	Cd vs PC
p	0.006	0.285

6.3.3 Interpretation

Recall that for women we selected the Ad model and for men we selected the AC model. It is important to remember that the plane has been chosen to allow all deviations from linearity

Figure 5: **Time effects, APC model of log BMI, men**



solid line = estimate; blue (red) dotted line = 1 (2) standard deviation

to begin and end with zero. As such it does not have a typical best-fit interpretation. However, it is still somewhat reassuring that the estimated intercept is within the plausible range for BMI (about 3 on the log scale, corresponding to a BMI of 20), and that the significant slope of the plane is positive, consistent with the known increase in mean BMI over time for the full population. Due to the identification problem it is impossible to say whether that is a period effect, or a result of the aging population combined with an age or cohort effect.

For both men and women, there are significant deviations from linearity. For women there is concavity in log BMI with a peak at age 50. It should be borne in mind that this concavity will appear over an age slope, which is indeterminate and may be either positive or negative. The concavity may be consistent with general metabolic effects or selection effects towards the end of life, as those with higher BMI die sooner (Hruby et al., 2016). Children may also be a factor, both due to the biological effect of child-bearing on weight and the impact of child-rearing on free time for personal healthcare. For men, there is curvature in both the age and cohort dimensions. The age non-linearity is not as significant as that for women, and it begins later, suggesting that child-bearing may be an important factor among women. The significance of cohort among men is more difficult to explain, but may be related to generational shifts in the nature of employment. We hypothesize that men from the central cohorts may have similar dietary habits to men of earlier cohorts, but have a more sedentary lifestyle and do less physical labour; whereas more recent cohorts eat a more varied diet with less heavy, traditional British fare. Such factors could affect men more

than women due to the long-standing social pressure on women to moderate their diets to "keep their figure". Further targeted research would be required to validate any of these hypotheses.

There is little in the way of period non-linearities; the only point at which the period effect attains significance is in 2010, where there is an unusual and so far unexplained spike in log BMI. As discussed earlier, we judge this spike to be non-informative about the evolution of log BMI.

The effects of the covariates are largely consistent with the literature. Black individuals have higher BMI than white individuals, on average, while those of other ethnicities have lower BMI (Ogden et al., 2015; An & Xiang, 2016). BMI and social class are negatively correlated (McPherson et al., 2007). Those with more education have lower BMI on average (Baum II & Ruhm, 2009; An & Xiang, 2016). Those who currently smoke have lower BMI on average, while those with a history of smoking have higher BMI on average, than those who have never smoked (Akbartabartoori et al., 2005). Non-drinkers and rare drinkers have higher BMI than casual drinkers (the reference group), who in turn have higher BMI than frequent drinkers. This may be explained by the mismatch between frequency and quantity of consumption (O'Donovan et al., 2018).

There are some sex differences in the covariates, primarily relating to significance. Black women and women of mixed ethnicity have significantly higher and lower BMI, respectively, than white women; whereas for men these ethnicities are not significant. Non-drinking men do not differ significantly from casual drinkers, and the effects of social class are not significant for men.

6.3.4 Robustness checks

A range of alternative specifications of the normal model were examined as robustness checks. Using the same data, we replaced the three-class NSSEC with the eight-class version. We considered a model with log weight as the dependent variable and log height as an explanatory variable; a model with log BMI as the dependent variable implicitly imposes a coefficient of two in this regression, and we wanted to evaluate whether this was restrictive. These models did not change our substantive findings.

We also considered different subsets of the original HSE data. To examine whether income yielded different results to the NSSEC, we tried a specification which replaced the NSSEC with inflation-adjusted household income (quadratic in logs) using two samples: first with all observations where income information was available, then for only observations where both income and NSSEC information was available. There was no substantial change to the non-linearities or the covariates. Given the apparent insensitivity of the estimated covariate coefficients to the APC specification, we decided that the time and covariate effects were largely orthogonal and tested a model which excluded the covariates. This gave us a much larger sample size due to less missing information. The substantive results were unchanged. Finally, to check whether the differences in sample size across years affected our results we randomly selected 2000 observations from each year and ran the original analysis on this smaller sample, using three different random seeds. The age non-linearities were robust to this check for both men and women.

In our final set of robustness checks we tested extensions of the age-cohort space. We

Table 6: **Covariate effects, log BMI models**

	Women, Ad			Men, AC		
	$\hat{\zeta}$	*se*	*p*	$\hat{\zeta}$	*se*	*p*
Ethnicity indicators (excl. white)						
Black	0.068	0.007	0.000	-0.008	0.006	0.208
Asian	-0.045	0.007	0.000	-0.039	0.005	0.000
Mixed ethnicity	-0.023	0.011	0.035	-0.008	0.011	0.428
Other ethnicity	-0.069	0.012	0.000	-0.033	0.011	0.004
Behaviour indicators (excl. never smoked, occasionally drink alcohol)						
Former smoker	0.024	0.002	0.000	0.026	0.002	0.000
Current smoker	-0.030	0.002	0.000	-0.045	0.002	0.000
Never drink alcohol	0.043	0.008	0.000	0.001	0.010	0.921
Rarely drink alcohol	0.037	0.002	0.000	0.014	0.002	0.000
Frequently drink alcohol	-0.032	0.003	0.000	-0.017	0.002	0.000
Education level indicators (excl. GCSE)						
Below GCSE	0.016	0.003	0.000	0.010	0.002	0.000
Some higher education	-0.012	0.003	0.000	0.002	0.002	0.393
University degree	-0.048	0.003	0.000	-0.026	0.003	0.000
3 level NSSEC indicators (excl. routine/manual)						
Intermediate occupations	-0.021	0.002	0.000	-0.001	0.002	0.696
Managerial/Professional	-0.009	0.003	0.003	-0.001	0.002	0.485
Other occupations	-0.013	0.007	0.064	-0.020	0.011	0.063

considered the original model (with NSSEC) but with the age range extended to be from 20-80, and the cohort range extended accordingly. This incorporated some cells in which perfect separation was present, but that is not a problem in the normal model. The main consequence of this was a strengthening of the significance of age non-linearities for men, with curvature in the early twenties that could be explained by particularly rapid growth in log BMI over those ages. The NSSEC was not recorded prior to 2001, but we have income information back to 1997, so we were able to consider the model with income over a longer period horizon. We were also able to evaluate a no-covariates model with data back to 1992. The estimated age effects remained similar to the original models throughout. With an extended period range, the period non-linearities become significant and exhibit concavity, which could be explained by a reduction in the rate of growth in log BMI after the 1990s.

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

6.4 Model for the binary outcome variable, an obesity indicator

We defined a binary outcome variable, `obese`, that takes the value 1 for $BMI \geq 30$ and 0 otherwise. This was analysed using a logit model as described in §3.2. The covariates used are the same as those for the model of log BMI from §6.3. Again, the sexes are considered separately. The first step in the analysis is to produce a table of model comparison statistics, to determine which APC sub-model is appropriate for the data.

For the women, the model comparison statistics are presented in table 7. Direct tests comparing plausible candidate models are seen in table 8. Both sets of tests favour the Ad model. The estimated Ad model is seen in figure 6 with the covariate coefficients in table 11.

For the men, the model comparison statistics are presented in table 9. In this case models A, P, C, and t are clearly rejected and thus omitted. Direct tests are shown in table 10. From these it is clear that the Cd model is favoured. This contrasts with the study of log BMI, where the AC model was chosen over the Cd model. The estimated Cd model effects are seen in figure 7 with the covariate coefficients in table 11.

The substantive patterns seen when using obesity as the outcome variable are broadly similar to those seen when using log BMI as the outcome variable. For women, the age deviations are similar, but in the logit model the curvature is less smooth. There are apparent kinks at ages 30 and 50. This contrasts with more gradual curvature in the log BMI model. One possible interpretation is that between the ages of 30 and 50, an increasing number of women gain weight but few pass the obesity threshold. This interpretation could be justified by the existence of a psychological effect of being classified "obese" which causes women to

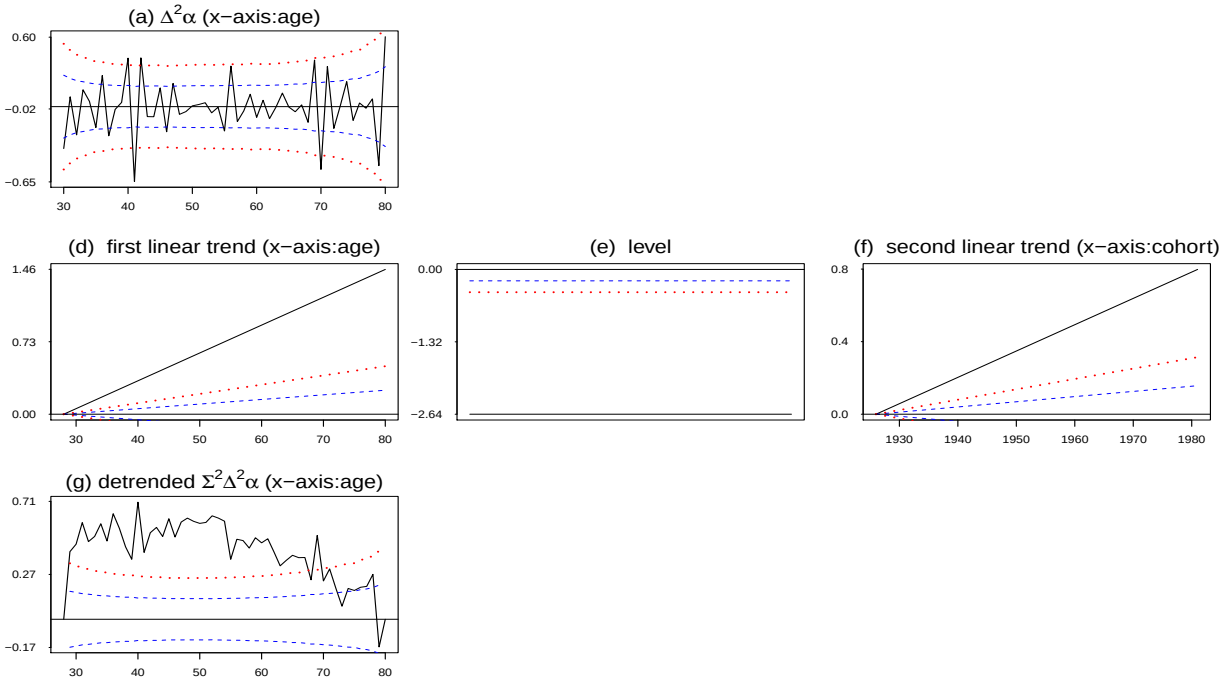
Table 7: **Model comparisons, obesity indicator, women**

	Against TS			Against APC				
	LR	df	p	LR	df	p	AIC	ℓ
TS							48708.92	-23627.46
APC	598.10	592	0.42				48123.02	-23926.51
AP	632.60	646	0.64	34.49	54	0.98	48049.51	-23943.76
AC	609.29	604	0.43	11.19	12	0.51	48110.21	-23932.10
PC	671.85	643	0.21	73.75	51	0.02	48094.76	-23963.38
Ad	643.83	658	0.65	45.72	66	0.97	48036.74	-23949.37
Pd	752.79	697	0.07	154.69	105	0.00	48067.71	-24003.85
Cd	683.41	655	0.21	85.31	63	0.03	48082.32	-23969.16
A	669.59	659	0.38	71.48	67	0.33	48060.50	-23962.25
P	809.36	698	0.00	211.26	106	0.00	48122.28	-24032.14
C	745.72	656	0.01	147.62	64	0.00	48142.63	-24000.32
t	763.97	709	0.07	165.86	117	0.00	48054.88	-24009.44

Table 8: **Model comparisons, obesity indicator, women**

Models compared	Ad vs AC	Ad vs AP	A vs Ad	t vs Ad	t vs A
p	0.98	0.51	0.00	0.00	0.00

Figure 6: **Time effects, Ad model of obesity indicator, women**



solid line = estimate; blue (red) dotted line = 1 (2) standard deviation

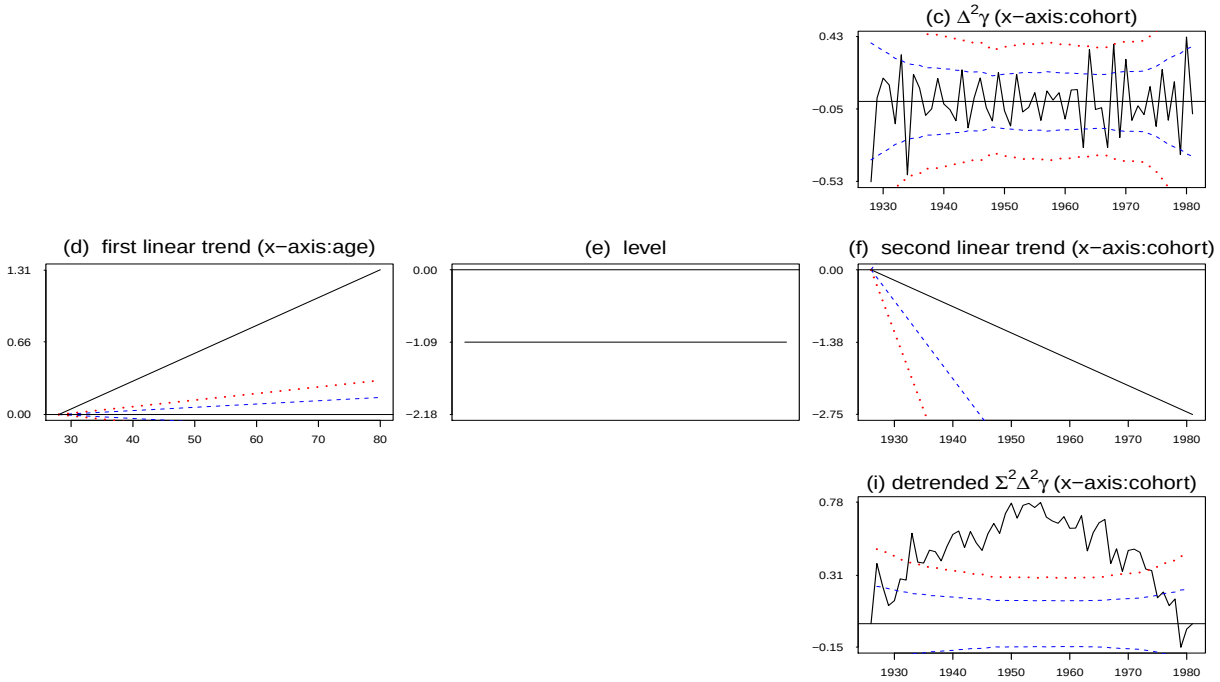
Table 9: **Model comparisons, obesity indicator, men**

	Against TS			Against APC				
	LR	df	p	LR	df	p	AIC	ℓ
TS							44563.50	-21554.75
APC	556.26	592	0.85				43935.76	-21832.88
AP	630.12	646	0.66	73.86	54	0.04	43901.62	-21869.81
AC	577.07	604	0.78	20.81	12	0.05	43932.57	-21843.29
PC	615.23	643	0.78	58.97	51	0.21	43892.74	-21862.37
Ad	651.07	658	0.57	94.81	66	0.01	43898.57	-21880.28
Pd	861.50	697	0.00	305.24	105	0.00	44031.00	-21985.50
Cd	635.40	655	0.70	79.14	63	0.08	43888.90	-21872.45

Table 10: **Model comparisons, obesity indicator, men**

Models compared	Cd vs AC	Cd vs PC	Ad vs AC
p	0.224	0.064	0.037

Figure 7: **Time effects, Cd model of obesity indicator, men**



solid line = estimate; blue (red) dotted line = 1 (2) standard deviation

avoid moving into the category, but does not affect weight loss.

For the men, in the APC model for obesity and in the AC model for log BMI, there is something like curvature in each of age and cohort. Upon reduction to the Cd model, it appears that the two are combined to result in a larger cohort curve with an earlier peak. This could be explained by the following mechanism: the group of men aged approximately 40-60 in 2001-2014 have higher mean BMI than those of other ages. Because of the limited period range of this dataset, the observation of 40-60-year-olds overlaps with the observation of cohorts born in 1940-1980; we do not observe middle-aged men from other cohorts, and we do not observe these cohorts at anything other than middle age. It is therefore impossible, even with good APC techniques, to separate the cohort and age influences for this group. A longer period range is needed.

Table 11: **Covariate effects, obesity models**

	Women, Ad			Men, Cd		
	$\hat{\zeta}$	*se*	*p*	$\hat{\zeta}$	*se*	*p*
Ethnicity indicators (excl. white)						
Black	0.658	0.077	0.000	-0.034	0.095	0.721
Asian	-0.593	0.105	0.000	-0.588	0.089	0.000
Mixed ethnicity	-0.190	0.149	0.203	-0.170	0.172	0.326
Other ethnicity	-0.628	0.178	0.000	-0.343	0.188	0.068
Behaviour indicators (excl. never smoked, occasionally drink alcohol)						
Former smoker	0.201	0.027	0.000	0.310	0.027	0.000
Current smoker	-0.216	0.030	0.000	-0.356	0.033	0.000
Never drink alcohol	0.507	0.097	0.000	0.347	0.143	0.015
Rarely drink alcohol	0.428	0.025	0.000	0.201	0.029	0.000
Frequently drink alcohol	-0.332	0.035	0.000	-0.158	0.029	0.000
Education level indicators (excl. GCSE)						
Below GCSE	0.194	0.031	0.000	0.160	0.035	0.000
Some higher education	-0.109	0.033	0.001	-0.008	0.034	0.811
University degree	-0.452	0.040	0.000	-0.280	0.040	0.000
3 level NSSEC indicators (excl. routine/manual)						
Intermediate occupations	-0.239	0.029	0.000	-0.036	0.033	0.267
Managerial/Professional	-0.084	0.032	0.009	-0.117	0.031	0.000
Other occupations	-0.078	0.086	0.363	0.015	0.162	0.924

7 Conclusion

In this paper, we developed a framework for studying the effects of age, period, and cohort on an outcome variable, using repeated cross-sectional data. Due to the impossibility of identifying linear effects of age, period, and cohort, we focused on the non-linear effects, which we isolated using a parametrization developed by Kuang et al. (2008). We extended

the methodological tools available to researchers by showing how this parametrization can be used with repeated cross-sectional data and covariates, and by introducing a test of this parametrization against a more general model. We illustrated these tools with an application to the study of obesity in the England.

Our analysis of data from the Health Survey for England demonstrated clear non-linear age and cohort effects that were robust to a range of specifications. For women, the only significant deviation from linearity is concavity in the age dimension, with a kink at middle age. We suggest metabolic changes, child-bearing, and child-rearing as potential reasons for this. For men, there is significant concavity in the cohort dimension, and, for log BMI only, in the age dimension. The concavity in cohort is a novel finding in the literature on English obesity. We suggest shifting dietary patterns, with later cohorts consuming less of heavy, traditional British fare, as a potential reason for the concavity. For both genders the impact of the covariates is largely consistent with existing literature.

To estimate these effects, we employed the parametrization of Kuang et al. (2008) for identification of the non-linear effects of age, period, and cohort. In employing this parametrization, we made several methodological contributions. We used a generalized linear modelling framework to introduce this parametrization at the individual level, via a normal model for the continuous measure, log BMI, and a logit model for the binary measure, obesity. To assess the adequacy of the age-period-cohort model, in particular the assumption of no interaction effects between age, period, and cohort, we introduced the idea of testing against a "time-saturated" model. We developed an algorithm for estimation of the time-saturated model that exploits its mutually exclusive indicator variable structure. Additionally, we showed in §5 that some other age-period-cohort identification strategies yield the same coefficient estimates for the covariates as the Kuang et al. (2008) parametrization; however, their estimates of the age, period, and cohort effects depend on untestable assumptions. Therefore these strategies are suitable for inference on the covariate coefficients only. Overall, this paper extends the range of options available for analysis of age, period, and cohort effects from repeated cross-sections.

That said, there is plenty of work yet to do. The existing framework could be expanded to allow for mixture models, interaction terms between time effects and covariates, and heteroskedastic errors. This would enable us to address some of the mis-specification concerns with the log BMI analysis. While standard methods exist for mixture models, interaction terms, and heteroskedasticity, it is not clear how they would be affected by the collinearity in age, period, and cohort effects. Therefore, care is needed when introducing these methods to our framework. A clear limitation of our framework is that the implications of missing age-cohort cells within the age-cohort array have not been thought through; in the present application, this meant we had to drop all ages below 28 due to perfect separation in one age-cohort cell at age 27. It would also be of interest to test parametric models for the deviations from linearity, for example quadratic models.

Acknowledgements

Funding was received from ESRC grant ES/J500112/1 (Fannon), ERC grants 681546, FAMSIZEMATTERS (Monden), and 694262, DisCont (Nielsen).

References

Agresti, A. (2013). Categorical Data Analysis (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Akbartabartoori, M., Lean, M. E. J., & Hankey, C. R. (2005). Relationships between cigarette smoking and body shape. International Journal of Obesity, 29, 236–243.

An, R. & Xiang, X. (2016). Age-period-cohort analyses of obesity prevalence in US adults. Public Health, 141, 163–169.

Baum II, C. L. & Ruhm, C. J. (2009). Age, socioeconomic status and obesity growth. Journal of health economics, 28, 635–648.

Carstensen, B. (2007). Age-period-cohort models for the Lexis diagram. 26, 3018–3045.

Clayton, D. & Schifflers, E. (1987). Models for tempral variation in cancer rates. II Age-period-cohort models. 6, 469–481.

Davidson, R. & MacKinnon, J. (1993). Estimation and Inference in Econometrics. Oxford: Oxford University Press.

Department of Health (2011). Healthy lives, healthy people: A call to action on obesity in England. Technical Report 16166, HM Government.

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

Fahrmeir, L. & Kaufmann, H. (1986). Asymptotic inference in discrete response models. Statistical Papers, 27, 179–205.

Fu, W. (2016). Constrained estimators and consistency of a regression model on a Lexis diagram. 111, 180–199.

Glenn, N. D. (2005). Cohort Analysis (2nd ed.)., volume 5 of Quantitative Applications in the Social Sciences. SAGE Publications, Inc.

Harnau, J. & Nielsen, B. (2017). Over-dispersed age-period-cohort models. Journal of the American Statistical Association, to appear.

Holford, T. R. (1983). The estimation of age, period and cohort effects for vital rates. Biometrics, 39, 311–324.

Howel, D. (2011). Trends in the prevalence of obesity and overweight in english adults by age and birth cohort, 1991–2006. Public health nutrition, 14(1), 27–33.

Hruby, A., Manson, J. E., Qi, L., Malik, V. S., Rimm, E. B., Sun, Q., Willet, W. C., & Hu, F. B. (2016). Determinants and consequences of obesity. AJPH Special Section: Nurses' Health Study Contributions, 106, 1656–1662.

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

Kupper, L. L., Janis, J. M., Karmous, A., & Greenberg, B. G. (1985). Statistical age-period-cohort analysis: A review and critique. Journal of Chronic Diseases, 38, 811–830.

Lean, M. E. J., Katsarou, C., McLoone, P., & Morrison, D. S. (2013). Changes in bmi and waist circumference in scottish adults: use of repeated cross-sectional surveys to explore mulitple age groups and birth-cohorts. International Journal of Obesity, 37, 800–808.

Martínez Miranda, M. D., Nielsen, B., & Nielsen, J. P. (2015). Inference and forecasting in the age-period-cohort model with unknown exposure with an application to mesothelioma mortality. Journal of the Royal Statistical Society, Series A, 178, 29–55.

Mason, K. O., Mason, W. M., Winsborough, H. H., & Poole, K. (1973). Some methodological issues in cohort analysis of archival data. American Sociological Review, 38, 242–258.

McPherson, K., Marsh, T., & Brown, M. (2007). Tackling obesities: Future choices - modelling future trends in obesity and the impact on health. Technical report, Government Office for Science.

Moody, A. (2016). Health Survey for England 2015 adult overweight and obesity. Technical report, Health and Social Care Information Centre.

NatCen Social Research, University College London. Department of Epidemiology and Public Health (2016). Health Survey for England 2014 [data collection]. Data retrieved from https://discover.ukdataservice.ac.uk/series/?sn=2000021.

Nielsen, B. (2014). Deviance analysis of age-period-cohort models. Discussion paper, Nuffield College.

Nielsen, B. (2015). apc: An R package for age-period-cohort analysis. The R Journal, 7, 52–64.

Nielsen, B. & Nielsen, J. P. (2014). Identification and forecasting in mortality models. The Scientific World Journal, 2014, Article ID 347043, 24 pages.

O'Donovan, G., Stamatakis, E., & Hamer, M. (2018). Associations between alcohol and obesity in more than 100 000 adults in england and scotland. British Journal of Nutrition, 119, 222–227.

Ogden, C. L., Carroll, M. D., Fryar, C. D., & Flegal, K. M. (2015). Prevalence of obesity among adults and youth: United States, 2011-2014. Technical Report 219, National Center for Health Statistics, Hyattsville, MD.

Poirier, D. J. (1998). Revising beliefs in nonidentified model. Econometric Theory, 14, 483–509.

Ramsey, J. B. (1969). Test for specification errors in classical linear least-squares regression analysis. Journal of the Royal Statistical Society, Series B (Methodological), 31, 350–371.

Reither, E. N., Hauser, R. M., & Yang, Y. (2009). Do birth cohorts matter? Age-period-cohort analyses of the obesity epidemic in the United States. Social Science & Medicine, 69, 1439–1448.

Scarborough, P., Bhatnagar, P., Wickramasinghe, K. K., Allender, S., Foster, C., & Rayner, M. (2011). The economic burden of ill health due to diet, physical inactivity, smoking, alcohol and obesity in the uk: an update to 2006-07 nhs costs. Journal of Public Health, 33(4), 527–535.

Smith, T. R. & Wakefield, J. (2016). A review and comparison of age-period-cohort models for cancer incidence. Statistical Science, 31, 591–610.

Wang, Y. C., McPherson, K., Marsh, T., Gortmaker, S. L., & Brown, M. (2011). Health and economic burden of the projected obesity trends in the usa and the uk. Lancet, 378, 815–825.

Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. Biometrika, 63, 27–32.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica, 48, 817–838.

Wooldridge, J. (2010). Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: Massachusetts Institute of Technology.

Yang, Y. (2008). Social inequalities in happiness in the United States, 1972 to 2004: An age-period-cohort analysis. American Sociological Review, 73, 204–226.

Yang, Y. & Land, K. C. (2006). A mixed models approach to the age-period-cohort analysis of repeated cross-section surveys, with an application to data on trends in verbal test scores. Sociological Methodology, 36, 75–97.

A Details of Kuang et al. (2008) parametrization

The classical APC model we wish to use to represent the data is given in equation (6) of §2.2, that is $\mu_{ik} = \alpha_i + \beta_j + \gamma_k + \delta$.

The reparametrization of this model is produced by introducing telescopic sums of the following form into the model equation:

$$\alpha_i = \alpha_1 + \sum_{t=2}^{i} \Delta\alpha_t \qquad\qquad \Delta\alpha_t = \Delta\alpha_2 + \sum_{s=3}^{t} \Delta^2\alpha_s, \tag{22}$$

where $\Delta\alpha_t = \alpha_t - \alpha_{t-1}$ and $\Delta^2\alpha_s = \Delta\alpha_s - \Delta\alpha_{s-1}$. This results in an equation of the general form

$$\begin{aligned}
\mu_{ik} &= \upsilon_o + (i-U)\upsilon_a + (k-U)\upsilon_c + A_i + B_j + C_k \qquad (23)\\
A_i &= 1_{(i<U)} \sum_{t=i+2}^{U+1} \sum_{s=t}^{U+1} \Delta^2\alpha_s + 1_{(i>U+1)} \sum_{t=U+2}^{i} \sum_{s=U+2}^{t} \Delta^2\alpha_s \\
B_j &= 1_{(L\,odd\,\&\,j=2U-2)} \Delta^2\beta_{2U} + 1_{(j>2U)} \sum_{t=2U+1}^{j} \sum_{s=2U+1}^{t} \Delta^2\beta_s \\
C_k &= 1_{(k<U)} \sum_{t=k+2}^{U+1} \sum_{s=t}^{U+1} \Delta^2\alpha_s + 1_{(k>U+1)} \sum_{t=U+2}^{k} \sum_{s=U+2}^{t} \Delta^2\alpha_s
\end{aligned}$$

Here L is the offset mentioned in (2), §2.1 and U is a central reference point calculated as the integer value of $(L+3)/2$. The three parameters υ_o, υ_a, and υ_c define the linear plane. The first, υ_o, is an intercept that combines structural level effects, calculated as $\upsilon_o = \mu_{UU} = \alpha_U + \beta_{2U-1} + \gamma_U$. The remaining two are slope parameters in age and cohort directions respectively that combine single-differences of age, period, and cohort:

$$\begin{aligned}
\upsilon_a =& \mu_{U+1,U} - \mu_{UU} = (\alpha_{U+1} - \alpha_U) + (\beta_{2U} - \beta_{2U+1}) \\
\upsilon_c =& \mu_{U,U+1} - \mu_{UU} = (\gamma_{U+1} - \gamma_U) + (\beta_{2U} - \beta_{2U+1}).
\end{aligned}$$

Note the symmetry in these definitions and the corresponding symmetry in the definitions of A_i and C_k.

Upon inspection it can be seen that the above may be written as $X_h'\xi$ for ξ given in equation (9), §2.3. X_i will contain an intercept, the two slopes, and cumulations of each

double-difference. Further details of this reparametrization including insight regarding the choice of L and U can be found in Nielsen (2014).

B Properties of ad hoc identification schemes

We show that the model $D_\lambda\psi$, estimated under the ad hoc identification scheme $\alpha_1 = \alpha_2 = \beta_J = \gamma_K = 0$ in equation (21) of §5 can be expressed in the form $\xi = Q\phi$ in (20) by finding Q and ϕ. The proofs appeal to the analysis of Nielsen & Nielsen (2014), henceforth NN14.

Some notation from linear algebra is needed, specifically the orthogonal complement. A matrix m has full column rank if $m'm$ is invertible. In this case the orthogonal complement $m_\perp$ is a matrix so $m'_\perp m = 0$ and $(m, m_\perp)$ is invertible.

Write the constraint in equation (21) as $L'\theta_\lambda = 0$, where L is the $(q \times 4)$-matrix that selects the coordinates in (21) and the subscript λ indicates that θ_λ is a constrained version of θ. Note that $L'L = I_4$.

The design matrix for the ad hoc identified model is D_λ. This is found by dropping the columns DL from D. That is $D_\lambda = DL_\perp$, where $L_\perp$ is the selection matrix for the remaining columns with the properties that $L'_\perp L_\perp = I_p$ and $L'_\perp L = 0$. Thus, the ad hoc identification gives $\mu = DL_\perp\psi$ where $\psi = L'_\perp\theta$ are the remaining elements of θ. Since $\mu = DL_\perp\psi$ as well as $D = XA'$ we get $\mu = XA'L_\perp\psi$. At the same time we have $\mu = X\xi$ implying $\xi = A'L_\perp\psi$. Here $Q = A'L_\perp$ and if it is invertible we can choose $\phi = \psi = Q^{-1}\xi$.

To verify that $(A'L_\perp)$ is invertible, recall from §2.3 and NN14 that the canonical parametrization can be expressed in terms of a $q \times p$ matrix A so $D = XA'$ and $A'\theta = \xi$. We note that the constraint satisfies the property that (A, L) is invertible. This is proved by writing out A which is implicitly defined in §2.3 and then finding an orthogonal complement $A_\perp$. An explicit expression for $A_\perp$ is given in NN14, equation 58. Next, it is checked that $A'_\perp L$ is invertible. This implies that (A, L) is invertible, see NN14, Lemma A.1. Lemma A.1 also shows that $A'L_\perp$ is invertible.

The difference between the two approaches is that with the canonical parametrization we focus on the estimable p-vector ξ whereas the ad hoc identification considers the q-vector $\theta_\lambda = L_\perp\psi$, but only the p-vector $L_\perp\theta_\lambda = \psi$ can be determined from data.

C Data and Design

The alcohol categories are: rare = drinks less than once a week, casual = drinks one to four times per week, frequent = drinks five or more times per week. Note this does not account for quantity of drinks per drinking event.

Table 12: **Descriptive statistics, women** ($N = 43077$)

Continuous variables	*Minimum*	*Mean*	*Median*	*Maximum*	
Age	28	51	50	80	
Period	2001	2007	2006	2014	
Cohort	1926	1956	1957	1981	
BMI	13.2	27.37	26.35	58.94	
Height (cm)	123.6	161.67	161.6	202	
Weight (kg)	28.4	71.49	69	164	
Categorical variables					
Ethnicity	*Black*	*White*	*Asian*	*Mixed*	*Other*
	762	41071	702	288	254
NSSEC (3 level)	*Routine*	*Intermediate*	*Professional*	*Other*	
	4585	11721	14302	748	
Education level	*Below GCSE*	*GCSE*	*Some higher*	*Degree*	
	13138	11847	9644	8448	
Alcohol	*Never*	*Rare*	*Casual*	*Frequent*	
	497	15703	19534	7343	
Smoking	*Never*	*Former*	*Current*		
	23037	10724	9316		

Table 13: **Descriptive statistics, men** ($N = 38316$)

Continuous variables	*Minimum*	*Mean*	*Median*	*Maximum*	
Age	28	52	51	80	
Period	2001	2007	2006	2014	
Cohort	1926	1955	1956	1981	
BMI	13.63	27.93	27.44	59.45	
Height (cm)	138.2	174.9	174.8	203.1	
Weight (kg)	34.2	85.52	84	203.4	
Categorical variables					
Ethnicity	*Black*	*White*	*Asian*	*Mixed*	*Other*
	600	36363	972	201	180
NSSEC (3 level)	*Routine*	*Intermediate*	*Professional*	*Other*	
	6972	7390	16363	201	
Education level	*Below GCSE*	*GCSE*	*Some higher*	*Degree*	
	10927	7865	10456	9068	
Alcohol	*Never*	*Rare*	*Casual*	*Frequent*	
	228	8330	19489	10269	
Smoking	*Never*	*Former*	*Current*		
	16692	13084	8540		