# cemmap

## centre for microdata methods and practice

# The age-period cohort problem: set identification and point identification

**Martin Browning**
**Ian Crawford**
**Marike Knoef**

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

**An ESRC Research Centre**

# THE AGE-PERIOD-COHORT PROBLEM: SET IDENTIFICATION AND POINT IDENTIFICATION

Martin Browning
University of Oxford

Ian Crawford
University of Oxford & IFS

Marike Knoef
Leiden University & CentERdata

**Abstract**

"Only entropy comes easily" - Anton Chekhov

Various methods have been used to overcome the point identification problem inherent in the linear age-period-cohort model. This paper presents a set-identification result for the model and then considers the use of the maximum-entropy principle as a vehicle for achieving point identification. We present two substantive applications (US female mortality data and and UK female labor force participation) and compare the results from our approach to some of the solutions in the literature.

Stata code to estimate the model described in this paper is available from the authors.

## 1 Introduction

The age-period-cohort (henceforth *apc)* model is used widely in number of disciplines such as sociology, demography, economics and epidemiology. It aims to separate, for some outcome of interest, those influences associated with the process of aging, from those influences associated with the date at which subjects are observed, from those influences associated with a subject's date of birth. That the three effects can be conceptualized as distinct was argued at least as far back as Ryder (1965); different age groups within the population are at different stages of life relating to education, work and fertility; at various dates individuals are exposed to different events which have population-wide effects such as famines, wars, recessions and

epidemics; successive date-of-birth cohorts experience different histories, institutions and peer-group socialization. Thus, it is argued, age, period and cohort all have distinctive influences on individuals and groups of individuals. The *apc* model aims to disentangle these influences.

Unfortunately, the linear *apc* model suffers from a well-known fundamental identification problem, namely that there is a perfect linear relationship between these effects: period equals year of birth plus age. It is impossible therefore to observe independent variation in these conditioning variables and hence standard linear regression techniques cannot separate them. The perfect multicollinearity in these variables makes point identification of the separate effects impossible.

Undeterred, researchers have held to the view that these influences are both important and distinct and so a large literature going back to the 1970s has grown up around the problem of identifying *apc* models; see, for example, Mason et al. (1973), Glenn (1976), Fienberg and Mason (1979), Kupper et al. (1985)[1]. At heart the *apc* model poses an ill-posed inverse problem and so most solutions proceed by using assumptions or prior information to transform it into a well-posed problem which can be solved by standard statistical methods. In this vein a number of solutions have been put forward: one is to re-specify the model and to either make it non-linear or to estimate it in differences; another is to impose a parameter restriction of some kind (one is enough); a fourth is to replace the dummies which capture one of the effects with data which reflect a causal mechanism. Recently an alternative generic - context independent - solution has been suggested by Yang et al. (2004, 2007, 2008) who introduced the *intrinsic estimator*. This approach does not use regularizing identifying assumptions in the usual sense, but does impose some requirements on the geometric orientation of the parameter vector in the parameter space.

---

[1] Examples of applied work include developments in social changes such as marriage (Hernes 1972), trends in mortality (Case 1965; Collins 1982; Cayuela et al. 2004), developments in income and wealth (Kapteyn et al. 2005), (female) labor force participation (Contreras et al. 2005; Euwals et al. 2011; Balleer et al. 2009), wages (Meghir and Whitehouse 1996; Kalwij and Alessie 2007), and savings behavior (Attanasio 1998).

This paper suggests another generic approach to *apc* models. We first show that in cases in which the range of the dependent variable is bounded in the population (for example, when it measures a binary event) the model is partially identified in the sense of Manski (2003): the parameters of the model can be shown to be confined to a closed, convex set.

For some purposes it may be sufficient to work with the identified set. Often, however, we seek a point within that set; we propose using a maximum entropy estimator to point identify the *apc* parameters. This approach has its roots in statistical mechanics and later information theory[2]. It proceeds by representing uncertainty about some object of interest (in this case the coefficient vector of the *apc* model) in terms of a probability distribution (where probability is interpreted as a measure of a state of knowledge rather than a limiting frequency). It then focusses on this probability distribution and considers how to arrange this distribution in a way that best represents the current state of knowledge as revealed by, and consistent with, the data at hand. The criterion function used to select the probabilities is an entropy measure which Shannon (1948) showed was identified (up to a constant) by the requirements that any measure of uncertainty should be continuous, symmetric (with respect to reordering of the outcomes), maximized when all events are equiprobable and additively decomposable (so that the total amount of uncertainty in a process is independent of how the process might be regarded as being divided into parts). The entropy measure we use is, in fact, a special case of the more general cross-entropy criteria suggested by Good (1963) and Kullback (1959) which includes prior information on the probability distribution. Both the maximum cross-entropy estimator and the maximum entropy estimators are therefore examples of shrinkage estimators (see for example Stein (1981) and Denzau et al. (1989)). Maximum entropy estimation methods are now quite well established (see Golan et al. 1996) but their use in this context is, as far as we know, novel. As we show in this paper these

---

[2] Boltzman (1872), Jaynes (1957a, 1957b), Shannon (1948).

methods are straightforward to apply and produce results which compare plausibly with those derived through other methods.

The paper is organized as follows. Section 2 explains the identification problem. Section 3 describes the partial identification of the *apc* coefficients when the dependent variable has a known support in the population and presents the maximum entropy based approach. Section 4 gives a brief overview of some existing strategies designed to deal with the identification problem. Section 5 describes the information-based approach to point identification. Section 6 offers two empirical examples and section 7 concludes.

## 2 The Problem

Suppose that we have panel data (or quasi-panels) on a group of subjects. For person $h$ in period $t$ we observe age (at the end of the survey year), denoted $a_{ht}$, from which we can construct a year of birth (cohort) variable, $c_h = t - a_{ht}$. We also observe some variable of interest $Y_{ht}$. It is often of interest to describe the evolution of this variable by a decomposition into additive age, period and cohort (apc) components[3]. Since all the right hand side variables are discrete, we can take a 'nonparametric'/local averaging approach:

$$Y_{ht} = \sum_a \alpha_a d_{ht}^a + \sum_t \pi_t d_t + \sum_c \gamma_c d_{ht}^c + \varepsilon_{ht} \tag{1}$$

where $d_{ht}^a$ is a dummy variable that is set to one if person $h$ is aged $a$ at the end of survey year $t$; $d_h^c$ is set to one if person $h$ was born in year $c$ and $d_t$ is set to one if $Y_{ht}$ was recorded for person $h$ in period $t$. The sums are taken over all possible values of the three variables. This formulation brings out clearly that the additivity imposes quite strong restrictions on the description of the evolution of the variable of interest since all 'cross terms' (for example, $d_{ht}^a d_t$) are dumped into the residual term $\varepsilon_{ht}$.

---

[3] In this note we take this as given. The strength of an additive decomposition is that it is purely statistical (or mechanical); the weakness (apart from the non-identification) is that it does not 'explain' the sources of any effects it uncovers.

In our discussion of the identification of *apc* effects we shall consider a simple sampling scheme in which two cohorts $(c = 1, 2)$ are followed for three periods $(t = 3, 4, 5)$ so that age ranges from 1 to 4 $(a = 1, 2, 3, 4)$. We have:

$$Y_{ht} = \alpha_1 d_{ht}^{a=1} + \alpha_2 d_{ht}^{a=2} + \alpha_3 d_{ht}^{a=3} + \alpha_4 d_{ht}^{a=4} + \pi_3 d_3 + \pi_4 d_4 + \pi_5 d_5 + \gamma_1 d_h^{c=1} + \gamma_2 d_h^{c=2} + \varepsilon_{ht} \quad (2)$$

Two sets of identification problems arise. The first is that each of the sets of dummies sums to one. This is easily dealt with by, for example, setting the coefficients for the youngest age and the most recent cohort $(\alpha_1, \gamma_2)$ to zero and interpreting the rest of the effects relative to this normalization.

$$Y_{ht} = \alpha_2 d_{ht}^{a=2} + \alpha_3 d_{ht}^{a=3} + \alpha_4 d_{ht}^{a=4} + \pi_3 d_3 + \pi_4 d_4 + \pi_5 d_5 + \gamma_1 d_h^{c=1} + \varepsilon_{ht} \quad (3)$$

Thus, the coefficient on the first period gives the expected outcome for the reference age-period-cohort which are those individuals who were born in period 2, observed in year 3, at the age of 1. The second identification issue is much more serious. Even with the normalization adopted here, the linear relationship between age, period and cohort (period equals year of birth plus age) imposes a linear relationship on the dummy variables in equation (3). In the present context this relationship is:

$$-d_{ht}^{a=2} - 2d_{ht}^{a=3} - 3d_{ht}^{a=4} + d_4 + 2d_5 + d_{ht}^{c=1} = 0 \quad (4)$$

Solving (4) for $d_{ht}^{a=2}$ and substituting into (3) gives the reduced form

$$Y_{ht} = b_1 d_{ht}^{a=3} + b_2 d_{ht}^{a=4} + b_3 d_3 + b_4 d_4 + b_5 d_5 + b_6 d_{ht}^{b=1} + \varepsilon_{ht} \quad (5)$$

where the matrix of regressors has full column rank and so the parameters of (5) can be estimated by ordinary least squares. The relationship between the reduced form parameters

and the parameters of interest may be written as

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \end{bmatrix} = \begin{bmatrix} -2 & 1 & 0 & 0 & 0 & 0 & 0 \\ -3 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \pi_3 \\ \pi_4 \\ \pi_5 \\ \gamma_1 \end{bmatrix} \tag{6}$$

or, more succinctly:

$$\mathbf{b} = \mathbf{A}\boldsymbol{\beta} \tag{7}$$

This describes a linear simultaneous equation system with seven unknowns and six equations and serves to emphasis the ill-posed inverse problem at the heart of the *apc* model. Whilst we may estimate the six reduced form parameters $\mathbf{b} = [b_1, ..., b_6]'$ we are unable, as it stands, to recover from them the seven parameters of interest $\boldsymbol{\beta} = [\alpha_2, \alpha_3, \alpha_4, \pi_3, \pi_4, \pi_5, \gamma_1]'$. The system is under-determined: there are an infinity of solutions given by

$$\boldsymbol{\beta} = \mathbf{A}^+\mathbf{b} - \left(\mathbf{I}_7 - \mathbf{A}^+\mathbf{A}\right)\mathbf{q} \tag{8}$$

where $\mathbf{A}^+$ is the Moore-Penrose inverse of $\mathbf{A}$ and $\mathbf{q}$ is an arbitrary seven-vector. We have no constructive basis for selecting, from this uncountably infinite feasible set, a particular solution vector for $\boldsymbol{\beta}$.

## 3 Set identification

We now show that although the parameters $\boldsymbol{\beta}$ are not point identified, they are set identified if the support of $Y$ is bounded. The *apc* model is a fully saturated regression model in which all of the regressors are dummy variables and sums of regression coefficients correspond to conditional (cell) means. Set identification follows since if the outcome variable has a bounded range in the population then by the law of iterated expectations all of its conditional means are necessarily bounded. To see this, without loss of generality, consider again our simple sampling scheme in which two cohorts are followed for three periods so that age ranges from

1 to 4 and consider our equation of interest (3):

$$Y_{ht} = \alpha_2 d_{ht}^{a=2} + \alpha_3 d_{ht}^{a=3} + \alpha_4 d_{ht}^{a=4} + \pi_3 d_3 + \pi_4 d_4 + \pi_5 d_5 + \gamma_1 d_h^{c=1} + \varepsilon_{ht}$$

Then sums of the *apc* coefficients correspond to conditional expectations for all observable and unobservable/counterfactual combinations of age, period and cohorts as follows:

$$
\begin{bmatrix}
E\left(Y|a=1,p=3,c=2\right) \\
E\left(Y|a=2,p=3,c=2\right) \\
E\left(Y|a=3,p=3,c=2\right) \\
E\left(Y|a=4,p=3,c=2\right) \\
E\left(Y|a=1,p=3,c=1\right) \\
E\left(Y|a=2,p=3,c=1\right) \\
E\left(Y|a=3,p=3,c=1\right) \\
E\left(Y|a=4,p=3,c=1\right) \\
E\left(Y|a=1,p=4,c=2\right) \\
E\left(Y|a=2,p=4,c=2\right) \\
E\left(Y|a=3,p=4,c=2\right) \\
E\left(Y|a=4,p=4,c=2\right) \\
E\left(Y|a=1,p=4,c=1\right) \\
E\left(Y|a=2,p=4,c=1\right) \\
E\left(Y|a=3,p=4,c=1\right) \\
E\left(Y|a=4,p=4,c=1\right) \\
E\left(Y|a=1,p=5,c=2\right) \\
E\left(Y|a=2,p=5,c=2\right) \\
E\left(Y|a=3,p=5,c=2\right) \\
E\left(Y|a=4,p=5,c=2\right) \\
E\left(Y|a=1,p=5,c=1\right) \\
E\left(Y|a=2,p=5,c=1\right) \\
E\left(Y|a=3,p=5,c=1\right) \\
E\left(Y|a=4,p=5,c=1\right)
\end{bmatrix}
=
\begin{bmatrix}
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 & 0 & 1 \\
0 & 0 & 1 & 1 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 \\
1 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 1 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 1 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 0 & 0 & 1 & 1
\end{bmatrix}
\begin{bmatrix}
\alpha_2 \\
\alpha_3 \\
\alpha_4 \\
\pi_3 \\
\pi_4 \\
\pi_5 \\
\gamma_1
\end{bmatrix}
$$

which we will write more compactly as $\mathbf{y} = \mathbf{B}\boldsymbol{\beta}$. We first show set identification of $\boldsymbol{\beta}$ if the conditional expectations are between zero and unity; that is $\mathbf{y} \in [0,1]^{24}$. The identification in the more general case in which $Y$ is bounded follows using the obvious transformation.

**Lemma.** If $Y \in [0,1]$ then $\beta_k \in [-1,1]$ for all $k$.

**Proof.** We can write the equality restrictions $\mathbf{y} = \mathbf{B}\boldsymbol{\beta} \in [0,1]^{24}$ as the matrix inequality $\mathbf{D}\boldsymbol{\beta} \leq \mathbf{C}$ where $\mathbf{D} = [\mathbf{B}' : -\mathbf{B}']'$ and $\mathbf{C}=[\mathbf{1}'_{24} : \mathbf{0}'_{24}]'$. We can also write the range restriction relating to $\boldsymbol{\beta} \in [-1,1]^7$ as $\mathbf{E}\boldsymbol{\beta} \leq \mathbf{F}$ where $\mathbf{E} = [\mathbf{I}_7 : -\mathbf{I}_7]'$ and $\mathbf{F} = [\mathbf{1}_{14}]$. We need to show that $\mathbf{D}\,\boldsymbol{\beta} \leq \mathbf{C}$ implies $\mathbf{E}\,\boldsymbol{\beta} \leq \mathbf{F}$. The condition under which this is the case is given in Rockafellar (1970, p. 199, Theorem 22.3) and it is that there exists a real non-negative matrix $\boldsymbol{\lambda}$ such that $\mathbf{D}'\boldsymbol{\lambda} = \mathbf{E}'$ and $\mathbf{C}'\boldsymbol{\lambda} \leq \mathbf{F}'$.

The problem of deciding whether a suitable $\boldsymbol{\lambda}$ vector exists can be determined by running Phase 1 of the Simplex algorithm which will find, in a finite number of steps, a feasible vector iff such a vector exists. In the appendix we construct a suitable $\boldsymbol{\lambda}$ matrix which satisfies the condition. Consequently we have $\mathbf{D}\,\boldsymbol{\beta} \leq \mathbf{C}.\blacksquare$

For cases in which the dependent variable is bounded in the population $Y \in [Y_{\min}, Y_{\max}]$ then we can transform it such that $\tilde{Y} = (Y - Y_{\min})/(Y_{\max} - Y_{\min})$ so that $\tilde{Y} \in [0, 1]$ and apply the above arguments. Transforming back using

$$Y = Y_{\min} + (Y_{\max} - Y_{\min})\,\tilde{Y} \tag{9}$$

we have the result:

**Corrolary** If $Y \in [Y_{\min}, Y_{\max}]$ then $\beta_k \in [Y_{\min} - Y_{\max}, Y_{\max} - Y_{\min}]$ for all $k$.

This partial identification result shows that, when the dependent variable has a known closed support in the population, the *apc* coefficient vector $\boldsymbol{\beta}$ lies in a closed, convex parameter space given by the hypercube $\mathcal{B}$, centered at zero with each side of length $2\,[Y_{\max} - Y_{\min}]$.

# 4    Some point identifying assumptions

Various solutions to the *apc* problem have been suggested. Most of them use assumptions or prior information to transform the underlying ill-posed inverse problem into a well-posed problem which is amendable to solution via standard maximum likelihood methods. One such is to abandon the 'nonparametric' model and to parameterize one or more of the explanatory variables with, for example, a polynomial and impose that (at least) one of the linear effects is zero (e.g. Fitzenberger et al. 2004). Another solution, along the same lines, is to proxy one of the variables with something meaningful in the context. For example, if the outcome of interest is household consumption and the time effects are there to reflect common macroeconomic effects then it has been suggested to proxy them with a macroeconomic time series.[4]    One

---

[4] This approach is, among others, used by Heckman and Robb (1985), Kapteyn et al. (2005), Winship and Harding (2005), Portrait et al. (2010), Euwals et al. (2011).

can also assume away one of the factors (for example, Firebaugh and Davis 1988; Myers and Lee 1998; Van der Schors et al. 2007). This will allow the others to be recovered but is vastly over-sufficient, since we only need *one* restriction to achieve identification no matter how many years, ages and cohorts we have as long as the additional restriction is not a linear combination of those already embodied in (7). Dropping one of the effects altogether generally results in over-identification which gives testable restrictions. Since we only need one restriction to identify all the parameters this suggests that we may be able to get away with weaker assumptions, for example, that the effects of two adjacent ages are the same. If we use this to augment $\mathbf{A}$ and $\mathbf{b}$ then $\mathbf{A}^{-1}$ exists and the system can be solved uniquely. This commonly used method was introduced by Mason et al. (1973). Another approach which also just identifies is the suggestion of Hanoch and Honig (1985) and Deaton and Paxson (1994) which involves detrending such that the period effect is orthogonal to a trend and sums to zero. However, it is important to note that, as Deaton (1997, p. 126), points out

> "This procedure is dangerous when there are few surveys, where it is difficult to separate trends from transitory shocks.... Only when there are sufficient years for trend and cycle to be separated can we make the decomposition with any confidence."

This points to a general issue, not one which pertains only to the Deaton and Paxson normalization: credible identifying assumption must be justified in each new context - there is no universally credible assumption. Furthermore, coefficient estimates are very sensitive to the choice of the identifying constraint, even if one chooses only one of these (to give a just identified model); we illustrate this below. For identifying the *apc* problem, the main problem is that the *a priori* information needed for reasonable identifying constraints is scarce (Glenn, 1976).

An alternative approach due to Fu (2000) and Yang et al. (2004, 2008) suggests the use of the 'intrinsic estimator'. This corresponds to the first term in the general solution (8). The

argument for focussing on this component appears to be that the second term is arbitrary (due to the random vector $\mathbf{q}$) so should not play a role whereas the first term is a deterministic (intrinsic?) component of all possible solutions. This deterministic component provides a solution for $\boldsymbol{\beta}$ which has minimum Euclidean norm among all solutions - it is in this sense the smallest solution, and this may be a desirable property (see Yang et al. (2008) for a better explanation of the intrinsic estimator, its properties and arguments for focussing on the solution with $\mathbf{q} = \mathbf{0}$). An application of the intrinsic estimator can be found in Yang (2008).

Another approach is due to Kuang et al. (2008). They reiterate the point that $\boldsymbol{\beta}$ cannot, in general, be point identified from the data alone and suggest, entirely sensibly, that researchers focus instead on those features of the model which *can* be identified from the data. They come up with a reparameterization of the model which allows the effects of second differences to be estimated and the model to be used for forecasting.

Our approach, whilst different from Kuang et al. (2008a,b), is in sympathy with their general philosophy: there is no way to point-identify the *apc* model, yet the data do tell us *something*, so we should therefore try to find ways of making best use of the information in the data.

# 5    An Information Based Approach

We propose using the maximum-entropy principle to address the problem. We stress that this is not really a *solution* to the *apc* problem. In our view the problem has no unique solution - there is not enough information in the data to provide one. Whilst the data do convey a certain amount of information about the *apc* decomposition, over and above this we must remain uncertain as to the precise solution. Rather than trying to solve the point identification problem directly, the maximum entropy principle[5] provides a framework within which we can

---

[5] For a full exposition of the maximum-entropy approach see Mittelhammer et al. (2000).

formalize this uncertainty.

The essential idea is due to Jaynes (1957a, 1957b) who suggested that one should reparameterize the object of interest (in this case the *apc* coefficient vector $\boldsymbol{\beta}$) in terms of a probability distribution over the set of possible solutions, and then select the probability distribution which reflects the level of one's uncertainty given the information available in the data. In order to do this, the method uses a measure of information entropy due to Shannon (1948) which is a special case of the Kullback-Leibler[6] distance. By choosing to use the probability distribution over the set of possible outcomes which has the maximum entropy allowed by the data, the argument goes, we are choosing the most uninformative distribution possible. To choose a distribution with lower entropy would be to assume information which we do not possess. To choose a distribution with higher entropy would violate the constraints provided by the information which we do possess from the data. Thus the maximum entropy distribution over the possible solutions best represents the current state of knowledge. That, at least, is the argument. Various authors[7] have attempted to show that choosing probabilities in order to maximize entropy (as measured by this class of functions) is the uniquely correct method of making inferences which satisfy the information in the data. Whether or not they have been entirely successful is an open question (see, for example, Uffink 1995). In our view the maximum entropy principle is plausible, but not completely compelling. Justifying the principle is, however, not the object of this paper. The purpose of this paper is to illustrate its application to a long standing and thorny inverse problem which has long troubled sociologists, demographers, epidemiologists and, occasionally, economists.

There is, however one important limitation: the maximum entropy approach is best applied in situations in which the set of possible solutions is bounded and this is not the case with

---

[6] Kullback and Leibler (1951), Kullback (1959).

[7] For example, Shore and Johnson (1980), Tikochinsky et al. (1984), Skilling (1989), Paris and Vencovska (1990), Csiszár (1991).

the *apc* problem in general. This limits the range of applications of the approach to those where we have a partial identification result like the one described in the previous section. Nonetheless, in many demographic, economic and sociological applications where the *apc* decomposition is of interest the outcome variable is naturally bounded and in these cases maximum entropy methods are very easy to apply. We therefore restrict our attention to this class of problems and we will describe the application of maximum entropy methods to the *apc* problem. After deriving the maximum entropy estimate we illustrate using two examples in which the dependent variable is suitably bounded: a demographic model of US mortality rates and a socioeconomic model of female labor force participation in the UK.

We begin by representing the identified set given by the hypercube $\mathcal{B}$ by:

$$\mathcal{B} = \{\boldsymbol{\beta}|\boldsymbol{\beta} = \mathbf{S}\mathbf{p}\} \tag{10}$$

where $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_J]$, the $\mathbf{s}_j$ are vectors representing the vertices of the cube and $\mathbf{p}$ is a vector of non-negative weights which sum to one and which are used to form all of the convex combinations of the vertices. Using this reparameterization means that $\boldsymbol{\beta}$ can be expressed as a convex combination of the extreme points of $\mathcal{B}$ we can insert (10) into (7) and rewrite it as

$$\mathbf{b} = \mathbf{A}\mathbf{S}\mathbf{p} \tag{11}$$

The problem of solving (7) for $\boldsymbol{\beta}$, is now reformulated as the problem of solving (11) for $\mathbf{p}$

Of course, whilst we have reparameterized the problem in (11), we have got no closer to solving it as the system is still under-determined (since $J > 6$). However, we are now able to reinterpret the problem in a way which allows us to use maximum entropy methods. Since the $p_j$'s have all of the necessary characteristics of probabilities (in this context probability is interpreted as a measure of a state of knowledge rather than a limiting frequency), we can treat the $\mathbf{p}$ vector as a discrete probability distribution over the $J$ multivariate outcomes represented by the columns of the matrix $\mathbf{S}$. In other words, the solution of (11) requires us

to supply a constructive principle for choosing one *probability distribution* over another rather than choosing one *parameter vector* over another. We therefore need a principle which allows us to claim that one probability distribution is "better" in some respect or other than another probability distribution.

The principle which we adopt was suggested by Jaynes (1957a, 1957b) and is an extension of LaPlace's *principle of insufficient reason*[8] to situations in which some information about the problem is available from data. The idea is that one should choose a distribution which does not unduly favor one outcome over another subject to the requirements that the probabilities are non-negative, sum to one and satisfy any data-based restrictions one might have (in this case the reduced form estimates). In the present context this means that the probability distribution must satisfy equation (11). A natural objective function which will achieve this is the entropy function

$$H\left(\mathbf{p}\right) = -\mathbf{p}' \ln \mathbf{p} \tag{12}$$

where in the case of elements of $\mathbf{p}$ with zero probability, $0 \ln\left(0\right) \equiv 0$. The function (12) was suggested by Shannon (1948) as a measure of uncertainty[9]. It is maximized when the probabilities are uniform (all outcomes are equally likely which is interpreted as being maximally uncertain), and it is minimized if the probability distribution is degenerate on a particular outcome (interpreted as perfect certainty about that outcome). The constrained optimization problem is therefore the maximum entropy problem:

$$\max_{\mathbf{p} \geq \mathbf{0}, \ \mathbf{1}'\mathbf{p}=1} -\mathbf{p}' \ln \mathbf{p} \text{ subject to } \mathbf{b} = \mathbf{ASp} \tag{13}$$

This is a straightforward nonlinear optimization problem with a unique solution. Because the $\mathbf{p}^*$ which solves this problem is interpretable as a vector of probabilities over support points,

---

[8]That, unless there is a reason for believing otherwise, each possible outcome should be regarded as equally likely. The PIR has, justifiably, had a bad press. The maximum entropy principle avoids many of the pitfall which beset the PIR. See Uffink (1995) for a discussion.

[9]The base of the logarithm used only affects the units. If $\log_2$ is used then the information uncertainty is measured in "bits", if the natural log is used it is measures in "nats".

and the expectation operator for a discrete random variable is a probability-weighted convex combination of support points the *apc* coefficients $\boldsymbol{\beta}^* = \mathbf{S}\mathbf{p}^*$ satisfy the data constraints in (7) exactly and can be interpreted as the expected value of a discrete multidimensional random variable consistent with the entropy maximizing choice of underlying probability distribution.

If the researcher has non-sample or pre-sample prior information on the probability distribution of the coefficient vector represented by the vector $\mathbf{q}$, then the objective function can be reformulated to minimize the probabilistic divergence between the probabilities which are consistent with the data $\mathbf{p}$, and the prior probabilities $\mathbf{q}$ (see Kullback and Leibler 1951; Kullback 1959; Good 1963). This modifies the objective function which becomes the Kullback-Leibler/cross-entropy measure

$$-\mathbf{p}'\ln\mathbf{p} - \mathbf{p}'\ln\mathbf{q} \tag{14}$$

The cross-entropy (14) can be interpreted as a measure of the new information on the coefficients provided by the data relative to the prior distribution. When the prior is uniform over the hypercube then $q_j = 1/J$ and (14) measures the additional information reflected in $\mathbf{p}$ relative to the maximally uninformative distribution and the resulting maximum cross-entropy solution is also then the maximum entropy solution. We also note that *a priori* restrictions (e.g. that certain effects may be monotone) can be included as additional constraints to the maximization equation (13) - linear, nonlinear, equality or inequality constraints can all be imposed easily.

To summarize: when the outcome variable is bounded in the population the *apc* problem can be reformulated in terms of an unknown probability distribution over a convex set. We use the maximum entropy principle to select a probability distribution which is as flat as possible over the set, subject to the informational constraints provided by the reduced form estimates and the requirement that the probabilities are non-negative and sum to one. We can then recover the expected values of the parameters of interest consistent with the maximum

entropy probability distribution. By construction, these values satisfy the data constraints (7) exactly whilst at the same time the associated probability distribution reflects, in a precise sense, the uncertainty which remains after the estimation of the reduced form.

## 6 Empirical Examples

In this section we present two applications. The first concerns an analysis of mortality data for females in the US, the second looks at labor force participation by women in the UK. It is important to understand that it is not possible to argue that the maximum entropy results correctly capture *the* true values of the *apc* coefficients. After all, the whole point of the *apc* problem is that a unique solution, true or not, *does not exist*. We can, and do, however argue that the maximum entropy approach represents a coherent and logically-consistent approach to the ill-posed inverse problem and we can compare the results it produces to those alternative solutions which the literature has suggested. In each case (with the exception of the intrinsic estimator) the existing solutions make an additional assumption which renders the problem well-posed and amenable to unique solution by standard statistical methods. The identifying assumption in each case might be somewhat arbitrary, but greater confidence about the resulting *apc* profiles is gained if the results from each approach exhibit common, plausible features (the claim would be that the identifying assumption, though arbitrary, "does not seem to matter much"). Similarly, the empirical plausibility of the maximum entropy partially depends on whether they also capture some of the same features which emerge from existing methods - the approach may be logically impeccable, but if the results are wildly dissimilar from those produced by standard methods with which researchers are more familiar and comfortable then it is unlikely to catch on.

## 6.1  US Female Mortality

For this first example we use mortality data for U.S. females between 1933 and 2007.[10]  The data contain cohorts born between 1823 and 2007. Figure 1 shows the raw data - in each panel we have plotted every tenth age/period/cohort so that the features of the data are easier to pick out.

As the figure shows, mortality rates increase with age (actually, the change between two ages in the bottom left figure are age-period effects, since age, period, and cohort effects cannot be identified in the figure). Furthermore, in general younger cohorts seem to have lower mortality rates than older cohorts. Typically the interpretation is that age effects relate to the biological process of aging, whereas period effects contain historical events such as wars, famines, and infectious diseases (Yang 2008). Improvements in the medical technology affect mortality and may also be regarded as period effects as far as these medical breakthroughs reduce mortality rates in all age groups. Cohort effects may reflect early life conditions that influence mortality later in life, such as the presence of a famine in early childhood, and different levels of hygiene. In addition, different cohorts accumulate different environmental and socioeconomic experiences during their life that may affect their mortality later in life.

We begin with the results which are produced by normalizing, in turn, the age, period and cohort effects to zero (figure 2)[11]. Normalizing age effects to zero is clearly a bizarre assumption in this context and it results in some implausible period effects where the mortality rates are increasing year-on-year. Nonetheless, and partly as a result of the period effects having this pattern, the cohort effects do not seem ridiculous with younger cohorts enjoying successively lower mortality rates. Relaxing the restriction on age effects (either by imposing it on the period or the cohort effects instead) results in much more plausible age effects which rise with

---

[10]These data can be downloaded from the Berkeley Human Mortality Database: www.mortality.org.

[11]We have omitted confidence intervals from the profile figures so that the patterns are easier to see. Full results are available from the authors.

age and which also show an increase at zero years consistent with elevated infant mortality rates relative to those of young children. Setting the period effects to zero effectively removes the possibility of population-wide effects related to common exposure to epidemics, medical improvements etc. It results in a patterns of cohort effects which indicate worsening mortality rates for cohorts born up until the mid 1800's in the US with improvements thereafter. This may actually be plausible as the mid 1800's were a bad time to be in the US and an especially bad time to be in the US and young - there were epidemics of cholera in North America in 1848-9, yellow fever in the US in 1850, influenza in North America in 1850-1 and further more localized epidemics of cholera and yellow fever in 1851-2 in population centres in Illinois, the Great Plains, Missouri and New Orleans. Removing these cohort effects by assumption produces the period effects seen in the final set of diagrams in figure 1. This shows a noisy but declining trend. Doubtless there were improvements in public health during the period covered (1933-2007) but Haines (2008), for example, argues that the most significant improvements in public health and sanitation and in particular the increased availability of clean water supplies and effective sewage treatments took place in the late 1800's and therefore pre-date this period. These major improvements may have influenced earlier date of birth cohorts but since such effects are excluded by assumption it is possible that the estimated effect on the period profile exaggerates this improving trend. To summarize: none of these three normalizations seem quite right. Zeroing out age effects is clearly inappropriate but dropping cohort effects when we know from the historical record that certain birth cohorts were exposed to epidemics of particularly nasty diseases in infancy seems implausible as well, whilst removing period effects also removes a role for the public health improvements which may have occurred over the period.

Normalizing a pair of adjacent cohort effects to be the same is a standard assumption in the *apc* analysis of mortality data (see Mason et al. 1973). On the face of it seems reasonable -

17

why, after all, should the mortality rates of those born a year apart be markedly different? The difficulty arises in deciding exactly which cohort-pairs to normalize. The effects of different choices (all equally plausible but none especially compelling) are illustrated in figure 3. The profiles (even for the age effect which is the most robust relatively) are all over the place: the period effects could be increasing, decreasing or approximately flat according to taste.

Figure 4 reports the results using the Hanoch and Honig/Deaton and Paxson normalization in which the period effects are constructed to be orthogonal to a linear trend. This makes the period effects cyclical and this probably accounts for this normalization's popularity with economists where time effects reflect common, population-wide macroeconomic effects - for example, "the business cycle". As noted above (Deaton 1997, p.126) this approach only works when the data cover a long enough span to include a full cycle. The cyclical period effects seen in figure 3 are therefore a necessary consequence of the normalization. In the present context it may not be appropriate to assume that period effects on female mortality in the US are essentially flat over the period - some improvement is to be expected. Nonetheless the age and cohort effects which the normalization produces do seem to be plausible - age effects are increasing after the subject survives infancy and cohort effects seem to fit what we know from the historic record.

The results from the intrinsic estimator of Yang et al. (2004) are shown in figure 5. Recall that this estimator uses the Moore-Penrose inverse in equation (8) with the random vector $\mathbf{q}$ set to zero. This estimator therefore does not rely on a normalization to regularize the *apc* model's inverse problem. Neither, therefore, does it force any particular properties on the age, period and cohort profiles which emerge. The profiles which are recovered by this method are interesting. The age profile shows what we, by now, take to be the robust pattern with a drop in mortality after the first year of life followed by year-on-year increases. The cohort profile also shows the oft-observed pattern wherein mortality rates peak in the mid-1800's.

The period effects look like they are following a cyclical, but slightly rising trend over the period. However the increase is not statistically significant and the vertical scale is much expanded so the period effects are really just small fluctuations around a flat profile - very similar to the Deaton and Paxson normalization.

Figure 6 plots the maximum entropy estimates. The age effects are not quite monotonic as they pick up the raised mortality rates in the first year of life, but thereafter mortality rates increase with each year of life. The period effects reflect the influences on mortality rates which bear equally on all ages and all data of birth cohorts. As can be seen, these effects show a certain amount of year-to-year variability but are essentially flat. These effects are probably what intuition would expect and seem to mirror what we see in the raw data when we plot mortality rates by age against both cohort and period. The cohort effects are rather interesting and echo what we have seen before: they too show the strong pattern in which mortality rates first rise with date of birth (so that older cohorts have better mortality rates) until around the 1850 cohort falling thereafter.

We see that the maximum entropy results mirror many of the features which crop up in other normalizations. In particular we see that the maximum entropy estimates are qualitatively most similar to the intrinsic estimator and the Deaton and Paxson normalization - the reason why the Deaton and Paxson method gives similar results may be that the identifying assumption that period effects are mean-zero fluctuations around a straight line is, if the maximum entropy results are to be believed, approximately correct for these data. As other studies routinely find; the intuitively fairly weak requirement that the cohort effects for two adjacent cohorts are equal can have profound influences on the estimated effects.

## 6.2 UK Female Labor Force Participation

Our second example is a model of female labor force participation based on UK data. These data are from the UK Family Expenditure Survey which is a long running cross sectional dataset. The data record the labor market status of women in the surveyed households over the 34 year period from 1974 to 2007. The raw data (once more, for a subset of the age/period/cohort groups in the interests of visibility) are plotted in figure 7. The sub-panel showing the labor force participation rates of different cohorts over the life-cycle (the top left figure) exhibits vertical differences between the lines measuring 'cohort-time' effects. We use this terminology to emphasize that it is not possible to disentangle age from cohort and time effects in this figure. The raw data show a characteristic hump-shaped pattern in the labor force participation over the life-cycle, with a marked drop at the retirement age across birth cohorts.

In the context of the female labor force participation, age effects may include life-cycle decisions such as the timing of education, children, and retirement. Period effects may include business cycle effects or policy changes that effect the female labor force participation. Finally, cohort effects may include the improved educational attainment and lower fertility rates of younger cohorts, and changed social norms.

Figure 8 shows the profiles produced by setting age, period and cohort effects to zero in turn. Once again omitting the age effects, this time in a model of labor force participation, seems to be the wrong thing to do. Once age effects are admitted to the model they display the expected "hump shape" where participation rates rise in early adulthood, level off in middle age and drop abruptly at retirement age (until recently the state retirement age was 60 for women in the UK). Close inspection of the age profiles in the second two rows might even suggest a hint of a levelling-off in the growth of participation rates in the late twenties consistent with women leaving work to have children before re-entering the labor market at a

20

slightly lower rate than before.

Omitting age effects produces estimates of period effects which decline over time whilst the cohort effects seem to take on the hump shape which one would expect of age effects. Note that the estimates for the youngest cohorts are unstable due to small numbers of observations for these cohorts. Omitting period effects whilst allowing for a plausible age profile seems to give declining cohort effects which is hard to reconcile with recent labor market history in the UK where women in younger date-of-birth cohorts are generally more likely to work (at least part time) rather than less likely. Again, the strategy of omitting either age, period or cohort effects entirely does not seem in each case to produce estimates which are entirely plausible

The next set of results reports some experiments with setting two adjacent age effects to be equal (figure 9). That two age groups separated by one year should have approximately the same rate of labor force participation seems plausible, but we are agnostic about which two ages to choose. We therefore have chosen three alternative normalizations at regular intervals. Interestingly the period and cohort effects are qualitatively similar in each case. The age effects are different however and only the 39=40 normalization appears to pick up the hump shape we would expect to see - possibly because the normalization in this case is more appropriate. In the other two normalizations, the retirement effect is only just discernible and the overall pattern of age effects is far from what we might reasonably expect. Furthermore the estimated effects in these two cases lie well outside the partial-identification bounds implying that counterfactual participation rates for some combinations of age, period and cohort far exceeds 100% in some cases and is much less than 0% in others.

Figure 10 shows the profiles which are derived through the Hanoch and Hoing/Deaton and Paxson normalization. The hump-shaped age profile is evident with a steep drop in participation around the state retirement age. The period effects are normalized to sum to zero and to be orthogonal to a linear trend. This results in a noisy cycle rather then the

declining period effects which all of the other normalizations have produced. This may be evidence that, in this context, the cyclicality assumption is not appropriate. The cohort effects seem to take on the declining pattern of the (omitted) period effects. Age effects aside, these profiles are qualitatively different from most of those produced by other normalizations - the most similar profiles are those of the "no period effects" model and this is, of course, because the Hanoch and Hoing/Deaton and Paxson normalization is an approximate version of the same idea.

The intrinsic estimator in figure 11 once again produces estimates which seem entirely plausible. The age profile reflects what we now, through weight of evidence from the normalization-dependent results, take as robust pattern. The cohort effects rise for successive date of birth cohorts and the period effects show the declining pattern displayed by every normalization which is able to allow for a profile which isn't either flat or approximately so.

The Maximum Entropy estimates are illustrated in figure 12. They show a hump shape in the age effects, and a retirement effect that is clearly present. The period effect shows a mild decline over time. The cohort effect (excluding the final few cohorts which are afflicted by small sample-sizes) shows a clear pattern in which women in younger cohorts participate more often in the labor market. These results seem to display a set of features which are common to the most plausible results derived from other methods based on normalizations and are extremely close, by not identical to, those from the intrinsic estimator.

## 7    Conclusions

We present a new approach to the *apc* problem based on a maximum entropy method which can be applied when the dependent variable has a finite support in the population. Whilst the approach would require further development before it can be applied to unbounded outcome data (see Golan et al. (1996) for a discussion of maximum entropy methods for outcomes with

unbounded support) it does, as it stands, cover a great many cases of applied interest. We also note that the method can be extended easily to include a prior probability distribution on the parameter values (where such a prior is available to the researcher) by using the cross-entropy criteria and further equality and inequality constraints on the moment conditions can be built into the estimation procedure very easily. We also provide two illustrations of the method applied to real data. These applications to female mortality in the US and female labor force participation in the UK are of substantive interest in themselves. Nonetheless our aim was principally to see whether or not the maximum entropy method could produce results in which researchers could be confident - arguments in favor of the theoretical coherence may or may not be successful but it seems to us to be essential that the estimator actual "works". To do this we compared the results from the maximum entropy approach to those produced by a number of methods based on various normalization as well as the intrinsic estimator of Yang et al. (2004, 2007, 2008). In both cases we found that the maximum entropy-based method reproduced patterns in the age, period and cohort profiles which agreed with the most plausible normalizations. It also agreed closely with the results from the intrinsic estimator - which does not rely on a normalization in the usual sense - and therefore perhaps lends further support to that particular method (and *vice versa*). We conclude that maximum entropy methods provide a coherent and useful approach to the *apc* problem.

$$\lambda = \frac{1}{12}
\begin{bmatrix}
0 & 0 & 0 & 12 & 0 & 0 & 0 & 2 & 2 & 2 & 0 & 0 & 0 & 1 \\
2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 2 & 2 & 0 & 0 & 0 & 0 \\
2 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 12 & 0 & 0 & 2 & 2 & 2 & 0 & 0 & 0 & 1 \\
2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 2 & 2 & 0 & 0 & 0 & 0 \\
2 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 12 & 0 & 2 & 2 & 2 & 0 & 0 & 0 & 1 \\
2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 2 & 2 & 0 & 0 & 0 & 0 \\
2 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
2 & 2 & 2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 12 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\
2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 1 \\
2 & 2 & 2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 12 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\
2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 1 \\
2 & 2 & 2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 12 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\
2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 1
\end{bmatrix}$$

# References

[1] Attanasio, Orazio. 1998. "Cohort Analysis of Saving Behavior by U.S. Households." *The Journal of Human Resources* 33(3): 575–609

[2] Balleer, Almut, Ramon Gomez-Salvador, and Jarkko Turunen. 2009. "Labour Force Participation in the Euro Area - A Cohort Based Analysis." Technical report 1049. *European Central Bank working paper series.*

[3] Boltzmann, Ludwig. 1872. "Further Studies on the Thermal Equilibrium of Gas Molecules." *Wiener Berichte* 66: 275–370

[4] Case, Robert Alfred Martin. 1956. "Cohort Analysis of Mortality Rates as an Historical or Narrative Technique." *Brit J Prev Soc Med* 10(4): 159–171.

[5] Cayuela, Aurelio, Susana Rodriquez-Dominquez, M. Ruiz-Borrego, and M. Gili. 2004. "Age-Period-Cohort Analysis of Breast Cancer Mortality Rates in Andalucia (Spain)." *Annals of Oncology* 15: 686–688

[6] Collins, James J. 1982. "The Contribution of Medical Measures to the Decline of Mortality from Respiratory Tuberculosis: An Age-Period-Cohort Model." *Demography* 19: 409–427

[7] Contreras, Dante, Esteban Puentes, and David Bravo. 2005. "Female Labour Force Participation in Greater Santiago, Chile: 1957-1997. A Synthetic Cohort Analysis." *Journal of International Development* 17: 169–186

[8] Csiszar, Imre. 1991. "Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems" *Annals of Statistics* 19(4): 2032–2066

[9] Deaton, Angus. 1997. *The Analysis of Household Surveys*, Johns Hopkins University Press.

[10] Deaton, Angus, and Christina Paxson. 1994. "Saving, Growth, and Aging in Taiwan," in D. Wise (ed) *Studies in the Economics of Aging*, pp. 331–362 National Bureau of Economic Research.

[11] Denzau, Arthur T., Patrick C. Gibbons, and Edward Greenberg. 1989. "Bayesian Estimation of Proportions with a Cross-Entropy Prior." *Communications in Statistics - Theory and Methods* 18(5): 1843–1861

[12] Euwals, Rob, Marike Knoef, and Daniel van Vuuren. 2011. "The Trend in Female Labour Force Participation: What can be Expected for the Future?" *Empirical Economics* 40(3): 729–753

[13] Fienberg, Stephen E., and William M. Mason. 1979. "Identification and Estimation of Age-Period-Cohort Models in the Analysis of Discrete Archival Data." *Sociological methodology* 10: 1–67

[14] Firebaugh, Glenn, and Kenneth E. Davis. 1988. "Trends in Antiblack Prejudice, 1972-1984: Region and Cohort Effects." *The American Journal of Sociology* 94(2): 251–272

[15] Fitzenberger, Bernd, Reinhold Schnabel, and Gaby Wunderlich. 2004. "The Gender Gap in Labor Market Participation and Employment: A Cohort Analysis for West Germany." *Journal of Population Economics* 17(1): 83–116

[16] Fu, Wenjiang J. 2000. "Ridge Estimator in Singular Design with Application to Age-Period-Cohort Analysis of Disease Rates." *Communications in Statistics–Theory and Method* 29(2): 263–278

[17] Glenn, Norval D. 1976. "Cohort Analysts' Futile Quest: Statistical Attempts to Separate Age, Period, and Cohort Effects." *American Sociological Review* 41: 900–904.

[18] Glenn, Norval D. 1989. "A Caution about Mechanical Solutions to the Identification Problem in Cohort Analysis: Comment on Sasaki and Suzuki." *American Journal of Sociology* 95: 754–761

[19] Golan, Amos, George Judge, and Jeffrey M. Perloff. 1996. "A Maximum Entropy Approach to Recovering Information from Multinomial Response Data." *Journal of the American Statistical Association* 91(434): 841–853

[20] Good, I.J. 1963. "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables." *Annals of Mathematical Statistics* 34(3): 911—934

[21] Haines, Michael. 2008. "Fertility and Mortality in the United States." EH.Net Encyclopedia, edited by Robert Whaples. March 19.

[22] Hanoch, Giora, and Marjorie Honig. 1985. "'True' Age Profiles of Earnings: Adjusting for Censoring and for Period and Cohort Effects." *Review of Economics and Statistics* 67(3): 383–94

[23] Harding, David J. 2009. "Recent Advances in Age-Period-Cohort Analysis. A Commentary on Dregan and Armstrong, and on Reither, Hauser and Yang." *Social Science and Medicine* 69: 1449–1451.

[24] Harding, David J., and Christopher Jencks. 2003. "Changing Attitudes Toward Premarital Sex: Cohort, Period, and Aging Effects." *Public Opinion Quarterly* 67(2): 211–226

[25] Hastings, Donald W., and J. Gregory Robinson. 1974. "Incidence of Childlessness for United States Women, Cohorts born 1891-1945." *Social Biology* 21: 178–184

[26] Heckman, James, and Richard Robb. 1985. "Using Longitudinal Data to estimate Age, Period and Cohort Effects in Earnings Equations." In Mason, William M. and Stephen E.

Fienberg: Cohort Analysis in Social Research. Beyond the Identification Problem. New York: Springer–Verlag 137–150

[27] Hernes, Gudmund. 1972. "The Process of Entry into First Marriage." *American Sociological Review* 37: 173–182

[28] Honig, Marjorie, and Giora Hanoch. 1985. "Partial Retirement as a Separate Mode of Retirement Behavior." *Journal of Human Resources* 20(1): 21–46

[29] Jaynes, Edwin Thompson. 1957a. "Information Theory and Statistical Mechanics." *Physical Review* 106–620

[30] Jaynes, Edwin Thompson 1957b. "Information Theory and Statistical Mechanics II" *Physical Review* 108–171

[31] Kalwij, Adriaan, and Rob Alessie. 2007. "Permanent and Transitory Wages of British Men, 1975-2001: Year, Age and Cohort Effects." *Journal of applied econometrics* 22(6): 1063–1093.

[32] Kapteyn, Arie, Rob Alessie, and Annamaria Lusardi. 2005. "Explaining the Wealth Holdings of Different Cohorts: Productivity Growth and Social Security." *European Economic Review* 49: 1361–1391

[33] Knoef, Marike, Rob Alessie, and Adriaan Kalwij. 2009. "Changes in the Income Distribution of the Dutch Elderly over the Years 1989-2020: a Microsimulation." *Netspar Discussion Paper,* 09/2009-0302009

[34] Kuang, Di, Bent Nielsen, and Jens Perch Nielsen. 2008. "Forecasting with the Age-Period-Cohort Model and the extended Chain-Ladder Model." *Biometrika* 95(4): 987—991

[35] Kullback, Solomon. 1959. *Information Theory and Statistics.* New York: John Wiley & Sons.

[36] Kullback, Solomon, and Richard A. Leibler. 1951. "On Information and Sufficiency." *The Annals of Mathematical Statistics* 22(1): 79–86

[37] Kupper, Lawrence L., Joseph M. Janis, Azza Karmous, and Bernard G. Greenberg. 1985. "Statistical Age-Period-Cohort Analysis: A Review and Critique." *Journal of Chronic Diseases* 38: 811–830

[38] Manski, Charles F. 2003. "Identification Problems in the Social Sciences and Everyday Life." *Southern Economic Journal* 70(1): 11–21

[39] Mason, Karen Oppenheim, William M. Mason, Halliman H. Winsborough, and W. Kenneth Poole. 1973. "Some Methodological Issues in Cohort Analysis of Archival Data." *American Sociological Review* 38: 242–258

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

[41] Mittelhammer, Ron C., George G. Judge, and Douglas Miller. 2000. *Econometric foundations, Volume 1.* Cambridge University Press.

[42] Myers, Dowell, and Seong Woo Lee. 1998. "Immigrant Trajectories into Homeownership: A Temporal Analysis of Residential Assimilation." *International Migration Review* 32(3): 593–625

[43] Nakamura, Takashi. 1986. "Bayesian Cohort Models for General Cohort Table Analyses." *Annals of the institute of Statistical Mathematics* 38: 353–370

[44] O'Brien, Robert M. 2000. Age Period Cohort Characteristic Models. *Social Science Research* 29: 123–139

[45] O'Brien, Robert M., and Jean Stockard. 2002. "Variations in Age-Specific Homicide Death Rates: a Cohort Explanation for Changes in the Age Distribution of Homicide Deaths." *Social Science Research* 31: 124–150

[46] Paris, Jeff B., and Alena Vencovska. 1990. "A Note on the Inevitability of Maximum Entropy." *International Journal of Approximate Reasoning* 4(3): 183—223

[47] Portrait, France, Rob Alessie, and Dorly Deeg. 2010. "Do Early Life and Contemporaneous Macroconditions Explain Health at Older Ages?". *Journal of Population Economics* 23(2): 617–642

[48] Rockafellar, R. Tyrrell. 1970. *Convex analysis.* Princeton NJ: Princeton University Press.

[49] Ryder, Norman B. 1965. The Cohort as a Concept in the Study of Social Change. *American Sociological Review* 30: 843–861

[50] Sasaki, Masamichi S., and Tatsuzo Suzuki. 1987. "Changes in Religious Commitment in the United States, Holland, and Japan." *American Journal of Sociology* 92: 1055–1076

[51] Sasaki, Masamichi S., and Tatsuzo Suzuki. 1989. "A Caution about the Data to be used for Cohort Analysis: Reply to Glenn." *American Journal of Sociology* 95: 761–765

[52] Shannon, Claude E. 1948. "A Mathematical Theory of Communication". *Bell System Technical Journal* 27: 379–423

[53] Shore, John E., and Rodney W. Johnson. 1980. "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy." *IEEE Transactions on Information Theory* 26(1): 26–37

[54] Skilling, John. 1989. "Classic Maximum Entropy." In: *Maximum Entropy and Bayesian Methods.* J. Skilling, editor. Kluwer Academic, Norwell, MA. 45–52.

[55] Smith, Herbert L. 2004. "Response: Cohort Analysis Redux." *Sociological Methodology* 34: 111–119

[56] Stein, Charles M. 1981. "Estimation of the Mean of a Multivariate Normal Distribution." *Annals of Statistics* 9(6): 1135—1151

[57] Tikochinsky, Y., Naftali Tishby and Raphael David Levine. "Alternative Approach to Maximum-Entropy Inference." *Physical Review A* 30(5): 2638-2644

[58] Uffink, Jos. 1995. "Can the Maximum Entropy Principle Be Explained as a Consistency Requirement?" *Studies in History and Philosophy of Science Part B* 26(3): 223-261.

[59] Van der Schors, Anna, Rob Alessie, and Mauro Mastrogiacomo. 2007. "Home and Mortgage Ownership of the Dutch Elderly: explaining Cohort, Time and Age Effects." *De Economist,* 155: 99–121

[60] Winship, Christopher, and David J. Harding. 2008. "A Mechanism-Based Approach to the Identification of Age Period Cohort Models." *Sociological Methods and Research* 36(3): 362–401

[61] Yang, Yang, Wenjiang Fu, and Kenneth C. Land. 2004. "A Methodological Comparison of Age-Period-Cohort Models: The Intrinsic Estimator and Conventional Generalized Linear Models." *Sociological Methodology* 34: 75–110

[62] Yang Yang. 2008. "Trends in U.S. Adult Chronic Disease Mortality, 1960—1999: Age, Period, and Cohort Variations." *Demography* 45(2): 387–416

[63] Yang, Yang, Wenjiang Fu, and Kenneth C. Land. 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

[64] Yang, Yang, Sam Schulhofer-Wohl, and Kenneth C. Land. 2007. "A Simulation Study of the Intrinsic Estimator for Age-Period-Cohort Analysis." Paper presented at the Methodology Paper Session at the Annual Meetings of the American Sociological Association in New York, August 2007.

Figure 1: US Female Mortality Data
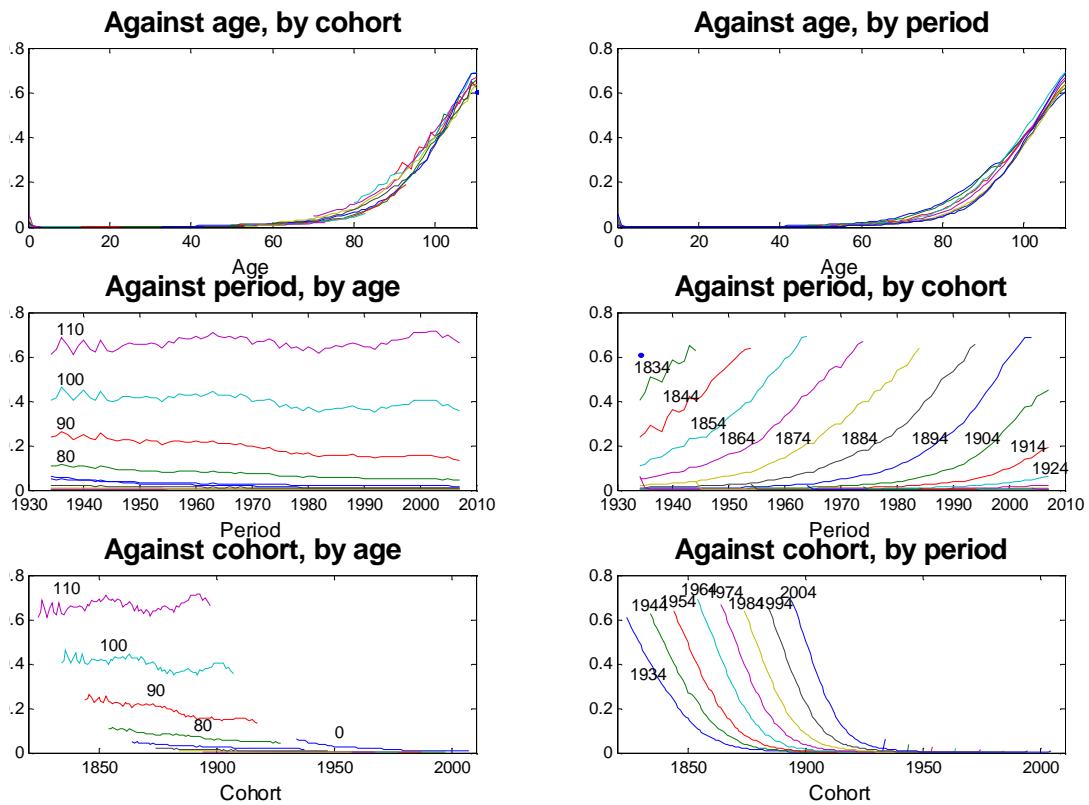
FIGURE 2: Setting age (top), period (middle), cohort (bottom) effects to zero
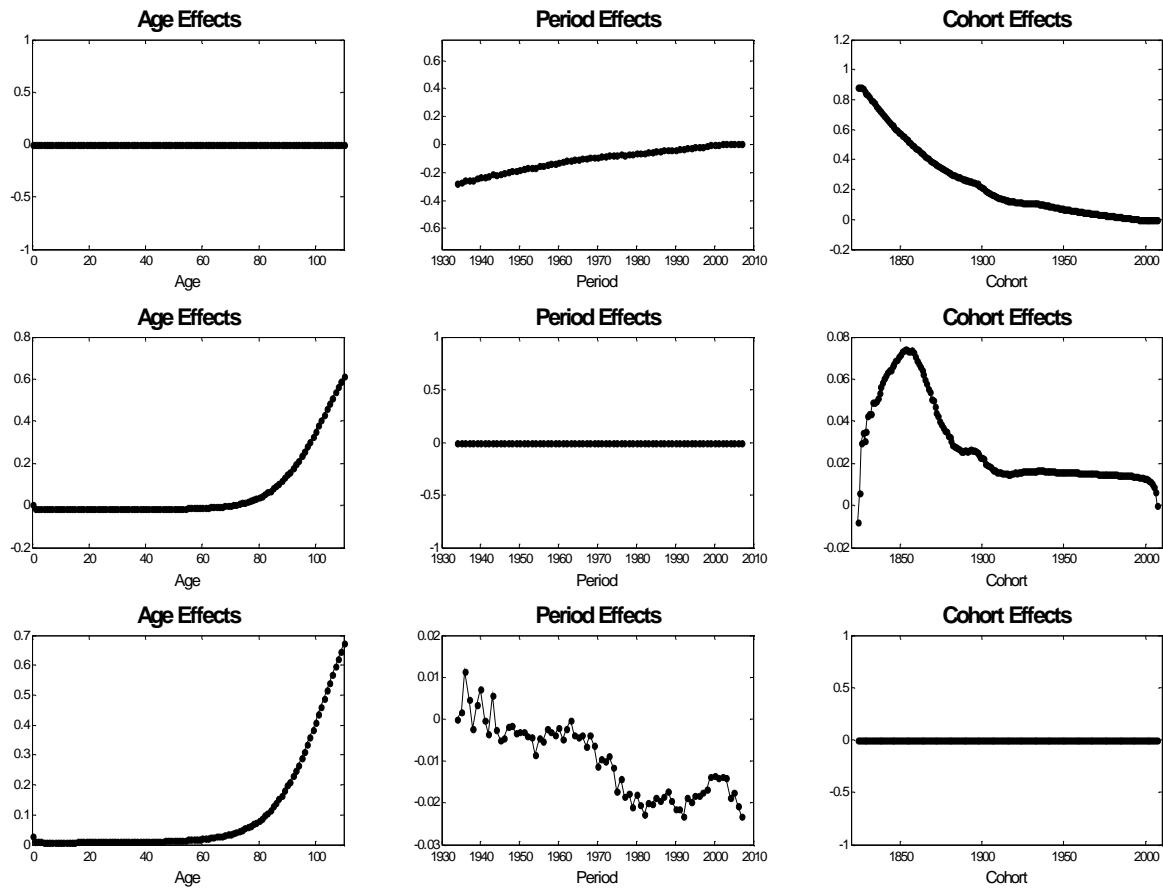
FIGURE 3: Setting adjacent cohorts to be equal
(top: 1849=1850, middle 1899=1900, bottom: 1949=1950)
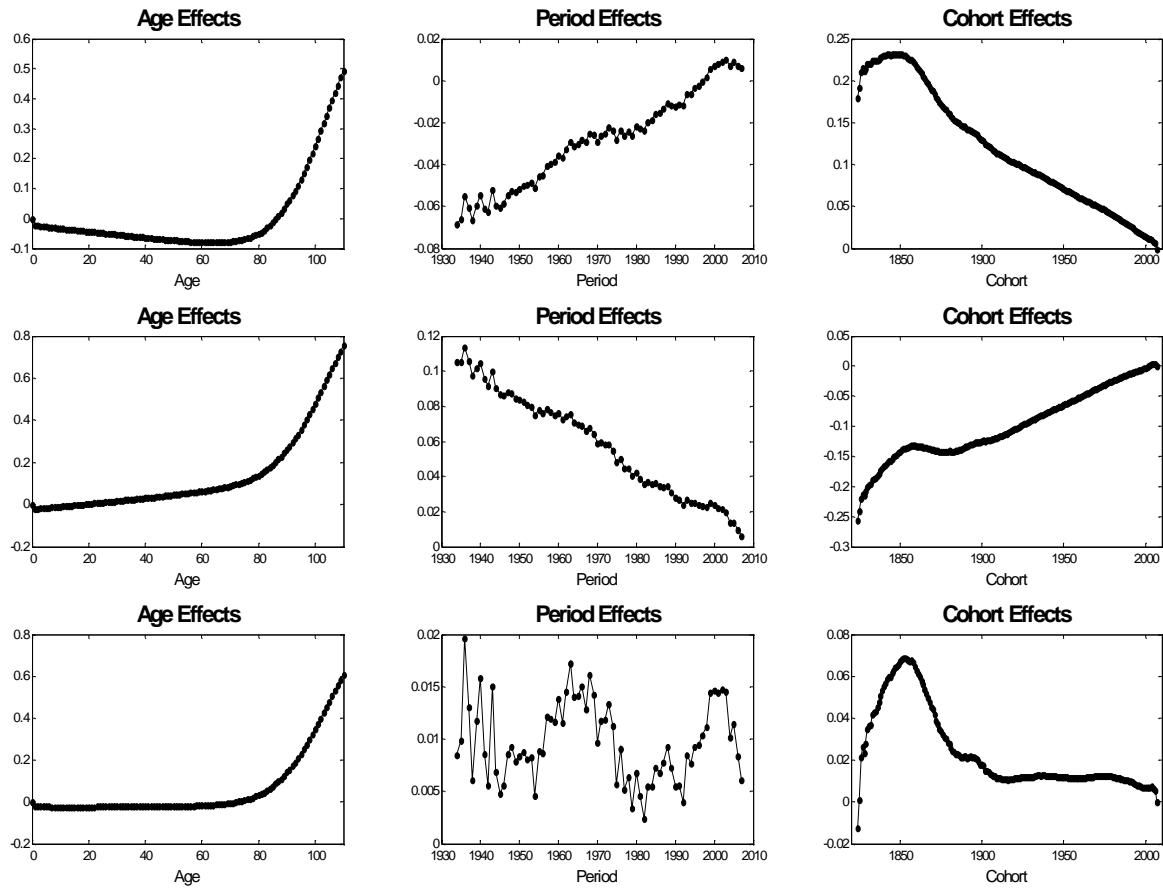
FIGURE 4: Hanoch and Honig/Deaton and Paxson normalization
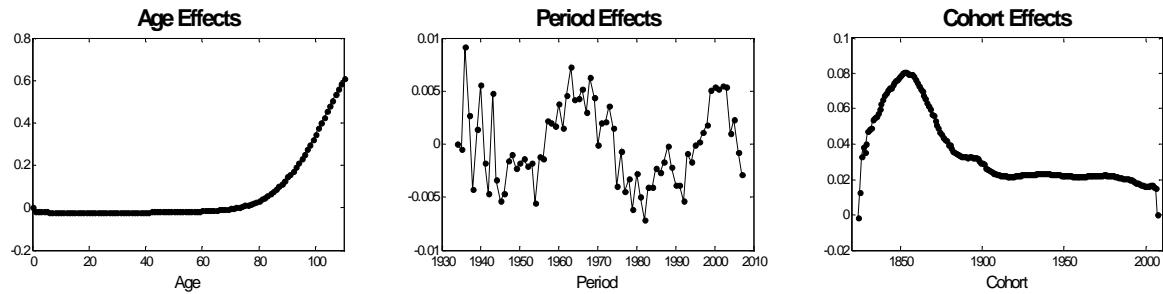


FIGURE 5: The Intrinsic estimator
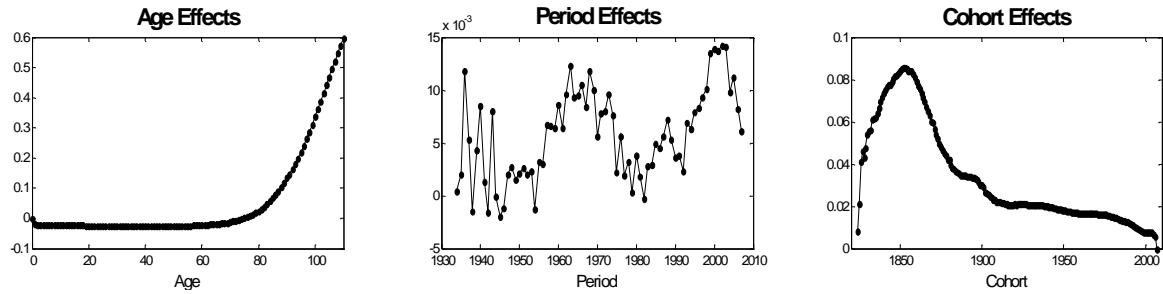


FIGURE 6: Maximum Entropy Estimates



36
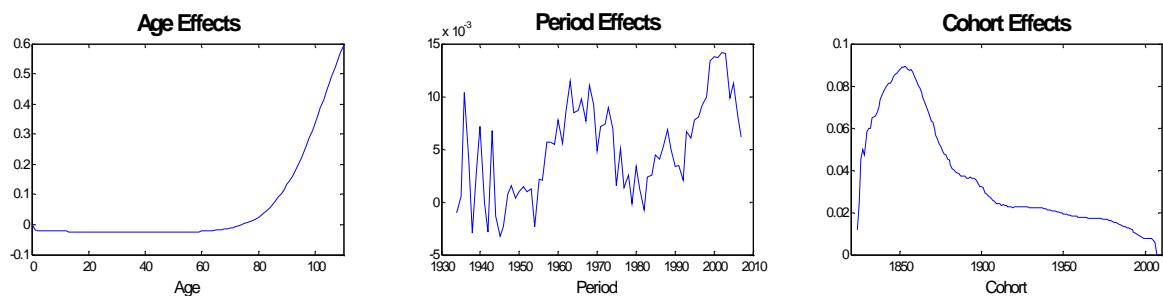
FIGURE 7: UK Female Labor Force Participation Data
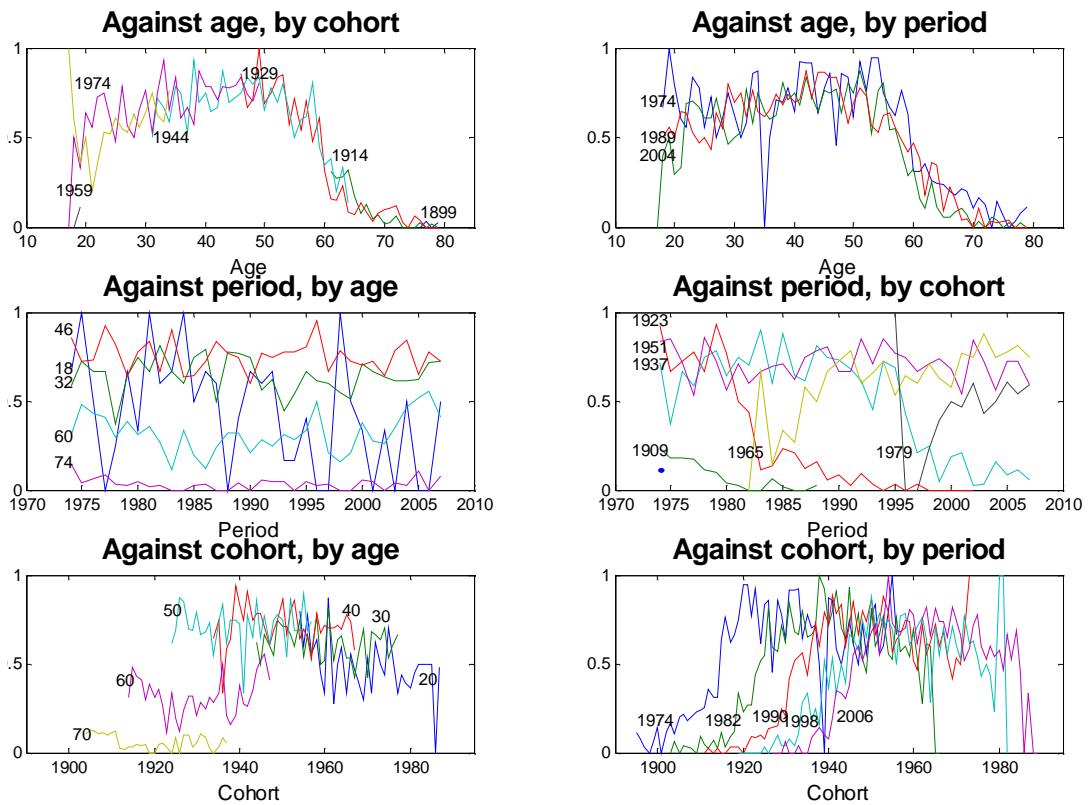
FIGURE 8: Setting age (top), period (middle), cohort (bottom) effects to zero
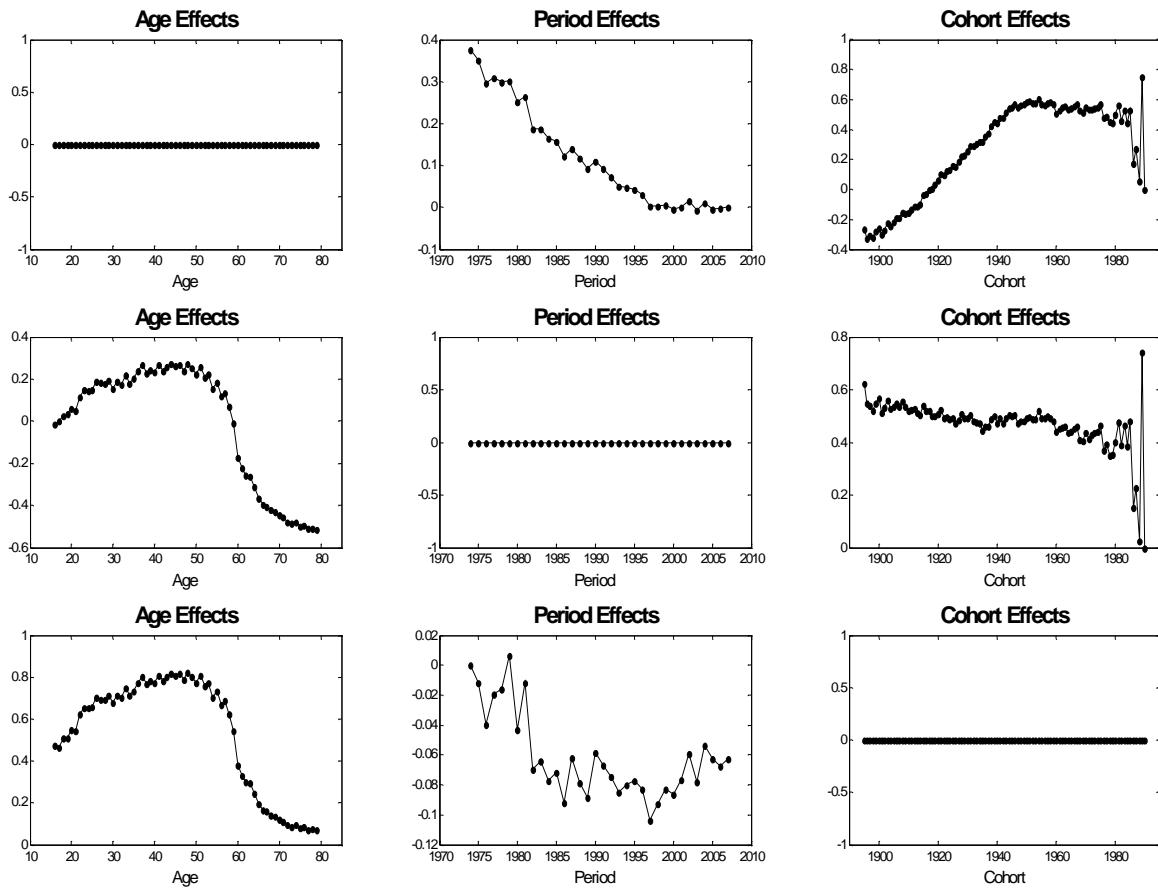
FIGURE 9: Setting adjacent ages to be equal
(top: 29=30, middle 39=40, bottom: 49=50)
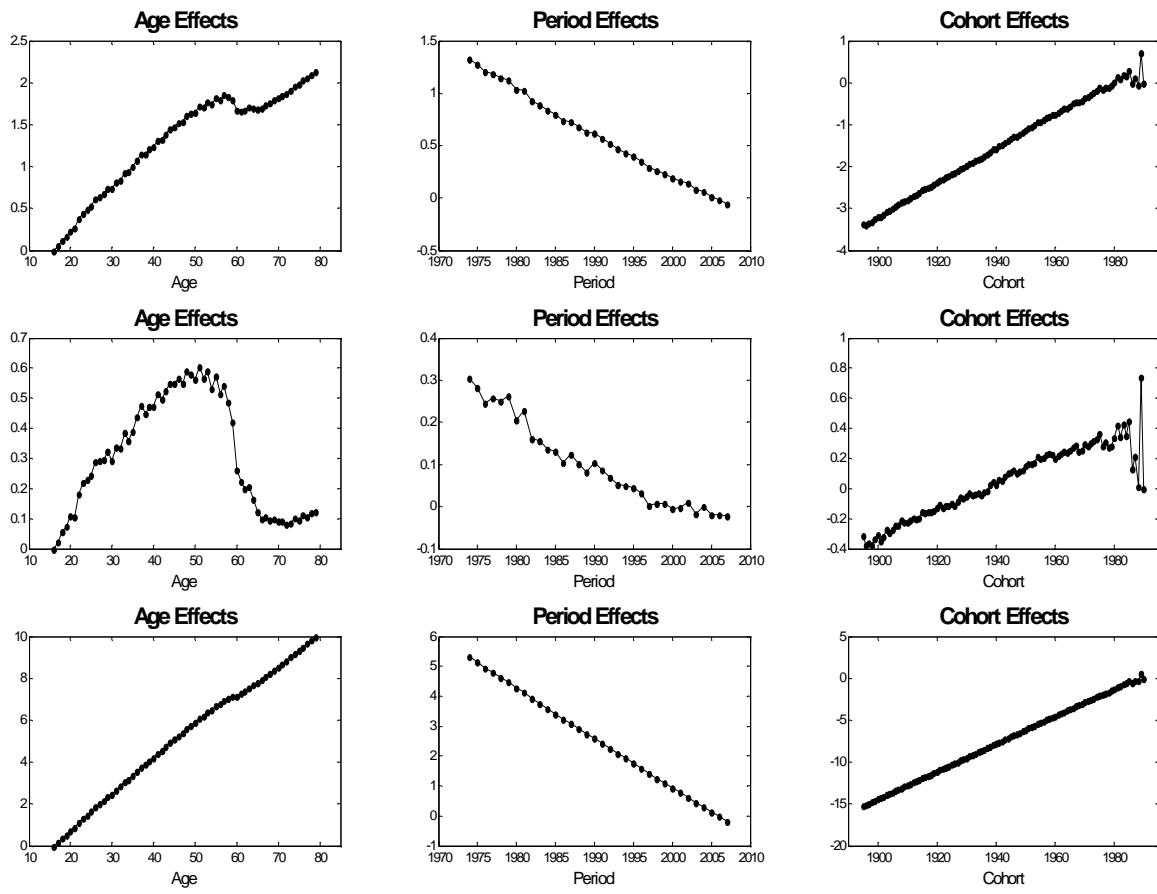
FIGURE 10: Hanoch and Hoing/Deaton and Paxson normalisation
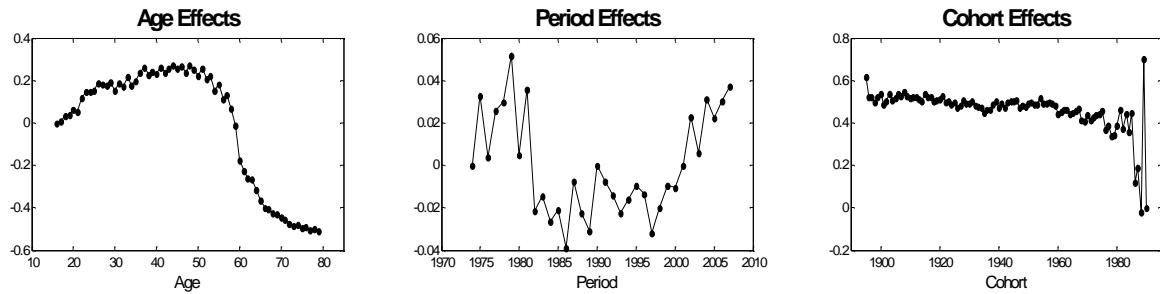


FIGURE 11: Intrinsic estimator
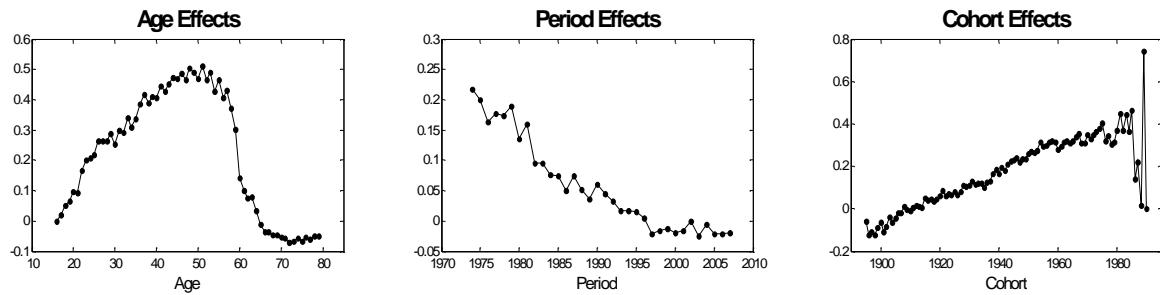


FIGURE 12: Maximum Entropy Estimates