The Orthogonal Estimator (OE): An Alternative to the Intrinsic Estimator (IE) for Estimating Age-Period-Cohort Effects

Abstract

Over the past decade the Intrinsic Estimator (IE) has become a popular way to model age-period-cohort (APC) trends in sociology, demography, epidemiology, and other fields. However, confusion exists over the fundamental properties of the IE and its proper use in applied research. In this paper we clarify the meaning of the IE, examine its strengths and weaknesses, and propose an alternative estimator. Despite some desirable statistical properties, the IE will give differing results depending on the number of age, period, and cohort groups as well as the size and direction of the nonlinearities in the data. Moreover, to achieve identification the IE uses a mathematical constraint that lacks a clear substantive interpretation. To overcome the limitations of the IE, we introduce the orthogonal estimator (OE), which uses a coding scheme that separately estimates the linear and nonlinear effects, unlike the IE's sum-to-zero effect (or deviation) coding. The OE has the same desirable statistical properties as the IE. However, in contrast to the IE, the OE clarifies the nature of the identification problem and provides a set of estimates that are invariant to the number of APC groups as well as the size and sign of the nonlinearities. Moreover, the OE can be interpreted as the simple average of three zero-linear-trend (ZLT) models, unlike the IE, which is the average of a rotated set of ZLT models. Like the IE, the OE in general will not recover the true data generating parameters.

Introduction

The intrinsic estimator (IE) has become a popular technique for estimating age-period-cohort (APC) effects[1] across a wide range of fields. In recent years, researchers have used the IE to examine APC trends in pornography use (Price et al. 2016), behavioral problems in adolescents (Keyes et al. 2017), heart disease mortality (Kramer, Valderrama, and Casper 2015), breast cancer mortality (Li, Yu, and Wang 2015), obesity in China (Fu and Land 2015), and social trust (Hu 2015), among other topics. As pointed out by various authors, the IE has several desirable statistical properties, producing results that are estimable, unbiased given the mathematical constraint imposed by the IE, and with minimum variance among estimators based on the same design matrix.[2] What is often not appreciated is that these properties hold for a host of other estimators that, like the IE, are based on the Moore-Penrose generalized inverse, but that employ different design matrices. As shown by Luo and colleagues (2016), estimators based on design matrices other than the zero-sum effect (or deviation) coding used by the IE can give dramatically different results.[3]

Although widely-used, a number of misconceptions exist about the IE. Most importantly, researchers frequently have incorrectly claimed or implied that the IE produces, without applying a constraint, what are known as the (true) underlying generating parameters for the data. To take one area of study, consider recent research on APC trends in social trust, long thought to include a substantial cohort effect. In their analysis of APC trends in social trust, Clark and Eisenstein (2013) assert the following: "The IE avoids the identification problem and makes it possible to estimate effects without imposing constraints on the data and thus, provides unbiased estimates of age, period, and cohort coefficients (365)." Likewise, in a study of APC tends in social trust and related variables, Schwadel and Stout (2012) state erroneously that the "APC intrinsic estimator models provide unbiased estimates of regression coefficients for age groups, time periods and birth cohorts" (238). Even when researchers do not directly claim that the IE provides unbiased estimates of the underlying APC trends, it is often implied. For instance, in an analysis of social trust in China, Hu (2015) approvingly cites Schwadel and Stout, stating that the "IE has been proved by many previous studies to be a practically workable strategy in singling out the unique effect of age, period, and cohort (e.g., Schwadel and Stout 2012) (238)."

[1] Following the convention in the APC literature, we use the term "effects" when referring to age, period, and cohort processes (e.g., Glenn 1981: 249; Mason, Mason, et al. 1973: 243; Fienberg and Mason 1979: 133; O'Brien 2015: 1; Yang and Land 2013c: 1-2). These "effects" need not refer to causal effects in the sense of parameters associated with well-defined potential outcomes (Morgan and Winship 2014: 37-76).

[2] In the context of the present paper, a set of parameters are *identifiable* if there is an estimator that would produce the true, underlying data generating parameters on a sample of infinite size. In contrast, *estimability* implies that an estimator gives some unique value whether or not the estimate is equal to the true underlying data generating parameters. Thus identification implies estimability, but estimability does not imply identification. For a succinct but clear discussion on defining identifiability and estimability, see Greenland (2005).

[3] Land and colleagues (2016) state in no uncertain terms that "only the sum-to-zero [effect] coding is used to define and estimate the IE (964)." However, we note that in the first published paper to introduce the IE, no mention is made of the necessity of zero-sum effect coding in formally defining the IE (see Fu 2000: 263-268; 276-277). Rather, the IE is presented as a general method for estimating effects from singular design matrices, of which the APC case is used as an example.

The literature on social trust is by no means unique in misrepresenting the properties of the IE.[4] For example, in a recent study of breast cancer mortality, Li and colleagues (2015) claim that "the IE algorithm could obtain the unique solution of parameters without any constrained conditions or new variable" and that "the estimated values are interpreted more intuitively and [are] unbiased (5)." Similarly, Price et al. (2016) falsely state in their analysis of temporal trends in pornography use that the IE requires "less restrictive constraints" than other estimation techniques (15). To applied researchers in a wide range of disciplines, it appears the IE provides unbiased estimates of the underlying data generating parameters with minimal or no constraints.

However, the developers and proponents of the IE, in particular Yang, Land, and Fu, have repeatedly suggested or stated that the IE is not intended to and does not estimate the underlying data generating parameters. For example, in an extended discussion of the IE, Fu, Land, and Yang (2011) assert that "the model generating parameters" are "not identifiable" (456). Likewise, in a full-length book detailing a variety of APC models, Yang and Land (2013c) state outright that "the objective of the IE is not to estimate the unidentifiable regression coefficient vector [i.e., the set of underlying data generating parameters] (119)." These points are repeated elsewhere in several publications (e.g., Yang, Fu, and Land 2004; Yang, Schulhofer-Wohl, et al. 2008), including a detailed working paper by Yang and Land (2013). Instead of recovering the true data generating parameters, the IE applies a particular mathematical constraint to produce one of an infinite number of estimates of the APC effects that are consistent with the data (Yang, Fu, and Land 2004: 84-85; Yang and Land 2013c: 82; Land et al. 2016: 965-966). The estimates given by the IE will not, in general, equal the true data generating parameters of interest to applied researchers.[5]

Why, then, do these critical misunderstandings about the IE persist? We believe these misunderstandings arise in part from confusion about what it means for a parameter in formal mathematical statistics to be estimable and an estimate of it to be unbiased. As well, even though the proponents of the IE have provided detailed, explicit discussions of the properties of the IE, much of the literature is highly technical, requiring facility in linear algebra, mathematical statistics, and the geometry of regression models. Moreover, although stating that the IE is not intended to recover the true data generating parameters, the proponents of the IE have, in various works, repeatedly claimed that the IE gives "sensible", "useful", "reliable", or "replicable" results (Yang, Schulhofer-Wohl, et al. 2008: 1711; Fu, Land, and Yang 2011: 456; Yang and Land 2013b: 75; 119-120). To a lay researcher, such statements can be easily misinterpreted as meaning that the IE yields the true data-generating parameters, which it does not.[6] In fact, the IE produces unbiased estimates of the

[4] For a lucid overview of other studies that misinterpret the IE, see Luo (2013: 1950-1951).

[5] That is, as we discuss in the following sections, the estimates produced by the IE and the true data generating parameters will equal each other only if the true data generating parameters by happenstance conform to the IE's very specific mathematical constraint. There is no way to determine whether this is true or not (see Luo 2013: 1951-1960).

[6] The likelihood of confusion is compounded by the fact that, in a rejoinder to criticisms by Luo et al. (2016), Land and colleagues (2016) state: "We think [the IE gives unbiased estimates of the true data-generating parameters] based both on the cited robustness studies and on our personal experiences with simulations studies and empirical applications of the IE (967)." As we show later in the text, it is trivial to construct simulations where the IE does a poor job of recovering the underlying data generating parameters, thereby contradicting this assertion (for example, see Tables 2, 4, and 5).

parameters given the IE's particular mathematical constraint, not unbiased (and unconstrained) estimates of the true data generating parameters. Applied researchers hoping to estimate the underlying temporal trends but lacking the requisite technical knowledge are likely to miss this crucial distinction and fail to recognize the inherent limitations of the IE.

As we demonstrate later, the IE suffers from three major weaknesses that are largely unacknowledged in the applied literature.[7] First, the estimates of the linear effects produced by the IE are a function of the number of age, period, and cohort categories in the data available to the researcher. Under the same underlying data generating process, the IE can produce dramatically different results simply based on the number of categories used. Second, estimates of the linear effects using the IE are not only a function of the number of categories used, but also the extent of the nonlinear effects in the data. Depending on the magnitude and direction of the nonlinearities, the IE will yield differing linear estimates, sometimes radically so, for the same underlying age, period, and cohort linear effects. Finally, because the IE's mathematical constraint differs depending on the number of age, period, and cohort categories as well as the extent of nonlinearities, it is complicated, non-intuitive, and highly variable from study to study.

The IE represents what might be called a *statistical solution* to the APC identification problem in that it uses a constraint that is statistically rather than theoretically motivated. For example, as noted previously, the IE is typically justified by referring to its desirable statistical properties; to wit, the IE is estimable, unbiased given its mathematical constraint, and has minimum variance among estimators based on the same design matrix. The IE can be viewed as imposing a statistical constraint in that for a sample of data, the IE gives the set of estimates, for that particular coding scheme that has minimum variance. This statistical constraint is equivalent to the mathematical constraint of choosing that point on the solution line closest to the origin. However, such statistical solutions are problematic because in general they do not provide estimates of the underlying unknown data generating parameters of interest to most substantive researchers. Accordingly, we strongly advocate using a *theoretical solution* to the identification problem instead, in which the researcher applies constraints that explicitly reflect theories about the underlying temporal processes (Fosse and Winship 2017; Winship and Harding 2008). For example, in examining APC trends in criminal behavior, one might constrain the age trend to monotonically decrease after early adulthood, reflecting the presence of an age-crime curve (Loeber and Farrington 2014), or construct a model specifying the mechanisms, such as marriage or military service, that are widely thought to cause desistance over the life course (Sampson 2005). Nevertheless, if one is going to use what we have termed a statistical solution to the identification problem, there is a superior estimator.

As an alternative to the IE, we propose what we call the Orthogonal Estimator (OE).[8] The OE, like the IE, is based on the Moore-Penrose generalized inverse and thus shares the IE's desirable statistical properties. However, the OE uses a design matrix that separately estimates the linear and

[7] For recent discussions on limitations of the IE in the methodological literature, see Luo (2013), Luo et al. (2016), O'Brien (2011; 2015), and Pelzer et al. (2014). For rejoinders, see Yang and Land (2013) as well as Land et al. (2016).

[8] The name of the OE is based on the fact that, for each of the temporal variables, the linear components are constructed to be orthogonal to the nonlinear components.

nonlinear effects, unlike the the zero-sum effect (or deviation) coding of the IE. As a result, the OE's mathematical constraint is invariant to the number of age, period, and cohort groups as well as the magnitude and direction of the nonlinearities in the data. Additionally, the OE can be interpreted as the simple average of three basic APC models in which one of the age, period, or cohort linear components is set to zero. In contrast, as we demonstrate later, the IE is an average of a rotated set of simpler models. Thus, although the IE also can be interpreted as a simple average, as a general rule it will not give results that are equal to the average of three zero-linear-trend (ZLT) models in which one of the linear trends of age, period, or cohort is set to zero. In short, the OE is an estimator that provides APC estimates with the same desirable statistical properties as the IE, but avoids the IE's sensitivity to the number of categories and the extent of the nonlinearities while also providing a transparent, easily-interpretable mathematical constraint, thereby facilitating the accumulation of knowledge.

The rest of the paper is organized as follows. First, we describe the basic identification problem in APC models. Specifically, we show how, regardless of the design matrix, for any APC model it is only the linear, not the nonlinear, effects that are unidentified. Second, we discuss the IE and its derivation, outlining its underlying mathematical constraint. Third, we discuss the desirable statistical properties of the IE, in particular the fact that it is estimable, unbiased given its constraint, and has minimum variance given its design matrix. Fourth, we discuss the limitations of the IE, presenting several simulations that demonstrate the sensitivity of the IE to the number of categories as well as the size and sign of the nonlinearities. In the fifth section we introduce the OE, which shares the IE's desirable statistical properties but overcomes the weaknesses of the IE. Moreover we show that the OE is equal to the simple average of three ZLT models, unlike the IE. We conclude with guidelines for applied researchers as well as a discussion of the strengths and weaknesses of using a statistical as opposed to a theoretical approach to identifying APC effects.

1 Nature of the Problem

Researchers have long sought to estimate age, period, and cohort effects for a wide set of social, biological, and cultural outcomes.[9] However, because of the exact linear relationship between age, period, and cohort, one cannot use conventional statistical methods to estimate the unique contributions of age, period, and cohort on a particular outcome (Mason, Mason, et al. 1973). The *APC identification problem*, as it has become known, is the result of the simple the fact that if we know a person's age in years and the year in which their outcome was measured, then we know their birth year. As such, holding say age and period constant, there is no independent variation in cohort

[9] For example, researchers have examined verbal ability (Alwin 1991; Hauser and Huang 1997; Wilson and Gove 1999), social trust (Clark and Eisenstein 2013; Putnam 1995; Robinson and Jackson 2001; Schwadel and Stout 2012), party identification (Ghitza and Gelman 2014; Hout and Knoke 1975; Tilley and Evans 2014), religious affiliation (Chaves 1989; Firebaugh and Harley 1991), drug use (Chen et al. 2003; Kerr et al. 2004; O'Malley, Bachman, and Johnston 1984; Vedøy 2014), obesity (Diouf et al. 2010; Fu and Land 2015), cancer (Clayton and Schifflers 1987; Liu et al. 2001), and mental health (Lavori et al. 1987; Yang 2008).

with which to estimate its effect. When the data consists of only three continuous (interval-coded) variables, the linear relationship is simply $\text{cohort} = \text{period} - \text{age}$. As we discuss later, when the data is expressed as categorical variables, the linear combination will appear in a different form since the data may be represented in various ways depending on the coding scheme. Regardless, due to the APC identification problem we can always express one or more of the APC variables as a linear combination of the other variables, since there is redundancy of information in the data.

1.1 The Classical APC (C-APC) Model

To clarify the discussion that follows, suppose we have categorically-coded age, period, and cohort data for a set of n respondents.[10] We let $i = 1, \ldots, I$ denote the unique age groups, $j = 1, \ldots, J$ the unique period groups, and $k = 1, \ldots, K$ the unique cohort groups with $k = j - i + I$ and $K = I + J - 1$.[11] We let n denote the number of respondents. The model we would like to run is

$$Y = \mu + \sum_{i=1}^{I}(\alpha_i)(\text{age}_i) + \sum_{j=1}^{J}(\pi_j)(\text{period}_j) + \sum_{k=1}^{K}(\gamma_k)(\text{cohort}_k) + \epsilon, \tag{1}$$

which we refer to as the Classical APC (C-APC) model, also known as the Multiple Classification or Accounting Model. More generally, we can represent the temporal effects in matrix notation as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \epsilon \tag{2}$$

where $\mathbf{y}$ is an $n \times 1$ outcome vector which, without loss of generality, we assume to be continuous; $\mathbf{X}$ is a design matrix of dimension $n \times p$; $\mathbf{b}$ is a $p \times 1$ parameter vector with elements corresponding to the age, period, and cohort groups; and ϵ is an $n \times 1$ vector of random errors.[12]

If there were no linear dependence in $\mathbf{X}$, then we could obtain a unique least-squares solution

$$\mathbf{b}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{3}$$

where the superscripted -1 indicates the regular inverse. However, due to linear dependence, $\mathbf{X}$ is rank deficient one and a regular inverse of $\mathbf{X}^T\mathbf{X}$ does not exist.[13] Thus, we cannot estimate $\mathbf{b}_{\text{OLS}}$ and any particular least-squares solution requires an additional constraint.

[10] For simplicity we also assume that the age and period categories are of equal width.

[11] Note that I is added to $j-i$ so that the cohort index begins at $k = 1$. This ensures that, for example, $i = j = k = 1$ refers to the first group for all three temporal scales. One could just as easily index the cohorts using $k = j - i$, but this identity would be lost.

[12] Without loss of generality, in the discussion that follows we assume that there are no disturbances, such that $\epsilon = \mathbf{0}$.

[13] There are other possible design matrices that are deficient by more than rank one. These would be matrices that do not provide a full set of terms for the linear and nonlinear effects of age, period, and cohort. For our purposes here such matrices are not a concern.

1.2 The Crux of the APC Identification Problem

The lack of a unique least-square solution reflects the fact it is only the nonlinear components of an APC model are individually identifiable, not the linear components.[14] For any design matrix there is an invertible $p \times p$ transformation matrix $\mathbf{T}$ that allows us to separate the linear from the nonlinear components (cf. Luo et al. 2016: 947-952). Using the transformation matrix $\mathbf{T}$, we can convert any design matrix $\mathbf{X}$ as well as its corresponding parameter vector $\mathbf{b}$ into their linear and nonlinear components (2016: 947):

$$\mathbf{X}\mathbf{b} = \big(\mathbf{X}\mathbf{T}^{-1}\big)\big(\mathbf{T}\mathbf{b}\big) = \mathbf{X}_{\text{O}}\mathbf{b}_{\text{O}} \tag{4}$$

where $\mathbf{X}_{\text{O}}$ is a design matrix of sum-to-zero orthogonal polynomial contrasts and $\mathbf{b}_{\text{O}}$ is a vector of the new, transformed estimates expressed in terms of the linear and nonlinear components.[15] Specifically, the first column of $\mathbf{X}_{\text{O}}$ is a set of 1's for the intercept and the remaining columns span the linear and nonlinear components for age, period, and cohort, respectively. The corresponding vector of parameters expressed in terms of sum-to-zero orthogonal polynomials is

$$\mathbf{b}_{\text{O}} = \big(\mu, \alpha, \pi, \gamma, \alpha^{i+1} \ldots, \alpha^{I-1}, \pi^{j+1}, \ldots, \pi^{J-1}, \gamma^{k+1}, \ldots, \gamma^{K-1}\big)^T, \tag{5}$$

where μ is the intercept; α, π, and γ denote the linear trends; and the nonlinear parameters are denoted with the appropriate superscripts.[16] In short, to express the original estimates $\mathbf{b}$ as sum-to-zero orthogonal polynomial effects we use $\mathbf{T}\mathbf{b} = \mathbf{b}_{\text{O}}$. To change the transformed estimates $\mathbf{b}_{\text{O}}$ back into their original coding scheme, we can simply multiply by the regular inverse such that $\mathbf{T}^{-1}\mathbf{b}_{\text{O}} = \mathbf{b}$. Consequently, regardless of the initial coding scheme used, we can always separate the APC effects into their linear and nonlinear components.

Once we have transformed the initial set of APC estimates as well as its design matrix, we can prove that only the nonlinearities and particular combinations of the linear components are identifiable.[17] Let $\mathbf{b}_{\text{O}}^{\dagger}$ denote the basis set of identifiable functions for the C-APC model. Then using the transformation matrix we know that

[14] Yang and colleagues (2008) state that the "conventional wisdom is that only the nonlinear, but not the linear, components of APC models can be estimable… however, there have been only numeric demonstrations, but no rigorous proofs, to support the idea that no estimable function exists (1703)." Although the linear trends are not individually identifiable, there are "estimable functions" in the sense that particular linear combinations of the linear trends are identifiable.

[15] As we show in the appendix, an alternative approach is to construct a design matrix of sum-to-zero effect coding that is orthogonal to the linear components for age, period, and cohort. Findings with such a design matrix will be substantively the same as those with orthogonal polynomial contrasts.

[16] For instance, α^2 gives the quadratic age effect, α^3 gives the age cubic effect, and so forth. These are *not* to be confused with, say, taking the age linear effect and squaring or cubing it.

[17] We provide a proof in the appendix.

$$\mathbf{b}_{\text{O}}^{\dagger} = \begin{cases} \mu \\ \alpha^2, \alpha^3, \ldots, \alpha^{I-1} \\ \pi^2, \pi^3, \ldots, \pi^{J-1} \\ \gamma^2, \gamma^3, \ldots, \gamma^{K-1} \\ \alpha + \pi \\ \gamma + \pi. \end{cases} \tag{6}$$

That is, while the intercept and the nonlinear components are identifiable, we can only identify particular linear combinations of the linear trends: $\alpha + \pi$ and $\gamma + \pi$. Note further that we also can identify $\gamma - \alpha$ since $(\gamma + \pi) - (\alpha + \pi) = \gamma - \alpha$ and $\alpha - \gamma$ since $(\alpha + \pi) - (\gamma + \pi) = \alpha - \gamma$. More generally, any function of the form $\alpha(\omega_1) + \pi(\omega_1 + \omega_2) + \gamma(\omega_2)$ for arbitrary values of ω_1 and ω_2 is identifiable (cf. Holford 1983: 314). In short, for any constrained C-APC model, including the IE, the nonlinearities and some linear combinations of the linear effects are identifiable, but *not* the individual linear trends α, π, and γ. In other words, as Fienberg (2013) has emphasized, the "APC problem is a linear effects problem (1982)." In the appendix, we provide a proof that it is only the linear effects that are not identified. A proof can also be found in Holford (1983).

1.3 The Solution Line

Due to the non-identifiability of the linear trends, the design matrix $\mathbf{X}$ for the C-APC exhibits linear dependence (that is, $\mathbf{X}$ is rank deficient one). In practice this means that at least one of the columns of $\mathbf{X}$ can be rewritten as a linear combination of the other columns. Formally, there is a non-trivial linear combination of the columns of $\mathbf{X}$ that results in a vector of zeros:

$$\mathbf{X}\mathbf{v} = \mathbf{0} \tag{7}$$

where $\mathbf{v}$ is a $p \times 1$ vector and $\mathbf{0}$ is a $p \times 1$ vector of zeros. The vector $\mathbf{v}$, referred to as the null vector, represents the null space of $\mathbf{X}$ and is unique up to multiplication by an arbitrary scalar s. Accordingly, Equation 7 generalizes to $\mathbf{X}s\mathbf{v} = \mathbf{0}$. The elements of the null vector differs depending on the design matrix, since each design matrix has a different representation of the linear dependency among the age, period, and cohort variables. However, the null vector of $\mathbf{X}_{\text{O}}$, the design matrix of the OE, has the simple representation

$$\mathbf{v}_{\text{O}} = (0, 1, -1, 1, 0 \ldots 0), \tag{8}$$

where the first zero corresponds to the intercept; the elements one, negative, and one correspond to the age, period, and cohort linear components; and the remaining zeros correspond to the $(I - 2) + (J - 2) + (K - 2)$ nonlinear components. That is, the null vector $\mathbf{v}_{\text{O}}$ encodes the basic linear relationship that a person's age minus their year of measurement plus their birth year equals zero. This is the fundamental linear identity of any C-APC model.

Geometrically, the data do not provide a point estimate but rather a one-dimensional line, or a *solution line*, in a multidimensional parameter space (Holford 1991; O'Brien 2011; O'Brien 2015: 59-91). Using the transformation matrix outlined in the previous section, we can make an important simplification: for any C-APC model, the solution line runs through just three dimensions defined by the range of the slopes for age, period, and cohort. The reason for this is that the nonlinearities are identified and as such their parameter estimates each equal a single point on their respective axis. For example, Figure 1 visualizes the solution line for an APC model with simulated data in which $\alpha = 1$, $\pi = -4$, and $\gamma = 6$. All possible estimates for the linear effects lie on this line in three-dimensional space.

To derive the solution line, we let $\mathbf{b}^*$ denote any specific constrained least-squares solution to the least-squares normal equations. For any particular constraint, we can construct a generalized inverse of $\mathbf{X}^T\mathbf{X}$ to find a corresponding solution (Mazumdar, Li, and Bryce 1980; O'Brien 2015: 27-28):

$$\mathbf{b}^* = (\mathbf{X}^T\mathbf{X})^*\mathbf{X}^T\mathbf{y} \tag{9}$$

where the superscript $*$ denotes the appropriate generalized inverse. The vector $\mathbf{b}^*$ is *a* least-squares solution to the normal equations, such that $\mathbf{X}^T\mathbf{b}^* = \mathbf{X}^T\mathbf{y}$. We can then write:[18]

$$\mathbf{b} = \mathbf{b}^* + s\mathbf{v}. \tag{10}$$

In other words, the true data generating parameter vector $\mathbf{b}$ equals the constrained estimates $\mathbf{b}^*$ plus the product of an unknown scalar s and the null vector $\mathbf{v}$. The solution line is then simply traced out by varying the values of s.

Two simplifications are possible that allow us to express the solution line in just three dimensions. Since $\mathbf{T}\mathbf{b} = \mathbf{b}_{\text{O}}$, we can re-write Equation 10 as:

$$\mathbf{b}_{\text{O}} = \mathbf{b}^*_{\text{O}} + s\mathbf{T}\mathbf{v}, \tag{11}$$

where $\mathbf{b}^*_{\text{O}}$ is a constrained set of estimates separated into nonlinear and linear components. Second, because $s\mathbf{T}\mathbf{v} = s\mathbf{v}_{\text{O}}$, we can write:

$$\mathbf{b}_{\text{O}} = \mathbf{b}^*_{\text{O}} + s\mathbf{v}_{\text{O}} \tag{12}$$

or, equivalently,

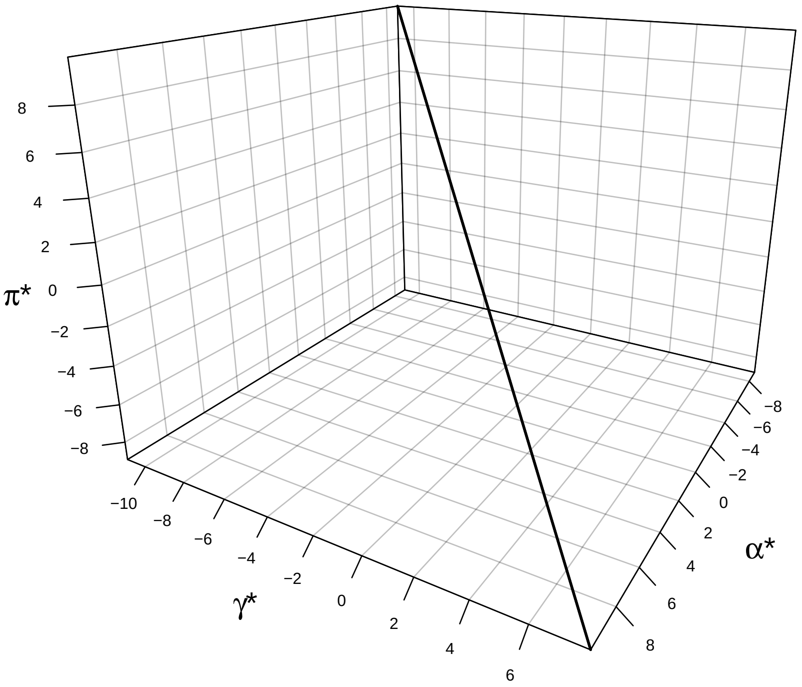
$$\begin{aligned} \alpha^* &= \alpha + s \\ \pi^* &= \pi - s \\ \gamma^* &= \gamma + s \end{aligned} \tag{13}$$

where α, π, and γ denote the true, unknown data generating slopes; s is an unknown scalar; and the

[18] Because $\mathbf{X}s\mathbf{v} = \mathbf{X}^T\mathbf{X}s\mathbf{v} = \mathbf{0}$, we know that $\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\mathbf{b}^* + \mathbf{X}^T\mathbf{X}s\mathbf{v}$. Rearranging we can thus write $\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}(\mathbf{b}^* + s\mathbf{v})$. Since $\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{X}(\mathbf{b}^* + s\mathbf{v})$, it follows that $\mathbf{b} = \mathbf{b}^* + s\mathbf{v}$.

asterisks denote any particular constrained set of estimates on the solution line. Since s can take on any real number, Equation 13 defines a three-dimensional solution line for the design matrix $\mathbf{X}_O$ and outcome vector $\mathbf{y}$. Importantly, because only the linear components are unidentified, the solution line for *any* constrained C-APC model runs through just three dimensions defined by the set of possible values of the linear slopes for age, period, and cohort.

Figure 1: The Solution Line of an APC Model



Notes: Based on data generated with values of $\alpha = 1$, $\pi = -4$, and $\gamma = 6$.

The fact that the linear effects are confounded with each other places strict limits on what can be known from the data using fit statistics and graphical techniques (Yang and Land 2013c: 125-153). Before using the IE, Land and colleagues contend that one should use fit statistics and graphical tools to determine whether or not all three time scales are operating. The IE, in their view, should only be used when such analyses "suggest that all three dimensions are operative" (Yang and Land 2013a: 1969). Otherwise, they argue, one should use just one or two of the temporal variables, since then "there is no identification problem (Yang and Land 2013a: 1969)."

Unfortunately, it is impossible to determine from the data alone whether or not all three temporal linear trends are operating. This can seriously misled researchers. For example, suppose the age and cohort variables have large nonlinearities, the period effects have no nonlinearities, and all three variables have large linear effects. Since the linear trends for age and cohort are confounded with the linear trend for period, and the period variable has no nonlinearities, fit statistics and visual techniques will suggest that only the age and cohort variables are operating. Fitting a twofactor APC model with only age and cohort effects will impose the identification assumption that the period slope is zero, even though the true period linear effect is large. This ZLT constraint for the period variable is external to the data, imposed by the researcher. Depending on the substantive application, it may or may not be reasonable to assume that, because the nonlinear effects of period are observed to be zero, its linear effect is also zero. However, to reiterate, this is an assumption that can only be justified by appealing to theory or the inclusion of additional data.

2 Defining and Interpreting the IE

There are multiple ways to derive the IE. It is closely related to principal components regression (Kupper, Janis, Salama, et al. 1983) as well as ridge regression (Fu 2000). For the present purposes it is useful to think of the IE as being defined by a design matrix based on zero-sum effect (or deviation) coding, $\mathbf{X}$, and the use of the Moore-Penrose generalized inverse (Land et al. 2016). Since any APC matrix $\mathbf{X}$ is deficient rank one, it does not have a regular inverse. The Moore-Penrose generalized inverse of $\mathbf{X}^T\mathbf{X}$ is used instead, giving, parallel to traditional OLS, the formula:

$$\mathbf{b}_{\text{IE}} = (\mathbf{X}^T\mathbf{X})^{+}\mathbf{X}^T\mathbf{y} \tag{14}$$

wher the superscript $(+)$ denotes the Moore-Penrose generalized inverse.

As with any constrained least-squares estimate for a particular design matrix $\mathbf{X}$ and outcome $\mathbf{y}$, we can use the IE to construct a multidimensional solution line:

$$\mathbf{b} = \mathbf{b}_{\text{IE}} + s\mathbf{v} \tag{15}$$

where, as previously, s is a scalar that can take on any real number.

The IE solution is unique, but in a very specific sense. For any particular $\mathbf{X}$ and $\mathbf{y}$ there is only one set of least-squares estimates with a minimum (Euclidean) length or, equivalently, minimum L^2 norm. In other words, the IE is an estimator that produces a *minimum norm least-squares* (MNLS) solution, where the norm is Euclidean. However, there are many different MNLS solutions, because we can construct the design matrix using any number of different coding schemes. To reiterate, among these coding schemes the IE, as defined by Land et al. (2016), is based *only* on sum-to-zero effect coding of a design matrix of categorical age, period, and cohort variables.

There are many ways to derive the Moore-Penrose generalized inverse estimates in Equation 14. An especially intuitive approach is to find the Moore-Penrose generalized inverse of $\mathbf{X}^T\mathbf{X}$ using a decomposition technique. We can write the following:

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T \tag{16}$$

where $\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$ is the spectral (or eigenvalue) decomposition of $\mathbf{X}^T\mathbf{X}$. Note that $\boldsymbol{\Lambda}$ is a $p \times p$ diagonal matrix of the eigenvalues of $\mathbf{X}^T\mathbf{X}$ in descending order $\lambda_1, \lambda_2, \ldots, \lambda_{r-1}$, with the rank of the matrix $r = p - 1$. It is straightforward to find the generalized inverse of $\mathbf{X}^T\mathbf{X}$. First, we find the tranpose of $\mathbf{\Lambda}$ and take the reciprocal of the nonzero eigenvalues along the diagonal, keeping the zero eigenvalues. This will give the generalized inverse of $\mathbf{\Lambda}$, denoted as $\mathbf{\Lambda}^+$. Second, we calculate $\mathbf{V}\mathbf{\Lambda}^+\mathbf{U}^T = (\mathbf{X}^T\mathbf{X})^+$. Finally, we use $(\mathbf{X}^T\mathbf{X})^+$ to find the Moore-Penrose generalized inverse estimates: $\mathbf{b}^+ = (\mathbf{X}^T\mathbf{X})^+\mathbf{X}^T\mathbf{y}$.[19]

With categorical variables, alternative coding schemes are possible that avoid the overparameterization in the C-APC due to the inclusion of the intercept (Yang, Fu, and Land 2004: 79-80). For example, one could fix to zero the first set of levels (e.g., $\alpha_{i=1} = \pi_{j=1} = \gamma_{k=1} = 0$) or the last set (e.g., $\alpha_{i=I} = \pi_{j=J} = \gamma_{k=K} = 0$). However, as emphasized by Land and colleagues, the IE requires that the age, period, and cohort groups are coded so that the parameters represent effects (or deviations) from the overall mean with the constraint that these effects add up to zero (Yang and Land 2013c: 79; Land et al. 2016: 964). Formally, we collect the IE parameters in the $p \times 1$ vector:

$$\mathbf{b} = (\mu, \alpha_i \ldots, \alpha_{I-1}, \pi_j, \ldots, \pi_{J-1}, \gamma_k, \ldots, \gamma_{K-1})^T \tag{17}$$

where μ is the intercept, representing the overall (or grand) mean, and the superscripted T denotes the transpose. The design matrix constrains the age, period, and cohort parameters to sum to zero, such that $\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \pi_j = \sum_{k=1}^{K} \gamma_k = 0$.[20]

In an identified model with a full rank design matrix, the choice of coding scheme is relatively trivial since the underlying values of the parameters remain unchanged. However, when using the Moore-Penrose generalized inverse on a rank deficient design matrix, the coding scheme is of utmost importance. As demonstrated by Luo and colleagues (2016), different coding schemes will result in estimators based on different constraints, or what Land and colleagues (2016) call "pseudo-IE" estimators (971), and these alternative estimators can lead to radically different estimates of the linear effects of age, period, and cohort. To reiterate, as defined by Land and colleagues, the IE is based *only* on a design matrix of sum-to-zero effect (or deviation) coding (Land et al. 2016: 964). Since the nonlinear effects are identified, the IE and these alternative estimators will produce identical parameter estimates for the nonlinear effects, but often quite different estimates of the linear effects.

2.1 Linear Dependency and the IE

It has been shown that with sum-to-zero effect coding, the elements of the null vector have the closed-form representation (Kupper, Janis, Karmous, et al. 1985: 829)

[19] Besides allowing for estimation of the IE, the decomposition of $\mathbf{X}^T\mathbf{X}$ reveals two crucial features of the data. First, the number of non-zero eigenvalues in $\mathbf{\Lambda}$ gives the rank of the matrix $\mathbf{X}^T\mathbf{X}$. Since the columns of the data are linearly dependent, there will always be a zero eigenvalue and thus the matrix is rank deficient one. Second, along with the zero eigenvalue in $\mathbf{\Lambda}$ there will always be a corresponding eigenvector in $\mathbf{\Lambda}$, which is the orthonormal basis for the null space of $\mathbf{X}^T\mathbf{X}$. This eigenvector is simply the normalized null vector $\widehat{\mathbf{v}}$.

[20] In the APC literature, this constraint is variously referred to as "parameter centralization" (Fu 2016: 182), "ANOVA normalization" (Smith 2004: 115), or the "usual" constraints (Yang and Land 2013c: 79).

$$\mathbf{v} = (0, \mathbf{v}_{\mathrm{A}}, -\mathbf{v}_{\mathrm{P}}, \mathbf{v}_{\mathrm{C}})^T \tag{18}$$

where

$$\mathbf{v}_{\mathrm{A}} = \left(i - \frac{I+1}{2}, \ldots, I - \frac{I+1}{2} \right) \tag{19}$$

$$\mathbf{v}_{\mathrm{P}} = \left(j - \frac{J+1}{2}, \ldots, J - \frac{J+1}{2} \right) \tag{20}$$

$$\mathbf{v}_{\mathrm{C}} = \left(k - \frac{K+1}{2}, \ldots, K - \frac{K+1}{2} \right). \tag{21}$$

Often the null vector is normalized when referring to the IE (e.g., Yang and Land 2013c: 77-78):

$$\widehat{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2} = \frac{\mathbf{v}}{(\mathbf{v}^T\mathbf{v})^{1/2}} \tag{22}$$

where the caret ($\widehat{.}$) indicates normalization and the double bars ($\|.\|_2$) denote the Euclidean norm (or length) of the vector. The normalized null vector still encodes the linear dependency in the design matrix, but unlike $\mathbf{v}$, it has the additional property that its Euclidean length is one, such that $\widehat{\mathbf{v}}^T\widehat{\mathbf{v}} = 1$. Since $\widehat{\mathbf{v}}$ and $\mathbf{v}$ differ only by multiplication by a scalar, for ease of exposition in this appendix we will refer to $\mathbf{v}$ unless necessary for computational purposes.

The linear dependence in $\mathbf{X}$ is a direct result of the fact that the age, period, and cohort linear components are perfectly linearly related. To illustrate this fact, it is informative to simplify the null vector in Equation 18 as

$$\mathbf{A}\mathbf{v}_{\mathrm{A}} - \mathbf{P}\mathbf{v}_{\mathrm{P}} + \mathbf{C}\mathbf{v}_{\mathrm{C}} = \mathbf{0} \tag{23}$$

where $\mathbf{A}$, $\mathbf{P}$, and $\mathbf{C}$ are the columns for age, period, and cohort, respectively. The terms $\mathbf{A}\mathbf{v}_{\mathrm{A}}$, $\mathbf{P}\mathbf{v}_{\mathrm{P}}$, and $\mathbf{C}\mathbf{v}_{\mathrm{C}}$ are identical to the linear contrasts for age, period, and cohort in a design matrix of orthogonal polynomial contrasts. Thus, the linear dependence in $\mathbf{X}$ is simply a re-expression of the fundamental linear identity $\mathrm{age} - \mathrm{period} + \mathrm{cohort} = 0$.

The fundamental point is that the APC identification problem reveals itself algebraically in the fact that at least one variable in an APC dataset can be written as a linear combination of the other variables.[21] Recall that a linear combination is any mathematical expression that entails adding a set of terms each multiplied by a constant, where the constant can include one. For example, the well-known relationship between temperature in degrees Fahrenheit (F) and Celsius (C) is a linear combination: $F = C \times \frac{9}{5} + 32$. When we convert from degrees Celsius to Fahrenheit, we are not changing the temperature; rather, we are simply recentering (by adding 32) and rescaling (by multiplying by $\frac{9}{5}$) the distribution of the temperature in degrees Celsius. Crucially, once we know

[21] An equivalent geometric interpretation is that any given solution for a particular APC dataset lies on a multidimensional solution line.

the temperature in degrees Celsius, we know the temperature in Fahrenheit since it is a simple linear combination. In a similar way, regardless of the coding scheme for a set of APC data, we can express at least one variable as a linear relationship of the other variables.

2.2 The IE's Mathematical Constraint

The mathematical constraint imposed by the IE will not, generally speaking, recover the underlying data generating parameters. The reason for this is that the estimates produced by the IE and the true data generating parameters will equal each other only if the true data generating parameters happen to conform to the IE's specific mathematical constraint. We know of no reason why this should ever be the case in any particular substantive application. Specifically, the IE will yield the underlying data generating parameters $\mathbf{b}$ if $s = 0$ in Equation 15 (see Yang and Land 2013c: 82; Land et al. 2016: 966):

$$\begin{aligned} \mathbf{b}_{\text{IE}} &= \mathbf{b}, \quad \text{when } s = 0. \\ \mathbf{b}_{\text{IE}} &\neq \mathbf{b}, \quad \text{when } s \neq 0. \end{aligned} \tag{24}$$

Rearranging Equation 15 and taking the expectation, we know that $\mathbb{E}(\mathbf{b}_{\text{IE}}) = \mathbf{b} - s\mathbf{v}$. Thus, the expected value of the IE estimate will not equal the true data generating parameter unless $s = 0$. Constraining $s = 0$ is equivalent to assuming that $\mathbf{b}\mathbf{v}^T = 0$, since if $s = 0$ in Equation 52 then becomes $\mathbf{b}_{\text{IE}} = \mathbf{b}$ (see Luo 2013: 1951-1953). Hence, $\mathbf{b}_{\text{IE}}\mathbf{v}^T = \mathbf{b}\mathbf{v}^T = 0$. We call this the *linear dependency constraint*. Similar to representation of the linear dependency in Equation 23, we can write the IE linear dependency constraint as

$$\begin{aligned} &(\mathbf{b}_{\mathbf{A}_{\text{IE}}})\mathbf{v}_{\text{A}}^T - (\mathbf{b}_{\mathbf{P}_{\text{IE}}})\mathbf{v}_{\text{P}}^T + (\mathbf{b}_{\mathbf{C}_{\text{IE}}})\mathbf{v}_{\text{C}}^T \\ &= (\mathbf{b}_{\mathbf{A}})\mathbf{v}_{\text{A}}^T - (\mathbf{b}_{\mathbf{P}})\mathbf{v}_{\text{P}}^T + (\mathbf{b}_{\mathbf{C}})\mathbf{v}_{\text{C}}^T = 0 \end{aligned} \tag{25}$$

where $\mathbf{b}_{\mathbf{A}_{\text{IE}}}$, $\mathbf{b}_{\mathbf{P}_{\text{IE}}}$, and $\mathbf{b}_{\mathbf{C}_{\text{IE}}}$ are the estimated IE coefficients for age, period, and cohort; $\mathbf{b}_{\mathbf{A}}$, $\mathbf{b}_{\mathbf{P}}$, and $\mathbf{b}_{\mathbf{C}}$ are the corresponding true data generating parameters; and $\mathbf{v}_{\text{P}}$, $\mathbf{v}_{\text{P}}$, and $\mathbf{v}_{\text{C}}$ are the respective null vector elements (see Equations 19-21). Again, we know of no reason why the linear dependency constraint in Equation 25 should ever hold in any particular substantive application in the social or biomedical sciences.

2.3 Example: The IE's Mathematical Constraint

To illustrate the fundamental properties of the IE, including its underlying mathematical constraint, we simulated the APC dataset with $I = 3$ age groups, $J = 3$ period groups, and $I+J-1 = K = 5$ cohort groups. The design matrix for this simulated dataset is shown in Table 1. For the dataset in Table 1, the underlying data generating process is based on an age slope of $\alpha = 3$, period slope of $\pi = 1$, and cohort slope of $\gamma = 2$. We also include nonlinearities in the data generating process, reflecting the fact that over-time trends consist of both linear and nonlinear components. The true data generating process in terms of sum-to-zero effects is displayed in the second column of Table 2. To aid in explaining the IE, without loss of generality we simulated a small dataset without random error.

Table 1: APC Data with Sum-to-Zero Effect Coding

Person	age_1	age_2	period_1	period_2	cohort_1	cohort_2	cohort_3	cohort_4	Y
1	1	0	1	0	0	0	1	0	1.625
2	0	1	1	0	0	1	0	0	3.500
3	−1	−1	1	0	1	0	0	0	2.375
4	1	0	0	1	0	0	0	1	3.125
5	0	1	0	1	0	0	1	0	5.250
6	−1	−1	0	1	0	1	0	0	5.625
7	1	0	−1	−1	−1	−1	−1	−1	6.625
8	0	1	−1	−1	0	0	0	1	9.000
9	−1	−1	−1	−1	0	0	1	0	9.625

Notes: The data generating process for this dataset is $Y = 5.000 + 3.000(\text{age}_L) + 0.500(\text{age}^2) + 1.000(\text{period}_L) - 0.750(\text{period}^2) + 2.000(\text{cohort}_L) + 1.000(\text{cohort}^2) + 0.500(\text{cohort}^3) - 0.250(\text{cohort}^4)$, where the L subscripts denote linear contrasts and the superscripts denote higher-order orthogonal polynomial contrasts. Total sample size is $n = 9$.

Using the simulated dataset in Table 1, the zero-sum effects model we would like to run is

$$\begin{aligned} \mathbf{y} &= \mathbf{Xb} \\ &= \mu + \alpha_1(\text{age}_1) + \alpha_2(\text{age}_2) + \pi_1(\text{period}_1) + \pi_2(\text{period}_2) + \\ &\gamma_1(\text{cohort}_1) + \gamma_2(\text{cohort}_2) + \gamma_3(\text{cohort}_3) + \gamma_4(\text{cohort}_4) \end{aligned} \tag{26}$$

where the partial slopes give deviations from the overall mean. Unfortunately, however, we cannot use OLS to estimate Equation 26 because of the linear dependence in our dataset.

The linear dependence is not immediately obvious from observing Table 6, but it is nonetheless present in the data. With effect coding we can write the linear dependence as

$$\begin{aligned} \mathbf{Xv} &= \mathbf{Av}_{\text{A}} - \mathbf{Pv}_{\text{P}} + \mathbf{Cv}_{\text{C}} \\ &= \big[(\text{age}_1)(-1) + (\text{age}_2)(0)\big] - \\ &\big[(\text{period}_1)(-1) + (\text{period}_2)(0)\big] + \\ &\big[(\text{cohort}_1)(-2) + (\text{cohort}_2)(-1) + (\text{cohort}_3)(0) + (\text{cohort}_4)(1)\big] \\ &= \mathbf{0} \end{aligned} \tag{27}$$

s where as before **0** is a vector of zeros. Since a variable times zero is zero, after multiplying terms we can simplify the linear dependency in Equation 27 as:

$$(\text{age}_1)(-1) + (\text{period}_1)(1) + \big[(\text{cohort}_1)(-2) + (\text{cohort}_2)(-1) + (\text{cohort}_4)(1)\big] = \mathbf{0}. \tag{28}$$

If we remove any of the variables in Equation 28 from our dataset, then we can estimate the APC model since this is equivalent to assuming that the corresponding parameter is zero in the population. This reflects the fact that the parameters corresponding to zero elements in the null vector are identified.

In Table 2 we show how the IE estimates differ from the parameters of the true, unknown data generating process. To distinguish between statistical bias discussed elsewhere in this paper, we define bias* as *the difference between the IE estimates and the true data generating parameters.* If bias* $= 0$ this would mean that the IE has estimated the true underlying data generating parameters. Large values of bias* indicate that the IE has done a poor job of estimating the data generating parameters. Our belief is that in most circumstance bias* is what APC researchers are most interested in, not statistical bias as discussed elsewhere in this paper. In other words, researchers want to know how close their estimates are to the true underlying data generating parameters, not whether their estimator would on average across an infinite number of samples give the same values if it were applied to the whole population of data. The value of bias* is large for those parameters corresponding to non-zero null vector elements. However, because their null vector elements are zero, the intercept as well as the middle groups for age, period, and cohort are all estimated with zero bias*.

Table 2: Comparison of IE Estimates with
Data Generating Parameters for a Simulated Dataset

Parameter	True DGP ($\mathbf{b}$)	IE Estimates ($\mathbf{b}_{\text{IE}}$)	Bias* ($\mathbf{b}_{\text{IE}} - \mathbf{b}$)	Null Vector
μ	5.000	5.000	0.000	0
α_1	-3.250	-1.297	-4.547	-1
α_2	0.500	0.500	0.000	0
π_1	-0.625	-2.578	-1.953	1
π_2	-0.750	-0.750	0.000	0
γ_1	-4.600	-0.844	3.756	-2
γ_2	-1.300	0.578	1.878	-1
γ_3	0.000	0.500	0.000	0
γ_4	2.300	0.172	-2.128	1

Notes: DGP is shorthand for "data generating process." The DGP for the simulated dataset is $Y = 5.000 + 3.000(\text{age}_L) + 0.500(\text{age}^2) + 1(\text{period}_L) - 0.750(\text{period}^2) + 2.000(\text{cohort}_L) + 1.000(\text{cohort}^2) + 0.500(\text{cohort}^3) - 0.250(\text{cohort}^4)$, where the L subscripts denote linear contrasts and the superscripts denote higher-order orthogonal polynomial contrasts. Total sample size is $n = 9$.

As discussed previously, the IE produces estimates that conform to the linear dependency constraint. For example, taking the IE estimates from the third column of Table 2 and multiplying by the null vector we obtain the following relationship among the estimates:

$$\begin{aligned}
\mathbf{b}_{\mathrm{IE}}\mathbf{v}^T &= (\mathbf{b}_{\mathbf{A}_{\mathrm{IE}}})\mathbf{v}_{\mathrm{A}}^T - (\mathbf{b}_{\mathbf{P}_{\mathrm{IE}}})\mathbf{v}_{\mathrm{P}}^T + (\mathbf{b}_{\mathbf{C}_{\mathrm{IE}}})\mathbf{v}_{\mathrm{C}}^T \\
&= (\widehat{\alpha}_1)(-1) + (\widehat{\pi}_1)(1) + \big[(\widehat{\gamma}_1)(-2) + (\widehat{\gamma}_2)(-1) + (\widehat{\gamma}_4)(1)\big] \\
&= (-1.297)(-1) + (-2.578)(1) + \big[(-0.844)(-2) + (0.578)(-1) + (0.172)(1)\big] \\
&= 0
\end{aligned} \tag{29}$$

where $\mathbf{b}_{\mathbf{A}_{\mathrm{IE}}}$, $\mathbf{b}_{\mathbf{P}_{\mathrm{IE}}}$, and $\mathbf{b}_{\mathbf{C}_{\mathrm{IE}}}$ are the estimated IE coefficients for age, period, and cohort; $\mathbf{v}_{\mathrm{P}}$, $\mathbf{v}_{\mathrm{P}}$, and $\mathbf{v}_{\mathrm{C}}$ are the respective null vector elements; and the caret ($\widehat{\cdot}$) denotes the IE estimates of the individual parameters.

The IE's particular mathematical constraint assumes that the true, unknown age, period, and cohort parameters are also linearly dependent among each other such that:

$$\begin{aligned}
\mathbf{b}\mathbf{v}^T &= (\mathbf{b}_{\mathbf{A}})\mathbf{v}_{\mathrm{A}}^T - (\mathbf{b}_{\mathbf{P}})\mathbf{v}_{\mathrm{P}}^T + (\mathbf{b}_{\mathbf{C}})\mathbf{v}_{\mathrm{C}}^T \\
&= \alpha_1(-1) + \pi_1(1) + \big[\gamma_1(-2) + \gamma_2(-1) + \gamma_4(1)\big] \\
&= 0,
\end{aligned} \tag{30}$$

where $\mathbf{b}_{\mathbf{A}}$, $\mathbf{b}_{\mathbf{P}}$, and $\mathbf{b}_{\mathbf{C}}$ are the true data generating parameters for the age, period, and cohort columns in the design matrix of zero-sum effects. The IE would only estimate the underlying data generating parameters if those parameters happen to satisfy Equation 30. To reiterate, there is no reason that this should ever be the case. For example, the data generating parameters in our example do not satisfy this constraint, differing substantially from the IE estimates. This can be seen by the large levels of bias* of the IE estimates in Table 2.[22] Researchers who use the IE in the hopes of uncovering the true data generating parameters are unlikely to attain their goal. If using the IE, one should state explicitly why it is believed that the model parameters in the population have the same relationship among each other as their corresponding columns in the design matrix. In general, we suspect that this will be difficult if not impossible to justify.

3 Strengths of the IE

As noted above and pointed out by Fu and his associates in multiple publications (Fu 2016; Fu, Land, and Yang 2011; Fu 2000; Yang, Fu, and Land 2004; Fu and Hall 2006; Yang, Schulhofer-Wohl, et al. 2008), the IE has several desirable statistical properties.[23] First, it is an estimable function, meaning that it produces a unique set of estimates for the effects of age, period and cohort (Yang, Schulhofer-Wohl, et al. 2008: 1708-1710; Fu, Land, and Yang 2011: 456-458; Yang

[22] The parameters with corresponding zero null vector elements are identifiable, so only those estimates with non-zero null vector elements exhibit bias* in Table 2.

[23] Like any constrained least-squares estimator of the C-APC model, the IE also correctly estimates the intercept, which gives the overall mean in the data, and the nonlinearities, which provide crucial information on temporal shifts or turning points in the outcome.

and Land 2013c: 84-85). Second, it is unbiased, meaning that the average of IE estimates over an infinite number of simple random samples will be equal to the IE values when it is applied to the full population data (Yang, Fu, and Land 2004: 101-102; 107; Yang, Schulhofer-Wohl, et al. 2008: 1709; Yang and Land 2013c: 86; 115-116). Finally, it has minimum variance among all possible estimators based on the specific zero-sum effect-coded design matrix being used (Yang, Fu, and Land 2004: 102-103; 108; Yang, Schulhofer-Wohl, et al. 2008: 1709; Yang and Land 2013c: 86; 116-117).

To say that a function is estimable means that when applied to data, it produces a unique set of estimates. Intuitively, it means that it is possible for the data to tell us what the function is equal to. Thus, if we have variables that are linearly dependent, then the standard OLS function is not estimable, because it is based on a regular inverse which is not well-defined (i.e., it does not exist) because of the linear dependence. The IE is estimable because the Moore-Penrose generalized inverse is well-defined (i.e., it exists) even when our variables are linearly dependent. In particular, it imposes a specific mathematical constraint on the estimates that will only be satisfied by a single point. How is this possible?

The IE estimate will equal that point on the solution line that is nearest the origin. That is, the IE is that particular set of estimates produced by the IE under the Euclidean distance metric (or, equivalently, those estimates corresponding to the minimum L_2 norm). The IE estimates are unique and estimable in the very specific sense that for a particular design matrix and outcome there is a set of points that are closest in Eculidean distance to the origin on the solution line. This implies that the IE is that set of estimates that minimizes the sum of the squared parameter estimates, or $(\mathbf{b} + s\mathbf{v})^T(\mathbf{b} + s\mathbf{v})$. This is the mathematical constraint that the IE imposes in order to achieve a set of estimates.

The second claim about the IE is that it produces estimates that are unbiased. A function is unbiased if when it is calculated for an infinite number of random samples, its average is equal to its value when it is calculated for on the population as a whole. In the context of the IE this means that if we had an infinite number of samples and applied the IE in each, the average of the estimates from the IE across these samples would be equal to the estimates produced by the IE using the whole population. In other words, the average estimates from the IE across an infinite number of samples will give the same point on the solution line as being the closest point to the origin as applying the IE to the data from the entire population.

Where researchers seem to become confused is that in saying a function is unbiased is equivalent to saying that it is an unbiased function for some specific property of the population. To clarify this point consider the following simple example. The function five times the mean of a variable is an unbiased estimator. If we were to calculate it over an infinite number of samples and take its average it will equal to five times the mean in the population. Formally, this is because five times the mean is a linear function. Obviously, five times the mean is not an unbiased estimator for the mean in the population itself. The fact that IE is an unbiased estimator (meaning its average across an infinite number of samples is equal to its value calculated in the population) does *not* mean that the IE provides an unbiased estimate of the parameters of the underlying model that generated the data.

The third property of the IE is that its estimates have minimum variance among the estimators based on the design matrix it uses. The minimum variance property is a consequence of being that point on the solution line that is closest to the origin. All the points on the solution line give the same predicted value for the outcome as the IE. As such, all estimates on the solution line give predicted values with the same variance. The IE, however, gives that set of estimates whose *parameters* have the smallest variance.

The statistical properties of the IE are not unique to it. Rather they are the result of its use of the Moore-Penrose inverse, not its definition in terms of a design matrix using zero-sum effect coding.[24] There are numerous design matrices one might use, with reference group coding being a common alternative to zero sum effect coding. Use of alternative design matrices with the Moore-Penrose inverse will also produce estimates, like the IE, that are estimable, unbiased, and minimum variance relative to other estimators based on that design matrix (Fu 2016). Thus the statistical properties of the IE are not unique to it. As shown by Luo et al. (2016) different design matrices can result in quite different estimates of the parameter effects. To be precise, since the nonlinear effects are identified, different design matrices will potentially produce different estimates of the linear trends, even trends that differ in sign from those of the IE. The appendix to Luo (2016) in fact shows that there is always some estimator that will produce any point on the solution line. In other words, one can choose that set of preferred parameter values that falls on the solution line and an estimator can be found that will produce those values.

4 Weaknesses of the IE

However, there are three main weaknesses with the IE. First, because the IE constraint is based on zero-sum effect coding, the constraint obscures the fact that it is only the linear components that are constrained. In this particular sense the IE's constraint is "hidden." Second, the IE is sensitive to the number of age, period, and cohort groups in the data (see Land et al. 2016: 963). With a different coding scheme the linear dependency in the data changes, reflecting a different number of groups. This can radically alter the estimated temporal effects with the IE. Finally, the estimates of the linear trends produced by IE are sensitive to the size and sign of the nonlinearities in the data.

4.1 The General Form of the IE Constraint

So far we have discussed the IE's mathematical constraint in terms of the zero-sum effect coding. However, this conceals the underlying nature of the constraint, since the estimate for each group is a

[24] This can be seen by simply examining the various proofs showing that the IE has these properties and noting that none depend on using a zero-sum effects-coded design matrix (for example, see Fu 2000: 263-268; 276-277; Yang, Fu, and Land 2004: 107-108; Yang and Land 2013c: 75-123).

combination of constrained linear components and identified nonlinearities. To clarify the nature of the IE constraint, it is useful to decompose each zero-sum effect into its linear and nonlinear components using orthogonal polynomials (see Mason and Fienberg 1985: 233-237).

Table 3: Decomposition of Sum-to-Zero Effects

		Decomposed Effect			
Sum-to-Zero Effect		Linear Component		Nonlinear Component	Null Vector Element
α_1	$=$	$(-1)\alpha$	$+$	$(1)\alpha^2$	-1
α_2	$=$	$(0)\alpha$	$+$	$(-2)\alpha^2$	0
π_1	$=$	$(-1)\pi$	$+$	$(1)\pi^2$	1
π_2	$=$	$(0)\pi$	$+$	$(-2)\pi^2$	0
γ_1	$=$	$(-2)\gamma$	$+$	$(1)\gamma^2 + (-1)\gamma^3 + (1)\gamma^4$	-2
γ_2	$=$	$(-1)\gamma$	$+$	$(-1)\gamma^2 + (2)\gamma^3 + (-4)\gamma^4$	-1
γ_3	$=$	$(0)\gamma$	$+$	$(-2)\gamma^2 + (0)\gamma^3 + (6)\gamma^4$	0
γ_4	$=$	$(1)\gamma$	$+$	$(-1)\gamma^2 + (-2)\gamma^3 + (-4)\gamma^4$	1

Notes: Decomposition based on $I = 3$ age, $J = 3$ period, and $K = I + J - 1 = 5$ cohort groups. The IE constraint derived from this table is $\alpha - \pi + 6\gamma - \alpha^2 + \pi^2 - 4\gamma^2 - 2\gamma^3 - 2\gamma^4 = 0$.

In Table 3 we decompose the effects for the simulated dataset in Tables 1 and 2. To express the IE constraint in terms of the linear and nonlinear components, we simply replace each effect with its decomposed counterpart, multiply the decomposed effects by the null vector, and rearrange terms. After we do these manipulations, we obtain the constraint

$$\alpha - \pi + 6\gamma = \alpha^2 - \pi^2 + 4\gamma^2 + 2\gamma^3 + 2\gamma^4 \tag{31}$$

for the dataset in Table 3. More generally, we can express the IE mathematical constraint as:

$$\omega_1\alpha - \omega_2\pi + \omega_3\gamma = \nu \tag{32}$$

where ω_1, ω_2, and ω_3 are weights for the age, period, and cohort slopes; and ν is a scalar, since the nonlinearities are identified. For example, based on our simulated dataset $\omega_1 = 1, \omega_2 = 1, \omega_3 = 6$, and $\nu = \alpha^2 - \pi^2 + 4\gamma^2 + 2\gamma^3 + 2\gamma^4$.

The general form of the IE constraint in Equation 32 reveals that the IE's estimates of the slopes vary depending on two aspects of the data: first, the number of APC groups in the dataset, which alter the weights for α, π, and γ as well as the value of the scalar ν; second, the size and direction of the nonlinearities, which shift the value of ν.[25] If there are zero nonlinearities in the dataset, then

[25] Note also that the IE constraint depends on the fact that $\text{age} - \text{period} + \text{cohort} = 0$, which is why the sign of

$\nu = 0$. Otherwise, the IE constrains the weighted sum of the slopes to equal some other arbitrary value of ν. As we show in the following sections, because the IE's constraint differs depending on the number of APC categories as well as the extent of nonlinearities, not only is it unlikely to recover the true data generating parameters in any particular application, but its underlying constraint is complicated, non-intuitive, and highly variable.

4.2 Sensitivity of Linear Estimates to Number of Groups

The general IE constraint in Equation 32 tells us that, mathematically, the IE is sensitive to the number of APC groups. Table 4 shows how the IE's estimates of the slopes changes as the number of period groups increases from $J = 3$ to $J = 1000$.[26] For all simulations in Table 4 we keep the data generating process the same at $Y = 3.000 - 2.000(\text{age}_L) + 4.000(\text{period}_L) + 1.000(\text{cohort}_L)$, where the subscripts denote the linear components. For simplicity we keep the number of age groups constant at $I = 3$ as we increase the number of period groups (and, accordingly, the number of cohorts). A different number of age groups does not alter our findings regarding the sensitivity of the IE to the number of APC groups.

We purposely constructed the data generating process so that it initially conforms to the IE's constraint. With $I = 3$ age groups and $J = 3$ period groups, as well as no nonlinearities, $\omega_1 = 1$, $\omega_2 = 1, \omega_3 = 6$, and $\nu = 0$ in Equation 32. By design we picked values values of α, π, and γ so that $\alpha - \pi + 6\gamma = -2 - (4) + 6(1) = 0$. This is indicated in the first row of Table 4, which is shaded.

As we increase the number of period groups, the values of the weights change thereby altering the IE's mathematical constraint. The IE constraints are shown in last column of Table 32. Since the data generating process contains no nonlinearities and remains the same as we increase the number of groups, the value of ν for each constraint is zero. However, the values of the ω's for the period and cohort slopes increase as we increase the number of period (and cohort) groups.[27] Although our initial dataset satisfies the IE constraint, as the number of period groups increases the IE estimates diverge dramatically from the true data generating parameters. In this particular case as we increase the number of period groups the IE's estimates of the age slope increases, the period slope decreases, and the cohort slope increases. The reason is that, as the number of period and cohort groups increase, their values of ω become very similar. With $J = 1000$ period groups, the weights for the period and cohort slopes are approximately the same.

The last column of Table 4 underscores the complicated, non-intuitive, and highly variable nature of the IE constraint, as well as the extreme difficulty of interpreting it substantively. For example, with $J = 10$ groups, the constraint in Table 4 is $\alpha - \left(\frac{249}{4}\right)\pi + \left(\frac{451}{4}\right)\gamma = 0$, with weights of $\omega_1 = 1, \omega_2 = \left(\frac{249}{4}\right)$, and $\omega_3 = \left(\frac{451}{4}\right)$. However, with $J = 15$ groups the IE constraint becomes $\alpha - 231\pi + 344\gamma$, with weights of $\omega_1 = 1$, $\omega_2 = 231$, and $\omega_3 = 344$. It is exceedingly difficult,

the period slope in Equation 32 differs from that for age and cohort.

[26] This reflects the fact that in actual studies conducted over time, the number of age groups are constant due to the limits of human lifespan but the number of period (and cohort) groups will increase.

[27] The weight for the age slope would similarly increase if we increased the number of age groups across the simulations.

if not impossible, to muster a reason these specific values of ω have any theoretical importance or relevance.

Table 4: Sensitivity of the IE to Number of Period (and Cohort) Groups

J Groups	α_{IE}	π_{IE}	γ_{IE}	IE Constraint
3	-2.000	4.000	1.000	$\alpha - \pi + 6\gamma \;=\; 0$
5	-1.731	3.731	1.269	$\alpha - 6\pi + 19\gamma \;=\; 0$
8	-1.368	3.368	1.632	$\alpha - \left(\frac{119}{4}\right)\pi + \left(\frac{249}{4}\right)\gamma \;=\; 0$
10	-1.214	3.214	1.786	$\alpha - \left(\frac{249}{4}\right)\pi + \left(\frac{451}{4}\right)\gamma \;=\; 0$
15	-0.990	2.990	2.010	$\alpha - 231\pi + 344\gamma \;=\; 0$
25	-0.798	2.798	2.202	$\alpha - 1156\pi + 1469\gamma \;=\; 0$
50	-0.650	2.650	2.350	$\alpha - \left(\frac{39{,}249}{4}\right)\pi + \left(\frac{44{,}251}{4}\right)\gamma \;=\; 0$
100	-0.575	2.575	2.425	$\alpha - \left(\frac{323{,}499}{4}\right)\pi + \left(\frac{343{,}501}{4}\right)\gamma \;=\; 0$
500	-0.515	2.515	2.485	$\alpha - (10,354,375)\pi + (10,479,375)\gamma \;=\; 0$
1000	-0.507	2.507	2.493	$\alpha - (83,083,750)\pi + (83,583,750)\gamma \;=\; 0$
DGP	$\alpha = -2.000$	$\pi = 4.000$	$\gamma = 1.000$	

Notes: DGP is shorthand for "data generating process." The DGP for the linear components is $Y = 3.000 - 2.000(\text{age}_L) + 4.000(\text{period}_L) + 1.000(\text{cohort}_L)$. Number of age groups is set at $I = 3$ for all simulations. Sample size for each simulation is $n = 100 \times (I \times J)$. Shaded row indicates that the IE constraint is satisfied for that particular simulated dataset. Due to rounding some IE constraints displayed here will not equal zero exactly.

4.3 Dependence of Linear Estimates on Size and Direction of Nonlinearities

The general form of the IE's mathematical constraint also clarifies that the IE's estimates depend on the size (e.g., large or small in absolute value) and sign (e.g., positive or negative) of the nonlinearities. For example, using the same data generating process as in the previous section, Table 5 shows how the IE's estimates alter depending on the magnitude and direction of the age nonlinearity. For all simulations we keep the number of age and period groups at $I = 3$ and $J = 3$, respectively. Again, since we specifically constructed a data generating process that conforms to the IE constraint when there are $I = 3$ age groups and $J = 3$ period groups as well as zero nonlinearities, the IE indeed recovers the true slopes when the age nonlinearity is zero. This row is shaded in Table 5.

Since the number of age groups is set at $I = 3$ for all simulations in Table 5, the age weight is set at $\omega_1 = 1$ and the value of ν changes directly with the age nonlinearity. As the age nonlinearity becomes more positive, the age and cohort slopes move towards positive infinity on the real number line, while the period slope moves towards negative infinity. In contrast, as the age nonlinearity becomes more negative, the age and cohort slopes move towards negative infinity on the real number line, while the period slope moves towards negative infinity. For the same underlying age, period, and cohort linear effects in the population, the IE will give radically different estimates of the slopes depending on the nonlinear effects.

Table 5: Sensitivity of the IE to Age Nonlinearities

α^2	α_{IE}	π_{IE}	γ_{IE}	IE Constraint		
-20.000	-4.500	6.500	-1.500	$\alpha - \pi + 6\gamma$	$=$	-20.000
-10.000	-3.250	5.250	-0.250	$\alpha - \pi + 6\gamma$	$=$	-10.000
-5.000	-2.625	4.625	0.375	$\alpha - \pi + 6\gamma$	$=$	-5.000
-1.000	-2.125	4.125	0.875	$\alpha - \pi + 6\gamma$	$=$	-1.000
-0.500	-2.062	4.062	0.938	$\alpha - \pi + 6\gamma$	$=$	-0.500
-0.250	-2.031	4.031	0.969	$\alpha - \pi + 6\gamma$	$=$	-0.250
-0.050	-2.006	4.006	0.994	$\alpha - \pi + 6\gamma$	$=$	-0.050
0.000	-2.000	4.000	1.000	$\alpha - \pi + 6\gamma$	$=$	0.000
0.050	-1.994	3.994	1.006	$\alpha - \pi + 6\gamma$	$=$	$+0.050$
0.250	-1.969	3.969	1.031	$\alpha - \pi + 6\gamma$	$=$	$+0.250$
0.500	-1.938	3.938	1.063	$\alpha - \pi + 6\gamma$	$=$	$+0.500$
1.000	-1.875	3.875	1.125	$\alpha - \pi + 6\gamma$	$=$	$+1.000$
5.000	-1.375	3.375	1.625	$\alpha - \pi + 6\gamma$	$=$	$+5.000$
10.000	-1.075	2.750	2.250	$\alpha - \pi + 6\gamma$	$=$	$+10.000$
20.000	0.500	1.500	3.500	$\alpha - \pi + 6\gamma$	$=$	$+20.000$
DGP	$\alpha = -2.000$	$\pi = 4.000$	$\gamma = 1.000$			

Notes: DGP is shorthand for "data generating process." The DGP for the linear components is $Y = 3 - 2.000(\text{age}_L) + 4.000(\text{period}_L) + 1.000(\text{cohort}_L)$. For all simulations, number of age, period, and cohort groups is set at $I = 3$, $J = 3$, and $K = I + J - 1 = 5$, respectively. Sample size for each simulation is $n = 1000 \times (I \times J) = 9,000$. Shaded row indicates that the IE constraint is satisfied for that particular simulated dataset.

The last column of in Table 5 again brings to light the complex and variable nature of the IE constraint. For instance, with $\alpha^2 = -1.000$ the constraint in Table 5 is $\alpha - \pi + 6\gamma = -1.000$ but with $\alpha^2 = +1.000$ the constraint becomes $\alpha - \pi + 6\gamma = +1.000$. Again, it is very difficult, if not impossible, to give a theoretical reason why this particular linear combination of the temporal slopes in the population should equal $\nu = +1.000$ rather than $\nu = -1.000$ simply because the quadratic age trend is positive one rather than negative one.

5 An Alternative: The Orthogonal Estimator (OE)

In light of the weaknesses of the IE, we propose the OE as an alternative estimator. The principal advantage of the OE over the IE is that the OE's identifying constraint will always have the same transparent form regardless of the number of APC groups or the form of the nonlinearities. This will in turnfoster the accumulation of knowledge and, we hope, the development of explicit theoretically- and empirically-based constraints by applied researchers.

Like the IE, the OE is based on the Moore-Penrose generalized inverse and applies sum-to-zero constraints on the age, period, and cohort parameters. However, unlike the IE, the OE decomposes each zero-sum effect into its linear and nonlinear components, resulting in a set of coefficients and statistical inferences for the identifiable nonlinearities. The OE's coefficients for the linear effects have the desirable property of transparency, with an easily-interpretable, invariant identifying constraint. In contrast to the IE, the estimated linear effects of the OE are not affected by the number of APC groups or the nature of the nonlinearities in the underlying data generating process. Finally, unlike the OE, the linear effects from the OE are always a simple average of ZLT models in which each of the age, period, or cohort slopes is set to zero.

5.1 OE's Mathematical Constraint

Due to the partitioning of the linear and nonlinear components, the linear dependency in the OE design matrix is always $\mathbf{X}_\mathrm{O}\mathbf{v}_\mathrm{O} = \mathrm{age}_L - \mathrm{period}_L + \mathrm{cohort}_L = \mathbf{0}$, where age_L, period_L, and cohort_L refer to the linear components in the OE design matrix. That is, the OE always has a null vector of the form:

$$\mathbf{v}_\mathrm{O} = (0, 1, -1, 1, 0 \ldots, 0) \tag{33}$$

where the first zero corresponds to the intercept; the elements positive one, negative one, and positive one correspond to the age, period, and cohort linear components; and the remaining zeros correspond to the $(I-2)+(J-2)+(K-2)$ nonlinear components. Adding more period groups or altering the size and sign of the nonlinearities will not affect the form of the linear dependence of the OE design matrix; that is, the null vector will always have the form in Equation 33. As a result, the OE's mathematical constraint will always be

$$\alpha - \pi + \gamma = 0 \tag{34}$$

regardless of the number of APC categories or the nature of the nonlinearities. In reference to the general form of the IE constraint in Equation 32, the OE always sets the ω's to one and ν to zero. This mathematical constraint on the data generating parameters is equivalent to assuming that $s = 0$ in the three-dimensional solution line given by the equation $\mathbf{b} = \mathbf{b}_\mathrm{OE} + s\mathbf{v}_\mathrm{O}$.[28]

5.2 The OE as a Simple Average

Interpreting the OE is facilitated by the fact that it equals the simple average of three ZLT models in which the linear effect of age, period, or cohort is fixed to zero (or, equivalently, its corresponding column in the design matrix is dropped). To illustrate this unique property of the OE, we simulated data with values of $\alpha = 3$, $\pi = 1$, and $\gamma = 2$ as well as a full set of nonlinearities for age, period,

[28] In contrast to the OE, the IE will generally set s to some other value and thereby locate the IE on some other point on the three-dimensional solution line.

and cohort.[29] The estimates from the three ZLT models as well as their simple average are shown in Table 6.

Table 6: Estimates from Three ZLT Models

	Slopes		
Variable Dropped	$\widehat{\alpha}$	$\widehat{\pi}$	$\widehat{\gamma}$
age_L	0	4	-1
period_L	4	0	3
cohort_L	1	3	0
Simple Average	$5/3$	$7/3$	$2/3$

Notes: The data generating process for this dataset is $Y = 5+3(\text{age}_L)+0.5(\text{age}^2)+1(\text{period}_L)-0.75(\text{period}^2)+2(\text{cohort}_L)+1(\text{cohort}^2)+0.5(\text{cohort}^3) - 0.25(\text{cohort}^4)$, where the L subscripts denote linear contrasts and the superscripts denote higher-order orthogonal polynomial contrasts. Total sample size is $n = 10,000$.

The rows of Table 6 display the estimated slopes for the different ZLT models. For example, as shown in the first row of the table, if we drop the age linear component (age_L) then we obtain slope estimates of zero, four, and negative one for age, period, and cohort, respectively. Likewise, as shown in the second and third rows, by dropping the period linear component (period_L) we obtain estimates of four, zero, three and by dropping the cohort linear component (cohort_L) we obtain estimates of one, three, and zero. In other words, the three ZLT models give us three naive estimates for each of the temporal slopes. Using these results we can calculate three simple averages

$$\begin{aligned} (0+4+1)\times(\tfrac{1}{3}) &= \tfrac{5}{3} &&= \widehat{\alpha}_{\text{OE}} \\ (4+0-3)\times(\tfrac{1}{3}) &= \tfrac{7}{3} &&= \widehat{\pi}_{\text{OE}} \\ (-1+3+0)\times(\tfrac{1}{3}) &= \tfrac{2}{3} &&= \widehat{\gamma}_{\text{OE}}, \end{aligned} \tag{35}$$

which equal the slope estimates produced by the OE. Formally, the OE is a convex combination (or weighted linear combination) of the estimated linear effects from the three ZLT models.[30] That is, we can express the OE estimates as

$$\begin{aligned} \widehat{\alpha}_1\lambda_1 + \widehat{\alpha}_2\lambda_2 + \widehat{\alpha}_3\lambda_3 &= \widehat{\alpha}_{\text{OE}} \\ \widehat{\pi}_1\lambda_1 + \widehat{\pi}_2\lambda_2 + \widehat{\pi}_3\lambda_3 &= \widehat{\pi}_{\text{OE}} \\ \widehat{\gamma}_1\lambda_1 + \widehat{\gamma}_2\lambda_2 + \widehat{\gamma}_3\lambda_3 &= \widehat{\gamma}_{\text{OE}} \end{aligned} \tag{36}$$

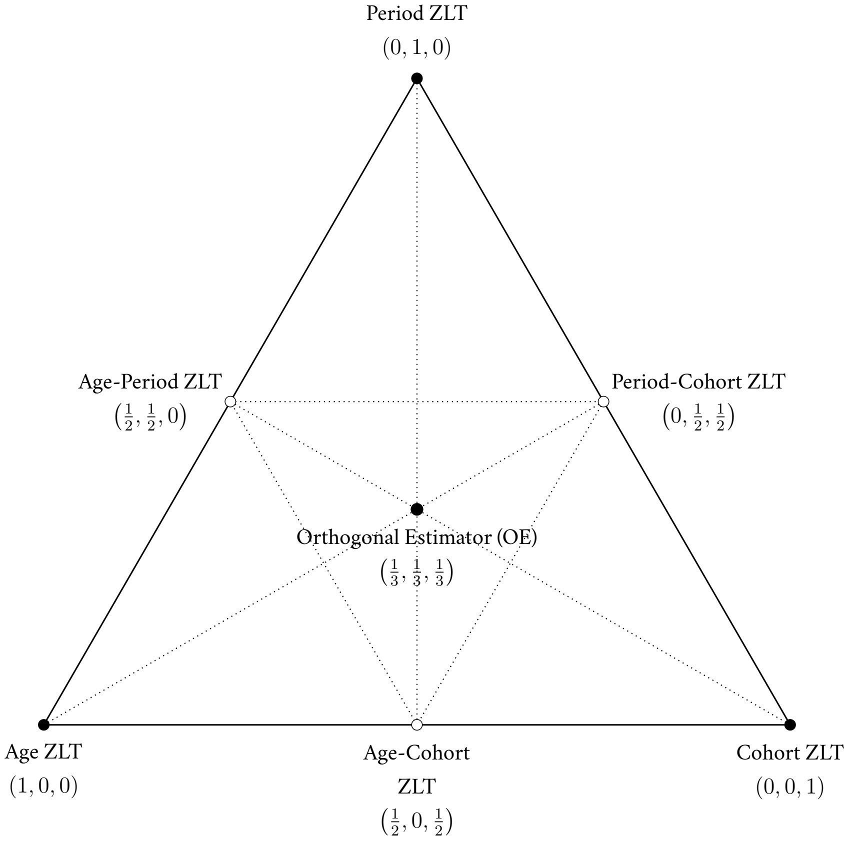
where the subscripts 1, 2, and 3 refer to the ZLT models for age, period, and cohort, respectively,

[29] Since the OE partitions the linear from the nonlinear effects, and the nonlinear effects are identifiable, the OE is a simple average of only the linear effects.

[30] In the appendix we show that the IE is also a convex combination of simpler models, and thus can also be interpreted as a weighted average of the estimates from a set of drop-one-variable models. However, the IE will not in general equal a simple average of three ZLT models.

with weights $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$.

Figure 2: OE and Related Models in Barycentric Coordinates


Period ZLT
$(0, 1, 0)$
Age-Period ZLT
$(\frac{1}{2}, \frac{1}{2}, 0)$
Period-Cohort ZLT
$(0, \frac{1}{2}, \frac{1}{2})$
Orthogonal Estimator (OE)
$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$
Age ZLT
$(1, 0, 0)$
Age-Cohort
ZLT
$(\frac{1}{2}, 0, \frac{1}{2})$
Cohort ZLT
$(0, 0, 1)$

Notes: Each point in the triangle (or two-simplex) is given by barycentric coordinates $(\lambda_1, \lambda_2, \lambda_3)$, with $\lambda_1 + \lambda_2 + \lambda_3 = 1$. Any particular slope estimate a is a convex combination of the vertex estimates a_1, a_2, and a_3, which correspond to the slope estimates from the ZLT age, period, and cohort models.

The fact that we can express the OE in terms of three simpler models suggests an informative geometric representation of the OE in terms of (normalized) barycentric coordinates. With barycentric coordinates, any particular point in a triangle (or two-simplex) is expressed as a weighted linear combination of the values of the vertices. These weights are the barycentric coordinates of the point. Specifically, let each vertex of a triangle have a value of a_1, a_2, or a_3. The position of any given point a within the triangle is a weighted linear combination of the values of the vertices:

$$a_1\lambda_1 + a_2\lambda_2 + a_3\lambda_3 = a \tag{37}$$

where the weights sum to one such that $\lambda_1 + \lambda_2 + \lambda_3 = 1$. Thus, once we know the values of the vertices (a_1, a_2, a_3), we can locate any point a within the triangle using barycentric coordinates (λ_1, λ_2, λ_3).In Figure 2 we display the OE and related estimators in terms of barycentric coordinates. Each vertex in Figure 2 corresponds to an estimate from a ZLT model and every point within the triangle equals some weighted linear combination of the three ZLT models. For example, the barycentric coordinates for the ZLT age model are $(1, 0, 0)$ and those for the ZLT period model are $(0, 1, 0)$. Likewise, as shown in Figure 2, the OE is the centroid of the triangle with barycentric coordinates of $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, thereby giving equal weight to each of the ZLT models.[31]

Intriguingly, the OE can be seen as a starting point for considering more realistic, theoretically-informed models by altering the barycentric coordinates in the triangle.[32] For example, suppose a researcher has strong theoretical grounds to prefer the ZLT period model and ZLT age model equally. The barycentric coordinates for such an age-period ZLT hybrid model are $(\frac{1}{2}, \frac{1}{2}, 0)$, which gives equal weight to the age and period ZLT models and no weight to the cohort ZLT model. Using the values of the vertices (i.e., the estimates from the three ZLT models), we obtain the slope estimates

$$\begin{array}{rcllll}
\widehat{\alpha}_1\lambda_1 + \widehat{\alpha}_2\lambda_2 + \widehat{\alpha}_3\lambda_3 & = & (0)(\frac{1}{2}) + (4)\frac{1}{2} + (1)(0) & = 2 & = \widehat{\alpha}^* \\
\widehat{\pi}_1\lambda_1 + \widehat{\pi}_2\lambda_2 + \widehat{\pi}_3\lambda_3 & = & (4)(\frac{1}{2}) + (0)\frac{1}{2} + (3)(0) & = 2 & = \widehat{\pi}^* \\
\widehat{\gamma}_1\lambda_1 + \widehat{\gamma}_2\lambda_2 + \widehat{\gamma}_3\lambda_3 & = & (-1)(\frac{1}{2}) + (3)\frac{1}{2} + (0)(0) & = 1 & = \widehat{\gamma}^*
\end{array} \qquad (38)$$

where the asterisk denotes the slopes estimates from the ZLT age-period model with weights (or barycentric coordinates) of $\lambda_1 = \frac{1}{2}$, $\lambda_2 = \frac{1}{2}$, and $\lambda_3 = 0$.

The OE is a natural starting point for applied researchers focused on understanding temporal effects, but it is not the end point. When using the OE, we recommend the following three-step procedure. First, one should estimate the OE model, separately reporting the linear and nonlinear effects. By presenting the actual estimates of the linear and nonlinear effects, other researchers can know which components of the temporal trends are identifiable and which are not. Second, the ZLT models for age, period, and cohort should be estimated, with the linear and nonlinear effects again reported separately. Finally, theory or additional data should be used to justify the weights applied to the three ZLT models. If theory or additional information suggests the OE is invalid (that is, unlikely to represent the true data generating slopes), then one should use a different set of weights on the ZLT models or weigh a different set of extremal models.[33] As with the IE, it is unlikely that the OE will happen to equal the true data generating parameters. However, by following the three-step procedure outlined above researchers will produce a range of estimates that can be transparently communicated to other researchers, who can then use these estimates to develop more realistic models of temporal effects.

[31] In Cartesian coordinates the OE corresponds to the point closest to the origin in terms of Euclidean distance for the OE's particular design matrix and outcome. In barycentric coordinates the OE corresponds to the geometric centroid of a triangle defined by the three ZLT models.

[32] Note, however, that the true data generating slopes may lie outside the triangle. This suggests a different of extremal models should be used.

[33] For example, one could rotate the ZLT models so that they represent another set of extremal models on the solution line. We show how this can be done in the following section.

5.3 The IE as a Rotated Average of ZLT Models

Using the OE, we can clarify several important properties of the IE. As discussed in the previous section, the OE correctly identifies the nonlinearities and applies a constraint on the linear components. With the OE and the transformation matrix, we can demonstrate that IE also applies a constraint on the linear components and identifies the nonlinearities. The solution line for the design matrix $\mathbf{X}_{\text{Ortho}}$ and outcome $\mathbf{y}$ is $\mathbf{b} = \mathbf{b}_{\text{OE}} + s\widehat{\mathbf{v}}_{\text{OE}}$, where $\mathbf{b}$ are the true, unknown data generating parameters expressed as orthogonal polynomial contrasts. All we know from the data is that these true parameters lie somewhere on a line in three dimensional space, since s can take on any real number. Thus, any estimate must entail an additional constraint, which is equivalent to setting a value for s. The OE is that particular set of estimates that minimizes the squared length of $\big(\mathbf{b}_{\text{OE}} + s\widehat{\mathbf{v}}_{\text{O}}\big)^T\big(\mathbf{b}_{\text{OE}} + s\widehat{\mathbf{v}}_{\text{O}}\big)$, which is when $s = 0 = \widehat{\mathbf{v}}_{\text{O}}^T\mathbf{b}_{\text{OE}}$.

Recall that the main difference between the IE (as defined by Land and colleagues) and the OE is that the former uses a design matrix of sum-to-zero effect contrasts, which is a transformation of the solution line defined previously. Specifically, the solution line for the design matrix $\mathbf{X}_{\text{Effect}}$ and same outcome $\mathbf{y}$ is $\mathbf{Tb} = \mathbf{T}(\mathbf{b}_{\text{OE}} + s\widehat{\mathbf{v}}_{\text{O}})$. Let $\mathbf{M} = \mathbf{T}^T\mathbf{T}$. The squared distance of $\mathbf{Tb}$ from the origin for the transformed solution line is

$$\mathbf{b}^T\mathbf{M}\mathbf{b} = \big(\mathbf{b}_{\text{OE}} + s\widehat{\mathbf{v}}_{\text{O}}\big)^T\mathbf{M}\big(\mathbf{b}_{\text{OE}} + s\widehat{\mathbf{v}}_{\text{O}}\big), \tag{39}$$

which is at its minimum when

$$s^* = -\frac{\Big(\widehat{\mathbf{v}}_{\text{O}}^T\mathbf{M}\mathbf{b}_{\text{OE}}\Big)}{\Big(\widehat{\mathbf{v}}_{\text{O}}^T\mathbf{M}\widehat{\mathbf{v}}_{\text{O}}\Big)} \tag{40}$$

where s^* is the scalar corresponding to the IE estimates. That is, the IE estimates expressed as orthogonal polynomial contrasts is:

$$\mathbf{b}_{\text{IE}} = \mathbf{b}_{\text{OE}} + s^*\widehat{\mathbf{v}}_{\text{O}}. \tag{41}$$

Because the OE null vector only has non-zero elements for the linear components, the IE is simply a different location on the solution line defined by the estimable combinations of the age, period, and cohort slopes. That is, the IE and OE only differ because of the non-zero elements in $\widehat{\mathbf{v}}_{\text{O}}$, which correspond to the linear components, and both estimators will recover the equivalent set of nonlinearities.

We can use the OE to also show that the IE can be interpreted as a simple average of three extremal models, but these will not equal the three ZLT models unless $s^* = 0$. Recall that the OE is a simple average of three different ZLT models, each setting the age, period, or cohort slope to zero, since these are the only non-zero elements in the OE null vector. The IE is the simple average of a rotated set of ZLT models, where the estimates are shifted by s^*:

$$\begin{aligned}(\widehat{\alpha}_1 + s^*)\lambda_1 + (\widehat{\alpha}_2 + s^*)\lambda_2 + (\widehat{\alpha}_3 + s^*)\lambda_3 &= \widehat{\alpha}_{\text{IE}} \\ (\widehat{\pi}_1 - s^*)\lambda_1 + (\widehat{\pi}_2 - s^*)\lambda_2 + (\widehat{\pi}_3 - s^*)\lambda_3 &= \widehat{\pi}_{\text{IE}} \\ (\widehat{\gamma}_1 + s^*)\lambda_1 + (\widehat{\gamma}_2 + s^*)\lambda_2 + (\widehat{\gamma}_3 + s^*)\lambda_3 &= \widehat{\gamma}_{\text{IE}}\end{aligned} \tag{42}$$

where the subscripts 1, 2, and 3 refer to the ZLT models for age, period, and cohort, respectively, with weights $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$. In other words, the IE is based on three extremal models, but it fixes each of the age, period, and cohort slopes to a non-zero value given by s^* instead of zero. This is a convenient mathematical result. However, providing a theoretical or substantive justification for any particular value of s^* imposed by the IE is likely to be difficult if not impossible.

6 Conclusion

Researchers have repeatedly attempted to estimate the unique contributions of age, period, and cohort processes on wide range of outcomes. In this paper we have proposed a new technique, which we call the orthogonal estimator (OE), that overcomes some, but not all of the limitations of the intrinsic estimator (IE). Specifically, we have made several main points.

First, we demonstrated that it is only the parameters associated with the linear components in age, period, and cohort that are unidentified. In contrast, the parameters associated with the nonlinear effects are identified. As a result, estimators based on different design matrices will give the same estimates of the nonlinear effects, but will in general yield differing estimates of the linear effects. We have also shown how to take the estimates based on any design matrix and decompose them into their linear and nonlinear components. This is critical for understanding the weaknesses of the IE.

Second, we presented the IE as defined by using a zero-sum effect-coded design matrix along with the Moore-Penrose generalized inverse, which allows for the inversion of singular matrices. We discussed the IE's statistical properties of being estimable, unbiased given its constraint, and with minimum variance. We pointed out that other estimators based on design matrices other than zero-sum effect coding, but that also use the Moore Penrose inverse, share these desirable statistical properties. However, in many cases these alternative estimators will produce estimates that are radically different than the IE's.

Third, we discussed several important weaknesses of the IE. Assuming a fixed number of age categories, the IE is sensitive to the number of period and cohorts of data available to the researcher. Simulations show that, even when keeping the data generating process constant, the IE can produce remarkably different estimates of the linear age, period, and cohort trends depending on how many periods and cohorts of data are available. Mathematically, these results reflect the fact that the non-zero null vector elements change in number, size, and magnitude as the number of period (and thus cohort) groups changes. We also demonstrated how the size and sign of the identified nonlinear effects influence the IE's linear effect estimates, sometimes radically so. Because of the complicated and highly-variable nature of the IE, it is often very difficult, if not impossible, for applied researchers to understand the consequences of the constraint imposed by the IE on actual empirical results.

The fact that IE is sensitive to both the number of APC categories as well as the extent of the nonlinearities in the data is a consequence of zero-sum effect coding. Each of the parameter estimates in a zero-sum effect design matrix is a combination of both the linear and nonlinear effects in the data. As such, they combine aspects of the model that are nonidentified (the linear effects) with aspects that are identified (the nonlinear effects). This mixture produces the undesirable sensitivity of the IE's estimates of the linear effects to both the number of periods and cohorts as well as the magnitude and direction of the nonlinear trends in the data.

Fourth, we outlined an alternative to IE, what we term the orthogonal estimator (OE). The OE uses the Moore-Penrose generalized inverse with a design matrix in which the linear components of the temporal variables are orthogonal to the nonlinear components. By keeping the linear and nonlinear parameter estimates distinct, the OE avoids the problems with the linear parameter estimates of the IE. Specifically, in contrast to the IE, the linear effects given by the OE are not sensitive to the number of periods and cohorts of data used by the researcher or to the size and sign of the nonlinear effects.

In addition, we demonstrated that the OE is equal to the simple average of three zero-linear-trend (ZLT) models in which either the age, period, or cohort linear effects are constrained to equal zero. Geometrically, the OE can be represented as the centroid in a system of barycentric coordinates, reflecting the fact that it is the average of three ZLT models. This offers a novel way of thinking about constraints on APC models in terms of a convex of combination of simpler models. In contrast to the OE, the IE is an average of three extremal models, but one in which the linear effects have been rotated relative to the three ZLT models. This latter fact adds to the difficulty in interpreting the IE.

For applied researchers, we suggest several guidelines when using the OE. When reporting the OE, one should also report the three ZLT models on which it is based. As well, one should report the full set of linear and nonlinear effects (rather than zero-sum effects) so that other researchers can clearly determine the size and direction of the identifiable nonlinearities as well as their statistical significance. Finally, one should consider using theory or insights from additional data to alter the weights on the OE. For example, one might prefer a ZLT period model, and thus examine the results when the model is weighted towards the ZLT period model.[34]

Although we think that the OE is preferable to the IE, this in no way implies that it is preferable to other approaches to analyzing APC data. As we have discussed, one can divide the set of methods for identifying APC effects into statistical and theoretical approaches. The strength of a statistical approach, such as the OE or IE, is that it is not based on any explicit theoretical or substantive assumptions that researchers may strongly disagree about. The weakness of a statistical approach is there is in general no reason to believe that it estimates the true parameters for the model that

[34] Alternatively, one may consider altering the set of extremal models on which the OE is based by rotating the three ZLT models.

generated the data. This is case for both the IE and the OE.

A theoretical approach will provide estimates of the underlying data generating parameters if the theoretical assumptions are valid. One example of a theoretical approach is to assume that the effect of any of age, period, or cohort effects is positive or negative over specific ranges of the variables. As Fosse and Winship (unpublished) show, this can be used to bound estimates of the data generating parameters, in some cases leading to quite narrow bounds despite weak assumptions. An alternative theoretical approach entails specifying the specific mechanisms through which age, period, and cohort impact the outcome. As Winship and Harding (2008) demonstrate, if one has variables measuring all the pathways through which one of the age, period, or cohort variables operate, then it is possible to identify the underlying data generating parameters. To be sure, an appreciable issue with any theoretical approach is that researchers may well disagree on the validity of particular assumptions. Although this can be a serious problem, at least when there is disagreement it will be clear why different approaches lead to different estimates.

Ultimately a theoretical approach is to be preferred to a purely statistical one. In most cases, researchers are interested in estimating the data generating parameters of their model. We appreciate that others see more value in a statistical approach. Nonetheless, if one is going to use a statistical approach to identification, there is merit in using an estimator such as the OE that provides separate estimates of the nonidentified linear and identified nonlinear parameters. As we have demonstrated, doing so produces estimates of the linear effects that are not affected by the number of periods and cohorts available to the researcher or to the size of the nonlinear effects.

Appendix

The technical literature on the IE is scattered across a range of publications, some of which are not widely available or are incomplete in presentation. To deal with this problem, in this appendix we provide the technical details of the IE as well as the OE. First we outline the APC identification problem and prove that the nonlinear effects and linear combinations of the linear effects are identified. Next, we formally define the IE, specify its design matrix, and show how it can be derived. Then we show how one can construct a transformation matrix to convert the IE zero-sum effects into estimates that partition the linear from the nonlinear components. Finally, we prove that both the IE and OE can be interpreted as a weighted average of simpler models in which one of the columns of the design matrix is dropped, with the OE always giving a simple average.

6.1 The APC Identification Problem

Typically in APC analysis the goal is to recover the underlying data generating parameters collected in vector $\mathbf{b}$ using a design matrix $\mathbf{X}$ and an outcome vector $\mathbf{y}$. Formally we have a linear system

$$\mathbf{X}\mathbf{b} = \mathbf{y} \tag{43}$$

with a corresponding system of normal equations $\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y}$.[35] Our goal is to solve this system of linear equations by estimating the best-fitting solution, which is defined as the solution that minimizes some measure of error (or misfit). By far the most common technique is to find the least-squares solution $\mathbf{b}_{\text{OLS}}$, which minimizes the L_2 norm (or Euclidean length) of the residual vector:

$$\mathbf{e} = \|\mathbf{X}\mathbf{b} - \mathbf{y}\|_2 = \sqrt{\sum_{i=1}^{n} \Big(y_i - (\mathbf{X}\mathbf{b})_i\Big)^2} \tag{44}$$

where $\mathbf{e}$ is an $n \times 1$ vector of residuals (or errors); the operator $\| \;.\; \|_2$ denotes the L_2 norm; and i indexes the observations (or rows) of the data set from $i = 1, \ldots n$.[36] If $\mathbf{X}$ were of full column rank, then the system $\mathbf{X}\mathbf{b} = \mathbf{y}$ would have a unique least-squares solution $\mathbf{b}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, where the superscripted -1 indicates the regular inverse. However, due to linear dependence, $\mathbf{X}$ is rank deficient and a regular inverse of $\mathbf{X}^T\mathbf{X}$ does not exist. Consequently, we cannot estimate $\mathbf{b}_{\text{OLS}}$ and any particular least-squares solution requires an additional constraint.

6.2 Proof of the Identifiability of the Nonlinear Effects and Linear Combinations of the Linear Effects

Given the linear model in Equation 2, $\mathbf{u}^T\mathbf{b}$ is identifiable (estimable) if and only if there exists an $n \times 1$ vector $\mathbf{q}$ such that $\mathbf{u}^T = \mathbf{q}^T\mathbf{X}$ (see Searle 1965: 486-487). The simple case occurs when $\mathbf{X}$ is of full rank. In this situation the rows of the $p \times n$ matrix $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ are the set of vectors, since $[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{X}\mathbf{b} = \mathbf{b}$, and thus all parameters of $\mathbf{b}$ can be estimated. However, in the case of the C-APC (see Equation 26), the design matrix is not of full rank due to linear dependence. To determine which parameters in $\mathbf{b}$ are identified, we can reduce the matrix $\mathbf{X}$ to row echelon form. However, as demonstrated in the previous section, regardless of the coding scheme we can convert $\mathbf{X}$ into orthogonal polynomial contrasts: $\mathbf{X}\mathbf{T}^{-1} = \mathbf{X}_{\text{O}}$. Thus, our goal simplifies to reducing the matrix $\mathbf{X}_{\text{O}}$ into row echelon form and then ascertaining which parameters are identifiable. For similar proofs in the APC literature, see Rodgers (1982) and Holford (1983).

We can easily transform $\mathbf{X}_{\text{O}}$ into row echelon form through a set of standard row operations conducted r times. These operations are performed until the upper $k \times p$ submatrix is in row echelon form with the remaining rows as null vectors, where $\text{rank}(\mathbf{X}_{\text{O}}) = k$. The row operations can be understood as premultiplying $\mathbf{X}_{\text{O}}$ by an $n \times k$ matrix of full rank. Replacing the second row of $\mathbf{X}_{\text{O}}$ with the difference between the second and first rows is equivalent to premultiplying $\mathbf{X}_{\text{O}}$ by a matrix $\mathbf{Q}_1$, where:

[35] These are referred to as normal equations because the residual vector is required to be normal (or orthogonal) to every vector in the span of $\mathbf{X}$.

[36] Equation 44 defines a measure of error (or misfit) that we want to minimize, but it is not the only measure we could use. Another measure of error is the L_1 norm, corresponding to the solution that minimizes $\mathbf{e} = \|\mathbf{X}\mathbf{b} - \mathbf{y}\|_1 = \sum_{i=1}^{n} |(y_i - (\mathbf{X}\mathbf{b})_i)|$.

$$\mathbf{Q}_1 = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 0 & \dots & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{pmatrix}. \tag{45}$$

Through a similar series of row operations conducted r times, the row echelon form may accordingly be expressed as $\mathbf{X}_{\text{O}}^{\dagger} = \mathbf{Q}\mathbf{X}_{\text{O}}$, where $\mathbf{X}_{\text{O}}^{\dagger}$ is the matrix of orthogonal polynomial contrasts converted into row echelon form and $\mathbf{Q} = \mathbf{Q}_r\mathbf{Q}_{r-1}\dots\mathbf{Q}_1$. The matrix $\mathbf{Q}$ is of full rank with dimensions $n \times n$; additionally, each of the $\mathbf{Q}_r$ to $\mathbf{Q}_1$ matrices is also of full rank with dimensions $n \times n$. Let the first k rows of $\mathbf{X}_{\text{O}}^{\dagger}$ be defined as $\mathbf{X}_{\text{O}_k}^{\dagger}$. The lower $(n-k) \times p$ submatrix of $\mathbf{X}_{\text{O}}^{\dagger}$ is a null matrix (i.e., it consists only of zeros). We can accordingly state that $\mathbf{X}_{\text{O}_k}^{\dagger}$ is a $k \times p$ full-rank matrix with the following form:

$$\mathbf{X}_{\text{O}_k}^{\dagger} = \begin{pmatrix} \mathbf{x}_1^{\dagger T} \\ \mathbf{x}_2^{\dagger T} \\ \vdots \\ \mathbf{x}_k^{\dagger T} \end{pmatrix}, \tag{46}$$

where $\mathbf{x}_i^{\dagger}$ is a $p \times 1$ vector for $i = 1, 2, \dots, k$. We can likewise define $\mathbf{Q}_k$ as a $k \times n$ full-rank matrix consisting of the first k rows of $\mathbf{Q}$:

$$\mathbf{Q}_k = \begin{pmatrix} \mathbf{q}_1^T \\ \mathbf{q}_2^T \\ \vdots \\ \mathbf{q}_k^T \end{pmatrix}, \tag{47}$$

where $\mathbf{q}_i$ is an $n \times 1$ vector for $i = 1, 2, \dots, k$. Because $\mathbf{q}_i^T\mathbf{X}_{\text{O}} = \mathbf{x}_i^{\dagger}$ for $i = 1, 2, \dots, k$, then $\mathbf{x}_i^{\dagger}\mathbf{b}$ is an identifiable function. Moreover, because $\mathbf{X}_{\text{O}_k}^{\dagger}$ is full rank with respect to k, we know that $\mathbf{X}_{\text{O}_k}^{\dagger}\mathbf{b}$ is a basis set of identifiable functions.

For the C-APC model in Equation 26, the row echelon form of $\mathbf{X}_{\text{O}}$ has the following form:

$$\mathbf{X}_{\text{O}}^{\dagger} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & 1 & 0 & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 1 & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 & 1 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{48}$$

This matrix indicates those components of the APC effects that are identifiable and which that are not. For example, suppose we want to know whether or not the jth parameter of $\mathbf{b}_{\text{O}}$ can be uniquely estimated. We first find the row of $\mathbf{X}_{\text{O}}^{\dagger}$ for which the jth entry is one and the first $j-1$ entries are zero. If the remaining $p-j$ entries are also zero, then we know the jth parameter of $\mathbf{b}_{\text{O}}$ can be identified. However, if the remaining of the entries are non-zero then the identifiable function is derived from the values of those row entries. In general, each row of $\mathbf{X}_{\text{O}}^{\dagger}$ is an identifiable function. Furthermore, because $\mathbf{X}_{\text{O}}^{\dagger}\mathbf{b}_{\text{O}}$ is a basis set, every function that can be identified can be generated from this set of functions. Let $\mathbf{b}_{\text{O}}^{\dagger}$ denote the basis set of identifiable functions for APC models. Then we know that:

$$\mathbf{b}_{\text{O}}^{\dagger} = \begin{cases} \mu \\ \alpha^2, \alpha^3, \ldots, \alpha^{I-1} \\ \pi^2, \pi^3, \ldots, \pi^{J-1} \\ \gamma^2, \gamma^3, \ldots, \gamma^{K-1} \\ \alpha + \pi \\ \gamma + \pi \end{cases} \tag{49}$$

where $\theta_1 = \alpha + \pi$ and $\theta_2 = \gamma + \pi$. More generally, any function of the form

$$\alpha(\omega_1) + \pi(\omega_1 + \omega_2) + \gamma(\omega_2) \tag{50}$$

for arbitrary values of ω_1 and ω_2 is identifiable (cf. Holford 1983: 314).

6.3 Defining the IE

In the case of the IE we can express the estimates as

$$\mathbf{b}_{\text{IE}} = (\mathbf{X}^T\mathbf{X})^{+}\mathbf{X}^T\mathbf{y} = \mathbf{X}^{+}\mathbf{y}. \tag{51}$$

where the superscript $+$ denotes the Moore-Penrose generalized inverse and again $\mathbf{X}$ is a design matrix of categorical age, period, and cohort variables with sum-to-zero effect coding. Equation 51 underscores that the IE is based on $(\mathbf{X}^T\mathbf{X})^{+}$ or, equivalently, $\mathbf{X}^{+}$.

Formally, $\mathbf{X}^{+}$ is defined as a generalized inverse meeting four Moore-Penrose conditions:

1. General Condition: $\mathbf{X}\mathbf{X}^{+}\mathbf{X} = \mathbf{X}$
2. Reflexive Condition: $\mathbf{X}^{+}\mathbf{X}\mathbf{X}^{+} = \mathbf{X}^{+}$
3. Normalized Condition: $(\mathbf{X}\mathbf{X}^{+})^T = \mathbf{X}^{+}\mathbf{X}$
4. Reverse Normalized Condition: $(\mathbf{X}^{+}\mathbf{X})^T = \mathbf{X}\mathbf{X}^{+}$

For any particular matrix $\mathbf{X}$, the generalized inverse always exists (i.e., it is well-defined) and unique (i.e., there is only one such generalized inverse that meets the conditions above). From the four conditions above, we can express the solution in terms of the normal equations. It can be shown that if $\mathbf{X}$ is of full rank then $\mathbf{X}^{+}\mathbf{y} = (\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{y}$.[37]

As with any constrained least-squares estimate for a particular design matrix $\mathbf{X}$ and outcome $\mathbf{y}$, we can use the IE to construct a solution line:

$$\mathbf{b} = \mathbf{b}_{\text{IE}} + s\mathbf{v} \tag{52}$$

where, as previously, s is a scalar that can take on any real number. By varying s, we slide up and down the solution line, resulting in an infinite number of constrained least-squares solutions. Among these values of s, the IE assumes $s = 0$ in Equation 52. Geometrically, the IE is the least-squares solution corresponding to the point on the line closest to the origin in terms of Euclidean distance. This point coincides with the minimium (Euclidean) length of $\mathbf{b}_{\text{IE}} + s\mathbf{v}$, which is at its minimum when $s = 0$.[38] Equivalently, the vector $\mathbf{b}_{\text{IE}}$ is the projection of $\mathbf{b}$ on the nonnull space of $\mathbf{X}$, which is orthogonal to the null space:

$$\mathbf{b}_{\text{IE}} = (\mathbf{I} - \widehat{\mathbf{v}}\widehat{\mathbf{v}}^{T})\mathbf{b} \tag{53}$$

where $\mathbf{I}$ is a $p \times p$ identity matrix and $\widehat{\mathbf{v}}$ is the normalized null vector so that $\widehat{\mathbf{v}}\widehat{\mathbf{v}}^{T} = 1$. That is, the IE separates the true unknown parameter $\mathbf{b}$ into two orthogonal components, the null vector $\mathbf{v}$ and the vector of IE estimates $\mathbf{b}_{\text{IE}}$. Since these two vectors are perpendicular to each other, $\mathbf{b}_{\text{IE}}\mathbf{v}^{T} = 0$ or, equivalently, $s = 0$ in the solution line defined by $\mathbf{b}_{\text{IE}} + s\mathbf{v}$.

Since the IE finds that particular set of least-squares estimates on the solution line for which the (Euclidean) length is minimized, it is a *minimum-norm least-squares* (MNLS) solution. In other words, the IE is fundamentally a two-stage minimization algorithm:

$$\begin{aligned}
&\text{Stage 1:} \quad \text{minimize} \quad \|\mathbf{Xb} - \mathbf{y}\|_2 && (54)\\
&\text{Stage 2:} \quad \text{minimize} \quad \|\mathbf{b}\|_2 \ \text{ among all solutions from Stage 1} && (55)
\end{aligned}$$

where again the operator $\|\ .\ \|_2$ denotes the L_2 norm and $\mathbf{X}$ is a matrix of sum-to-zero effect coding for categorical age, period, and cohort variables. In the first stage, the IE finds the least-squares solution to $\mathbf{Xb} = \mathbf{y}$. Since $\mathbf{X}$ is rank deficient one, there is no unique least-squares solution; rather, there are many such solutions lying on a line in multidimensional space. In the second stage, the IE obtains a particular solution by applying a minimum-norm constraint. Among the least-squares estimates on the solution line, the IE selects that particular set of estimates with the minimum (Euclidean) length or, equivalently, that is closest to the origin in terms of Euclidean distance.[39]

[37] The first and third conditions imply that $\mathbf{X}^{T}\mathbf{X}\mathbf{X}^{+} = \mathbf{X}^{T}$. Substituting $\mathbf{b} = \mathbf{X}^{+}\mathbf{y}$ into the normal equations $\mathbf{X}^{T}\mathbf{X}\mathbf{b} = \mathbf{X}^{T}\mathbf{y}$, we obtain $\mathbf{X}^{T}\mathbf{X}\mathbf{X}^{+}\mathbf{y} = \mathbf{X}^{T}\mathbf{y}$. If $\mathbf{X}$ is of full rank, then we can take the regular inverse to solve for $\mathbf{X}^{+}\mathbf{y}$, which results in $\mathbf{X}^{+}\mathbf{y} = (\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{y}$.

[38] Note, however, that the value of s differs depending on the construction of the solution line. For any particular constrained set of estimates $\mathbf{b}^{*}$, the IE minimizes the Euclidean length of $\mathbf{b}^{*} + s\mathbf{v}$ or, likewise, its squared length $(\mathbf{b}^{*} + s\mathbf{v})^{T}(\mathbf{b}^{*} + s\mathbf{v})$. Using calculus, this squared length is at its minimum when $s = -\mathbf{v}^{T}\mathbf{b}^{*}$.

[39] Combining these two stages, in the limit we can write the IE as $\text{minimize}\ \|\mathbf{Xb} - \mathbf{y}\|_2^2 + \lambda^2\|\mathbf{b}\|_2^2$, where λ is a

Various coding schemes are possible for the design matrix. However, as emphasized by Land and colleagues, the design matrix of the IE *must* be specified as sum-to-zero effect (or deviation) contrasts for the age, period, and cohort groups (Yang and Land 2013c: 79; Land et al. 2016: 964). The IE design matrix has the form

$$\mathbf{X} = \big(\mathbf{1}|\mathbf{A}|\mathbf{P}|\mathbf{C}\big) \tag{56}$$

where $\mathbf{1}$ is a vector of 1's for the intercept, $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_{I-1})$ is the set of age columns, $\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_{J-1})$ is the set of period columns, and $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_{K-1})$ is the set of cohort columns. Each of these columns is coded so that their parameters equal the difference between the mean of the particular group coded one and the overall (or grand) mean. For example, the coding for the ith age column is:

$$\mathbf{a}_i = \begin{cases} 1, & \text{if the } i\text{th age group,} \\ -1, & \text{if the } I\text{th age group,} \\ 0, & \text{otherwise.} \end{cases} \tag{57}$$

The matrix $\mathbf{A}$ accordingly gives the parameters α_i for $i = 1, \ldots i = I - 1$, with the Ith parameter given by $\alpha_I = -\sum_{i=1}^{I-1} \alpha_i$. The columns for period and cohort are similarly coded, with similar parameter definitions. Specifically, the period parameters for $\mathbf{P}$ are π_j for $j = 1, \ldots j = J - 1$, with the Jth parameter given by $\pi_J = -\sum_{j=1}^{J-1} \pi_j$, and the cohort parameters for $\mathbf{C}$ are γ_k for $k = 1, \ldots k = K - 1$, with the Kth parameter given by $\gamma_K = -\sum_{k=1}^{K-1} \gamma_k$. We collect these parameters in the $p \times 1$ vector

$$\mathbf{b} = (\mu|\alpha_i \ldots, \alpha_{I-1}|\pi_j, \ldots, \pi_{J-1}|\gamma_k, \ldots, \gamma_{K-1})^T \tag{58}$$

where μ is the intercept, representing the overall (or grand) mean, and the superscripted $\mathbf{T}$ denotes the transpose. Note that to avoid overparameterization due to the inclusion of the intercept, the design matrix constrains the age, period, and cohort parameters to sum to zero, such that $\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \pi_j = \sum_{k=1}^{K} \gamma_k = 0$.[40]

6.4 Deriving the IE

There are numerous ways to obtain the IE estimates, reflecting the fact that there are many techniques for computing the Moore-Penrose generalized inverse of a matrix. A particularly intuitive and revealing approach is to decompose $\mathbf{X}$ into component matrices that clarify the structure of the data and facilitate manipulation. With singular value decomposition (SVD), we can factorize

regularization parameter and $\lambda \to 0$. As λ approaches zero, both the IE and ridge (or Tikhonov) regression solutions converge to that set of estimates with the smallest Euclidean norm.

[40] Alternatively, one could fix the parameters at one of the levels to zero. For example, researchers often fix the first set of levels (e.g., $\alpha_{i=1} = \pi_{j=1} = \gamma_{k=1} = 0$) or the last set (e.g., $\alpha_{i=I} = \pi_{j=J} = \gamma_{k=K} = 0$). However, as outlined by (Yang and Land 2013c: 79), the IE is based *only* on a design matrix with sum-to-zero constraints and effect (or deviation) coding (see also Land et al. 2016: 964).

$\mathbf{X}$, which is of dimension $n \times p$, into the product of three distinct matrices:

$$\mathbf{X} = \mathbf{U\Sigma V}^T \tag{59}$$

where $\mathbf{U}$ is an $n \times n$ orthogonal matrix with columns that are unit basis vectors spanning the data space (that is, the columns are orthonormal); $\mathbf{\Sigma}$ is a $p \times p$ diagonal matrix in which the diagonal elements are the singular values of $\mathbf{X}$ in descending order: $\sigma_1, \sigma_2, \ldots \sigma_r$, with the rank of the matrix $r = p - 1$; and $\mathbf{V}$ is a $p \times p$ orthogonal matrix with columns that are basis vectors spanning the parameter (or model) space. Since $\mathbf{X}^+\mathbf{y} = \mathbf{b}_{\text{IE}}$, we can obtain the IE estimates by computing the Moore-Penrose generalized inverse of $\mathbf{X}$. There are three main steps. First, we tranpose $\mathbf{\Sigma}$ and then take the reciprocal of the nonzero singular values along the diagonal, retaining those with a value of zero.[41] This will give the generalized inverse of $\mathbf{\Sigma}$, denoted as $\mathbf{\Sigma}^+$. Second, we use $\mathbf{\Sigma}^+$ to calculate $\mathbf{V\Sigma^+U^T} = \mathbf{X}^+$. Finally, we find the IE estimates by calculating $\mathbf{b}_{\text{IE}} = \mathbf{X}^+\mathbf{y}$.

Alternatively, we can derive the IE estimates by taking the Moore-Penrose generalized inverse of $\mathbf{X}^T\mathbf{X}$. For applied researchers accustomed to linear regression, this is likely to be a more intuitive approach than finding the generalized inverse of $\mathbf{X}$. Using the SVD for $\mathbf{X}$, we can write the following:

$$\begin{aligned} \mathbf{X}^T\mathbf{X} &= \big(\mathbf{V\Sigma}^T\mathbf{U}^T\big)\big(\mathbf{U\Sigma V}^T\big) \\ &= \mathbf{V\Sigma}^2\mathbf{V}^T \\ &= \mathbf{V\Lambda V}^T \end{aligned} \tag{60}$$

where the last line is the spectral (or eigenvalue) decomposition of $\mathbf{X}^T\mathbf{X}$. Note that $\mathbf{\Lambda}$ is a $p \times p$ diagonal matrix of the squared non-zero singular values (that is, the eigenvalues) of $\mathbf{X}^T\mathbf{X}$ in descending order $\sigma_1^2, \sigma_2^2, \ldots \sigma_r^2 = \lambda_1, \lambda_2, \ldots, \lambda_r$, with the rank of the matrix $r = p - 1$. It is straightforward to find the generalized inverse of $\mathbf{X}^T\mathbf{X}$, since we follow a similar set of steps as before. First, we find the tranpose of $\mathbf{\Lambda}$ and take the reciprocal of the nonzero eigenvalues along the diagonal, keeping the zero eigenvalues. This will give the generalized inverse of $\mathbf{\Lambda}$, denoted as $\mathbf{\Lambda}^+$. Second, we calculate $\mathbf{V\Lambda^+U}^T = (\mathbf{X}^T\mathbf{X})^+$. Finally, we use $(\mathbf{X}^T\mathbf{X})^+$ to find the IE estimates: $\mathbf{b}_{\text{IE}} = (\mathbf{X}^T\mathbf{X})^+\mathbf{X}^T\mathbf{y}$.[42]

Using spectral decomposition, we can show how the IE is an orthonormal transformation of principal components regression(e.g., Yang and Land 2013c: 79). After factorizing $\mathbf{X}^T\mathbf{X}$, we select the columns of $\mathbf{V}$ that correspond to the non-zero eigenvalues of $\mathbf{\Lambda}$, since these columns are an

[41] In practice it is common to specify a threshold (or tolerance) value such that any singular values less than this (typically very small) value are set to zero. The number of singular values above this specified tolerance is the effective rank.

[42] Besides allowing for estimation of the IE, the decomposition of $\mathbf{X}^T\mathbf{X}$ reveals two crucial features of the data. First, the number of non-zero eigenvalues in $\mathbf{\Lambda}$ gives the rank of the matrix $\mathbf{X}^T\mathbf{X}$. Since the columns of the data are linearly dependent, there will always be a zero eigenvalue and thus the matrix is rank deficient one. Second, along with the zero eigenvalue in $\mathbf{\Lambda}$ there will always be a corresponding eigenvector in $\mathbf{\Lambda}$, which is the orthonormal basis for the null space of $\mathbf{X}^T\mathbf{X}$. This eigenvector is simply the normalized null vector $\widehat{\mathbf{v}}$.

orthonormal set of basis vectors for the range of $\mathbf{X}^T\mathbf{X}$.[43] We call these non-null eigenvectors $\mathbf{V}^*$. Next, we obtain the principal components by multiplying the design matrix $\mathbf{X}$ by the matrix of non-null eigenvectors $\mathbf{V}^*$. We call this matrix $\mathbf{C} = \mathbf{X}\mathbf{V}^*$. Then we run OLS using this new matrix $\mathbf{C}$, denoting the coefficients as $\mathbf{b}_{\text{PCR}}$ since principal components are the inputs in the model:

$$\mathbf{b}_{\text{PCR}} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{X}^T\mathbf{y}. \tag{61}$$

Importantly, by omitting the null eigenvector we are constraining its coefficient to equal zero in Equation 61. To obtain the coefficients on the original scale, we extend $\mathbf{b}_{\text{PCR}}$ to include the zero coefficient for the null eigenvector and then multiply the extended vector by the entire matrix of eigenvectors: $\mathbf{Q}\mathbf{b}_{\text{PCR}_{\text{Extended}}} = \mathbf{b}_{\text{IE}}$. From this derivation, the IE can be understood as a technique that reduces the dimensionality of the data by fixing the coefficient of the null eigenvector to zero.

6.5 Defining the OE

As described in the main text, there are any number of estimators using the Moore-Penrose generalized inverse, each with a different coding of the design matrix. Rather than applying the Moore-Penrose generalized inverse to a design matrix of effect (or deviation) coding, what we call the orthogonal estimator (OE) uses a matrix in which the nonlinearities of age, period, and cohort are orthogonal to their respective linear components. That is, the OE estimates are given by

$$\mathbf{b}_{\text{OE}} = (\mathbf{X}_{\text{O}}^T\mathbf{X}_{\text{O}})^{+}\mathbf{X}_{\text{O}}^T\mathbf{y} \tag{62}$$

where the superscript $+$ indicates the Moore-Penrose generalized inverse and the subscript O denotes a design matrix of nonlinearities orthogonal to the linear components for age, period, and cohort. The design matrix can be set up by using zero-sum orthogonal polynomial contrasts (Draper and Smith 2014: 461-472). Equivalently, we can construct matrix of zero-sum effects orthogonal to the linear components. Here we present the latter approach.

Let our design matrix $\mathbf{X}_{\text{O}}$ be represented as the following:

$$\mathbf{X}_{\text{O}} = \left(\mathbf{1}|\mathbf{a}|\mathbf{p}|\mathbf{c}|\mathbf{A}^*|\mathbf{P}^*|\mathbf{C}^*\right) \tag{63}$$

where the first column is a vector of ones for the intercept; $\mathbf{a}$, $\mathbf{p}$, and $\mathbf{c}$ are $n \times 1$ vectors encoding the linear trends; and $\mathbf{A}^*$, $\mathbf{P}^*$, as well as $\mathbf{C}^*$ are matrices of dimensions $n \times (I-2)$, $n \times (J-2)$, and $n \times (K-2)$ representing the nonlinear components of age, period, and cohort. The columns for the age, period, and cohort linear trends can be represented in a straightforward manner. For the ith age group, we can represent the linear trend as $\mathbf{a}_i = i - (I+1)/2$. Likewise, the period and cohort vectors can be coded as $\mathbf{p}_j = j - (J+1)/2$ and $\mathbf{c}_k = k - (K+1)/2$, respectively.

To construct the nonlinear contrasts we generate a set of columns orthogonal to the constant

[43] In contrast, the orthonormal basis for the null space of $\mathbf{X}^T\mathbf{X}$ is the column (or eigenvector) in $\mathbf{V}^T$ corresponding to the zero eigenvalue of $\mathbf{\Lambda}$.

and linear trend for age, period, and cohort, respectively. For example, to construct the nonlinear contrasts for age we specify a matrix $\mathbf{A}_{\text{Effect}}$ as a matrix of age levels with effect (or deviation) coding (see Equation 57). We denote the constant and linear coding for age as $(\mathbf{1}|\mathbf{a})$. To obtain the coding for the remaining columns we project the columns of $\mathbf{A}_{\text{Effect}}$ onto the orthogonal complement of $(\mathbf{1}|\mathbf{a})$. The projection matrix for age is given by

$$\mathbf{A}_{\text{Proj}} = (\mathbf{1}|\mathbf{a})\Big[(\mathbf{1}|\mathbf{a})^T(\mathbf{1}|\mathbf{a})\Big]^{-1}(\mathbf{1}|\mathbf{a})^T. \tag{64}$$

where $\mathbf{A}_{\text{Proj}}$ is of dimension $n \times n$. We then obtain the coding matrix of nonlinear age components $\mathbf{A}^*$ as follows:

$$\mathbf{A}^* = (\mathbf{I} - \mathbf{A}_{\text{Proj}})\mathbf{A}_{\text{Effect}} \tag{65}$$

where $\mathbf{I}$ is an identity matrix of order $n \times (I-2)$ and $(\mathbf{I} - \mathbf{A}_{\text{Proj}})$ is the orthogonal complement of $(\mathbf{1}|\mathbf{a})$. We similarly set up matrices $\mathbf{P}^*$ and $\mathbf{C}^*$ for period and cohort.

6.6 The Transformation Matrix

To illustrate some of the properties of the IE, it is useful to transform its estimates into its linear and nonlinear components. In general, the transformation matrix $\mathbf{T}$ is a block diagonal matrix in which the main diagonal blocks are square matrices and off-diagonal blocks are zero matrices (Luo et al. 2016: 947). Let $\mathbf{A}$, $\mathbf{P}$, and $\mathbf{C}$ denote the original contrast matrices for age, period, and cohort with dimensions $I \times (I-1)$, $J \times (J-1)$, and $K \times (K-1)$, respectively. Although sum-to-zero effect (or deviation) coding is the most frequently used in the APC literature, these matrices may be coded with any number of schemes without loss of generality. Let $\mathbf{A}_{\text{O}}$, $\mathbf{P}_{\text{O}}$, and $\mathbf{C}_{\text{O}}$ denote the corresponding sum-to-zero orthogonal polynomial contrast matrices for age, period, and cohort (Draper and Smith 2014: 461-472). Each of these contrast matrices has full column rank, but not full row rank. Hence we can construct three left inverses:

$$\begin{aligned}
\mathbf{A}_{\text{O}}^L &= (\mathbf{A}_{\text{O}}^T\mathbf{A}_{\text{O}})^{-1}\mathbf{A}_{\text{O}}^T \\
\mathbf{P}_{\text{O}}^L &= (\mathbf{P}_{\text{O}}^T\mathbf{P}_{\text{O}})^{-1}\mathbf{P}_{\text{O}}^T \\
\mathbf{C}_{\text{O}}^L &= (\mathbf{C}_{\text{O}}^T\mathbf{C}_{\text{O}})^{-1}\mathbf{C}_{\text{O}}^T
\end{aligned} \tag{66}$$

where the superscript L denotes a left inverse. We are now in position to construct the transition matrix $\mathbf{T}$, which has the generic form

$$\mathbf{T} = 1 \oplus \mathbf{A}_{\text{O}}^L\mathbf{A} \oplus \mathbf{P}_{\text{O}}^L\mathbf{P} \oplus \mathbf{C}_{\text{O}}^L\mathbf{C} \tag{67}$$

or, equivalently,

$$\mathbf{T} = \begin{pmatrix} 1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{\text{O}}^{L}\mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{P}_{\text{O}}^{L}\mathbf{P} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{C}_{\text{O}}^{L}\mathbf{C} \end{pmatrix} \tag{68}$$

where $\oplus$ is the direct sum.

6.7 The MNLS Solution as a Weighted Average

As shown previously, the both the IE and OE provide a particular solution by estimating $(\mathbf{X}^T\mathbf{X})^+$, the Moore-Penrose generalized inverse of the $p \times p$ matrix $\mathbf{X}^T\mathbf{X}$. For both the IE and OE, this gives the minimum-norm least-squares (MNLS) solution. In what follows, we prove that both estimators are a weighted average of simpler "drop-one-variable" models. In the case of the OE these drop-one-variable models correspond to three ZLT models that are weighted equally. Thus, the OE can be interpreted as a simple average, unlike the IE.

Generally speaking, we denote the rank of $\mathbf{X}^T\mathbf{X}$ as $p-1$, since it is rank deficient one. However, although $\mathbf{X}^T\mathbf{X}$ is rank deficient, it consists of a number of submatrices that are of full rank. To keep track of the submatrices of $\mathbf{X}^T\mathbf{X}$, we introduce some notation. Let S_I denote the set of all full row-rank submatrices, each of which corresponds to an element of an index set I. That is, the set S_I consists of elements $(\mathbf{X}^T\mathbf{X})_{i*}$, where the asterisk is a placeholder and $i \in I$. Likewise, let S_J indicate the set of all full column-rank submatrices, each of which corresponds to an element of an index set J. In other words, the set S_J consists of elements $(\mathbf{X}^T\mathbf{X})_{*j}$, where the asterisk is a placeholder and $j \in J$. Finally, let S_N denote the set of all full column-rank and row-rank submatrices, each of which corresponds to an element of an index set N. That is, the set S_N consists of elements $(\mathbf{X}^T\mathbf{X})_{ij}$, where $i \in I$ and $j \in J$.

To clarify the index set notation, suppose $p = 3$ and the rank is $p - 1 = 2$. Then we have the index sets $I = \{\{1,2\},\{1,3\},\{2,3\}\}$ and $J = \{\{1,2\},\{1,3\},\{2,3\}\}$, with N consisting of all combinations of I and J. The set S_I consists of three 2×3 full row-rank submatrices indexed by I, while the set S_J consists of three 3×2 full column-rank submatrices indexed by J. Finally, the set S_N indexed by N consists of nine 2×2 submatrices that are of full rank. For example, the submatrix $(\mathbf{X}^T\mathbf{X})_{\{1,2\}\{2,3\}}$ in set S_N is indexed by $i = \{1,2\}$ and $j = \{2,3\}$ in set N. In other words, this notation informs us that by subsetting $\mathbf{X}^T\mathbf{X}$ to rows $\{1,2\}$ and columns $\{2,3\}$ (or, equivalently, removing the third row and first column), we obtain the submatrix $(\mathbf{X}^T\mathbf{X})_{\{1,2\}\{2,3\}}$. This submatrix is of full rank, so it can then be inverted using a regular inverse.

Using the index set notation defined above, we can express the Moore-Penrose generalized inverse as a weighted linear combination of all regular inverses $(\mathbf{X}^T\mathbf{X})^{-1}_{ij}$ in $\mathbf{X}^T\mathbf{X}$ (Berg 1986):

$$(\mathbf{X}^T\mathbf{X})^+ = \sum_{(ij)\in N} \lambda_{ij} (\mathbf{X}_0^T\mathbf{X}_0)^{-1}_{ij} \tag{69}$$

where the λ_{ij}'s are the weights and the subscript 0 denotes that the submatrix $(\mathbf{X}_0^T\mathbf{X}_0)_{ij}$ is padded with zeros in the appropriate rows and columns after it is inverted. The weights λ_{ij} sum to one and are proportional to the squares of the determinants of the $(\mathbf{X}^T\mathbf{X})_{ij}$ submatrices:

$$\lambda_{ij} = \frac{\left|(\mathbf{X}^T\mathbf{X})_{ij}\right|^2}{\sum_{(ij)\in N}\left|(\mathbf{X}^T\mathbf{X})_{ij}\right|^2} \tag{70}$$

where $|\ .\ |^2$ denotes the squared determinant. [44]

To show that the MNLS solution is a weighted average of drop-one-variable models, we re-express $(\mathbf{X}^T\mathbf{X})^+$ by summing over the rows $i \in I$ for each combination of the columns of $\mathbf{X}^T\mathbf{X}$ (Ben-Israel and Greville 2003: 104-151). For each column we can find the generalized inverse

$$(\mathbf{X}^T\mathbf{X})_{*j}^+ = \sum_{i\in I}\lambda_{ij}(\mathbf{X}_0^T\mathbf{X}_0)_{ij}^{-1} \tag{71}$$

where the weights λ_{ij} are defined as before in Equation 70. These submatrices $(\mathbf{X}^T\mathbf{X})_{*j}^+$ are a weighted sum equal to the Moore-Penrose generalized inverse of the entire matrix $\mathbf{X}^T\mathbf{X}$:

$$(\mathbf{X}^T\mathbf{X})^+ = \sum_{j\in J}\lambda_{*j}(\mathbf{X}_0^T\mathbf{X}_0)_{*j}^+ \tag{72}$$

where the weights are now defined as

$$\lambda_{*j} = \frac{\sum_{i\in I}\left|(\mathbf{X}^T\mathbf{X})_{*j}\right|^2}{\sum_{(ij)\in N}\left|(\mathbf{X}^T\mathbf{X})_{ij}\right|^2}. \tag{73}$$

That is, for each j submatrix in J we calculate the weights λ_{*j} by summing the squared determinants for the submatrices indexed by I and then dividing by the sum of the squared determinants for all submatrices indexed by N.[45] Multiplying both sides of Equation 72 by $\mathbf{X}^T\mathbf{y}$, we have the following relationship:

$$\begin{aligned}(\mathbf{X}^T\mathbf{X})^+\mathbf{X}^T\mathbf{y} &= \sum_{j\in J}\lambda_{*j}(\mathbf{X}_0^T\mathbf{X}_0)_{*j}^+\mathbf{X}^T\mathbf{y} \Rightarrow \\ \mathbf{b}_{\text{MNLS}} &= \sum_{j\in J}\lambda_{*j}\mathbf{B}_{*j}\end{aligned} \tag{74}$$

where $\mathbf{B}_{*j}$ is the $m \times p$ matrix of estimates in which each one of the variables (that is columns) of $\mathbf{X}$ is dropped, where m is the number of drop-one-variable models. That is, each row is a different drop-one model of $\mathbf{X}$ and the columns refer to each of the parameter estimates p. Note that $p = m$

[44] Because determinants have an interpretation as volume, the weights can be interpreted as ratios of volumes, with the square root of the numerator giving the volume of each of the submatrices $(\mathbf{X}^T\mathbf{X})_{ij}$ and the square root of the denominator giving the volume of the entire matrix $\mathbf{X}^T\mathbf{X}$.

[45] Similar to Equation 70, the square root of the numerator gives the volume of each of the submatrices $(\mathbf{X}^T\mathbf{X})_{*j}$, while the square root of the denominator is the volume of $\mathbf{X}^T\mathbf{X}$.

in Equation 74, since we are dropping each column of $\mathbf{X}$.

We can make two further simplifications. First, for both the IE and the OE the weights in Equation 73 are equal to the squared elements of the null vector divided by the sum of the squared elements of the null vector:

$$\lambda_m = \frac{\mathbf{v}^2}{\sum_{m=1}^{M} \mathbf{v}^2} \tag{75}$$

where the elements in the null vector are indexed from $m = 1, \ldots m = M$, with $M = p$ since each parameter in $\mathbf{b}_{\mathrm{OE}}$ has a corresponding null vector element. Second, we only place a just-identifying constraint on the parameter vector $\mathbf{b}$ by dropping the columns of $\mathbf{X}$ that correspond to the non-zero elements of the null vector. To put it another way, we only weigh the estimates of the drop-one columns when that particular column has a null vector value other than zero. Thus, we can present a simplified formula for the MNLS solution as a weighted average of drop-one-variable models:

$$\mathbf{b}_{\mathrm{MNLS}} = \sum_{k=1}^{K} \lambda_k \mathbf{B} \tag{76}$$

where k indexes the non-zero null vector elements in $\mathbf{v}$ and corresponding columns in $\mathbf{X}$, with $\lambda_k = \frac{\mathbf{v}^2}{\sum_{k=1}^{K} \mathbf{v}^2}$. To reiterate, $\mathbf{B}$ is a $k \times p$ matrix of coefficients. Each row of $\mathbf{B}$ is a different drop-one-variable model, where we only drop those columns of $\mathbf{X}$ with corresponding non-zero null vector elements. Since the IE and OE are both MNLS solutions, they both can be interpreted as a weighted average of estimates from drop-one-variable models. The OE reduces to a simple average because its null vector only has three non-zero elements of equal magnitude corresponding to the age, period, and cohort linear components.

References

Alwin, Duane F. 1991. "Family of Origin and Cohort Differences in Verbal Ability". *American Sociological Review* 56.5, pp. 625–638.

Ben-Israel, Adi and T. N. E. Greville. 2003. *Generalized Inverses: Theory and Applications*. 2nd ed. CMS books in mathematics 15. New York: Springer. 420 pp.

Berg, Lothar. 1986. "Three Results in Connection with Inverse Matrices". *Linear Algebra and its Applications* 84, pp. 63–77.

Chaves, Mark. 1989. "Secularization and Religious Revival: Evidence from U.S. Church Attendance Rates, 1972-1986". *Journal for the Scientific Study of Religion* 28.4, pp. 464–477.

Chen, Xinguang et al. 2003. "Secular Trends in Adolescent Never Smoking from 1990 to 1999 in California: An Age-Period-Cohort Analysis". *American Journal of Public Health* 93.12, pp. 2099–2104.

Clark, April K. and Marie A. Eisenstein. 2013. "Interpersonal Trust: An Age–period–cohort Analysis Revisited". *Social Science Research* 42.2, pp. 361–375.

Clayton, D. and E. Schifflers. 1987. "Models for Temporal Variation in Cancer Rates. II: Age-Period-Cohort Models". *Statistics in Medicine* 6.4, pp. 469–481.

Diouf, Ibrahima et al. 2010. "Evolution of Obesity Prevalence in France: An Age-Period-Cohort Analysis". *Epidemiology* 21.3, p. 360.

Draper, Norman R. and Harry Smith. 2014. *Applied Regression Analysis*. John Wiley & Sons.

Fienberg, Stephen E. 2013. "Cohort Analysis' Unholy Quest: A Discussion". *Demography* 50.6, pp. 1981–1984.

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, pp. 1–67.

Firebaugh, Glenn and Brian Harley. 1991. "Trends in U.S. Church Attendance: Secularization and Revival or Merely Lifecycle Effects?" *Journal for the Scientific Study of Religion* 30.4, pp. 487–500.

Fosse, Ethan and Christopher Winship. 2017. "Bounding Analyses of Age-Period-Cohort Effects".

Fu, Qiang and Kenneth C. Land. 2015. "The Increasing Prevalence of Overweight and Obesity of Children and Youth in China, 1989-2009: An Age-Period-Cohort Analysis". *Population Research and Policy Review* 34.6, pp. 901–921.

Fu, Wenjiang. 2016. "Constrained Estimators and Consistency of a Regression Model on a Lexis Diagram". *Journal of the American Statistical Association* 111.513, pp. 180–199.

Fu, Wenjiang J. 2000. "Ridge Estimator in Singular Design with Application to Age-Period-Cohort Analysis of Disease Rates". *Communications in Statistics - Theory and Methods* 29.2, pp. 263–278.

Fu, Wenjiang J. and Peter Hall. 2006. "Asymptotic Properties of Estimators in Age-Period-Cohort Analysis". *Statistics & probability letters* 76.17, pp. 1925–1929.

Fu, Wenjiang J., Kenneth C. Land, and Yang Yang. 2011. “On the Intrinsic Estimator and Constrained Estimators in Age-Period-Cohort Models”. *Sociological Methods & Research* 40.3, pp. 453–466.

Ghitza, Yair and Andrew Gelman. 2014. *The Great Society, Reagan's Revolution, and Generations of Presidential Voting.*

Glenn, Norval D. 1981. “The Utility and Logic of Cohort Analysis”. *The Journal of Applied Behavioral Science* 17.2, pp. 247–257.

Greenland, Sander. 2005. “Identifiability”. *Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd.

Hauser, Robert M. and Min-Hsiung Huang. 1997. “Verbal Ability and Socioeconomic Success: A Trend Analysis”. *Social Science Research* 26.3, pp. 331–376.

Holford, T. R. 1991. “Understanding the Effects of Age, Period, and Cohort on Incidence and Mortality Rates”. *Annual Review of Public Health* 12.1, pp. 425–457. pmid: 2049144.

Holford, Theodore R. 1983. “The Estimation of Age, Period and Cohort Effects for Vital Rates”. *Biometrics* 39.2, pp. 311–324.

Hout, Michael and David Knoke. 1975. “Change in Voting Turnout, 1952–1972”. *Public Opinion Quarterly* 39.1, pp. 52–68.

Hu, Anning. 2015. “A Loosening Tray of Sand? Age, Period, and Cohort Effects on Generalized Trust in Reform-Era China, 1990–2007”. *Social science research* 51, pp. 233–246.

Kerr, William C. et al. 2004. “Age, Period and Cohort Influences on Beer, Wine and Spirits Consumption Trends in the U.S. National Alcohol Surveys”. *Addiction* 99.9, pp. 1111–1120.

Keyes, Katherine M. et al. 2017. “Age, Period, and Cohort Effects in Conduct Problems among American Adolescents from 1991 through 2015”. *American Journal of Epidemiology*.

Kramer, Michael R., Amy L. Valderrama, and Michele L. Casper. 2015. “Decomposing Black-White Disparities in Heart Disease Mortality in the United States, 1973–2010: An Age-Period-Cohort Analysis”. *American Journal of Epidemiology* 182.4, pp. 302–312.

Kupper, Lawrence L., Joseph M. Janis, Azza Karmous, et al. 1985. “Statistical Age-Period-Cohort Analysis: A Review and Critique”. *Journal of Chronic Diseases* 38.10, pp. 811–830.

Kupper, Lawrence L., Joseph M. Janis, Ibrahim A. Salama, et al. 1983. “Age-Period-Cohort Analysis: An Illustration of the Problems in Assessing Interaction in One Observation per Cell Data”. *Communications in Statistics - Theory and Methods* 12.23, pp. 201–217.

Land, Kenneth C. et al. 2016. “Playing With the Rules and Making Misleading Statements: A Response to Luo, Hodges, Winship, and Powers”. *American Journal of Sociology* 122.3, pp. 962–973.

Lavori, Philip W. et al. 1987. “Age-Period-Cohort Analysis of Secular Trends in Onset of Major Depression: Findings in Siblings of Patients with Major Affective Disorder”. *Journal of psychiatric research* 21.1, pp. 23–35.

Li, Chunhui, Chuanhua Yu, and Peigang Wang. 2015. “An Age-Period-Cohort Analysis of Female Breast Cancer Mortality from 1990–2009 in China”. *International Journal for Equity in Health* 14, p. 76.

Liu, S. et al. 2001. "Increasing Thyroid Cancer Incidence in Canada, 1970–1996: Time Trends and Age-Period-Cohort Effects". *British Journal of Cancer* 85.9, p. 1335.

Loeber, Rolf and David P. Farrington. 2014. "Age-Crime Curve". *Encyclopedia of Criminology and Criminal Justice*. Ed. by Gerben Bruinsma and David Weisburd. Springer New York, pp. 12–18.

Luo, Liying. 2013. "Assessing Validity and Application Scope of the Intrinsic Estimator Approach to the Age-Period-Cohort Problem". *Demography* 50.6, pp. 1945–1967.

Luo, Liying et al. 2016. "The Sensitivity of the Intrinsic Estimator to Coding Schemes: Comment on Yang, Schulhofer-Wohl, Fu, and Land". *American Journal of Sociology* 122.3, pp. 930–961.

Mason, Karen Oppenheim, William M. Mason, et al. 1973. "Some Methodological Issues in Cohort Analysis of Archival Data". *American Sociological Review* 38.2, pp. 242–258.

Mason, William M. and Stephen E. Fienberg, eds. 1985. *Cohort Analysis in Social Research*. New York: Springer.

Mazumdar, Sati, Ching Chun Li, and G. Rex Bryce. 1980. "Correspondence Between a Linear Restriction and a Generalized Inverse in Linear Model Analysis". *The American Statistician* 34.2, pp. 103–105.

Morgan, Stephen L. and Christopher Winship. 2014. *Counterfactuals and Causal Inference*. Cambridge, UK: Cambridge University Press.

O'Brien, Robert. 2015. *Age-Period-Cohort Models: Approaches and Analyses with Aggregate Data*. Boca Raton, FL: CRC Press. 220 pp.

O'Brien, Robert M. 2011. "Constrained Estimators and Age-Period-Cohort Models". *Sociological Methods & Research* 40.3, pp. 419–452.

O'Malley, Patrick M., Jerald G. Bachman, and Lloyd D. Johnston. 1984. "Period, Age, and Cohort Effects on Substance Use among American Youth, 1976-82." *American Journal of Public Health* 74.7, pp. 682–688.

Pelzer, Ben et al. 2014. "The Non-Uniqueness Property of the Intrinsic Estimator in APC Models". *Demography* 52.1, pp. 315–327.

Price, Joseph et al. 2016. "How Much More XXX Is Generation X Consuming? Evidence of Changing Attitudes and Behaviors Related to Pornography Since 1973". *The Journal of Sex Research* 53.1, pp. 12–20.

Putnam, Robert D. 1995. "Tuning in, Tuning out: The Strange Disappearance of Social Capital in America". *PS: Political Science & Politics* 28 (04), pp. 664–683.

Robinson, Robert V. and Elton F. Jackson. 2001. "Is Trust in Others Declining in America? An Age–period–cohort Analysis". *Social Science Research* 30.1, pp. 117–145.

Rodgers, Willard L. 1982. "Estimable Functions of Age, Period, and Cohort Effects". *American Sociological Review* 47.6, pp. 774–787.

Sampson, R. J. 2005. "A Life-Course View of the Development of Crime". *The ANNALS of the American Academy of Political and Social Science* 602.1, pp. 12–45.

Schwadel, Philip and Michael Stout. 2012. "Age, Period and Cohort Effects on Social Capital". *Social Forces* 91.1, pp. 233–252.

Searle, S. R. 1965. "Additional Results Concerning Estimable Functions and Generalized Inverse Matrices". *Journal of the Royal Statistical Society. Series B (Methodological)* 27.3, pp. 486–490.

Smith, Herbert L. 2004. "Response: Cohort Analysis Redux". *Sociological Methodology* 34.1, pp. 111–119.

Tilley, James and Geoffrey Evans. 2014. "Ageing and Generational Effects on Vote Choice: Combining Cross-Sectional and Panel Data to Estimate APC Effects". *Electoral Studies* 33, pp. 19–27.

Vedøy, Tord F. 2014. "Tracing the Cigarette Epidemic: An Age-Period-Cohort Study of Education, Gender and Smoking Using a Pseudo-Panel Approach". *Social Science Research* 48, pp. 35–47.

Wilson, James A. and Walter R. Gove. 1999. "The Age-Period-Cohort Conundrum and Verbal Ability: Empirical Relationships and Their Interpretation: Reply to Glenn and to Alwin and McCammon". *American Sociological Review* 64.2, pp. 287–302.

Winship, Christopher and David J. Harding. 2008. "A Mechanism-Based Approach to the Identification of Age–Period–Cohort Models". *Sociological Methods & Research* 36.3, pp. 362–401.

Yang, Yang. 2008. "Social Inequalities in Happiness in the United States, 1972 to 2004: An Age-Period-Cohort Analysis". *American Sociological Review* 73.2, pp. 204–226.

Yang, Yang Claire and Kenneth C. Land. 2013a. "Misunderstandings, Mischaracterizations, and the Problematic Choice of a Specific Instance in Which the IE Should Never Be Applied". *Demography* 50.6, pp. 1969–1971.

— 2013b. "The Statistical Properties of the Intrinsic Estimator for Age-Period-Cohort Analysis".

Yang, Yang, Wenjiang J. Fu, and Kenneth C. Land. 2004. "A Methodological Comparison of Age-Period-Cohort Models: The Intrinsic Estimator and Conventional Generalized Linear Models". *Sociological Methodology* 34.1, pp. 75–110.

Yang, Yang and Kenneth C. Land. 2013c. *Age-Period-Cohort Analysis: New Models, Methods, and Empirical Applications*. Boca Raton, FL: CRC Press. 338 pp.

Yang, Yang, Sam Schulhofer-Wohl, et al. 2008. "The Intrinsic Estimator for Age-Period-Cohort Analysis: What It Is and How to Use It". *American Journal of Sociology* 113.6, pp. 1697–1736.