5-2017

# Do Data Structures Matter? A Simulation Study for Testing the Validity of Age-Period-Cohort Models

Sun Young Jeon
*Utah State University*

DO DATA STRUCTURES MATTER? A SIMULATION STUDY FOR

TESTING THE VALIDITY OF AGE-PERIOD-COHORT MODELS

by

Sun Young Jeon

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Sociology

Approved:

| | |
|---|---|
| _____ | _____ |
| Eric N. Reither, Ph.D.<br>Major Professor | Sojung Lim, Ph.D.<br>Committee Member |
| _____ | _____ |
| Erin Hofmann, Ph.D.<br>Committee Member | John R. Stevens, Ph.D.<br>Committee Member |
| _____ | _____ |
| Kenneth Land, Ph.D.<br>Committee Member | Mark R. McLellan, Ph.D.<br>Vice President for Research and<br>Dean of the School of Graduate Studies |

UTAH STATE UNIVERSITY
Logan, Utah

2017

# ABSTRACT

Do Data Structures Matter? A Simulation Study for Testing the Validity of Age-Period-Cohort Models

by

Sun Young Jeon, Master of Science

Utah State University, 2017

Major Professor: Dr. Eric N. Reither
Department: Sociology, Social Work & Anthropology

For the last four decades, scholars have made significant efforts to develop statistical techniques to estimate the independent contributions of three temporal dimensions (i.e. age, period, and cohort) to population health and other outcomes. These efforts have been challenged by the "identification (ID) problem," a statistical conundrum that occurs due to an algebraic dependence between the three temporal terms. Hierarchical Age-Period-Cohort (HAPC) modeling and Intrinsic Estimator (IE) methods, which are two of the most recent and important innovations in Age-Period-Cohort (APC) analysis, provide unique model specifications to address the ID problem. However, recent critiques have questioned the validity of these two methods in properly addressing the ID problem by presenting evidence that both have limitations, including potentially invalid estimation of age, period, and cohort effects. In this dissertation, I test the argument advanced by proponents of HAPC and IE methods that each of them provides unbiased estimation of parameter values when the data structure satisfies model assumptions. In Chapters 2 and 3, I conduct a series of simulation analyses to assess the validity of these

claims, as well as the usefulness of preliminary analyses (i.e., descriptive and model selection statistics) in identifying data structures that are compatible with APC models. In Chapter 4, I provide a step-by-step demonstration of the HAPC method to empirical data to study how age, period and cohort contribute to educational inequalities in health in United States. The results from these analyses indicate that descriptive and model selection statistics are useful in identifying temporal data structures prior to the application of HAPC and IE models, and that these methods tend to provide unbiased estimates when the data structures are three-dimensional. Furthermore, even when the data structures and corresponding "best models" were ambiguous, it was possible to utilize APC methods by cross-validating nested models.

(162 pages)

# PUBLIC ABSTRACT

Do Data Structures Matter? A Simulation Study for Testing the Validity of Age-Period-Cohort Models

Sun Young Jeon

Age, period, and cohort are three temporal dimensions that can make unique contributions to social and epidemiological changes that occur in populations over time. However, while the theoretical underpinnings for each temporal dimension are well established, the statistical techniques to assess the distinctive contributions of age, period and cohort are controversial. Unless questionable assumptions are imposed on the data, traditional linear regression models are incapable of estimating the independent contribution of each temporal dimension due to the linear dependence between age, period and cohort (A=P-C). Two recently developed methods, Hierarchical Age-Period-Cohort (HAPC) and Intrinsic Estimator (IE) models, enable researchers to estimate how all three temporal dimensions contribute to an outcome of interest without resorting to such assumptions. However, some simulation studies suggest that these new methods provide biased estimates of each temporal dimension. In this dissertation, I investigated whether practitioners can avoid biased results by first understanding the structure of the data. In Chapters 2 and 3, I examined whether visual plots of descriptive statistics and model selection statistics could identify various types of data structures through a series of simulation analyses. The results showed that preliminary data analysis is useful for identifying data structures that are compatible with the assumptions of HAPC and IE models. Moreover, when the data satisfied assumptions such as three-dimensionality and

slight deviations from perfect functional forms, both HAPC and IE models tended to provide unbiased estimates of age, period and cohort effects. In Chapter 4, I provided a step-by-step demonstration for applying HAPC models by investigating the unique contributions of age, period and cohort to educational inequalities in the health of a large sample of U.S. adults. This study found that age and cohort effects contribute most to variability in health, and also that cross-validation is a useful way to incorporate HAPC models when preliminary analyses do not definitively show that the data structure is three dimensional.

# ACKNOWLEDGMENTS

Foremost, I would like to thank my major professor Dr. Eric Reither for his training and advising over the last seven years. Working with Dr. Reither in graduate school has been the luckiest event in my career. I have no doubt that my journey in graduate school could never have been the same without him. I also want to thank another major professor of mine, Dr. John R. Stevens for fully supporting the concurrent degree plan in Statistics, and helping me to finish my training in the department of Mathematics and Statistics. I am sure that I could not have accomplished some work in this dissertation without those wonderful experiences in statistics. Also, I would like to acknowledge that it has been my great honor to work with Dr. Kenneth Land, and to have his insightful comments and suggestions for my work on age-period-cohort models. I express my special gratitude to Dr. Sojung Lim for her warm advice and care, in addition to the wonderful opportunities to be part of her research projects. I am grateful to Dr. Erin Hofmann for her amazing classes in demography and comments, feedback and recommendations on this dissertation. Besides my committee professors, I am grateful to the faculty, staff, and graduate students in the Department of Sociology.

I give my special thanks to my parents, sister Jee Young, brother-in-law Seunggeun, and two adorable nephews for their encouragement and support from 6,000 miles away. Lastly, I would like to thank Pedro, who is the dearest partner of my life and academic journey, for his support and belief in me.

Sun Y. Jeon

CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER I

# INTRODUCTION

Since Norman Ryder's (1965) pioneering work establishing the concept of cohort, studies in sociology, demography, and epidemiology have explored the differential influences of three temporal units on time-related changes to the phenomena of interest. Those three units are age, period, and birth cohort, which conceptually make distinct contributions to the changes that occur over time. All individuals experience a biological and psychological aging process that happens within an individual as time passes, and the term *age effects* refers to the influence of such processes on the phenomena of interest. Apart from individual aging, society at the macro level—wherein the individuals reside—is transformed over time by societal, cultural, and epidemiological changes. The term *period effects* indicate impacts of such temporal variations of the society on the outcome of its interests, which are made evenly and simultaneously to all members at all ages. Lastly, as individuals make "fresh" contacts with emerging facets of society at different ages, the impacts may differ across generations who experience these new realities at different stages of the life course. The impacts caused by the distinct experiences of the contemporary society depending on age-at-exposure are defined as *cohort effects*, and they influence the phenomena of interest over individuals' life courses.

This work is a significant conceptual advancement in the understanding of temporal dimensions and their contributions to time-related changes. Unfortunately, it has not been accompanied by robust measurement techniques because independent effects of

age, period, and cohort are hard to operationalize in the traditional regression model setting. This is the case because the concepts of the three terms are algebraically dependent on their definitions. That is, birth year, which determines the membership of the birth cohort (C), is defined by subtracting the age (A) from the period (P) in which the data were taken (C=P-A). Due to this perfect collinearity, a model including all three terms is not estimable. To further illustrate, suppose that a researcher specifies a linear regression model that estimates the independent effects of age, period, and cohort on the outcome of interest as follows:

$$Y = \alpha + \beta_1 Age + \beta_2 Period + \beta_3 Cohort + \varepsilon, \quad (1)$$

where Y is the outcome variable, $\alpha$ is the intercept; $\beta_1$, $\beta_2$, and $\beta_3$ are the partial slopes of age, period, and cohort; and $\varepsilon$ is the random error term (Mason et al. 1973). As it is, Eq (1) cannot be used to estimate the coefficients $\alpha$, $\beta_1$, $\beta_2$, $\beta_3$, and $\varepsilon$. The perfect collinearity between the three regressors in the model causes the design matrix (X) of Eq (1) to be singular with one less-than-full column rank (Kupper et al. 1985). Consequently, the ordinary least squares (OLS) estimator, which is identical to the maximum likelihood (MLE) estimator with the normally distributed error terms in this model setting, does not exist. As a result, Eq (1) has infinite numbers of coefficient sets and cannot identify which of them correspond to the true age, period, and cohort effects. This is called the identification (ID) problem, which is a well-known conundrum in the age-period-cohort (APC) analysis.

For decades, APC analysis, which encompasses systematic studies that attempt to measure the independent contributions of age, period, and cohort to temporal changes in a wide range of outcomes, has mainly been developed to determine how to address this intractable methodological issue. Two conventional solutions to this problem in the regression framework are 1) constraint-setting approaches (Fienberg and Mason 1979, Mason and Fienberg 1985); and 2) nonlinear parametric transformations (Yang 2008b). The former imposes one additional constraint, which is usually an equality constraint that supposes adjacent age, period, or cohort groups have equal-size effects (i.e., $\beta_n = \beta_{n+1}$). By doing this, the design matrix X is enforced to be non-singular, and the OLS/MLE estimator of the linear regression can be obtained. The latter uses continuous age, period, or cohort terms and specifies polynomial patterns for one or more of the three dimensions (Yang 2008b). This allows at least one of the age, period, and cohort terms to have a nonlinear relationship to the others, breaks the perfect algebraic dependence between the three terms, and yields a model that is just-identified.

These two approaches make the model estimable. However, both approaches have one critical limitation in common. They need a priori knowledge to guide selection of the "correct" constraint or polynomial function in order to estimate the true effects of age, period, and cohort. Unfortunately, such knowledge (sometimes referred to as "strong theories") is often not available or sufficiently solid (Reither et al. 2015b). When this is the case, the model still cannot be identified, or an arbitrary constraint or functional form may be selected. As a previous study showed, the estimates of age, period, and cohort effects can be highly sensitive to the selection of the constraint or the functional form (Yang, Fu and Land 2004), meaning a researcher should always be aware that these

approaches carry the risk of obtaining biased estimates. In real-world research, there is no way to assess whether the imposed constraint or functional form is the right one.

Consequently, APC scholars have searched for another statistical approach that can identify the set of true age, period, and cohort effects without requiring a priori knowledge. Such specification of a model should solely rely on the model setting and the data to obtain the model estimator. Then, the model will have very little room for bias, which is caused by the arbitrary selection of constraint or functional form. In the past decade, a new class of such methodologies in APC research has emerged, providing innovative strategies for addressing the ID problem. Two of the most important APC innovations are Hierarchical APC (HAPC) modeling and Intrinsic Estimator (IE) methods (Yang and Land 2006, Yang 2008a, Yang and Land 2013a). Applied to repeated cross-sectional survey data and tabular rate data, respectively, these two methods provide unique model specifications that address the ID problem. Unlike the conventional approaches to the ID problem, both methods identify the single set of age, period, and cohort effects without relying on any external information. With this desirable feature, both are endorsed by some scholars as non-arbitrary methods of defining the unique and true set of age, period, and cohort effects.

Since their development, these two methods have been widely used in studies for understanding the distinct contributions of age, period, and cohort on important social, political, and epidemiological issues. However, recent critiques have presented evidence that both models have limitations, including potentially invalid estimation of age, period, and cohort effects (Bell and Jones 2013, Bell and Jones 2014b, Held and Riebler 2013, Luo 2013a). In these critiques, the authors simulated APC data and showed that the

estimates of the HAPC and IE methods did not match the true age, period, and cohort effects that generated the outcome data. By interpreting these results as consequences of the ID problem that was not resolved by those two methods, the critiques questioned the validity of the HAPC and IE methods as well as the findings of the previously published studies that had employed either of those methods. Further, the authors recommended that researchers stop using the methods.

Proponents of innovative APC methodologies have interpreted results of these simulation studies with skepticism. In rejoinders to the critiques, they argue that it was premature to say that the HAPC and IE methods failed to address the ID problem, because results obtained in the simulation studies might have been caused by an incorrect application of the models. The "wrong" application here includes simulating unrealistic data that violate important model assumptions, a lack of understanding of the simulated data, and application of the HAPC and IE to datasets that are not ideal candidates for the full three-dimensional APC model. The proponents of new APC methods have further argued that scholars must avoid mechanical application of APC models, because there is no such universal method in APC analysis that will work in every instance.

According to Yang and Land (2013a), key to avoiding the misuse of the APC model is understanding the data structure and finding the best nested fit or the full APC model for the given data structure. Here, "data structure" is used to represent how much the effects of the age, period, and cohort independently contribute to the changes of Y over time. When the contribution of a term is negligible, the data structure is reduced for that term. In such a case, a reduced model rather than the full three-dimensional model, such as HAPC or IE, should be applied. Applying the model with more variables than

necessary is statistically inadvisable because inference with such a model may cause poor precision and identification of effects (Burnham and Anderson 2004). Similarly, applying a reduced model to fully three-dimensional data, wherein all three terms' contributions are significant, may lead to biased estimates (Burnham and Anderson 2004). Therefore, it is essential to carefully select a final model that can strike a balance between over- and under-fitting, to obtain unbiased estimates. When conducted along with those preliminary analyses, Yang and Land (2013a) argue that HAPC and IE are reliable estimators.

How, then, do we know the data structure (i.e., how much A, P, and C each contribute to changes in Y over time) prior to estimating an APC accounting model such as HAPC or IE? Yang and Land (2013a) propose a three-step procedure, providing a systematic method for understanding the data structure and applying the best-fitting model. It can be conducted as follows: Step 1 involves conducting descriptive analysis and visualizing the patterns of temporal variations to gain some qualitative understanding of the data; Step 2 involves fitting the nested models of age (A), period (P), and cohort (C) effects (i.e., A, P, C, AP, PC, AC, and APC), and comparing the goodness of fits by using measures such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). These two steps can help a researcher to better understand the structure of the given data and learn the best analytic model by which to gain unbiased results. Then, only for the cases for which Steps 1 and 2 suggest that the full three-dimensional APC model is a better fit than the other reduced models for the given data, the analysis moves on to Step 3: applying the HAPC or IE to estimate the effects of A, P, and C. With data for which a reduced model fits better than the full three-dimensional model, a researcher should not move on to Step 3. However, Yang and Land (2013a) argue that as

a result of ignoring this important procedure and lacking an understanding of the structure of the simulated datasets, the critics have made crucial errors in applying the full three-dimensional model to the data, which better fits a reduced model. In their rejoinders, Yang and Land (2013b) demonstrate that when applying a reduced model suggested by the preliminary step, the true effects of age, period, and cohort were well-captured by the simulation.

On the other hand, the critics still express doubt about these efforts. They are skeptical about the possibility of understanding the true data structure by using descriptive and model-fit statistics (Bell and Jones 2015). The critics argue that no statistical model can lend much help in understanding the true data structure, because the confounding by the linearity between the three dimensions is completed during the data-generating process (DGP). Then, at the stage of statistical model application, the data with different structures already appear identical due to the completed confounding, and no statistical model can yield different results when applied to identical data. The critics go on to say that the most we can understand from the preliminary stages of the three-step procedure is whether there exists significant nonlinear variation in each of the three dimensions, and moreover, when any linear trend is present in a data structure, the preliminary analysis cannot reveal much about the true data structure.

The debate on the HAPC and IE has not been settled. The proponents and critics hold fundamentally different understandings of APC data and the ID problem, and those different perspectives lead to diverging interpretations of simulation results. Consequently, scholars in these opposing camps routinely reach conflicting conclusions on the validity of the models. This dissertation aims to provide clarification to the

ongoing debates on the HAPC and IE. Chapters 2 and 3 start by reviewing different perspectives and interpretations of the ID problem and APC data in the contexts of HAPC and IE, and move on to test the validity of the HAPC and IE along with the suggested three-step procedure through sets of simulation analysis. To simulate plausible data structures, some temporal trends are borrowed from previously published studies on obesity (Reither, Hauser and Yang 2009) and suicide mortality in South Korea (Jeon, Reither and Masters 2016). Combining the empirically well-known age trends with possible patterns of period and cohort effects in multiple scenarios, I simulate a variety of realistic data structures in the first two chapters. Then, following the three-step procedure, the HAPC (Chapter 2) and the IE (Chapter 3) are applied to the simulated datasets. Through this focus, Chapters 2 and 3 of this dissertation aim to answer the two following questions:

1. Does the three-step procedure suggested by Yang and Land (2013a) detect the true data structure and inform users of whether to use or not use the HAPC or IE across a variety of datasets?

2. Do the HAPC or IE capture the true data structures when they do not violate model assumptions?

In Chapter 4, I will demonstrate the step-by-step application of the HAPC in empirical research using existing data. To be specific, I will employ the National Health Interview Survey (NHIS) with the goal of understanding age, period, and cohort effects on disparities in self-rated health across levels of educational attainment in the U.S. Following the three-step procedure, this chapter will demonstrate how a researcher can conduct APC analysis with enough caution to valid estimates of all three temporal dimensions. The procedure will include conducting the preliminary analysis, applying the HAPC, interpreting the results, and comparing the results to those of other extant studies.

Furthermore, this chapter is intended to provide an understanding of how real data could be different from simulated APC data that have been widely used in debates about innovative methodologies. By learning more about real APC data structures, the third chapter will add to insights provided in earlier chapters regarding guidelines for simulation analysis and empirical model application, which can contribute to further refinement of APC methodologies in the future.

# CHAPTER II

# PRELIMINARY ANALYSES TO IDENTIFY DATA STRUCTURES IN THE APPLICATION OF HIERARCHICAL AGE-PERIOD-COHORT MODELS

## Introduction

### *Two Understandings of the Identification Problem*

The heart of the identification (ID) problem in age-period-cohort (APC) analysis is that "If age, period and cohort are treated as continuous variables, then it will be impossible to estimate all parameters in a model (of the linear regression form) (Mason, Mason et al. 1973, p 243)." This problem occurs due to a unique property of APC data: the three temporal dimensions are defined in an algebraically linear relationship (A=P-C). Because of this linear dependency, the design matrix of a linear regression model that includes all three temporal terms as regressors is singular with one less than full column rank. The estimator of the linear regression, $(X^TX)^{-1}X^TY$, cannot be obtained, as the term $(X^TX)^{-1}$ does not exist for the singular design matrix. Therefore, this regression model has an infinite number of coefficient sets and cannot identify which one of them corresponds to the true effects of age, period, and cohort. Most scholars agree on this description of the ID problem (Bell and Jones 2014b, Kupper et al. 1985, Yang and Land 2013a).

However, scholars have different perspectives when it comes to the possibility of addressing the ID problem. That is primarily because scholars have different understandings about the nature of the ID problem itself. One group of scholars exclusively focuses on the linear dependency between the three temporal terms when

studying the ID problem and the structure of APC data (Bell and Jones 2014a, Bell and Jones 2014b). Of course, there is no doubt that the linear dependency is a fundamental property of the ID problem. This dependency indeed exists in all APC data as long as the three temporal terms are defined in the linear relationship. In other words, the ID problem is inherent to the APC data, not a statistical model applied to the APC data after confounding between the effects of the three terms is completed.

To better understand this argument, suppose that the following data structure is simulated:

$$Y = (1*Age) + (1*Period) + (1*Cohort) \quad (4)$$

Using the linear relationship between age, period, and cohort (A=P-C), Eq (4) can be converted into the two following equations:

$$Y = 2*Age + 2*Cohort \quad (5)$$

$$Y = 2*Period \quad (6)$$

When researchers have APC data, it means they have the left hand side of the equation (i.e., some outcome variable, Y). The goal of APC analysis is to estimate the coefficients on the right side of the equations by applying the model. However, the three aforementioned equations will produce identical data even though the coefficients of age, period, and cohort on the right hand sides of Eq (4) are not the same as those of Eq (5) and Eq (6). The problem is that a statistical model cannot distinguish between the differences on the right hand sides of Eq (4) – (6) when only the left side is given as APC data. In other words, there is no statistical model that can yield distinguishing results when applied to the same data. According to this understanding, advancement or

modification of the statistical model does not aid in solving the ID problem since it is an intractable feature of the data.

These scholars argue that the only possibility for identifying a single set of coefficients is to find a correct constraint by relying on a solid theory (Bell and Jones 2013). It is not an ultimate solution for the ID problem, but it does produce unique estimates for the coefficient set in the regression model. Imposing a constraint creates a modification to the design matrix X and forces it to be non-singular with full rank. Then the term $(X^TX)^{-1}$ exists and the coefficient vector can be obtained. Although this approach may be useful for estimating the true age, period, and cohort effects, the usage is limited to circumstances in which the solid theory exists. When the theory is not available or sufficiently solid, there is a possibility of obtaining biased outcomes by choosing the wrong constraint (Yang, Fu and Land 2004).

A second group of scholars understands the ID problem from a different perspective. They point out that not one but rather two conditions must be simultaneously met to induce the ID problem: the three temporal effects are (1) linearly related to each other (A=P-C), and (2) linearly related to the outcome of interest (Y=A+P+C) (Mason et al. 1973). The first group of scholars focuses exclusively on the first condition, assuming that it is sufficient to induce the ID problem. The second group of scholars argues that this is not a complete understanding of the ID problem. These scholars argue that while there is no doubt that the linear dependency between A, P, and C is inherent to all APC data, this special property of APC data is only one facet of the ID problem. Indeed, these scholars point out that the first condition becomes problematic because it causes a singular design matrix in the linear regression (i.e., the form of linear models, which

assume the additivity of the independent variables and treat them as fixed effects that are independent from each other), and the $(X^TX)^{-1}$ term for the estimator $(X^TX)^{-1}X^TY$ cannot be obtained. In other words, if we can specify a model wherein the estimator no longer needs the $(X^TX)^{-1}$ term, it may be possible to determine an estimable model specification, even though the linearity between age, period, and cohort still exists in the APC data. Therefore, for these scholars, the ID problem is a model-specific matter, rather than a data-specific one. For example, mixed-models use a different estimator, $(X^TVX)^{-1}X^TV^{-1}Y$, where $V=ZGZ^T+R$, Z contains the predictors of random effects (i.e., intercept and slope), G is a covariance matrix for random effects, and R is a covariance matrix for error terms. In this type of model, even though the linearity between A, P, and C remains, the estimator no longer requires the $(X^TX)^{-1}$ term found in a simple linear model.

*What is the Hierarchical APC Model?*

The Hierarchical APC (HAPC) model was developed by scholars who understand the ID problem as a model-specific matter (Yang and Land 2006, Yang and Land 2008). The model is specifically intended for repeated cross-sectional survey data. Unlike classical tabular rate data presented in the form of an age-by-period matrix, repeated cross-sectional survey data provide a more flexible APC data structure. Whereas a single observation consists of each element of the age-by-period matrix in the tabulated rate data, multiple observations belong to each element of the period-cohort matrix in the repeated cross-sectional data. In this structure, the temporal widths of time periods and birth cohorts are not required to be equal to the temporal width of age, and this allows flexibilities on the exact algebraic relationship between the three temporal terms (Reither

et al. 2015b). Taking advantage of this flexibility, the HAPC has two additional properties in model specification for addressing the second condition of the ID problem: (1) a multi-level setting in which three effects are not assumed to be linear and additive at the same level, and (2) an addition of an age-squared term that estimates a nonlinear function for age and does not assume a linearly additive relationship between the three temporal terms and Y. The model specification of the HAPC cross-classified random effects modeling (CCREM) is as follows:

$$\text{Level 1: } Y_{ijk} = \beta_{0jk} + \beta_1 Age_{ijk} + \beta_2 Age^2_{ijk} + e_{ijk} \text{ with } e_{ijk} \sim N(0, \sigma^2) \quad (7)$$

$$\text{Level 2: } \beta_{0jk} = \gamma_0 + u_{0j} + v_{0k} \text{ with } u_{0j} \sim N(0, \tau_u),\ v_{0k} \sim N(0, \tau_v) \quad (8)$$

$\beta_{0jk}$ is an intercept term, $\beta_1$ and $\beta_2$ are fixed regression slopes, and $e_{ijk}$ is the random individual effect, which is assumed to be normally distributed with mean 0 and variance $\sigma^2$. The intercept term $\beta_{0jk}$ in level 1 is differentiated in Level 2 where $\gamma_0$ is the grand mean of the outcome variable across all individuals; $u_{0j}$ is the period effect for period j averaged over all birth cohorts, which is assumed to follow the normal distribution with mean 0 and variance $\tau_u$; and $v_{0k}$ is the cohort effect for cohort k averaged over all periods, which is assumed to follow the normal distribution with mean 0 and variance $\tau_v$.

According to this model specification, the three effects are not assumed to be additive or linear at the same level. Age is treated as a fixed effect at the individual level, and period and cohorts as random effects that pertain to entire groups of individuals. This allows the level 1 intercept to randomly vary by period and cohort. This multilevel setting is also a conceptually elegant alternative to the conventional linear additive model. The

model specifies that secular changes and cohort membership influence the lives of individuals through "level 2" contextual effects. The fixed effect of age at level 1 can reflect biological and social aging processes and their effects on the outcome of interest over an individual's life course. The random effect of period at level 2 can represent the influences of historical events, demographic or epidemiological transitions, public health interventions, and new technologies on all individuals in the entire population. The random effect of cohort at level 2 can reflect a unique experience of historical or social events at a certain life stage, shared by individuals belonging to the cohort (Reither, Hauser and Yang 2009).

Most importantly, this model identifies a unique set of age, period, and cohort effects without relying on any external information. With these conceptual justifications and statistical advantages, the HAPC has been widely used to study issues such as temporal trends in happiness (Yang 2008b), obesity (Reither et al. 2015b), religious service attendance (Schwadel 2010), cannabis use prevalence (Piontek et al. 2012), and voter turnout volatility (Dassonneville 2013).

*Why Do HAPC Models Sometimes Fail?*

Despite these purported advantages, recent critiques emphasize that estimates from HAPC models do not match the true effects of A, P, and C in data generated through various simulation exercises. Through responses to those critiques and a series of rejoinders, the following has been alleged:

- The HAPC model only works if there are no linear trends in period and cohort (Bell and Jones 2014a).
- The HAPC model fails if there is no period effect (Bell and Jones 2014a).

- The HAPC model fails when descriptive plots of age and cohort effects are mirror images of each other (Reither et al. 2015a).

Depending on the understanding of the ID problem, the interpretation of these results differs. Scholars who believe that the ID problem is data-specific argue that the results are evidence that the HAPC framework fails to address the ID problem. According to them, the conversion of the data structures of Eq (4) into Eq (5) and Eq (6) is still valid even with the introduction of the quadratic age term. As a result, even though the HAPC model identifies a single set of coefficients, it may nevertheless select an arbitrary set of estimates rather than the true set of age, period, and cohort effects.

Proponents of HAPC modeling disagree with this interpretation. For example, Reither et al. (2015a) point out that although the conversion between the data structures is algebraically correct, this holds only with respect to the expected values (population means) of the HAPC model. Thus, the left side of Eq (4) – Eq (6) should be modified to be E(Y), where E(*) is the expectation operator. The HAPC model specifies a level-1 equation for Y, for individuals in a repeated cross-sectional research design where the level-1 equation is a stochastic equation (e.g., a logistic probability function or with an error term that is specified as being distributed according to some probability distribution [e.g., normal]). Then, the exact algebraic identity at the individual respondent level has, practically speaking, a probability of zero. In other words, the way that the critics interpret the results is erroneous, but that is not the reason why the HAPC has failed in their simulation analysis.

Why, then, does the model fail sometimes? Proponents of the model first highlight that it is not surprising that the HAPC fails in some situations. Yang and Land (2013b) state:

> *….in all cases, APC analysis should be approached with great caution and awareness of its many pitfalls. There will never be such a "final" or "universal" solution within the confines of conventional linear models that necessarily beget the ID problem (pg. 1971).*

According to these proponents, the critics of HAPC models engineered failure into their simulation procedures by violating key model assumptions—and subsequently applying HAPC models to data for which they were never designed. Just like any other statistical model, HAPC models are based on statistical assumptions that should be met to make the model specifications valid. For instance, the structure of the data should be three-dimensional, in which all the three temporal terms have obvious contributions to the outcome variable over time (Yang and Land 2013b). When the given data turns out to have a reduced structure (i.e., at least one of the temporal dimensions has essentially no contribution to the outcome of interest), the HAPC model may not work, producing biased results. In addition, none of the three temporal terms should not have perfect functional forms. Given the nature of the APC data, this can cause one dimension to be expressed as a functional form of another dimension, and confounding between dimensions may occur (Yang and Land 2013b). In such cases, the HAPC model may not work, failing to differentiate the confounded effects. This is highly improbable in "real world" data structures, but it has been overlooked in previous simulation procedures.

However, when a proper procedure is followed to ensure that application of a three-dimensional model is appropriate, the HAPC can be a reliable estimator for the true A, P, and C effects. Yang and Land (2013b) suggest that understanding the structure of

the data and checking the key assumptions, by conducting descriptive analysis and obtaining model-fit statistics, is essential prior to the implementation of APC models. By ignoring these important preliminary steps, simulation analyses in the critiques mistakenly applied three-dimensional HAPC models to data that cannot support this. Reither et al. (2015b) showed that these preliminary steps could have helped understanding the data structure in the critiques and avoiding the misuse of the HAPC model.

*What Should Be Studied From Here?*

In summary, scholars agree that the HAPC model is not a magic bullet that can always yield unbiased estimates for repeated cross-sectional data. However, they disagree on whether the potential failure of HAPC models can be predicted by understanding the data structure. Whereas one group of scholars believes it is impossible to understand the true structure of APC data with any statistical method, the other group recommends that practitioners follow a simple process to gather a qualitative and quantitative understanding of the data structure, which will help determine if the given data are suitable for HAPC modeling. When the data satisfy basic assumptions, HAPC models can provide valid and reliable estimates of the true age, period, and cohort effects.

To add evidence to this debate, I simulate various APC data structures and test the validity of HAPC models along with the suggested method for understanding the data structure. The aim of this study is two-fold. First, I will test the validity of the descriptive and model-fit statistics as preliminary steps to prevent the misapplication of the HAPC models by identifying data structures that violate model assumptions. Some of the

simulated data in this study are designed to violate APC model assumptions by having essentially no period or/and cohort effects, or by having linear trends of period and cohort effects at the same time. Such data structures should be distinguishable through the preliminary steps, enabling researchers to avoid biased estimates that can occur when full-three dimensional models are used inappropriately. Second, this study will examine if HAPC models provide valid and reliable estimates when the preliminary steps suggest that the data structure is three-dimensional. Some of simulated data in this study are clearly three-dimensional and do not violate important model assumptions. In such cases, preliminary steps should ascertain the three-dimensional structure, and the estimates of HAPC models should match the true effects that generated the data. By conducting this analysis, this chapter will attempt to answer the following two questions.

- Are the descriptive and model-fit statistics reliable methods for understanding the true data structure in APC analysis?
- When the data is eligible for the HAPC according to the descriptive and model-fit statistics, does the HAPC capture the true age, period, and cohort effects?

## Methods

### *Simulation Design*

The first step in this study is to develop data generating processes (DGPs) that simulate repeated cross-sectional data on obesity among U.S. adults belonging to different ages, periods of observation and birth cohorts. Altogether, thirteen different DGPs were created to produce a range of APC patterns. Some of these patterns are intended to produce large and clearly distinct age, period and cohort effects – situations where HAPC models are expected to capture the true APC patterns in the DGP. Other

patterns are intended to produce less distinct effects or to violate important model assumptions – situations where HAPC models might reasonably be expected to "fail." Descriptions of all thirteen scenarios are presented in Table 1.

The A and B-sets in Table 1 each include three scenarios in which the data structures begin as obviously three dimensional and are gradually shifted toward two dimensional models, as either cohort effects (A-set) or period effects (B-set) are diminished. For the three scenarios in the A-set, there were three different degrees of cohort effects, holding the age and period patterns constant. The coefficients for age and period effects used in the DGP are borrowed from Reither, Hauser and Yang (2009), which used HAPC models to estimate the contribution of age, period and cohort effects to the predicted probabilities of obesity among U.S. adults. The age pattern has a quadratic functional form peaking around the age of 60 (Table 2), and the period trend increases in a monotonic fashion (Figure 1-(a), Table 2). Variations in the cohort effects are also derived from the results presented in Reither, Hauser and Yang (2009). In this study, the magnitude of cohort effects varied substantially across race-sex subgroups, but the pattern was generally similar. This general pattern of cohort effects was a left-skewed V-shape, with the size of the V larger in some groups than others. The scenario in A-1 displays a similar but more obvious V-pattern of cohort effects than the one found by Reither, Hauser and Yang (2009) for any subgroup. Scenario A-2 used the exact cohort coefficients found by Reither, Hauser and Yang (2009) in the overall sample, and A-3 used a similar but less obvious pattern of cohort effects, which roughly follows the flatter pattern detected among non-black males.

For the B-set, another three scenarios are tested, for which the dimensions of the data structure vary by period effects, holding the age and cohort patterns constant. The pattern of age effects is borrowed from a critique of HAPC models (Bell and Jones 2014b), which also has a quadratic functional form that peaks in the late 70s (Table 1). The pattern of cohort effects is set to be the same as that in A-1, which is the most obvious left-skewed V-shape (Figure 1-(b)). Then, holding these conditions for the age and cohort effects constant, three different degrees of the period pattern were established (Figure 1-(b)). Scenario B-1 has the most obvious monotonically increasing period trend, with the coefficients spanning a wide range. In B-2, other settings stay the same, but the period pattern is set to be less obviously increasing. Lastly, B-3 has essentially no period effects, which is the exact same pattern of period effects that Bell and Jones (2014b) simulated in their critique of the HAPC model; it is only re-parameterized to make the sum of the coefficients equals to zero. In this critique, Bell and Jones skipped preliminary analyses recommended by Yang and Land (2013a), concluding that an HAPC model failed to detect the true data structure. I will examine whether conducting preliminary analyses might have helped Bell and Jones understand the true data structure, thus avoiding the sort of misapplication of HAPC models that Yang and Land warn against.

In the C-set, the same procedure is repeated for an even more reduced data structure. The C-set contains only one scenario, in which both period and cohort effects are almost negligible. Thus, the data structure is essentially one-dimensional. The age pattern is the same as that used in the B-set. The period pattern is the same as that in B-3, and the cohort pattern is the same as that in A-1, which are the least obvious period and cohort patterns in the A and B-sets, respectively, with a very low level of variation

allowed. I will examine whether the descriptive and model-fit statistics are useful for understanding the data structure when it is one-dimensional, and if they can aid in selecting the age-only model in particular for this specific data structure in the C-set.

Lastly, six scenarios were designed in the D-set to test what occurs when two of the three temporal dimensions display near-linear trends. One critique by Bell and Jones (2015) was that the preliminary analysis would not be useful for detecting the true data structure when near-linear trends are present in the data structure. To test this critique, I set a near-linear increasing trend for the period effects and allowed for substantial variation in the slope and direction of the near-linear trend for the cohort effects. The coefficients for age are the ones used in the B and C-sets. When both period and cohort effects display linear patterns, one can express the other in a functional form. This is likely to cause serious confounding between period and cohort effects and make it challenging to detect the true data structure in the preliminary analysis. For example, when the period has a linearly increasing trend and the linear cohort trend is obviously decreasing (i.e., D-6), the data is truly three-dimensional, as all of the three temporal dimensions make clear contributions to the outcome of interest. However, the offset between the two dimensions due to confounding may occur, and the descriptive and model-fit statistics may recognize the data structure as one-dimensional data in which only age effects are present. To provide clarification on this concern, I investigate how the descriptive and model-fit statistics react to such data structures.

All the period and cohort patterns shown in the scenarios are simulated in the DGPs as *near* linear (or polynomial) functional forms. Reither et al. (2015b) pointed out an important error in a previous critique (Bell and Jones 2014b), which was failing to

permit some deviation from *perfect* functional forms. Because perfect functional forms generate data structures that make the APC model specification invalid (Yang and Land 2013b) and are highly improbable in any data created by actual social and historical phenomena, I avoided random effects aligned on exact functional forms, and allowed enough deviations to make the data structures realistic.

*Simulation Procedure*

To start, I use Stata 13 to simulate data following the DGPs presented in the previous section. For each scenario in the A-set, I simulate 50 datasets containing one million observations, which is a similar sample size to that used in Reither, Hauser and Yang (2009). For such a large sample size, little variance in the descriptive statistics (i.e. prevalence) and the model estimates across the simulated datasets are expected. Therefore, I simulated only 50 datasets, which is half the number of the datasets generated in a previous HAPC simulation study (Reither et al. 2015a). For each scenario in the B, C, and D-sets, 100 datasets containing 50,000 observations are simulated. The sample size is set to be relatively larger than what was used (i.e. 20,000 or 30,000) in previous critiques of HAPC models and rejoinders to those critiques (Bell and Jones 2014a, Bell and Jones 2014b, Reither et al. 2015a, Reither et al. 2015b). I expect the larger sample size to provide more consistent results across simulated datasets compared to previous studies, and also provides a more robust rationale to decide the best fitting model when goodness-of-model fit statistics suggest different models. Considering this sample size is relatively smaller than the one used in the A-set, I simulated twice as many datasets in the B, C, and D-sets than in the A-set for encompassing wider ranges of

variances in estimates across datasets, which is also consistent with the number of dataset simulated in the previous study (Reither et al. 2015a).

Each simulated dataset has a continuous term for age ranging from 18 to 92 and 27 periods of observation ranging from 1976 to 2002. For each observation, I randomly assign age and period within these ranges. Then, birth year is calculated by subtracting age from period, which generates 18 5-year birth cohorts ranging from 1890 to 1980. Finally, age is centered at 25 to reduce the association between age and age-centered terms (Reither, Hauser and Yang 2009).

*Three-step Procedure*

The three-step procedure suggested by Yang and Land (2013) is applied to the simulated data using R 3.2.2. In Step 1, I create descriptive plots of obesity prevalence across all three temporal dimensions to gather an initial impression of variability across age groups, periods of observation and birth cohorts. The prevalence of obesity is calculated by dividing the frequency of Y=1 (obese) by the total frequency in each age, period and cohort group. To ease the interpretation, these proportions are converted into percentages, and median percentages are estimated across all 50 datasets in A-set and 100 data sets in B and C-sets. In Step 2, goodness-of-fit statistics, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), are calculated for the 7 nested models (i.e. A, P, C, AP, AC, PC, APC) for each dataset, and the best fitting model is tallied across all simulated datasets for each scenario. In their critique on the three step procedure of the HAPC application, Bell and Jones (2015) point out that "different model fit statistics give different answers (p. 332)". The inconsistency between AIC and BIC is

not surprising, and it is often observed in empirical research settings. This is mainly because those two methods penalize the number of parameters included in the model using different weights. AIC uses 2 as the weight by which to penalize the number of parameters, while BIC uses the logged sample size, which is usually heavier than 2. Note that the sample sizes in this study are set to be as big as one million and 50,000. In such cases, the difference in the number of parameter between the nested and the full models is amplified by the logged sample size (i.e., ln(1,000,000)=13.8, ln(5,000)=10.82), which has a pronounced influence on BIC. Due to this significant penalty, BIC tends to favor reduced models using fewer parameters with large sample sizes, which can lead to the omission of temporal dimensions that are present in the data structure and potentially important from a substantive point of view. For this reason, I will rely more on the AIC than the BIC in this study when those two model fit statistics are not consistent. However, I still present the BIC along with the AIC to show how those two criteria may differ with respect to identifying the best fitting model in the context of large sample sizes. In Step 3, the HAPC is applied only when all three dimensions appear to be active in the results from the descriptive analysis and goodness-of-fit tests. To evaluate the validity of the HAPC, the coefficients of age, period and cohort effects are estimated and compared to the true data structure that is used to generate the simulated dataset.

## Results

### *A-Set*

Figure 2 presents results of the descriptive analysis for the scenarios in the A-set, which includes substantial variability in cohort effects but holds age and period effects

constant. First, the descriptive plots by age and period maintain highly similar patterns regardless of the variation in the cohort effects. For all three scenarios, the descriptive age plots are concave, facing downward, with very similar curvatures and peaks ranging from 50 to 60. The pattern of the period plots is a monotonic increase displaying similar rate ranges for all three scenarios. The differences between the scenarios are revealed by the descriptive statistics on cohort, especially with respect to cohorts born after 1960. For all three scenarios, the prevalence increases from the earliest cohorts to the cohorts born in the 1930s, and then fall off until the cohort born in the 1960s. Then, for A-1 and A-2, which have more obvious cohort patterns than A-3, the prevalence climb thereafter. Of those two, the climb is steeper in A-1, which has an even more obvious cohort pattern. On the other hand, in A-3, which has essentially no cohort trend, the prevalence keeps decreasing after the birth cohort born in 1960, but at a gentler pace than the decreasing trend of earlier cohorts born between 1930 and 1960.

Note that descriptive statistics do not perfectly disentangle the effects of the three temporal terms. The prevalence of obesity by age, period, and cohort is a raw statistic, which does not control for the other terms' influence. In APC studies, descriptive plots for age and cohort are often nearly mirror images of each other, because people in the later age groups consist of earlier birth cohorts. Because of this property, we can learn one important thing if we compare descriptive age and cohort plots: If they have obviously different shapes, it is likely that age and cohort make unique contributions to the outcome of interest. Looking at the results of the A-set, the descriptive plots for those two terms are not close to mirror images. This implies that both age and cohort dimensions are creating their own patterns, so a model that includes both temporal

dimensions could be selected. On the other hand, as the cohort effects become less obvious from A-1 to A-3, the shape of the cohort plot begins to approach that of the age plot. When such thing is observed, either of those dimensions is likely to be inactive in the true data structure.

As all three dimensions in the A-set scenarios clearly have their own patterns, it is reasonable for a researcher to guess here that the data is fully three-dimensional. The model-fit statistics can add clarification to this guess. Table 3 presents how many simulated datasets out of 50 make the best fit for each of the seven nested APC models, when the model fits are estimated by AIC and BIC. For A-1, all 50 simulated datasets make the best fit with the fully three-dimensional model, according to AIC and BIC. Interpreted along with the descriptive statistics, this is strong evidence that A-1's data structure is three-dimensional. Therefore, the analysis can move on to the last step of the three-step procedure, applying the HAPC model to the data.

For A-2, which has a less obvious cohort pattern than A-1, the AIC points toward the full three-dimensional structure for all 50 datasets. However, BIC points toward the AP model for all 50 datasets. The difference in the goodness-of-fit between the AP and APC models may be smaller in A-2 than in A-1, and the penalty on the number of parameters that the BIC imposes is too heavy in that it makes the AP model preferable to the APC model. Considering the very large sample size, I rely more on the AIC than the BIC. In summary, it is a reasonable guess here that the data structure of A-2 is three-dimensional based on the AIC along with the descriptive plots.

For A-3, in which the pattern of the cohort effects was the least obvious in the A set, AIC still finds that 45 out of 50 datasets are three-dimensional, while 5 are two-

dimensional with active A and P dimensions. As the true data structure has very small cohort effects, it seems that the random variation imposed during the DGP produces the 10% of the simulated dataset that has further reduced structures than the rest. From the results, we can learn that the data structure is on the border between the three- and two-dimensions. Since the AIC points toward the three-dimensional model for the majority of the simulated datasets, I attempt to move on to Step 3 and applied the HAPC to the 50 datasets in A-3.

The results of the HAPC estimates in comparison with the true DGPs are presented in Figure 3. Overall, the HAPC models detect the true data structure quite well. The HAPC models catch the age trend accurately, and they also perform very well with the true period and cohort trends. Although the latest periods and earliest cohorts in A-1 tend to have wider ranges of coefficients, the patterns in the DGPs are well captured by the HAPC estimates. In particular, even for A-3 where the cohort effects are the least obvious, the HAPC successfully estimates the DGP of all three dimensions.

What if the given data is one of those five, for which the AIC suggested AP instead of the three-dimensional model in A-3? Could the AP model still estimate the true age, period and cohort effects? To test that, I applied hierarchical AP models for those five data sets, and the results are presented in Figure 3 (part d). These “HAP” models tend to underestimate the age effects, especially for those in the middle ages, but the overall patterns matches that of the true data structure. The period effects are perfectly captured by the HAP model. Based on this result, a researcher would conclude that the data has quadratic age effects, monotonically increasing period effects, and no significant cohort effects. In other words, the conclusion drawn by using the AP model is not

substantively different from that made by using the APC model, and these conclusions well reflect the true data structure. Also, this result confirms that a researcher can use whichever model the AIC suggests.

*B-Set*

With the B-set, I imposed variations on the period effects while holding age and cohort effects constant. As with the A-set, the descriptive period plots well reflect the DGP in the B-set (Figure 4). The increase in obesity prevalence is most obvious for B-1, less obvious for B-2, and almost flat for B-3, corresponding to pattern in each scenario. Descriptive plots of age and cohort are not heavily affected by variation in period effects. Obesity prevalence among cohorts born after 1960 shows some differences across the three scenarios, but the overall pattern of age and cohort effects does not change.

It is worth comparing B-1 with A-1. Those two scenarios have a very similar DGP: quadratic age effects, monotonically increasing period trends, and left-skewed V-shaped cohort effects. However, the patterns of age and cohort plots are very different. The age and cohort plots in B-1 are close to mirror images of each other, while those in A-1 are not. The only major difference between the two scenarios is the age coefficients. Note that the coefficient for the age term used in DGP for A-1 was smaller (i.e., 0.059) than the one used for B-1 (i.e., 0.1). That means the relative contribution of the cohort effects to the age effects may be smaller in B-1 than in A-1. This implies that as either age or cohort effects overpower the other, the descriptive plots of age and cohort may become closer to mirror images. This occurs because the dominant dimension of those two effects primarily drives the changes to the prevalence. Consequently, in APC data

where age and cohort have an inverse relationship (i.e., the older people consist of the earlier cohorts, and the younger people consist of the later cohorts), the prevalence of the minor dimension of those two factors is not largely different from a mirror image of the major dimension.

For B-1, researchers may think that the actual data structure is two-dimensional (i.e., AP or CP here), as descriptive plots of two dimensions are close to mirror images. Otherwise, they may think that it is still three-dimensional because the descriptive cohort plot has a flattening right tale, which the age plot does not have. In such an ambiguous situation, a researcher should obtain supplementary information from the model-fit statistics to gain a clearer idea of the data structure before applying any model.

The model-fit statistics calculated for the scenarios in the B-set are presented in Table 4. Although both AIC and BIC are calculated for the scenarios in the B-set, I would rely more on AIC, considering that 50,000 is still quite a large sample size and may create too heavy of a penalty for the number of parameters. For B-1, in which the pattern of the period was the most obvious in the B-set, the AIC shows that all of the 100 datasets are three-dimensional as expected. For B-2, which has a less obvious pattern of period effect than B-1 does, the AIC still suggests that all of the 100 datasets are three-dimensional. Lastly, for B-3, which has almost no contribution from period, the AIC swings between AC and APC models. The AIC points toward the AC model for 78 of 100 datasets, while 22 are still considered to have the full APC structure. This is good evidence that the data structure is on the border between the full and a reduced dimensions, and a reduced model (i.e. AC) could be considered for some of the datasets. Based on the descriptive and model-fit statistics, it is a reasonable choice for a researcher

to move on to apply the HAPC to B-1 and B-2. For B-3, the decision is not very simple. If a given data is one of those 78 datasets, for which the AIC suggested AC, a researcher is likely to choose a reduced model (i.e., AC), and this decision also makes sense with the descriptive plots. However, if a given data is one of those 22 datasets, for which the AIC suggested the full three-dimensional model (i.e., HAPC), there is a conflict between the results of the descriptive and model-fit statistics. In such a case, it is not recommended to move on to Step 3 because the data is likely to be on the border line of the two structures, and there is no guarantee that the HAPC can provide unbiased estimates.

The results of the HAPC models are presented in Figure 5. For B-1 and B-2, HAPC models capture the DGP for most of the age, period, and cohort groups. While these models tend to slightly underestimate the DGP for later ages, later periods, and earliest cohorts, these discrepancies are very small. For B-3, the AC model works quite well. Although the model slightly underestimates the age effects over all the ages, the pattern of the effects are very well captured. Estimates for the cohorts are very close to the DGP, except for the first four cohorts wherein the model slightly underestimates the DGP.

*C-Set*

The scenario of C-1 has an obvious age pattern and minimal period and cohort patterns. This data structure is essentially one-dimensional. The descriptive plot accurately reflects the temporal patterns of obesity prevalence embedded within the DGP (Figure 6). An important finding is that the descriptive period plots tend to reflect the true DGP of the period effects, not becoming seriously confounded by age and cohort effects

across all of the scenarios in the A, B, and C-sets. In addition, the descriptive plots of age and cohort are almost perfect mirror images of each other in the B and C-sets. As noted in B-1, when one of those two dimensions overpowers the other, the descriptive plots are likely to be mirror images. As the C-1 DGP is dominated by age with almost no cohort effects, it is not surprising that those mirror images between age and cohort show up in the descriptive plots. At this point, a researcher may suspect that the data are one-dimensional, but cannot be certain whether age or cohort dominates the data structure. At that stage, a researcher can move onto Step 2: obtaining the model-fit statistics.

According to the AIC, 65 of 100 datasets fit the best with the A model, implying that those datasets are essentially one-dimensional. Eight and 27 fit best with the AC and AP models, respectively. It is unclear why the AIC detects C and P dimensions, but it seems likely that it stems from deviations assigned to random effects during the DGP. In any case, a researcher should not apply the HAPC here, but rather should consider selecting a reduced model. As suggested for the majority of cases, when the A model is applied to all the 100 dataset, the model captures the true trend of age very well (Figure 7). Also when the AP model is fitted to those particular 27 dataset as the AIC suggested, the AC model captures the true pattern of age effects quite closely, and the results correctly detect that the data has almost no period effects. In other words, whichever model is applied, the conclusion on the contribution of age, period and cohort would not significantly change, as far as it is the model suggested by the AIC.

*D-Set*

Figure 8 shows the visualized descriptive statistics of the six scenarios in the D-set. They have linear patterns of cohort effects with different slopes and directions, as well as constant age and period patterns. When the period and cohort have linear patterns in the same positive direction (D-1 to D-3), the descriptive period plots pick up the pattern of the true coefficients quite well. The variations to the cohort effects are shown by the changes in the descriptive cohort plots. As the slope of the linear trend of the cohort coefficients becomes flatter from D-1 to D-3 and the relative contribution of cohort becomes less than that of age, the descriptive plots of age and cohorts become closer to mirror images of each other.

When period and cohort have linear patterns in opposite directions (D-4 to D-6), descriptive statistics of all three dimensions are affected by the variation of the cohort effects. First, the descriptive period pattern becomes flatter as the negative cohort effects have larger slopes. This is different from what was observed with the scenarios in the A, B, and C-sets. In the A, B, and C-sets, the period plots pick up the true pattern of the coefficients very well, without being greatly affected by variation of the cohort effects. However, these last three scenarios in the D-set show that the descriptive period pattern can be confounded by the other terms under certain conditions. As the cohort effects have obvious negative linear patterns that may offset the positive linear patterns of period effects, the period plot becomes closer to a flat line even though the true period effects still have an obvious increasing trend. In addition, as the cohort effects have negative patterns that are offset by the effects of period, the descriptive cohort pattern remains similar to the mirror images of age patterns. As a result, descriptive plots look similar

with those in which there are no obvious cohort effects although dominant age effects are present. At this point, a researcher may conclude that the data are one-dimensional, although it is actually three-dimensional, with period and cohort effects in opposite directions offsetting each other.

Do the model-fit statistics help identify the true data structure, which the descriptive statistics do not capture here? The answer is no. For the first five scenarios (i.e. D-1 to D-5), the AIC suggests using AP and AC models, rather than the full three-dimensional model. In other words, the AIC tends to detect only one of those two linear dimensions active, and is incapable of distinguishing between the two linear dimensions. This is especially true for D-6, in which the cohort has the most obvious decreasing trend and the AIC suggests using the single-factor A model in most cases. These results show that confounding occurs when two dimensions exhibit clear linear patterns, and the AIC does not help for distinguishing those different data structures after the DGP is completed.

Two important things can be learned from D-set. First, when the descriptive plots of all the three dimensions have distinct patterns, which are not a perfect mirror image of another dimension's plot or flat line (i.e. D-1, D-2, and D-3), a researcher may guess that the given data is three-dimensional, as learned in A-set. However if model selection statistics suggest a reduced model over the three dimensional-model, which does not match to the results of the descriptive analysis, s/he can suspect that some of the three dimensions may have been seriously confounded due to their linear trends. The HAPC modeling approach should never be applied in this case. For example, from D-1, D-2, and D-3, period and cohort have linear effects in the same direction, and they are not

offsetting each other. However, the confounding between these two dimensions may still occur and HAPC models fail to capture the true data structures although the data are three-dimensional as shown in Figure 9. In other words, to ensure that the given data structure is three-dimensional, *both* descriptive and model fit statistics should indicate so. Second, when the descriptive plots and the model fit statistics point to the same model, and when it is a reduced model such as D-4 and D-5 for which the AIC suggests the AP model, and D-6 for which the AIC suggests the A model, researchers should not mechanically conclude that the data really has a reduced structure. A researcher should always be aware that certain three-dimensional data in which two linear dimensions are confounded could look like reduced–structural data from the descriptive and model selection statistic.

**Discussion**

The first goal of this study is to test the validity of the descriptive and model fit statistics for identifying data structures that violate the assumptions of the HAPC model, and are not suitable for the HAPC analysis. For example, Yang and Land (2013a) argued that data that has a reduced structure is not applicable to the HAPC models. In this study, I tested how the descriptive and model fit statistics react to reductions in the dimensions of the data derived from little contribution of one or two dimensions. The results from A, B, and C-sets show that those two statistics are very useful and effective tools to understand the true data structure, unless more than one of the three dimensions have linear effects at the same time.

In particular, the descriptive period plots reflect the trend of the true period effects very well across the scenarios in the first three sets, and this was an important clue that assisted in deciding whether the period is an active dimension in the given data. When it comes to age and cohort trends, comparing the patterns of their descriptive plots was important. When descriptive plots of age and period have dissimilar shapes, it is likely that each dimension has obvious and nonlinear effects. When they are close to mirror images of each other, it is likely that only one of those two dimensions has a major contribution to the outcome variable.

By obtaining general idea on the data structure from the descriptive plots, a researcher should be able to narrow down the probable structures of the given data. At this stage, checking those data structures with the model-fit statistics could give further clarification. When the data structure was on the border between different dimensions due to very little effects of one or two dimensions (i.e. A-3, B-3, and C-1), the AIC tended to swing between several different nested models. In this case, the results from several scenarios in this chapter (i.e. A-3, B-3, and C-1) confirmed that selecting the best model to apply among them did not significantly change the conclusion that a researcher would make on the contributions of the three temporal dimensions.

In particular, the results of C-set showed how one-dimensional data turned out in the results of descriptive and model fit statistics. By observing the descriptive period plot which appears to be a flat line, and the age and cohort plots which are exact mirror images of each other, a researcher should be quite sure that the model is actually very close to single dimensional. The AIC clarified which of the dimensions is active and suggested the best fitting model to apply. When reduced models suggested by the AIC

(i.e. A and AP) were applied to the simulated dataset, both models nicely captured the true structure. As the period effects are very minor and the AP model can estimate them correctly, the conclusion on each dimension's contribution does not change by the selection between A and AP models. However, a researcher should never apply the full-three dimensional model in this case since none of the descriptive and model fit statistics showed evidence of the three-dimensional structure.

The second goal of this study is to test whether HAPC models provide unbiased estimates of age, period and cohort effects when preliminary steps indicate that all three dimensions are present in the data. That means, when the data has obvious contribution of all three dimensions and no confounding by having the exact linear dimensions, the preliminary analysis should be able to detect that such data is eligible for the HAPC models, and the model estimate should capture the true DGP. The results of A-1, A-2, B-1, and B-2 confirm that when *both* descriptive and model-fit statistics clearly suggest the three-dimensional model, the HAPC is a very useful estimator for obtaining true age, period and cohort effects. For the vast majority of the simulated datasets in those scenarios, both the descriptive plots and the AIC distinctly indicated that all three dimensions were active. For those data, the HAPC performed great, and most of the DGPs are captured closely by the estimates for the simulated data sets.

The results of B-set particularly highlight the importance of using descriptive and model fit statistics together. The descriptive age and period plots of the scenarios in B-set were somewhat close to the mirror images of each other, but not perfectly matched due to the small variations on the right tale of the cohort plots caused by the later birth cohorts. In this case, a researcher may not be sure if both dimensions are active. The AIC

provided supplementary information necessary in such a situation for confirming the active dimensions. For B-1 and B-2, the AIC pointed at the full-factor APC model, indicating that all the three dimensions are active. The HAPC performed great for those two scenarios. For B-3, the AIC suggested AC and APC models, which represent that period may be an inactive dimension for some of the simulated data sets. As the period descriptive plot look flat supporting this guess, it is not recommended to apply the HAPC in this case. Instead, when AC model was applied, the estimates of age and cohort effects match up with the true structures. While A, B, and C-sets clarify the scope of the application of the HAPC using the three-step procedure, the results of D-set revealed a limitation of the current application procedure of the HAPC. According to the results, when more than one dimension have linear effects at the same time, confounding between those two dimensions' effects may occur. In particular, when those linear effects have opposite directions, an offset between those two effects should be expected. Consequently, the results of the preliminary analysis look similar to those of reduced-dimensional data even though the data is actually three-dimensional. Although the current procedure is not capable of distinguishing those two different data structures, practitioners should be aware that in either case the HAPC model should not be applied. If applied, the HAPC model will fail to estimate the true age, period and cohort effects.

Overall, the results of Chapter 2 support Yang and Land (2013) argument that the HAPC is still a reliable estimator when used with enough understanding on the data structure. When *both* the descriptive and model fit statistics point to the three-dimensional model for the given data, the HAPC successfully captures the DGP. If either or both the descriptive and model fit statistics point to the reduced model, a conclusion on

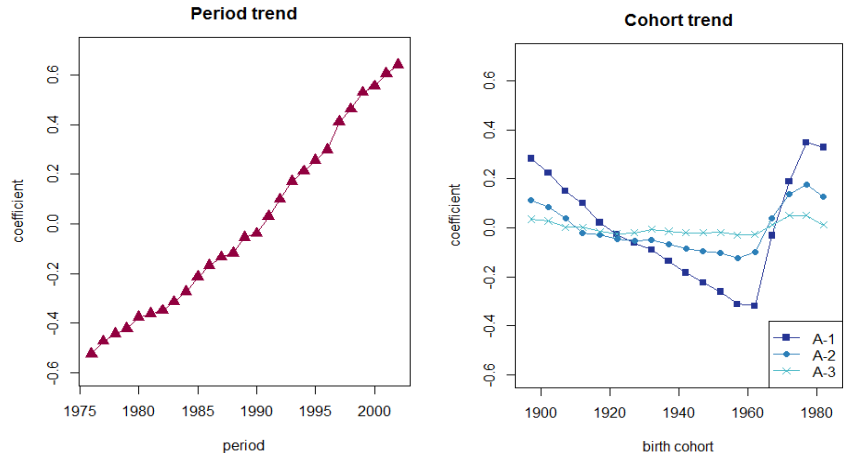
the data structure should be deferred, and the HAPC should never be applied. Using the current procedure of the APC application, it is hard for a researcher to know whether the data really has the reduced structure, or the data is actually three-dimensional but looks like reduced-dimensional from the preliminary analysis due to the confounding between two or three linear dimensions. For either, it is meaningless to attempt to apply the HAPC in this situation because the preliminary steps do not suggest the full three-dimensional models, and the HAPC is destined to fail.

The simulated data by Bell and Jones (2014b) had such structures. For example, they simulated dataset that has reduced dimensions (i.e. no period trend) or dataset that may look like reduced dimensional from the preliminary analysis due to the confounding between dimensions (i.e. correlated age and period trends). If they conducted the preliminary analysis with these dataset, the results would have suggested a reduced model rather than the three-dimensional model (Reither et al. 2015a, Reither et al. 2015b). The relevant procedure at this stage would be to defer the conclusion on the data structure, rather than arbitrarily assuming the data is three-dimensional and moving on to the application of the HAPC.

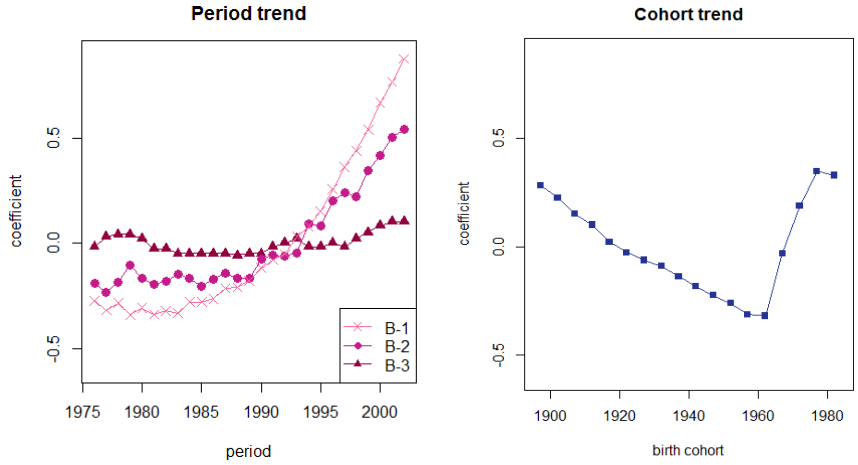
Lastly, according to the results of this chapter, one of the arguments of Bell and Jones (2014a) that the HAPC fails when period and cohort have linear trends is true. However, this should not be used as evidence to say that the HAPC fails to address the ID problem, because the HAPC does perform great in some situations, and those situations can be predicted by the preliminary analysis. Researchers still can fully take advantage of reliable estimates of the HAPC in their empirical studies as far as results of the preliminary analysis indicate that the given data is three-dimensional.

## Tables and Figures

(a) A-Set




(b) B-Set




(c) C-Set

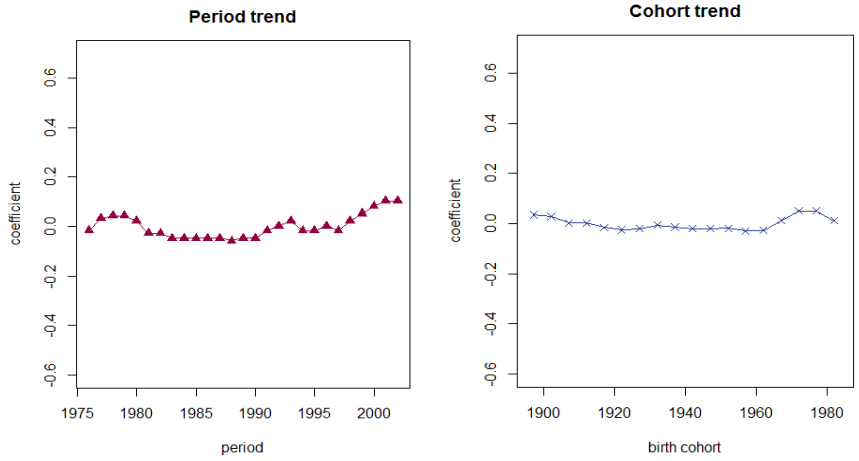



Figure 1. The Simulation Scenarios from A-Set to D-Set

(d) D-Set

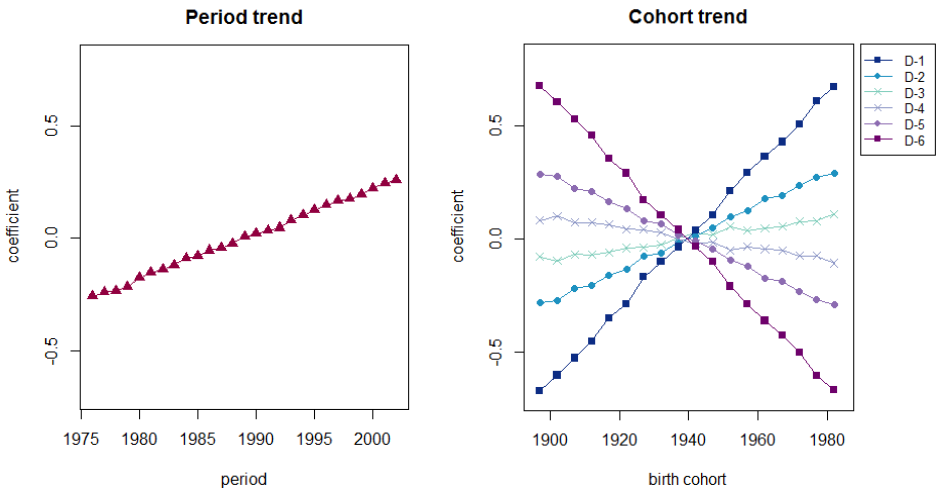



Figure 1 (cont'd). The Simulation Scenarios from A-Set to D-Set

Table 1. Description of Age, Period and Cohort Effects Simulated for Each Scenario

| Set | Scenario | Age Effects | Period Effects | Cohort Effects |
|---|---|---|---|---|
| A | A-1 | Quadratic | Linear increase (very obvious) | V-shape (very obvious) |
| | A-2 | Quadratic | Linear Increase (very obvious) | V-shape (less obvious) |
| | A-3 | Quadratic | Linear increase (very obvious) | V-shape (not obvious, essentially no trend) |
| B | B-1 | Quadratic | Increase (very obvious) | V-shape (very obvious) |
| | B-2 | Quadratic | Increase (less obvious) | V-shape (very obvious) |
| | B-3 | Quadratic | Not obvious, essentially no trend. (Coefficients from Bell and Jones 2015) | V-shape (very obvious) |
| C | C-1 | Quadratic | Essentially no trend | Essentially no trend |
| D | D-1 | Quadratic | Linear Increase | Linear Increase (Steepest Slope) |
| | D-2 | Quadratic | Linear Increase | Linear Increase (Less Steep Slope) |
| | D-3 | Quadratic | Linear Increase | Linear Increase (Least Steep Slope) |
| | D-4 | Quadratic | Linear Increase | Linear Decrease (Least Steep Slope) |
| | D-5 | Quadratic | Linear Increase | Linear Decrease (Less Steep Slope) |
| | D-6 | Quadratic | Linear Increase | Linear Decrease (Steepest Slope) |

Table 2. DGPs for Each Scenario

(1) A-Set

| | |
|---|---|
| A-1 | Logit[Pr(Y=1)] =-1.9876 + (.0589*Age)+(-.001*Age$^2$)+(-.5243*P1)+(-.4740*P2)+(-.4430*P3)+(-.4228 *P4)+(-.3767*P5)+(-.3618*P6)+(-.3492*P7)+(-.3143*P8)+(-.2747*P9)+(-.2143*P10)+(-.1685*P11)+(-.1346*P12)+(-.1174*P13)+(-.0544*P14)+(-.0390*P15)+(.0292*P16)+(.0992*P17)+(.1720*P18)+(.2118*P19)+(.2559*P20)+(.2989*P21)+(.4103*P22)+(.4625*P23)+(.5291*P24)+(.5549*P25)+(.6053*P26)+(.6400*P27)+(.2827*C1)+(.2255*C2)+(.1501*C3)+(.1003*C4)+(.0216*C5)+(-.0270*C6)+(-.0623*C7)+(-.0889*C8)+(-.1365*C9)+(-.1833*C10)+(-.2234*C11)+(-.2613*C12)+(-.3122*C13)+(-.3192*C14)+(-.0314*C15)+(.1881*C16)+(.3484*C17)+(.3288*C18)+$U_c$+$U_p$,<br>$U_c$~N(0, .01), $U_p$~N(0, .01) |
| A-2 | Logit[Pr(Y=1)] =-1.9876 + (.0589*Age)+(-.001*Age2)+(-.5243*P1)+(-.4740*P2)+(-.4430*P3)+(-.4228 *P4)+(-.3767*P5)+(-.3618*P6)+(-.3492*P7)+(-.3143*P8)+(-.2747*P9)+(-.2143*P10)+(-.1685*P11)+(-.1346*P12)+(-.1174*P13)+(-.0544*P14)+(-.0390*P15)+(.0292*P16)+(.0992*P17)+(.1720*P18)+(.2118*P19)+(.2559*P20)+(.2989*P21)+(.4103*P22)+(.4625*P23)+(.5291*P24)+(.5549*P25)+(.6053*P26)+(.6400*P27)+(.1127*C1)+(.0855*C2)+(.0401*C3)+(-.0203*C4)+(-.0284*C5)+(-.0470*C6)+(-.0523*C7)+(-.0489*C8)+(-.0665*C9)+(-.0833*C10)+(-.0934*C11)+(-.1013*C12)+(-.1222*C13)+(-.0992*C14)+(.0386*C15)+(.1381*C16)+(.1784*C17)+(.1288*C18)+Uc+Up,<br>Uc~N(0, .01), Up~N(0, .01) |
| A-3 | -1.9876 + (.0589*Age)+(-.001*Age2)+(-.5243*P1)+(-.4740*P2)+(-.4430*P3)+(-.4228 *P4)+(-.3767*P5)+(-.3618*P6)+(-.3492*P7)+(-.3143*P8)+(-.2747*P9)+(-.2143*P10)+(-.1685*P11)+(-.1346*P12)+(-.1174*P13)+(-.0544*P14)+(-.0390*P15)+(.0292*P16)+(.0992*P17)+(.1720*P18)+(.2118*P19)+(.2559*P20)+(.2989*P21)+(.4103*P22)+(.4625*P23)+(.5291*P24)+(.5549*P25)+(.6053*P26)+(.6400*P27)+(.0355*C1)+(.0283*C2)+(.0029*C3)+(.0031*C4)+(-.0156*C5)+(-.0242*C6)+(-.0195*C7)+(-.0061*C8)+(-.0137*C9)+(-.0205*C10)+(-.0206*C11)+(-.0185*C12)+(-.0294*C13)+(-.0264*C14)+(.0114*C15)+(.0509*C16)+(.0512*C17)+(.0116*C18)<br>+Uc+Up,<br>Uc~N(0, .01), Up~N(0, .01) |

Table 2 (cont'd). DGPs for Each Scenario

(2) B-Set

| | |
|---|---|
| B-1 | Logit[Pr(Y=1)] =-2 + (.1*Age)+(-.001*Age2)+(.0174*P1)+(.0326*P2)+(.0426*P3)+(.0426*P4)+(.0226*P5)+(-.0274*P6)+(-.0274*P7)+(-.0474*P8)+(-.0474*P9)+(-.0474*P10)+(-.0474*P11)+(-.0474*P12)+(-.0574*P13)+(-.0474*P14)+(-.0474*P15)+(-.0174*P16)+(.0026*P17)+(-.0226*P18)+(-.0174*P19)+(-.0174*P20)+(.0026*P21)+(-.0174*P22)+(.0226*P23)+(.0526*P24)+(.0826*P25)+(.1026*P26)+(.1026*P27)+(.2827*C1)+(.2255*C2)+(.1501*C3)+(.1003*C4)+(.0216*C5)+(-.0270*C6)+(-.0623*C7)+(-.0889*C8)+(-.1365*C9)+(-.1833*C10)+(-.2234*C11)+(-.2613*C12)+(-.3122*C13)+(-.3192*C14)+(-.0314*C15)+(.1881*C16)+(.3484*C17)+(.3288*C18)+Uc+Up, Uc~N(0, .01), Up~N(0, .01) |
| B-2 | Logit[Pr(Y=1)] =-2 + (.1*Age)+(-.001*Age2)+(-.1875*P1)+(-.2298*P2)+(-.1836*P3)+(-.1038*P4)+(-.1639*P5)+(-.1959*P6)+(-.1812*P7)+(-.1444*P8)+(-.1669*P9)+(-.2023*P10)+(-.1707*P11)+(-.1425*P12)+(-.1666*P13)+(-.1662*P14)+(-.0759*P15)+(-.0572*P16)+(-.0606*P17)+(-.0438*P18)+(.0914*P19)+(.0839*P20)+(.2022*P21)+(.2406*P22)+(.2198*P23)+(.3449*P24)+(.4171*P25)+(.5009*P26)+(.5422*P27)+(.2827*C1)+(.2255*C2)+(.1501*C3)+(.1003*C4)+(.0216*C5)+(-.0270*C6)+(-.0623*C7)+(-.0889*C8)+(-.1365*C9)+(-.1833*C10)+(-.2234*C11)+(-.2613*C12)+(-.3122*C13)+(-.3192*C14)+(-.0314*C15)+(.1881*C16)+(.3484*C17)+(.3288*C18)+Uc+Up, Uc~N(0, .01), Up~N(0, .01) |
| B-3 | Logit[Pr(Y=1)] =-2 + (.1*Age)+(-.001*Age2)+(-.2713*P1)+(-.3188*P2)+(-.2828*P3)+(-.3391*P4)+(-.3063*P5)+(-.3373*P6)+(-.3193*P7)+(-.3312*P8)+(-.2775*P9)+(-.2779*P10)+(-.2619*P11)+(-.2137*P12)+(-.2085*P13)+(-.1764*P14)+(-.1160*P15)+(-.0796*P16)+(-.0500*P17)+(.0317*P18)+(.0800*P19)+(.1525*P20)+(.2558*P21)+(.3628*P22)+(.4393*P23)+(.5390*P24)+(.6664*P25)+(.7656*P26)+(.8742*P27)+(.2827*C1)+(.2255*C2)+(.1501*C3)+(.1003*C4)+(.0216*C5)+(-.0270*C6)+(-.0623*C7)+(-.0889*C8)+(-.1365*C9)+(-.1833*C10)+(-.2234*C11)+(-.2613*C12)+(-.3122*C13)+(-.3192*C14)+(-.0314*C15)+(.1881*C16)+(.3484*C17)+(.3288*C18)+Uc+Up, Uc~N(0, .01), Up~N(0, .01) |

(3) C-Set

| | |
|---|---|
| C-1 | Logit[Pr(Y=1)] =-2 + (.1*Age)+(-.001*Age2)+(-.0174*P1)+(.0326*P2)+(.0426*P3)+(.0426*P4)+(.0226*P5)+(-.0274*P6)+(-.0274*P7)+(-.0474*P8)+(-.0474*P9)+(-.0474*P10)+(-.0474*P11)+(-.0474*P12)+(-.0574*P13)+(-.0474*P14)+(-.0474*P15)+(-.0174*P16)+(.0026*P17)+(-.0226*P18)+(-.0174*P19)+(-.0174*P20)+(.0026*P21)+(-.0174*P22)+(.0226*P23)+(.0526*P24)+(.0826*P25)+(.1026*P26)+(.1026*P27)+(.0355*C1)+(.0283*C2)+(.0029*C3)+(.0031*C4)+(-.0156*C5)+(-.0242*C6)+(-.0195*C7)+(-.0061*C8)+(-.0137*C9)+(-.0205*C10)+(-.0206*C11)+(-.0185*C12)+(-.0294*C13)+(-.0264*C14)+(.0114*C15)+(.0509*C16)+(.0512*C17)+(.0116*C18) +Uc+Up, |

Table 2 (cont'd). DGPs for Each Scenario

(4) D-Set

| | |
|---|---|
| D-1 | Logit[Pr(Y=1)] =-1.9876 + (.0589*Age)+(-.001*Age$^2$)+ (-0.2567*P1)+(-0.2403*P2)+(-0.2332*P3)+(-0.2176*P4)+(-0.1771*P5)+(-0.1535*P6)+(-.1396*P7)+(-.1183*P8)+(-.0904*P9)+(-.0808*P10)+(-.0539*P11)+(-.0438*P12)+(-.0232*P13)+(0.0073*P14)+(0.022*P15)+(.0342*P16)+(.0466*P17)+(.0796*P18)+(.1019*P19)+(.1244*P20)+(.1475*P21)+(.1676*P22)+(.1767*P23)+(.1937*P24)+(.2234*P25)+(.2432*P26)+(.26*P27) +( -0.6734*C1)+(-0.6051*C2)+(-0.5293*C3)+(-0.4544*C4)+(-0.3521*C5)+(-0.2886*C6)+(-0.1710*C7)+(-0.1033*C8)+(-0.0380*C9)+(0.0359*C10)+(0.1040*C11)+(0.2115*C12)+(0.2909*C13)+(0.3635*C14)+(0.4271*C15)+(0.5040*C16)+(0.6070*C17)+(0.6714*C18) |
| D-2 | Logit[Pr(Y=1)] =-1.9876 + (.0589*Age)+(-.001*Age$^2$) + (-0.2567*P1)+(-0.2403*P2)+(-0.2332*P3)+(-0.2176*P4)+(-0.1771*P5)+(-0.1535*P6)+(-.1396*P7)+(-.1183*P8)+(-.0904*P9)+(-.0808*P10)+(-.0539*P11)+(-.0438*P12)+(-.0232*P13)+(0.0073*P14)+(0.022*P15)+(.0342*P16)+(.0466*P17)+(.0796*P18)+(.1019*P19)+(.1244*P20)+(.1475*P21)+(.1676*P22)+(.1767*P23)+(.1937*P24)+(.2234*P25)+(.2432*P26)+(.26*P27)+(-0.2820*C1)+(-0.2750*C2)+(-0.2225*C3)+(-0.2079*C4)+(-0.1633 *C5)+(-0.1339*C6)+(-0.0771*C7)+(-0.0638*C8)+(-0.0155*C9)+(0.0128*C10)+(0.0471*C11)+(0.0960*C12)+(0.1228*C13)+(0.1771*C14)+(0.1884*C15)+(0.2350*C16)+(0.2715*C17)+(0.2903*C18) |
| D-3 | Logit[Pr(Y=1)] =-1.9876 + (.0589*Age)+(-.001*Age$^2$) + (-0.2567*P1)+(-0.2403*P2)+(-0.2332*P3)+(-0.2176*P4)+(-0.1771*P5)+(-0.1535*P6)+(-.1396*P7)+(-.1183*P8)+(-.0904*P9)+(-.0808*P10)+(-.0539*P11)+(-.0438*P12)+(-.0232*P13)+(0.0073*P14)+(0.022*P15)+(.0342*P16)+(.0466*P17)+(.0796*P18)+(.1019*P19)+(.1244*P20)+(.1475*P21)+(.1676*P22)+(.1767*P23)+(.1937*P24)+(.2234*P25)+(.2432*P26)+(.26*P27)+(-0.0803*C1)+(-0.0988*C2)+(-0.0707*C3)+(-0.0716*C4)+(-0.0614*C5)+(-0.0420 *C6)+( -0.0390*C7)+(-0.0270*C8)+(0.0005*C9)+(0.0229*C10)+(0.0170*C11)+(0.0530*C12)+(0.0356*C13)+(0.0468*C14)+(0.0525*C15)+(0.0763*C16)+(0.0771*C17)+(0.1089*C18) |
| D-4 | Logit[Pr(Y=1)] =-1.9876 + (.0589*Age)+(-.001*Age$^2$) + (-0.2567*P1)+(-0.2403*P2)+(-0.2332*P3)+(-0.2176*P4)+(-0.1771*P5)+(-0.1535*P6)+(-.1396*P7)+(-.1183*P8)+(-.0904*P9)+(-.0808*P10)+(-.0539*P11)+(-.0438*P12)+(-.0232*P13)+(0.0073*P14)+(0.022*P15)+(.0342*P16)+(.0466*P17)+(.0796*P18)+(.1019*P19)+(.1244*P20)+(.1475*P21)+(.1676*P22)+(.1767*P23)+(.1937*P24)+(.2234*P25)+(.2432*P26)+(.26*P27)+(0.0803*C1)+(0.0988*C2)+(0.0707*C3)+(0.0716*C4)+(0.0614*C5)+(0.0420 *C6)+(0.0390*C7)+(0.0270*C8)+(-0.0005*C9)+(-0.0229*C10)+(-0.0170*C11)+(-0.0530*C12)+(-0.0356*C13)+(-0.0468*C14)+(-0.0525*C15)+(-0.0763*C16)+(-0.0771*C17)+(-0.1089*C18) |
| D-5 | Logit[Pr(Y=1)] =-1.9876 + (.0589*Age)+(-.001*Age$^2$) + (-0.2567*P1)+(-0.2403*P2)+(-0.2332*P3)+(-0.2176*P4)+(-0.1771*P5)+(-0.1535*P6)+(-.1396*P7)+(-.1183*P8)+(-.0904*P9)+(-.0808*P10)+(-.0539*P11)+(-.0438*P12)+(-.0232*P13)+(0.0073*P14)+(0.022*P15)+(.0342*P16)+(.0466*P17)+(.0796*P18)+(.1019*P19)+(.1244*P20)+(.1475*P21)+(.1676*P22)+(.1767*P23)+(.1937*P24)+(.2234*P25)+(.2432*P26)+(.26*P27)+(0.2820*C1)+(0.2750*C2)+(0.2225*C3)+(0.2079*C4)+(0.1633 *C5)+(0.1339*C6)+(0.0771*C7)+(0.0638*C8)+(0.0155*C9)+(-0.0128*C10)+(-0.0471*C11)+(-0.0960*C12)+(-0.1228*C13)+(-0.1771*C14)+(-0.1884*C15)+(-0.2350*C16)+(-0.2715*C17)+(-0.2903*C18) |

Table 2 (cont'd). DGPs for Each Scenario

| | |
|---|---|
| D-6 | Logit[Pr(Y=1)] =-1.9876 + (.0589*Age)+(-.001*Age$^2$)+ (-0.2567*P1)+(-0.2403*P2)+(-0.2332*P3)+(-0.2176*P4)+(-0.1771*P5)+(-0.1535*P6)+(-.1396*P7)+(-.1183*P8)+(-.0904*P9)+(-.0808*P10)+(-.0539*P11)+(-.0438*P12)+(-.0232*P13)+(0.0073*P14)+(0.022*P15)+(.0342*P16)+(.0466*P17)+(.0796*P18)+(.1019*P19)+(.1244*P20)+(.1475*P21)+(.1676*P22)+(.1767*P23)+(.1937*P24)+(.2234*P25)+(.2432*P26)+(.26*P27)<br>+(0.6734*C1)+(0.6051*C2)+(0.5293*C3)+(0.4544*C4)+(0.3521*C5)+(0.2886*C6)+(0.1710*C7)+(0.1033*C8)+(0.0380*C9)+(-0.0359*C10)+(-0.1040*C11)+(-0.2115*C12)+(-0.2909*C13)+(-0.3635*C14)+(-0.4271*C15)+(-0.5040*C16)+(-0.6070*C17)+(-0.6714*C18) |

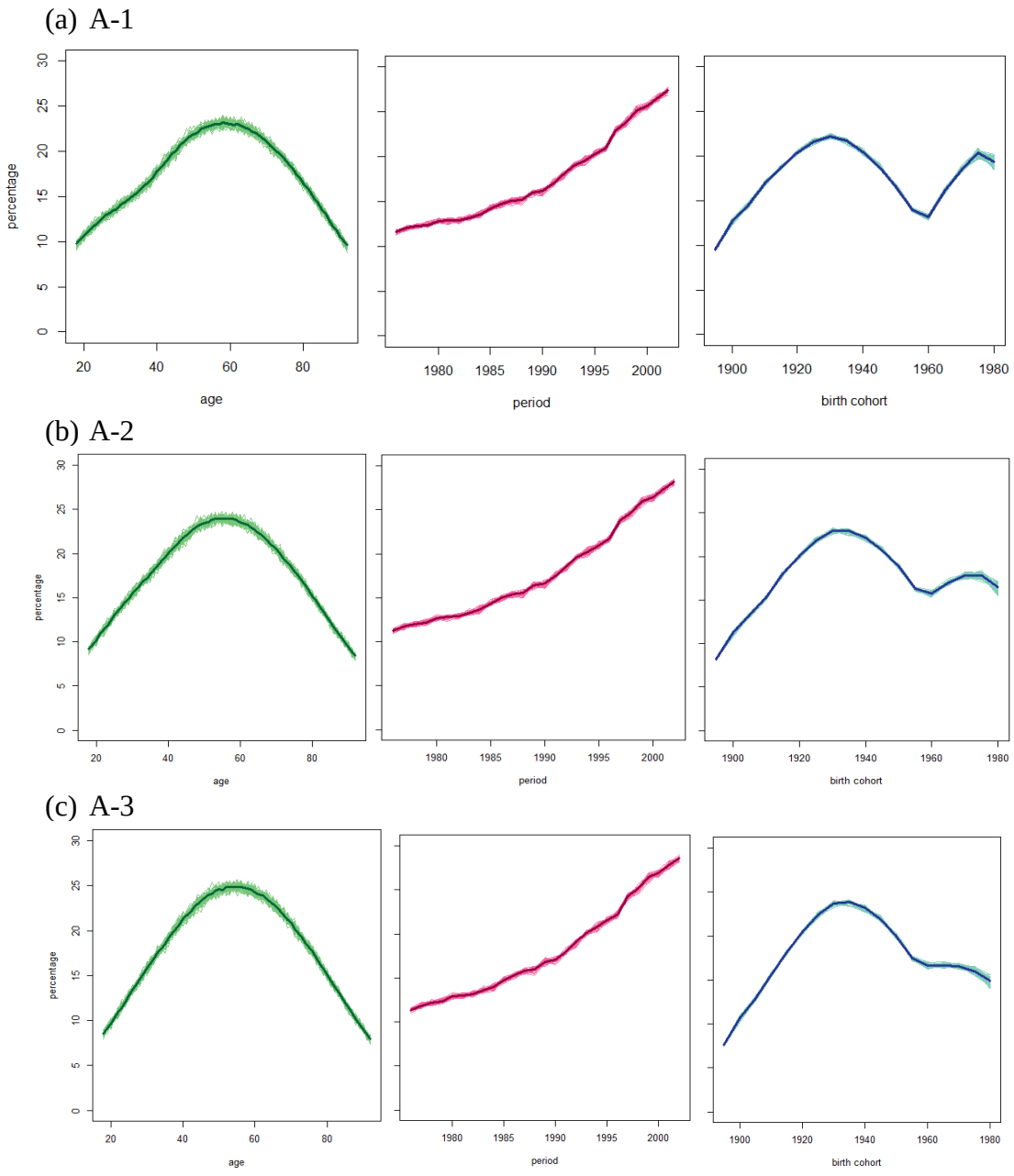



Figure 2. Visualized Descriptive Statistics of A-Set
*Each solid line represents the descriptive statistics from each simulated dataset, and the bold line represents the median of the 50 simulated dataset.

Table 3. Model Fit Statistics for Nested APC Models Fitted to 50 Simulated Data in A-Set

| | | A | P | C | A+C | A+P | P+C | APC |
|---|---|---|---|---|---|---|---|---|
| A-1 | AIC | 0 | 0 | 0 | 0 | 0 | 0 | 50 |
| | BIC | 0 | 0 | 0 | 0 | 0 | 0 | 50 |
| A-2 | AIC | 0 | 0 | 0 | 0 | 0 | 0 | 50 |
| | BIC | 0 | 0 | 0 | 0 | 50 | 0 | 0 |
| A-3 | AIC | 0 | 0 | 0 | 0 | 5 | 0 | 45 |
| | BIC | 0 | 0 | 0 | 0 | 50 | 0 | 0 |
| | DF | 3 | 27 | 18 | 29 | 20 | 44 | 46 |
| | N | 1million | 1million | 1million | 1million | 1million | 1million | 1million |

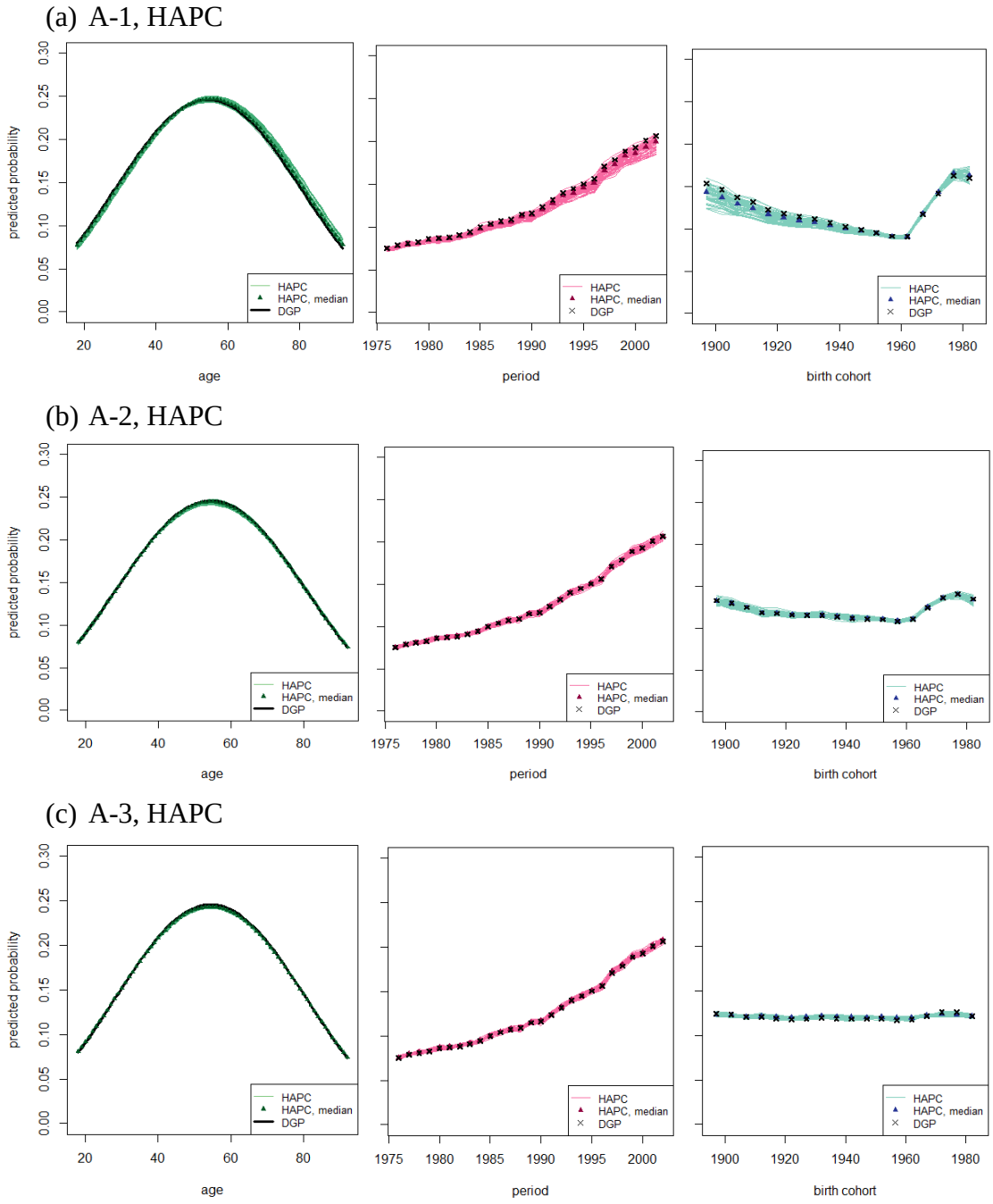



Figure 3. HAPC-CCREM and a Reduced Model Fitted to 50 Simulated Data in A-Set
*Each solid line in the figure represents the estimate of the HAPC for each simulated dataset.

(d) A-3, AP

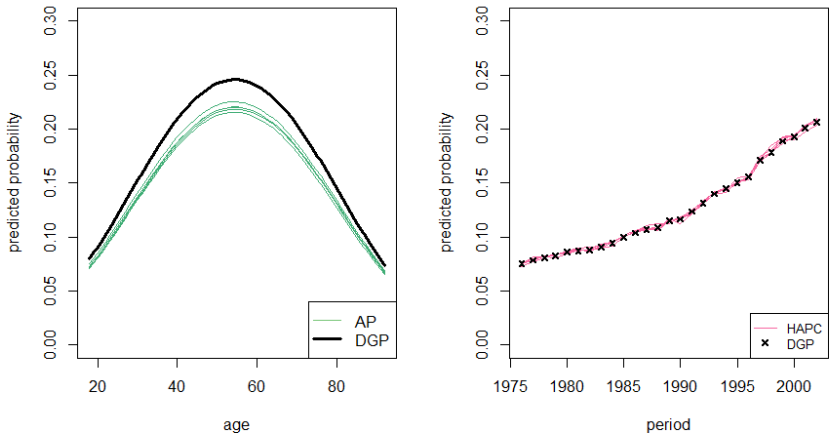



Figure 3 (cont'd). HAPC-CCREM and a Reduced Model Fitted to 50 Simulated Data in A-Set

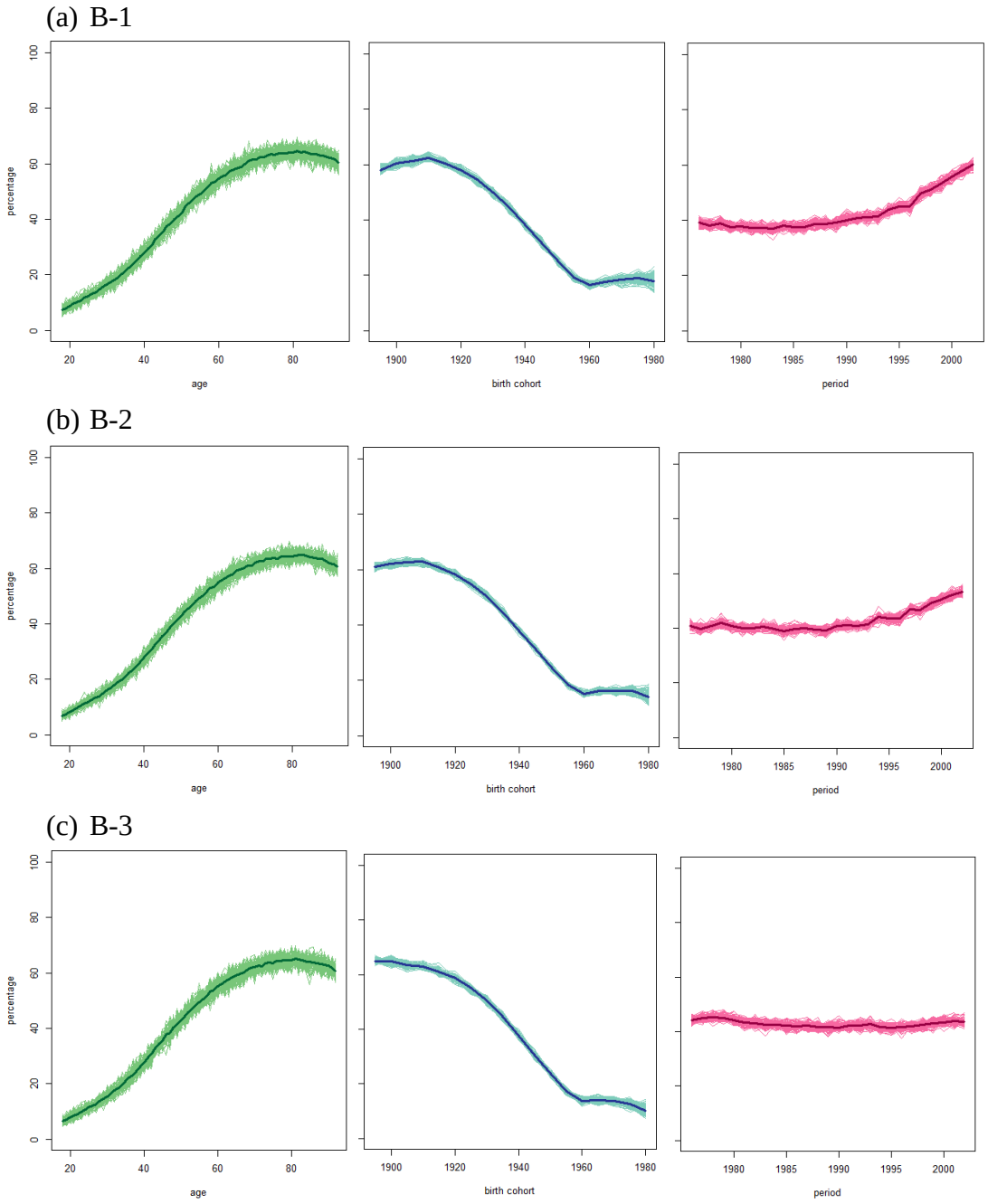
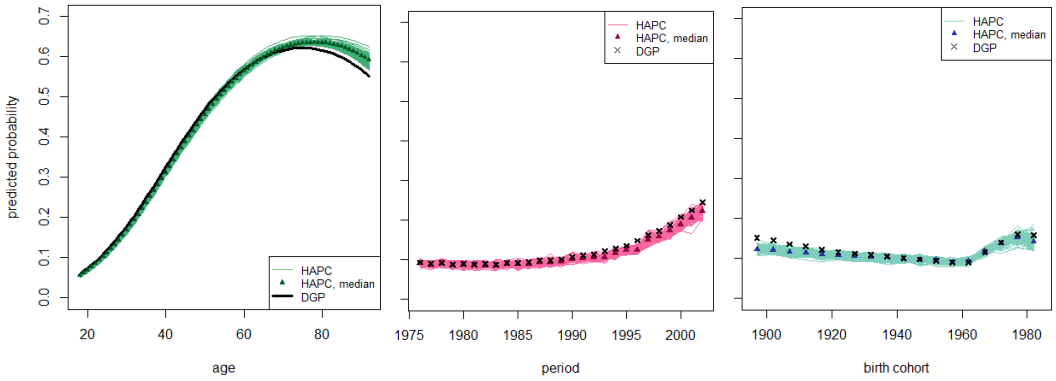



Figure 4. Visualized Descriptive Statistics of B-Set
*Each solid line represents the descriptive statistics from each simulated dataset, and the bold line represents the median of the 50 simulated dataset.

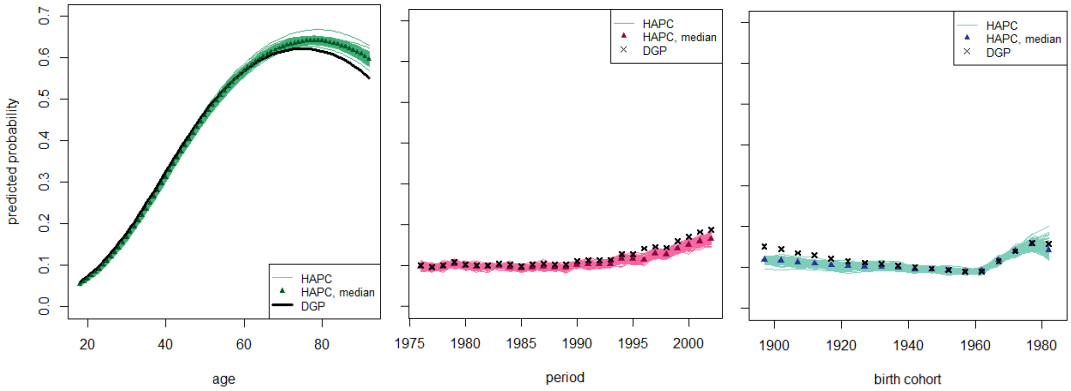
Table 4. Model Fit Statistics For Nested APC Models Fitted to 100 Simulated Data in B-Set

| | | A | P | C | A+C | A+P | P+C | APC |
|---|---|---|---|---|---|---|---|---|
| B-1 | AIC | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| | BIC | 0 | 0 | 0 | 8 | 92 | 0 | 0 |
| B-2 | AIC | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| | BIC | 0 | 0 | 0 | 99 | 1 | 0 | 0 |
| B-3 | AIC | 0 | 0 | 0 | 78 | 0 | 0 | 22 |
| | BIC | 99 | 0 | 0 | 1 | 0 | 0 | 0 |
| | DF | 3 | 27 | 18 | 29 | 20 | 44 | 46 |
| | N | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 |

(a) B-1, HAPC




(b) B-2, HAPC




(c) B-3, AC

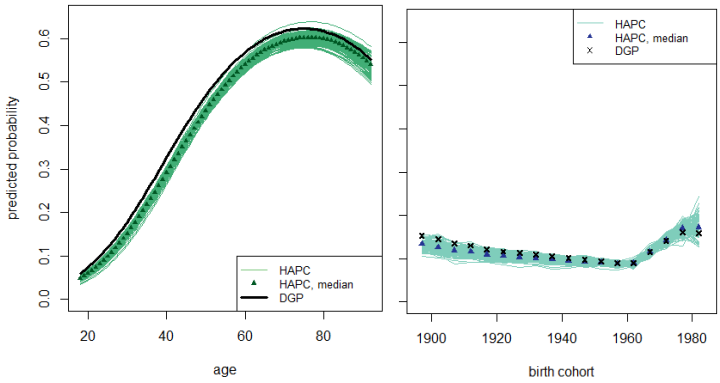



Figure 5. HAPC-CCREM and a Reduced Model Fitted to 100 Simulated Data in B-Set

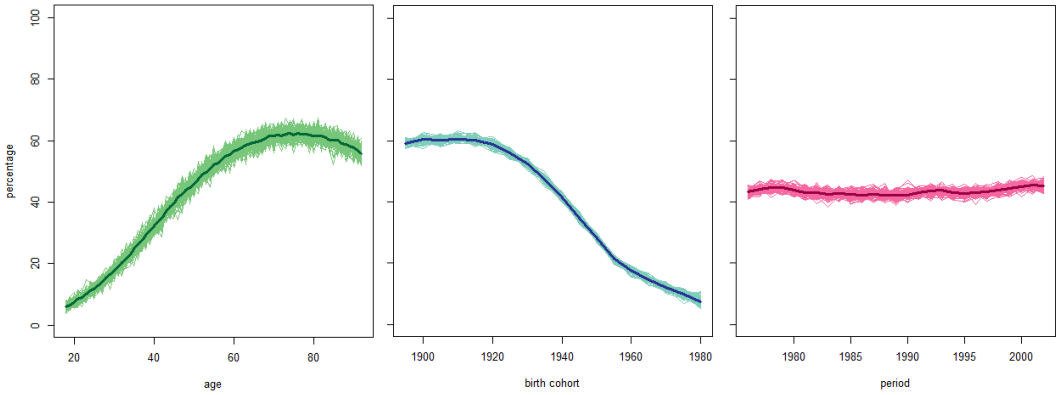
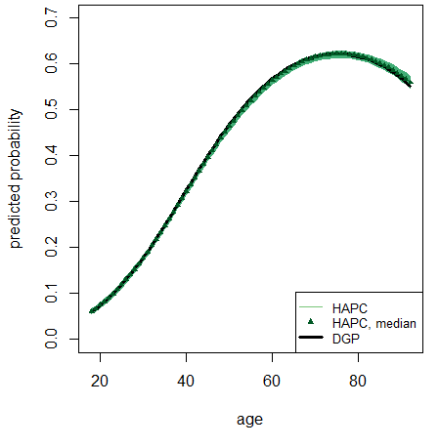



Figure 6. Visualized Descriptive Statistics of C-Set

Table 5. Model Fit Statistics for Nested APC Models Fitted to 100 Simulated Data in C-Set

| | | A | P | C | A+C | A+P | P+C | APC |
|---|---|---|---|---|---|---|---|---|
| C-1 | AIC | 65 | 0 | 0 | 8 | 27 | 0 | 0 |
| | BIC | 100 | 0 | 0 | 0 | 0 | 0 | 0 |
| | DF | 3 | 27 | 18 | 29 | 20 | 44 | 46 |
| | N | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 |

(a) C-1, A




(b) C-1, AP

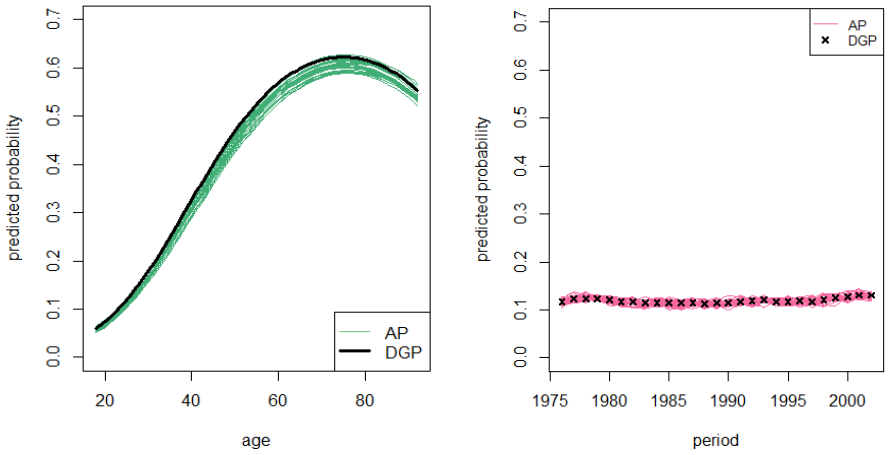



Figure 7. A Reduced Model (A and AP) Fitted to 100 Simulated Data in C-Set

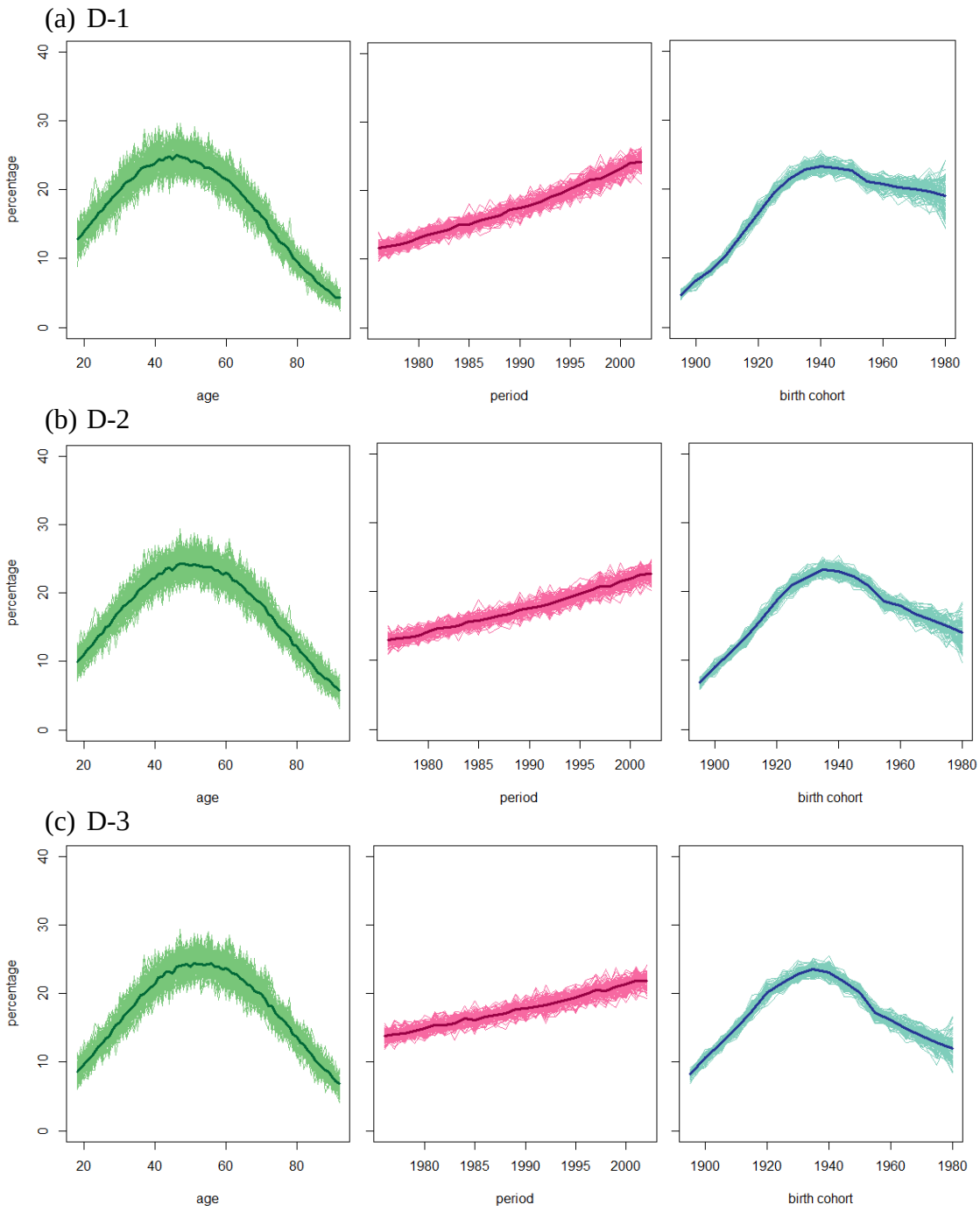



Figure 8. Visualized Descriptive Statistics of D-Set

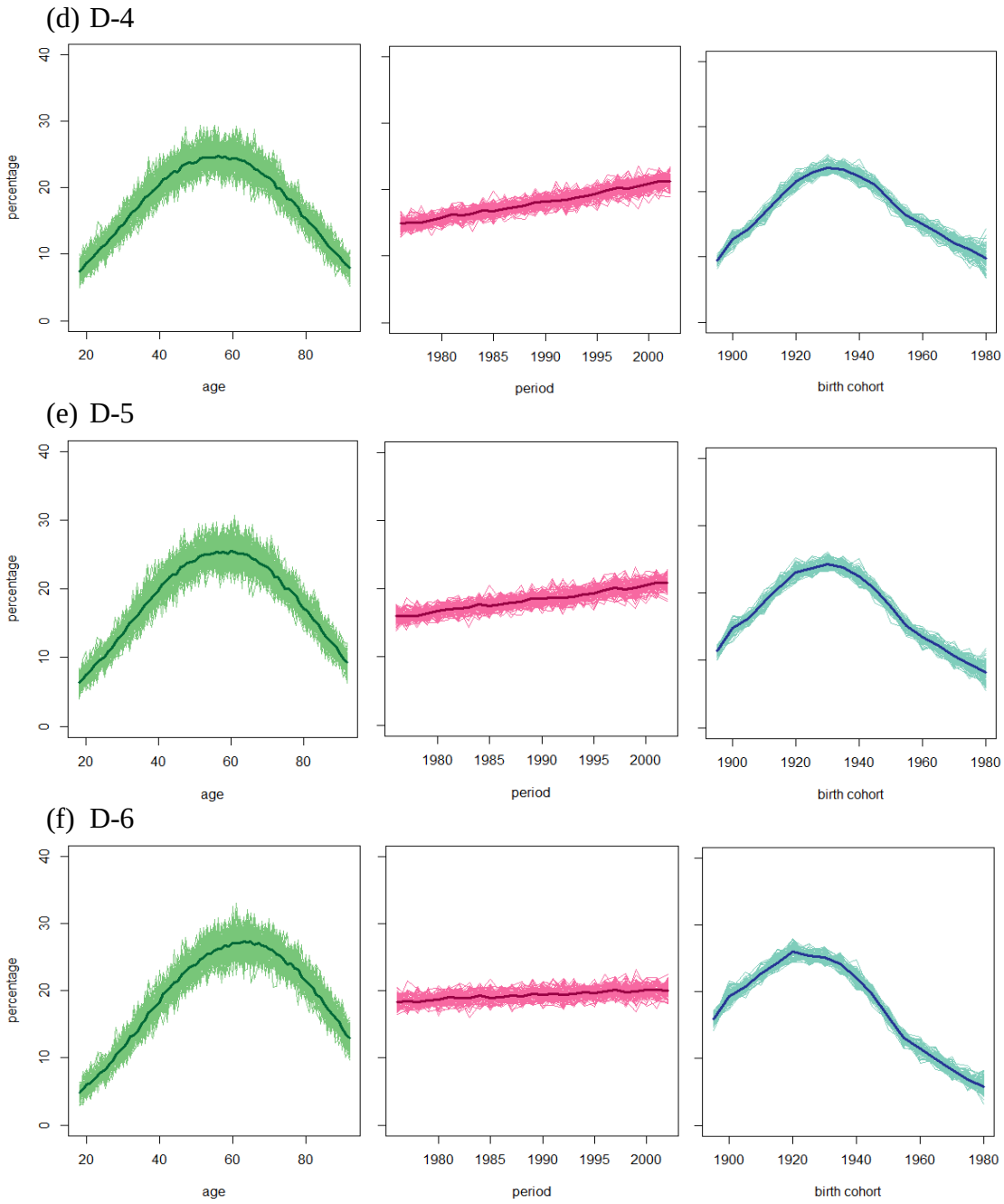
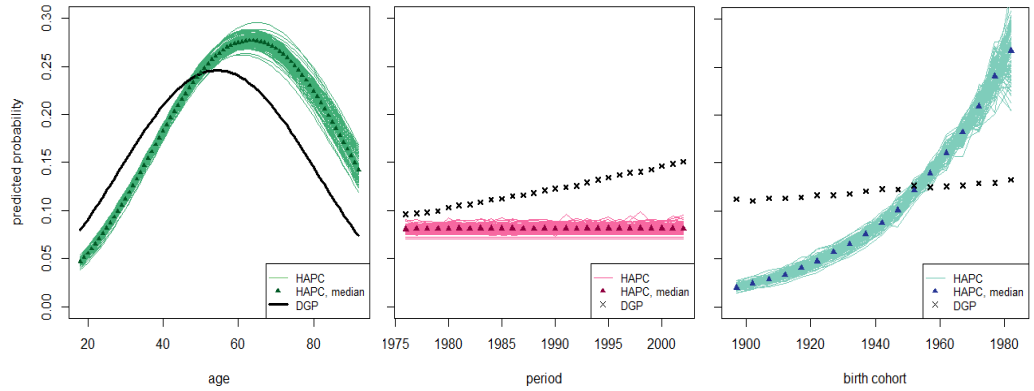



Figure 8 (cont'd). Visualized Descriptive Statistics of D-Set

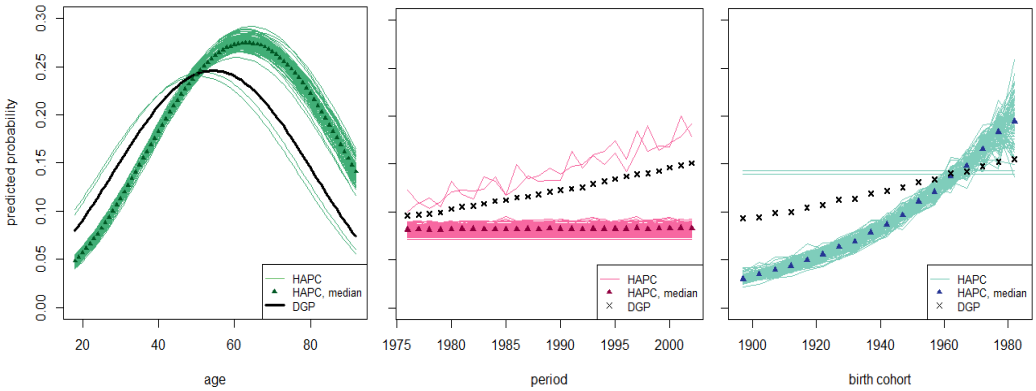
Table 6. Model fit Statistics for Nested APC Models Fitted to 100 Simulated Data in D-Set

| | | A | P | C | A+C | A+P | P+C | APC |
|---|---|---|---|---|---|---|---|---|
| D-1 | AIC | 0 | 0 | 0 | 73 | 27 | 0 | 0 |
| | BIC | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| D-2 | AIC | 0 | 0 | 0 | 68 | 32 | 0 | 0 |
| | BIC | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| D-3 | AIC | 0 | 0 | 0 | 74 | 26 | 0 | 0 |
| | BIC | 8 | 0 | 0 | 92 | 0 | 0 | 0 |
| D-4 | AIC | 0 | 0 | 0 | 71 | 29 | 0 | 0 |
| | BIC | 96 | 0 | 0 | 4 | 0 | 0 | 0 |
| D-5 | AIC | 0 | 0 | 0 | 74 | 26 | 0 | 0 |
| | BIC | 100 | 0 | 0 | 0 | 0 | 0 | 0 |
| D-6 | AIC | 84 | 0 | 0 | 15 | 0 | 0 | 1 |
| | BIC | 100 | 0 | 0 | 0 | 0 | 0 | 0 |
| | DF | 3 | 27 | 18 | 29 | 20 | 44 | 46 |
| | N | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 |

(a) D-1, HAPC




(b) D-2, HAPC




(c) D-3, HAPC

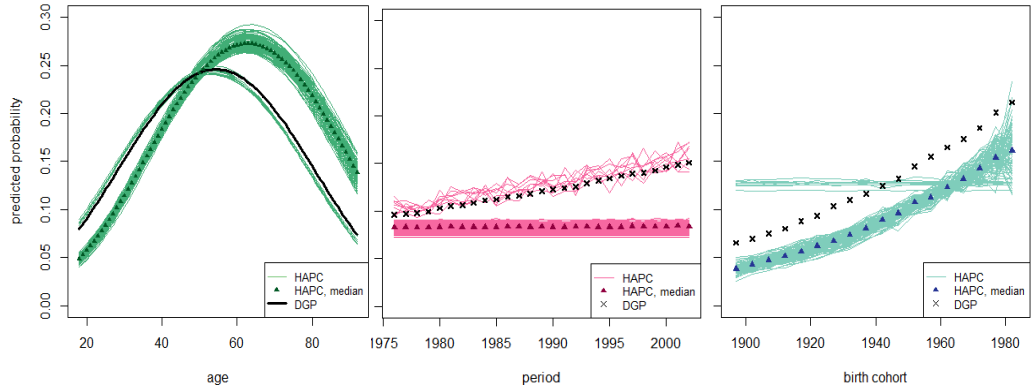
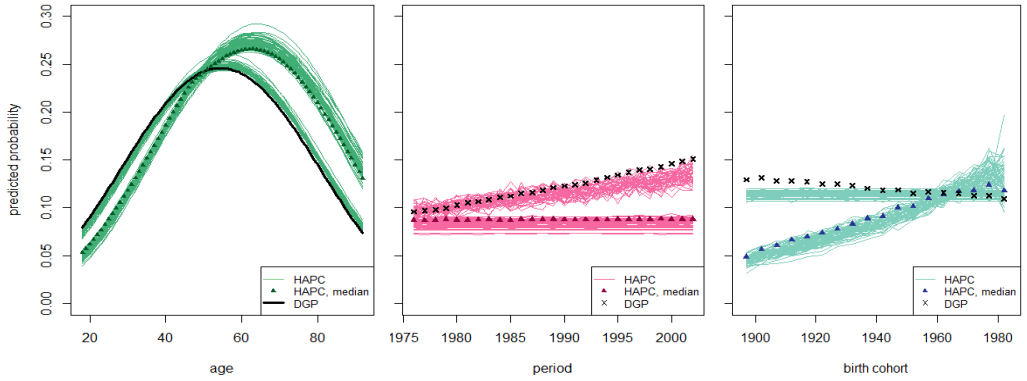
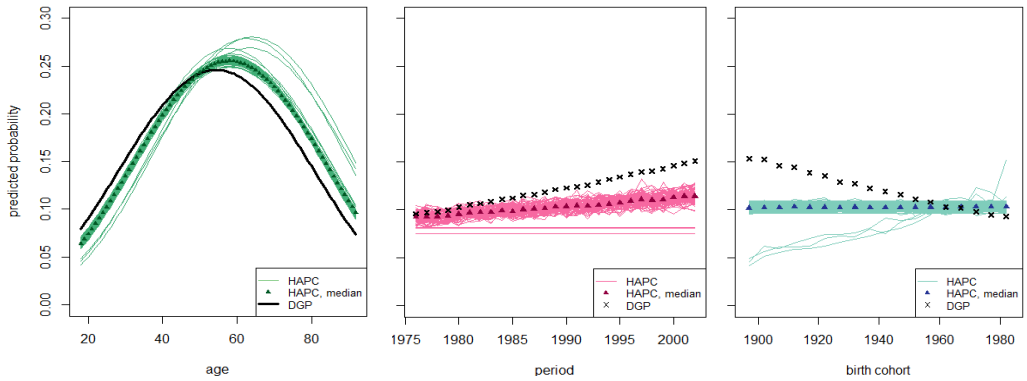



Figure 9. HAPC-CCREM Fitted to 100 Simulated Data in D-Set

(d) D-4 HAPC




(e) D-5 HAPC




(f) D-6 HAPC

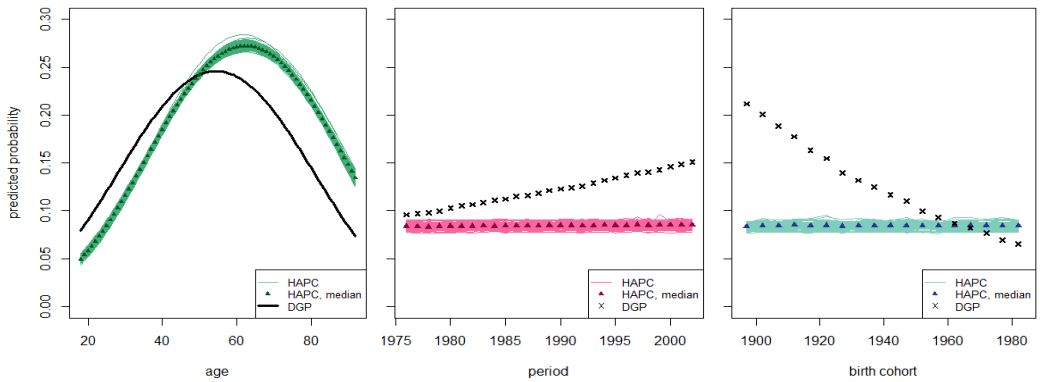



Figure 9 (cont'd). HAPC-CCREM Fitted to 100 Simulated Data in D-Set

# CHAPTER III

# METHODS TO EVAULATE DATA STRUCTURES PRIOR TO THE APPLICATION OF INTRINSIC ESTIMTAOR AGE-PERIOD-COHORT MODELS

## Introduction

### *The ID Problem Explained in Matrix Form*

Conventional age-period-cohort data are presented in a two-way table. The rows of this table represent age groups and the columns represents periods of observation. Each cell of the table contains age-specific rates for a particular time period. In this layout, the diagonals refer to birth cohorts, as they indicate the rates for those who were born in the same years and, therefore, age together. When such tabular-rate data has $a$ different age groups and $p$ different periods of observation, $a+p-1$ birth cohorts are defined.

Compared to data from repeated cross-sectional surveys, as described in Chapter 2 on HAPC analysis, tabular-rate data are much more restricted in structure. That is because the temporal widths of time periods and birth cohorts are always enforced to be equal to the temporal width of age, and only a single observation (e.g., a mortality rate) exists in each cell. Consequently, the approach employed in HAPC analysis, which takes advantage of a flexible data structure to help address the ID problem, does not work for conventional tabular-rate APC data.

For tabular-rate APC data, the APC accounting model is expressed in a linear regression form (Mason et al. 1973):

$$\frac{D_{ij}}{P_{ij}} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ij} \quad (1)^1$$

where $D_{ij}$ denotes the number of events (i.e., suicide deaths in all of the following exercises) observed for the *i-th* age group at the *j-th* observation period for $i=1,\ldots,a$ and $j=1,\ldots,p$; $P_{ij}$ denotes the size of the estimated population in the *ij-th* group exposed to risk of death, and therefore the left side of the equation refers to the rate of death observed for the *ij-th* group; $\mu$ denotes the intercept; $\alpha_i$ is the age effect in the *i-th* age group; $\beta_j$ is the period effect in the *j-th* observation period; $\gamma_k$ is the cohort effect on the *k-th* diagonal where $k=a-i+j$; $\varepsilon_{ij}$ is a random error term which is assumed to have the expected value of 0 (Kupper et al. 1985, Yang, Fu and Land 2004).

When the count of events follows the Poisson distribution, it is conventional to convert Eq (1) into a log-linear form by taking a log-link function. Such a model can be written as follows (Yang and Land 2013a):

$$\log(E_{ij}) = \log(P_{ij}) + \mu + \alpha_i + \beta_j + \gamma_k \quad (2)$$

where $\log(E_{ij})$ denotes the logged number of events (e.g., deaths) expected in the *i-th* age group at the *j-th* observation period for $i=1,\ldots,a$ and $j=1,\ldots,p$; $\log(P_{ij})$ denotes the logged term of the size of the population exposed to the risk of death in the *ij-th* group; it also serves as adjustment for the log-linear model. In this model, the coefficients on the right side are treated as fixed-effects, which means that the age effect ($\alpha_i$) is constant for the *i-th* row, the period effect ($\beta_j$) is constant for the j-th column, and the cohort effect ($\gamma_k$) is constant for the *k-th* diagonal in the tabular-rate data layout. Therefore, after

[1] The left-hand side of this equation is technically known as occurrence/exposure rates (For more details, see Land, Kenneth C., Yang Yang and Yi Zeng. 2005. "Mathematical Demography." Pp. 659–717 in *Handbook of Population*: Springer.). Note that occurrence/exposure rates are universal and tends to be so taken for granted in demography that it is called "demographic rates".

reparamaterizing the variables to deal with over-parameterization as follows, the model can be seen as a standard two-way ANOVA model (Yang, Fu and Land 2004):

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{p} \beta_j = \sum_{k=1}^{a+p-1} \gamma_k = 0 \quad (3)$$

This centering of the reparametrization does not yield any distortion in the patterns of estimated coefficients (Kupper et al. 1983, Kupper et al. 1985), but it expresses the coefficients in term of a deviation from the mean. After this reparametrization, equation (1) can be re-expressed in the matrix form:

$$Y = Xb + \varepsilon \quad (4)$$

where **Y** is the vector with a length of *ap,* whose components are age-period specific suicide mortality rates; **X** is the regression design matrix with a dimension of *ap* by *2(a+p)-3*; **b** is a parameter vector with a length of *2(a+p)-3* that can be written as {1; $\alpha_1$,.. $\alpha_{a-1}$; $\beta_1$,…, $\beta_{p-1}$; $\gamma_1$,…, $\gamma_{a+p-2}$}; and $\boldsymbol{\varepsilon}$ denotes a vector of random errors with a distribution centered on zero and a length of *ap*. Note that the equation (3) works as a restriction so that the last terms of age, period, and cohort parameters are excluded from the analysis as reference categories and estimated not by the regression model but by:

$$\alpha_a = -\sum_{i=1}^{a-1} \alpha_i \,;\; \beta_p = -\sum_{j=1}^{p-1} \beta_j \,;\; \gamma_{a+p-1} = -\sum_{k=1}^{a+p-2} \gamma_k \quad (5)$$

As described in the previous chapter, solving Eq (4) by using the ordinary least square (OLS), or maximum likelihood (MLE), method incurs the ID problem due to the exact linear relationship (age=period-cohort). The exact algebraic equality yields a singular design matrix with less than one full column rank, and a unique parameter vector $\widehat{\boldsymbol{b}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y}$ cannot be identified as $(\boldsymbol{X^T X})^{-1}$ does not exist. As it is, there exist infinite different sets of $\widehat{\boldsymbol{b}}$.

To understand more details on how the singular design matrix incurs the problem during the parameter estimation, it is worth looking more into the right side of Eq (4). When the design matrix X is less than full column rank, the design matrix has the following property:

$$Xb_0 = 0 \tag{6}$$

Eq (6) indicates that there exists a non-zero vector $b_0$ such that the product of the design matrix and the vector is 0. Here, $b_0$ is the null eigenvector—corresponding to an eigenvalue of zero—that represents the null space of the design matrix X. The elements of $b_0$ are:

$$b_0 = (0; \alpha_1,.., \alpha_{a-1}; \beta_1,.., \beta_{p-1}; \gamma_1,\ldots, \gamma_{a+p-2})^T \tag{7}$$

where

$$\alpha_i = i - \frac{a+1}{2},\ \beta_j = -j + \frac{p+1}{2},\ \gamma_k = k - \frac{a+p}{2} \tag{8}$$

(Kupper et al. 1985).

Then, how is the $\boldsymbol{b_0}$ derived? $\boldsymbol{b_0}$ is obtained when $\boldsymbol{b}$ in Eq (4) is decomposed into two components, which represent non-null and null spaces of the design matrix $\boldsymbol{X}$ as follows:

$$b = b_1 + s\, b_0 \tag{9}$$

where $b_1$ is the projection of $\boldsymbol{b}$ on the non-null space of X, $\boldsymbol{b_0}$ is the projection of $\boldsymbol{b}$ on the null space of X, and s is an arbitrary scalar. In order to understand the ID problem in matrix form, the fact that s can be any scalar is key. According to Eq (9), the age, period, and cohort effects, which are the elements of $\boldsymbol{b}$, change as $\boldsymbol{s}$ has a different value. However, a different value of $\boldsymbol{s}$ does not cause any change to $\boldsymbol{Y}$, as shown in the following equation.

$$Y = Xb + e = X(b_1+sb_0) = Xb_1 + s\,Xb_0 \quad (10)$$

The term $\boldsymbol{sXb_0}$ is zero because $\boldsymbol{Xb_0}$ is zero from Eq (6). That means $\boldsymbol{Y}$ is decided solely by $\boldsymbol{Xb_1}$ and not affected by $\boldsymbol{sXb_0}$. In other words, for the same $\boldsymbol{Y}$, we can have an infinite number of different values for $\boldsymbol{s}$, and consequently an infinite number of different sets of age, period, and cohort effects (i.e., $\boldsymbol{b}$). Note that $\boldsymbol{Y}$ represents the data—that is, the information given to researchers. When a statistical model is applied to $\boldsymbol{Y}$, it is impossible to identify which $\boldsymbol{b}$ out of those infinitely different $\boldsymbol{b}$s is the true one. This is the ID problem in the APC analysis shown in matrix form.

The conventional approach to address the ID problem is imposing an additional constraint (Mason et al. 1973). This approach, called Constrained Generalized Linear Models (CGLIM), assumes that adjacent age, period, or cohort categories have equal effects on the outcome. This assumption makes the design matrix $\boldsymbol{X}$ non-singular and Eq (6) no longer valid. Then, a unique $\boldsymbol{b}$ can be identified from Eq (4). Unfortunately, in spite of deriving an estimable model, one critical limitation of this method is that the approach needs a *strong priori* assumption to select a theoretically valid constraint. When such an assumption is not available, the approach should not be applied as the estimates of the model could be very sensitive to the selection of the constraint. In a real-world research setting, there is no way to confirm if the obtained coefficients from the imposed constraints correspond to the true effects of age, period, and cohort. Therefore, a researcher runs the risk of believing that the biased estimates derived from a "wrong" constraint are true.

*What is The Intrinsic Estimator (IE)?*

Due to this limitation of the conventional approach, researchers have sought another statistical technique to address the ID problem that is not dependent upon the availability of external information. As one of those efforts, the *intrinsic estimator (IE)* model was developed by Fu (2000) and Yang, Fu and Land (2004). Starting from Eq (9), the model focuses on $\mathbf{b_1}$ as a projection of unconstrained $\boldsymbol{b}$ on the non-null subspace of $\boldsymbol{X}$**.** The essential feature of the IE is to regress the outcome of interest only in the non-null subspace by removing the null space—in other words, removing the eigenvector $\boldsymbol{b_0}$ by setting $\boldsymbol{s}$ equal to zero. Note that $\boldsymbol{s}$ can be an arbitrary scalar in Eq (9), but $\boldsymbol{b_0}$ is not. Rather, $\boldsymbol{b_0}$ is fixed by the design matrix $\boldsymbol{X}$, and both $\boldsymbol{X}$ and $\boldsymbol{b_0}$ are independent from $\boldsymbol{Y}$ but determined by the number of age and period categories (i.e., *a* and *p)*. Setting $\boldsymbol{s}$ equal to zero is also algebraically equivalent to:

$$b^T b_0 = 0 \qquad (11)$$

That means that the projection of unconstrained b on the null subspace of X is zero.

The implementation of the IE borrows from the principal component analysis (PCA) technique. Where eigenvectors and eigenvalues of $\mathbf{(X^TX)}$ are $\boldsymbol{V}=[v_1, v_2,..,v_{2(a+p)-4}]$ [2] and $\boldsymbol{\Lambda}=\{\lambda_1, \lambda_2,.., \lambda_{2(a+p)-4}\}$ respectively, $\boldsymbol{X}\cdot\boldsymbol{V}$ yields a new design matrix in the PC space. Note that the transferring procedure itself does not resolve the ID problem of the APC model. When transferring $\boldsymbol{X}$ to the PC space, the IE uniquely addresses the identification problem by reducing the dimension of the analysis to one less dimension by excluding the null eigenvector ($b_0$) from $\boldsymbol{V}$. Since the PC formed by the eigenvector

[2] The dimension of original design matrix $\boldsymbol{X}$ is *ap* by *2(a+p)-**3***. Here for transferring to PC space, the column of the intercept term is excluded, so that the dimension of $\boldsymbol{X}$ becomes *ap* by *2(a+p)-**4***, from which $\boldsymbol{(X^TX)}$ has *2(a+p)-**4*** of eigenvectors and eigenvalues.

corresponding to the zero eigenvalue accounts for no variance[3], there is no loss of any explanatory power derived from this exclusion. Then the new design matrix $\boldsymbol{U}$ in the non-null subspace is:

$$\boldsymbol{U = X \cdot V'} \quad (12)$$

where $\boldsymbol{V'}$ has a dimension of $2(a+p)-4$ by $2(a+p)-\boldsymbol{5}$, free from the null eigenvector; $\boldsymbol{U}$, the new design matrix in the PC space, has a dimension of $ap$ by $2(a+p)-\boldsymbol{5}$, which does not incur the identification problem in the PC regression. Then, IE regresses the outcome of interest on $2(a+p)-\boldsymbol{5}$ PCs and returns $2(a+p)-\boldsymbol{5}$ coefficients. Since these coefficients are not interpretable by A, P, and C terms, the estimates are re-transferred from the PC space to the original space, and this re-transferred vector contains the coefficients for age, period, and cohort effects. Here the dimension of the vector is recovered from $2(a+p)-\boldsymbol{5}$ to $2(a+p)-\boldsymbol{4}$ by filling an element corresponding to the null eigenvector with zero.

According to the IE developers, this algorithm of the IE has some advantageous properties (Yang, Fu and Land 2004). First, IE is an unbiased estimator in its finite time-period, which does not depend on *a priori* assumptions, but rather only on the number of age ($a$) and period ($p$) categories. Therefore, even when relevant theory is not available or strong enough to guarantee a selection of the "true" constraint, the model is estimable. Second, the IE has a smaller variance than the CGLIM. Therefore, even in unusual situation where the "true" constraint can be selected based on strong a priori theory, the IE and CGLIM will have close estimates in such instances, but the IE is still preferable because it a more efficient estimator than CGLIM (Yang, Fu and Land 2004). In addition,

[3] The proportion of variance that any single principal component accounts for is $\frac{\lambda_i}{\sum \lambda_i}$. Consequently, the PC derived from an eigenvalue of zero accounts for no variance Dunteman, George H. 1989. "Principal Component Analysis. Quantitative Applications in the Social Sciences Series (Vol. 69)." Thousand Oaks, CA: Sage Publications..

although the entire algorithm of the IE is borrowed from the ideas of principal component analysis, the re-transferring from the principal component space to the original space at the last step eases the interpretation of the coefficients using usual age, period, and cohort terms. Due to these desirable properties, some scholars have endorsed the IE method as a non-arbitrary way to constrain the model and identify a unique set of APC estimates (Fu, Land and Yang 2011).

*Why Does the IE Sometimes Fail? – Critiques of the IE*

Despite these potential advantages, recently published critiques have claimed that the IE is no less arbitrary than other APC methods because it has the tendency to yield biased estimates of age, period, and cohort effects (Luo 2013a, Luo 2013b). According to these scholars, the IE can fail because it does not properly address the ID problem.

One critique argues that IE estimates can change substantially through different categorizations of age and period groups (Luo 2013a). Although the null eigenvector $\boldsymbol{b_0}$ is not affected by $\boldsymbol{Y}$, it is influenced by the numbers of age and period categories (i.e., $a$ and $p$). Then the "linear" constraint derived from Eq (11) changes accordingly. Consequently, the data structure that satisfies the new constraint also changes, ultimately altering the IE estimates of age, period, and cohort effects. Unfortunately, there is no way to verify that a particular categorization of age and period effects will accurately reflect the truth. Although this argument is not logically wrong, it is worth noting that dependence of APC estimates of the coefficients on the categorization of age and period groups is true for all approaches of estimation, not necessarily a unique issue of the IE estimates.

Another critique focuses on that the IE uses the last groups of age, period, and cohort as reference categories by default. When other reference categories are used, the estimated pattern of age, period, and cohort effects can rotate. Sometimes this rotation is significant enough to flip the entire pattern of the effects (Held and Riebler 2013). Also by default, the IE uses an ANOVA-type effect coding ("sum-to-zero" coding) scheme. When a dummy-coding scheme is used instead, it can significantly alter the estimates of age, period, and cohort effects (Luo et al. 2014). A rejoinder to this criticism can be found in Land et al. (2016).

The critique of Luo (2013) states that Eq (11) still imposes an arbitrary constraint on the data—and the estimates from an IE analysis will be biased when the data do not satisfy this constraint. She argues that a researcher can never be sure that the estimates are true age, period and cohort effects since in an empirical research setting, it is impossible to confirm that the given data meets this constraint. This critique that IE yields biased estimates is supported by the results from a simulation analysis. Simulated data in Luo (2013a) are normally distributed where the mean of the response can be expressed as $10 + k_a age_i + k_p period_j + k_c cohort_{ij}$, and the standard deviation of the error term is 0.1. Using different values of $k_a$, $k_p$, and $k_c$, three sets of data are simulated to test the validity of the IE estimates. Among the three scenarios tested in these simulations, only the case that satisfies "IE's linear constraint ($b \cdot b_0 = k_a - k_p + 6k_c = 0$, where $a=3$, $p=3$)" produces unbiased estimates of APC coefficients. The other two scenarios, which do not meet the IE's constraint, lead to APC estimates that are largely biased.

Note that the data structure and the IE's linear constraint in this critique are derived from a single assumption—that the age, period and cohort effects follow linear

trends. Using different $k_a$, $k_p$, and $k_c$ indicates using different slopes, but exact linear increases for all three terms. That is, the age coefficients are assumed to lie on a single linear line, and so do the period and cohort coefficients. The linear constraint of the IE is also derived upon this assumption so that the parameter vector b can be expressed as

$$\mathrm{b} = (\mu, i_a, i_a + k_a, i_p, i_p + k_p, i_c, i_c + k_c, i_c + 2k_c, i_c + 3k_c)^T,$$

where $i_a$, $i_p$, and $i_c$ are the intercepts for the age, period, and cohort effects.

In their response to this critique, Yang and Land (2013b) highlight the importance of understanding data structure in the application of the IE. This is a similar argument with the one that was brought up during the debate on the HAPC in Chapter 2. The reason why the IE produces biased estimates is not because of some inherent flaw, but rather because the IE is applied to a data structure that is not suitable for a fully three-dimensional analysis. Yang and Land (2013b) pointed out that the data structure simulated in Luo (2013a) is highly improbable—and one where a three-dimensional model should not be applied. That is because, with the assumption of the "exact linear trend" of age, period, and cohort effects, one temporal dimension has a perfect correlation with the others and can be expressed as a functional form of the other dimensions. However, this is not a valid assumption because APC accounting models, including the IE, assume that all three dimensions independently contribute to the outcome variable. Yang and Land (2013b) argue that:

> *If these model selection tests indicate that one or two of the three temporal dimensions of the APC model are sufficiently collinear with the other dimensions that they do not contribute significantly to the outcome variable, then the analyst should not specify the full three-way APC model but rather a reduced model with one or two of the temporal variables. (pg. 8)*

To avoid such misapplication, Yang and Land (2013b) recommend that future researchers follow the three-step procedure outlined in Chapter 2 to make sure that the given data is eligible for the IE.

Despite this clarification, critics remain skeptical about the usefulness of preliminary analyses (i.e., obtaining descriptive and model fit statistics) as a tool for understanding the structure of tabular-rate APC data. Luo (2013b) states:

> *It is unclear how and why a model fit statistics, a piece of internal information derived from the data set itself, can be used as an external criterion to assure appropriate applications of IE. Moreover, even for data sets that meet Yang and Land's criterion for "full-blown APC models," IE estimates are not reliable or valid. (pg. 1986)*

Which of these statements is true? In this debate, it is an important to test if the suggested preliminary analysis can be a useful tool for checking the data structure before estimating IE models. If preliminary steps can tell us whether the data structure is eligible for a three-dimensional IE model—and if the IE successfully estimates the true age, period, and cohort effects from such data—then the IE can be a very useful research tool. However, if preliminary steps do not provide enough evidence on the eligibility of the given dataset for the IE model, then researchers will risk obtaining biased estimates of the age, period, and cohort effects from the IE.

### *Research Goal of This Chapter*

No previous studies have focused on the validity of the preliminary analysis suggested by Yang and Land (2013b) as a tool for understanding the structure of conventional tabular-rate APC data. Chapter 3 of this dissertation first evaluates the performance of step 1 (visual plots of descriptive statistics) and step 2 (model fit

statistics) of the three-step procedure as tools to understand a data structure that is two-dimensional as a result of exact linear trends in two dimensions. According to Yang and Land's argument (2013), such a data structure is not suitable for the IE, and the first two steps should provide enough evidence for researchers to avoid the application of a three-dimensional model. Next, I simulate additional data sets with some deviations imposed on either or both of the linear trends, which make the data structure more plausible. I observe how step 1 and 2 react to those changes.

In the real world, data structures can also be reduced by having one or more temporal dimensions that contribute little to the outcome variable. In additional simulation analyses, I test if the descriptive and model-fit statistics can properly detect the structure of data when a temporal dimension makes essentially no contribution to the outcome variable. From there, I gradually alter the structure of the data to become increasingly three-dimensional by imposing more obvious and distinct trends for each temporal dimension, and see if the preliminary steps can detect these incremental changes in the data structure.

Finally, to test Luo (2013b)'s argument that "*even for data sets that meet Yang and Land's criterion for "full-blown APC models, IE estimates are not reliable or valid* (p.1986)*",* I introduce scenarios where the simulated data have very distinct three-dimensional structures. To verify if the three-step procedure is valid for IE models, I first examine if the preliminary steps accurately detect the three-dimensional data structures, and then whether application of IE models produces unbiased estimates of the "true" age, period and cohort effects.

## Methods

### *Simulation Design*

In this chapter, thirteen different scenarios were tested in four sets. The outcome variable in all instances is the rate of suicide mortality; temporal dimensions are borrowed from Jeon, Reither and Masters (2016) for certain scenarios to enhance the plausibility of the data generating process (DGP). Each of the four sets was designed to reflect a unique situation that can occur with respect to each temporal dimensions in the data. The descriptions of the scenarios in each set are presented in Table 7.

The first set (hereafter referred to as "A-set") tests Yang and Land's argument (2013) that preliminary analyses are capable of detecting a reduced data structure when two of the three dimensions have exact linear trends. As the authors noted, one temporal dimension can be written as a function of the other dimension due to the high correlation in such a scenario. I will test if the descriptive and model fit statistics detect such properties of the data structure, and if so, I will give evidence that the three-dimensional model should not be applied. Across all scenarios in the A-set, I hold cohort effects constant but introduce variations to the age and period trends. To create a plausible data structure rather than arbitrarily assigning an unrealistic one, I borrow cohort trends in suicide mortality rates in South Korea, previously reported by Jeon, Reither and Masters (2016). Next I simulate data in scenario A-1 with age and period trends that are both exactly linear. Thus, those two dimensions are perfectly correlated and one can be written in a functional form of the other. Scenario A-2 retains these features but introduces small deviations in the linear trends for age and period effects. In this setting, those two dimensions still have a perfect correlation, but the trends are no longer *exactly* linear due

to the given deviations. A-3 introduces additional deviation in the period trend but maintains an exactly linear functional form for the age trend. A-4 is the opposite of A-3; it introduces deviation in the age trend but constrains the period trend to be exactly linear. In A-3 and A-4 the correlation is not perfect but still as high as r=0.95.

The B-set tests how alteration of the age effects impacts the temporal dimensionality of the data structure. In these simulations, I hold period and cohort trends constant and assign different degrees of curvature to the age trend. The period trend is set to be the same as the one used in A-2 (i.e., *near*-linear), and the cohort trend is the same as the one used in the A-set. The shape of the age pattern is similar to the one found in Jeon et al. (2016) for suicide mortality in South Korea. I retain the general shape of age effects in these models, but alter the curvature of age effects to become increasingly obvious across models: B-1 shows the least obvious trend, B-2 a modest trend, and B-3 has the most obvious trend.

The C-set tests how altering period trends affects data dimensionality, holding age and cohort trends at a constant. The age trend is held the same as the one used in B-3 (i.e., the most obvious trend), and the cohort trend is retained from the A and B-sets. Period trends are *near*-linear in the C-set, but they have a different slope in each scenario. The period trend in C-1 is least obvious, followed by a modest period trend in C-2, and a relatively steep period trend in C-3, which has the most obvious trends for all three dimensions in all the scenarios tested in this chapter.

Finally, the D-set tests how changes to cohort effects impact the data structure, holding age and period trends constant. The age trend is retained from the C-set, and the period trend is borrowed from C-3 (i.e., the most obvious trend). All three scenarios in

the D-set have similar shapes of cohort trends, but the curvatures are different. In D-1, the cohort trend is very modest, or almost flat. D-2 has a more obvious cohort trend than D-1, but it is still moderate. D-3 has the most obvious cohort trend. Note that D-3 is essentially the same scenario as C-3 but presented one more time in the D-set for comparison with the other two scenarios in the D-set.

*Simulation Procedure*

To start, I use Stata 13 to simulate data that reflect the aforementioned age, period and cohort trends. For each scenario in all four sets, I simulate 100 datasets containing 50,000 observations. Each dataset has ten age groups, ten periods of observation, and nineteen birth cohorts. For each dataset, I randomly assign a continuous term for age ranging from 10 to 59. Then I divide the ages into ten 5-year age groups. In the same way, I randomly assign a continuous term for periods of observation ranging from 1965 to 2015. Then I divide the years into ten 5-year period groups. By subtracting the continuous age term from the continuous period term, I derive birth years ranging 1906 to 2005. Then I break the birth years into nineteen 5-year birth cohorts. Using these age, period, and cohort groups and data generating processes (DGPs) that follow the temporal trends for each scenario (Figure 10), I create the outcome variable. The random error term is set to follow the normal distribution with the mean of zero and a variance of 2.

*Three-step Procedure*

I applied the three-step procedure suggested by Yang and Land (2013) to each simulated dataset using R 3.2.2. In Step 1, I conduct a descriptive analysis by obtaining

the rate of the outcome variable (suicide mortality) for each age group, period of observation, and birth cohort. The suicide mortality rate for each age group is calculated by averaging across all periods and birth cohorts; similarly, period rates are calculated by averaging across all ages and birth cohorts, and cohort rates are calculated by averaging across all age groups and periods. To ease interpretation, rates are converted into the number of suicide deaths per 100,000 person-years lived. The median estimates of the rate across 100 data sets are also calculated. In Step 2, the goodness-of-fit statistics discussed in Chapter 2 (AIC and BIC) are calculated for the 7 nested models (i.e., A, P, C, AP, AC, PC, APC) for each dataset, and the best fitting model is tallied across all 100 simulated datasets.

The results from the descriptive analysis and goodness-of-fit tests will be checked against the DGP to understand if preliminary analyses can (1) detect the true dimensionality of the data structure and (2) appropriately suggest the model that correctly matches the true DGP. In cases where the full three-dimensional model is suggested, the analysis moves on to Step 3: applying the IE to estimate age, period and cohort effects. As one of the recent critiques argued, IE estimates can significantly change or rotate when different reference categories are employed. Therefore, I estimate two sets of coefficients to see the degree of variability or rotation in the estimates—one that uses the final age, period, and cohort groups as reference categories, and another that the first age, period, and cohort groups as reference categories instead. Consequently, for the 100 datasets in each scenario, 200 sets of coefficients are obtained. To understand if changing reference categories produces meaningfully different results, I will present separate median estimates from those sets.

## Results

### *A-Set*

Figure 11 presents the descriptive statistics for each scenario in the A-set. Some scenarios in the A-set reflect the fact that the DGP imposes an *exact* linear trend on age and/or period dimensions. Again, this is an unrealistic assumption that is not observed in real-world APC data because the probability that the effects of a temporal dimension lie on an exact linear line is extremely improbable. While age, period, and cohort processes that generate the outcome can take on any form (O'Brien 2013), it is not plausible to expect that they should ever precisely follow a mathematical function.

The cohort trend is held constant across all four scenarios in the A-set. The descriptive cohort plots reflect the shape of the original DGP well, showing very little change across scenarios. When some deviations are applied to the age and/or period trend, the descriptive plots of those two terms reflect this. In A-1, in which age and period have perfect linear trends, the descriptive age and period plots also look linear. The age plot has a slightly steeper slope than the period plot, although those two terms have the same slope in the DGP. When both age and period are given deviations from the exact linear pattern but the two dimensions still have a perfect correlation by being given the exact same deviations in A-2, the descriptive plots of both dimensions reflect those deviations. However, the patterns of the plots were not perfectly identical. When a deviation is introduced to the linear trend for age in A-3 or period in A-4, the descriptive plots also successfully pick up the change. Although the correlation between the two dimensions is high, the dimension upon which the deviation was imposed reflects this the

most (i.e., period in A-3, and age in A-4) while the other dimension remains with a nearly linear descriptive plot.

Overall, the descriptive plots in the A-set reflect the modification imposed on the DGP very well. By looking at the descriptive plots, a researcher might reasonably guess that all the dimensions make contributions to the outcome variable, as no descriptive plot is really flat. However, this alone is not sufficient evidence to warrant the application of an IE model to these data. That is because age and period plots have quite similar monotonic increasing patterns and there is a possibility of confounding between these two dimensions. If that is the case, the data may not be fully three-dimensional, and the IE may not be the best option for these data. At this stage, the model fit statistics can give some supplementary information to aid in appropriate model selection.

As shown in Table 7, there is some inconsistency in the "best" models, as suggested by AIC and BIC across the different scenarios. Considering that the sample size is as large as 50,000 in this study, I will rely primarily on the AIC for choosing the best fitting model when there are contradictions between model selection statistics. (For a more detailed rationale on this point, please see Chapter 2). For scenario A-1, the AIC fails to distinguish between the perfectly linear and identical patterns of age and period effects. The AIC considers only one of those two dimensions active, suggesting reduced-dimension models, especially AC and PC models. Not one of the 100 simulated datasets was suitable for a complete three-dimensional model. In such cases, the IE should not be applied.

In scenarios A-3 and A-4, either age or period effects are allowed to deviate slightly from a perfectly linear functional form—but at least one dimension is still

constrained to fit this mathematical function. In these scenarios, the vast majority of the 100 datasets are still not suitable for three-dimensional models. The age or period dimension that is forced into an exactly linear trend is generally considered inactive by the AIC. In A-3, where deviations were introduced for period effects but age effects remain exactly linear, the AIC for 96% of the simulated datasets detects that period is an active dimension, but age is not. Similarly, in A-4, where age has some deviations while period is exactly linear, the AIC for 94% of the simulated datasets finds that the age is an active dimension, but period is not.

Only for scenario A-2, where both age and period deviate from a perfect functional form, does the AIC indicate that the data are eligible for a three-dimensional model. As both the descriptive plots and model selection statistics suggest that all three dimensions are active in scenario A-2, the analysis can proceed to Step 3—applying the IE model. According to the results (Figure 12), IE performs well, capturing the true age and period trends. The IE estimates for the youngest and oldest cohorts are slightly biased, leading to rotation in the patterning of cohort effects relative to the true trends shown by the DGP. However, contrary to the concerns expressed by Held and Reibler (2013), the degree of rotation is not sufficient to lead to an incorrect substantive interpretation of the cohort patterns.

*B-Set*

Figure 13 presents visualized descriptive statistics for the three scenarios in the B-set. Scenarios in the B-set were set to have constant patterns of period and cohort effects, with variations imposed on the age effects. As expected, age plots have apparent

differences across the scenarios. As the DGP introduces increasingly obvious age patterns from B-1 through B-3, the descriptive age plots accordingly have more obvious curvatures across these three scenarios. The descriptive period plots stay the same across all three scenarios, being impacted very little by variations in the age pattern. Although cohort effects were not altered, descriptive cohort plots nevertheless show slight changes across the scenarios. For B-3, where the age pattern was set to be the most obvious among the three scenarios, the descriptive cohort plot has an apparent decreasing trend, and the bump between the 1930 and the 1950 cohorts is more evident than the ones in B-1 and B-2. The more obvious age pattern in B-3 causes the older age groups to have much higher suicide rates than younger age groups, which in turn leads to higher suicide rates among earlier cohorts that consist of older people. Because of this inverse relationship between age and cohort (i.e., younger people make up later cohorts, older people make up earlier cohorts), descriptive age and cohort plots are partly intertwined.

Because the descriptive plots for B-1 through B-3 show distinct patterns for each temporal dimension, it seems likely that each dimension makes a unique contribution to suicide mortality. Therefore, a researcher might reasonably speculate that all scenarios in the B-set are suitable for a three-dimensional analysis. The model fit statistics for the B-set (Table 8) support such speculation. As expected, the AIC statistic points to a full, three-dimensional model for all 100 datasets in all three scenarios in the B-set. Importantly, the results of B-1 indicate that the AIC is sufficiently sensitive to detect relatively modest age effects in scenarios where period and cohort effects are active.

Based on these results, the IE is applied to all three scenarios (Figure 14). As shown, the large sample size (N=50,000) used in the simulation yields small variations to

the estimates of the APC effects across the data sets. However, the IE performs very well, especially for capturing the true age and period effects. Although the IE tends to slightly underestimate the effects of the earliest cohorts and overestimate the effects of the latest cohorts, the degree of inaccuracy is trivial and the point intervals mostly catch the true DGP in all three scenarios.

*C-Set*

In the C-set, the age and cohort patterns are held constant while variation is imposed on the period pattern. The descriptive age and period plots reflect the DGPs in the three scenarios quite well (Figure 15). The age plots retain their shape, and the period plots detect changes assigned to the DGP across scenarios C-1 through C-3. However, although the cohort DGP was constant across scenarios, the descriptive cohort plots are affected by variation in the period effects. For scenario C-1, the cohort pattern exhibits a decreasing trend, with a couple of "bumps" for the 1945-49 and 1975-79 cohorts. Scenario C-2 has a quite similar pattern, but the cohort trends before 1930-34 and between 1950 and 1969 are flat rather than decreasing. Finally, scenario C-3 shows an overall increasing trend with similar bumps, with the exception of the latest two cohorts where there is a steep drop.

Based on the obvious and distinctive patterns of these descriptive plots, a researcher is likely to guess that age and cohort are active dimensions in all three scenarios. The key to estimating the data structure in the C-set is whether the monotonically increasing period effects are sufficiently obvious enough to warrant modeling this dimension. With the help of model-fit statistics, a clearer picture begins to

emerge (Table 9). In scenarios C-1 and C-2, the AIC suggests a reduced AC model for some of the 100 simulated datasets, as the period effects have less obvious trends. For C-1, the AIC suggests a reduced model (i.e., AC) for almost half of the simulated datasets. The other half in C-1 turned out to have the best fit with the full, three-dimensional model. For C-2, in which the period pattern is more obvious than C-1, 88 out of 100 simulated datasets appear to have the three-dimensional structure. The other 12 datasets have the best fit with the AC model. AIC suggest the full, three-dimensional model for all 100 datasets in C-3. This is not surprising, as all three temporal dimensions exhibited obvious patterns through the descriptive statistics.

As AIC suggests the three-dimensional model to 88 out of 100 datasets in C-2, I applied the IE to those 88 datasets (Figure 16(a)). The IE captured the less obvious trend of period effects well. The patterns of age and cohort effects in the DGP were also accurately estimated by the IE, with slight rotation of effects noted in the cohort dimension.

Based on the preliminary steps, all 100 datasets are eligible for estimation of an IE model in scenario C-3 (Figure 16(b)). Overall, the IE does a good job catching the temporal dimensions embedded in the data structure, although biases for the youngest and oldest cohorts cause more noticeable rotation in cohort effects than in scenario C-2. The IE captures the age DGP quite well overall, despite slight overestimation at the later ages. The true pattern of period effects is also accurately estimated by the IE, with minor overestimations for the earliest periods and underestimations for the latest periods.

*D-Set*

Finally, the scenarios in the D-set have fixed age and period trends with varied cohort patterns. Consistent with this DGP, the descriptive plots for age and period effects are unaffected by changes imposed on the cohort trend, showing little difference across the scenarios (Figure 17). The only easily noticeable difference is the cohort pattern for D-3, which has much sharper peaks than the cohort plots for D-1 or D-2.

One thing worth noting in the D-set is that when the cohort has essentially no trend (i.e., D-1), the descriptive cohort plots may nevertheless appear to have a distinctive pattern. This is quite different from what was observed with age and period descriptive plots in the B and C-sets. The true patterns of age and period effects were accurately reflected in their descriptive plots and not substantially distorted by changes in the other dimensions. However, descriptive plots for the cohort dimension are strongly impacted by variations introduced for the other dimensions. As a result, the descriptive plot in D-1 gives the appearance that cohorts independently contribute to suicide mortality, even though the DGP for cohort effects is actually flat. If a researcher were to consider the cohort as an active dimension, exclusively relying on the descriptive statistics, this would lead to an erroneous methodological decision in which a model were selected that includes the cohort as an active dimension. To avoid such a mistake, the model-fit statistics must always be given consideration as well.

The AIC statistic shows that cohort is not an active temporal dimension 99 of the 100 simulated datasets, indicating that the AP model best fits these data for D-1. As the cohort trend has increasingly obvious patterns in D-2 and D-3, the AIC suggests that the

three-dimensional model is best for 94 out of 100 simulated datasets in D-2 and all 100 simulated datasets in D-3.

Following the three-step procedure, the analysis can proceed to Step 3 for scenario D-3. As noted before, D-3 is the same scenario as C-3; those IE results are not reiterated here. As the AIC suggests the three-dimensional model for 94 datasets out of 100 in D-2, the IE is applied to those 94 datasets. As presented in Figure 18, the IE captures the true age and period effects very well. The estimates for the cohort effects are slightly rotated compared to the DGP. However, considering the data in D-2 has an age- and period-dominant structure, this minor rotation has minimal influence the overall conclusions about the contributions of these temporal dimensions to suicide mortality.

**Discussion**

In her recent critique of the IE, Luo (2013a) simulated age, period, and cohort effects on exact linear lines. In their response to this critique, Yang and Land (2013b) emphasized that this simulation design violates a key modeling assumption—namely that three-factor APC models are not estimable when all three temporal factors have exact, linear relationships with the outcome. Aside from being highly improbable and therefore unrealistic, the problem with such a scenario is that one temporal dimension can be expressed as an exact, linear function of the other two. Yang and Land go on to argue that such a clear violation of the modeling assumptions is detectable by conducting preliminary analyses. Note that if exact linear functional forms are not assumed, the "linear constraint" that Luo (2013a) derived does not hold true.

Although a data structure with two or more exact linear temporal trends is extremely unlikely in any data structure generated by real social phenomena, I

nevertheless examined the claim that misapplication of the IE to such data could be prevented by conducting preliminary analyses. Overall, the results of the A-set support Yang and Land (2013b)'s argument. In all four scenarios that comprise the A-set, age and period dimensions have either *near* or *exact* linear effects at the same time. When at least one dimension has an exact linear effect (i.e., A-1, A-3, and A-4) and one has a near linear effect, the descriptive statistics reflect this fact quite well, and the AIC suggests reduced models. This result implies that when confounding between two linear dimensions occurs during the DGP, the resulting data will have fewer than three dimensions. In such cases, three-dimensional IE models are not capable of disentangling the effects of those confounded dimensions. Therefore, the IE should never be applied to such data. Because preliminary analyses provide important clues about the presence of confounding, it is essential to conduct these analyses to prevent misapplication of a three-dimensional model.

However, when the data structure becomes more plausible by allowing some deviations from the exact linear trends for age and period effects (i.e., scenario A-2), the results of preliminary analyses correctly pointed to a full, three-dimensional data structure. When applied to such data, the IE performs very well. Based on these findings from the A-set, we now know three important features of preliminary analyses and the IE itself: First, imposing unrealistic functional forms during the DGP will lead to data structures that violate model assumptions, rendering such exercises meaningless with respect to the performance of the IE or any other APC model. Second, the preliminary steps outlined by Yang and Land (2013b) can indeed help detect data structures that are ineligible for three-factor APC models, such as the IE. Third, when preliminary analyses

point to a more realistic data structure with variability across all three temporal dimensions, the IE performs well.

In the B, C, and D-sets, I tested how well the descriptive and model selection statistics detected changes in the dimensionality of the data. In all three sets, I introduced changes to age, period, and cohort effects that induce a reduction in the data structure. Then, I observed how the descriptive and model fit statistics react to those changes and whether they reflect the true data structures well. The descriptive plots in sets B-D accurately reflected changes to the age and period patterns across different scenarios. However, the descriptive plots for cohort effects were not as straightforward to interpret and could at times be misleading. For example, plots of cohort effects were sensitive to changes in the age patterns (B-set) and period patterns (C-set). Conversely, descriptive period and age plots did not change substantially in response to changes in cohort patterns. Moreover, even though descriptive cohort plots responded to cohort variation in the DGP (D-set), they do not resemble the DGP as with age and period plots. Instead, cohort plots have distinct shapes that appear to be sensitive to all three temporal dimensions. Even where the cohort DGP is nonexistent (D-1), the descriptive plots do not look flat or indistinct. In summary, descriptive plots for tabular-rate APC data provide a good indication about the structure of age and period dimensions, but cohort patterns are difficult to discern from descriptive plots only. Therefore, it is imperative that researchers use model selection statistics along with descriptive plots to help determine whether all three dimensions are making unique contributions to an outcome.

Model selection statistics—most notably the AIC in my analyses—provide additional evidence to help researchers adjudicate between different possible data

structures. When I modified a data structure to become two-dimensional or on the border between two and three dimensions, the AIC tended to swing between two different models. However, in cases where the AIC pointed to a three-dimensional model, the IE performed quite well. Even in the cases where a particular temporal dimension had few obvious effects, the IE successfully captured the true data structure as long as the AIC indicated that this dimension is active.

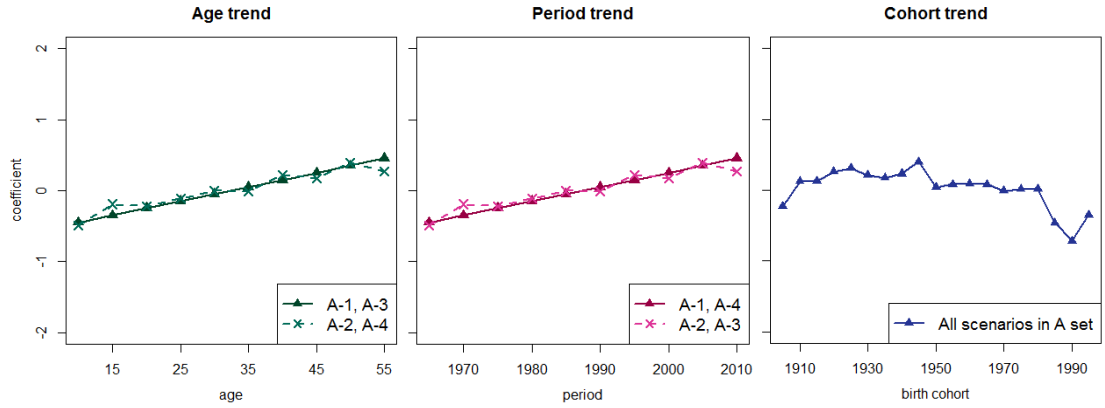
By integrating the results, I answer the two research questions of this chapter as follows: First, can the two preliminary steps detect the true data structures? The answer is yes. However, when two or more dimensions are perfectly correlated by having the same functional form, a three-dimensional dataset may appear to be reduced-dimensional due to the confounding between those dimensions during the DGP. In this case, as the model-selection statistics would not point to three active dimensions, a researcher should never apply the IE. If applied, the IE is likely to produce biased estimates of the true coefficients. Second, when the preliminary steps suggest the data are three-dimensional, does the IE perform well? The answer to this question is also yes. For plausible data structures, the preliminary steps are useful tools to understand the active dimensions. When the descriptive and the model-fit statistics verify that the given data is suitable for the three-dimensional models, the IE captures the DGP well.

In summary, my analyses provide little evidence that that the IE is an arbitrary estimator that researchers should avoid. The reason why the IE failed in the critique offered by Luo (2013a) is that she applied it to data that are not eligible for this model. The simulated data *are* three-dimensional in Luo's study, but not suitable for the IE because perfectly linear effects of more than two dimensions lead to confounding

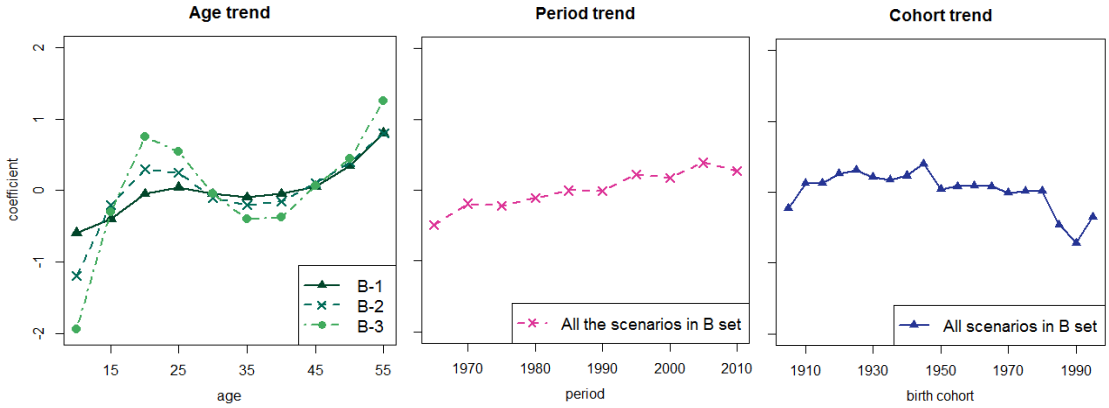
between them. This is a highly unrealistic assumption that induces the data to look and act like reduced-dimensional data at the completion of the DGP. Such data are not in the scope of the application of the IE, and researchers can avoid such misapplications of the IE by conducting preliminary analyses, as suggested by Yang and Land (2013b).

## Tables and Figures

(a) A-Set




(b) B-Set




(c) C-Set

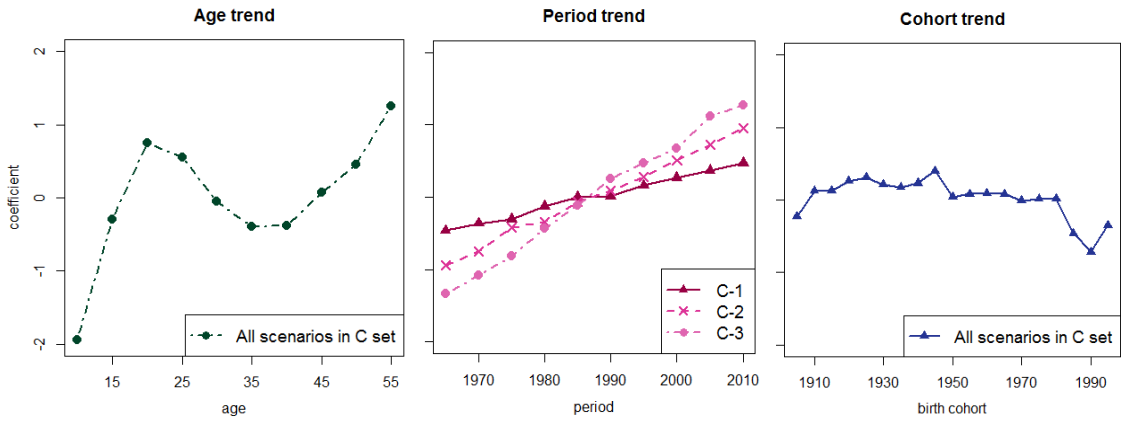



Figure 10. Visualized Simulation Scenarios of A, B, C, and D-Sets

(d) D-Set

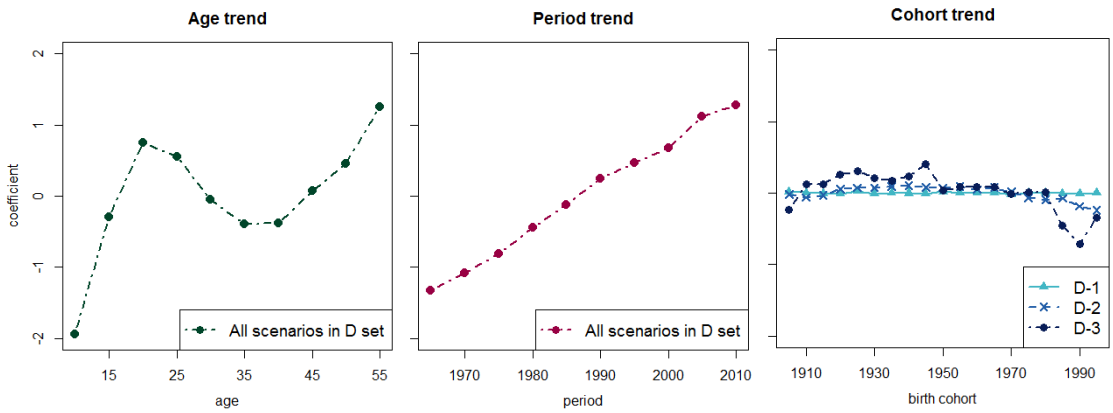



Figure 10 (cont'd). Visualized Simulation Scenarios of A, B, C, and D-Sets

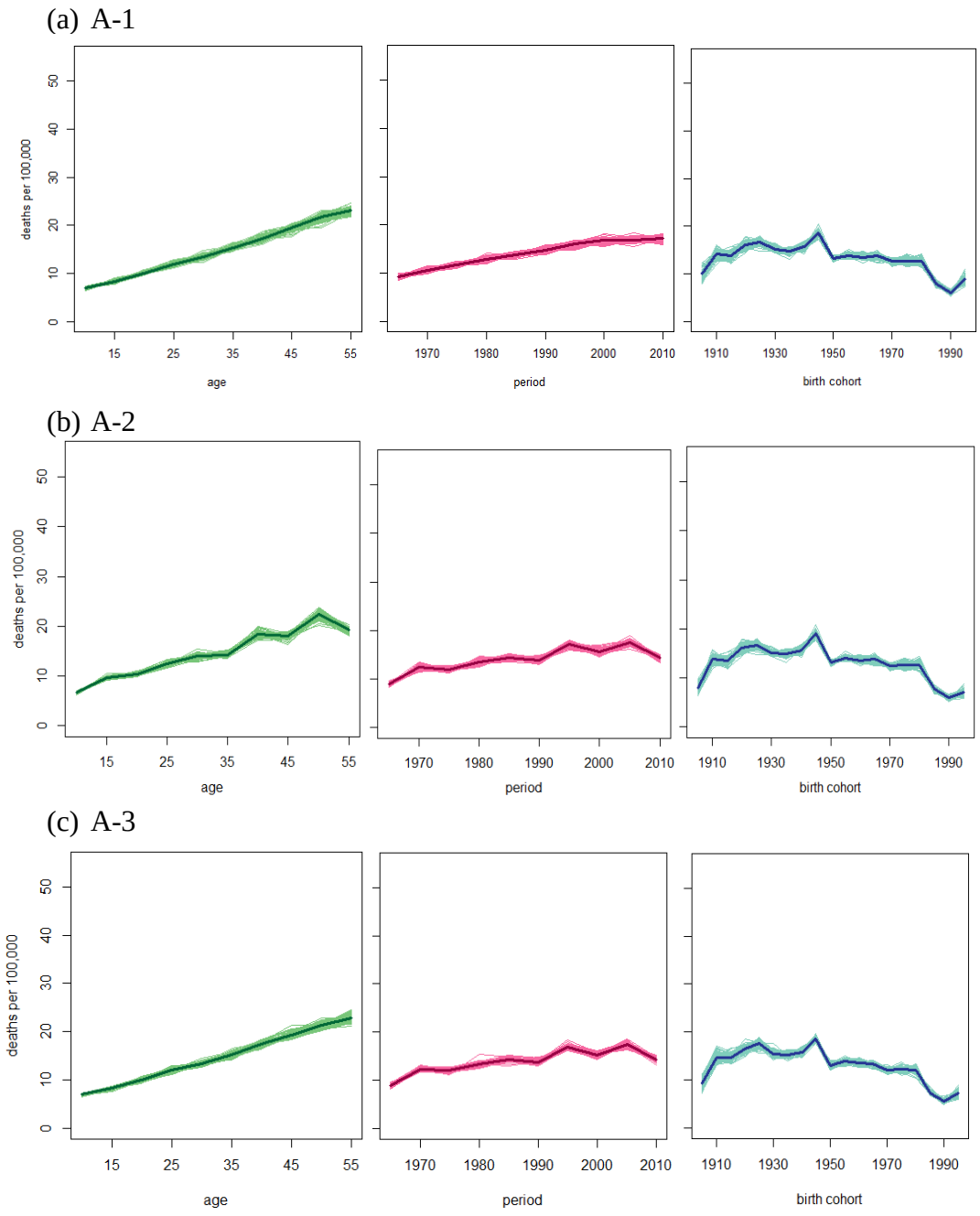



Figure 11. Visualized Descriptive Statistics of Scenarios in A-Set

(d) A-4

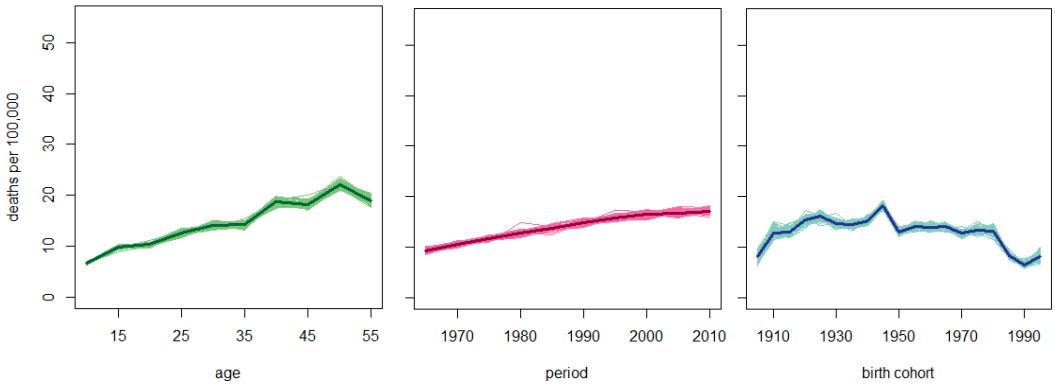



Figure 11 (cont'd). Visualized Descriptive Statistics of Scenarios in A-Set

Table 7. Model Fit Statistics for Nested APC Models Fitted to 100 Simulated Data in A-Set

| | | A | P | C | A+C | A+P | P+C | APC |
|---|---|---|---|---|---|---|---|---|
| A-1 | AIC | 0 | 0 | 0 | 52 | 0 | 48 | 0 |
| | BIC | 0 | 0 | 0 | 52 | 0 | 48 | 0 |
| A-2 | AIC | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| | BIC | 0 | 0 | 0 | 49 | 0 | 35 | 16 |
| A-3 | AIC | 0 | 0 | 0 | 0 | 0 | 96 | 4 |
| | BIC | 0 | 0 | 0 | 0 | 0 | 100 | 0 |
| A-4 | AIC | 0 | 0 | 0 | 94 | 0 | 6 | 0 |
| | BIC | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| | DF | 10 | 10 | 19 | 28 | 19 | 28 | 36 |
| | N | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 |

(a) A-2, IE

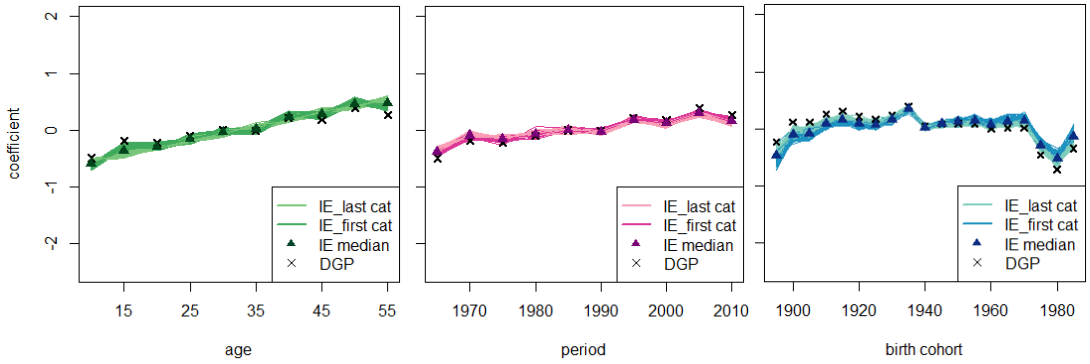



Figure 12. Estimates of the IE for 100 Simulated Data in A-Set

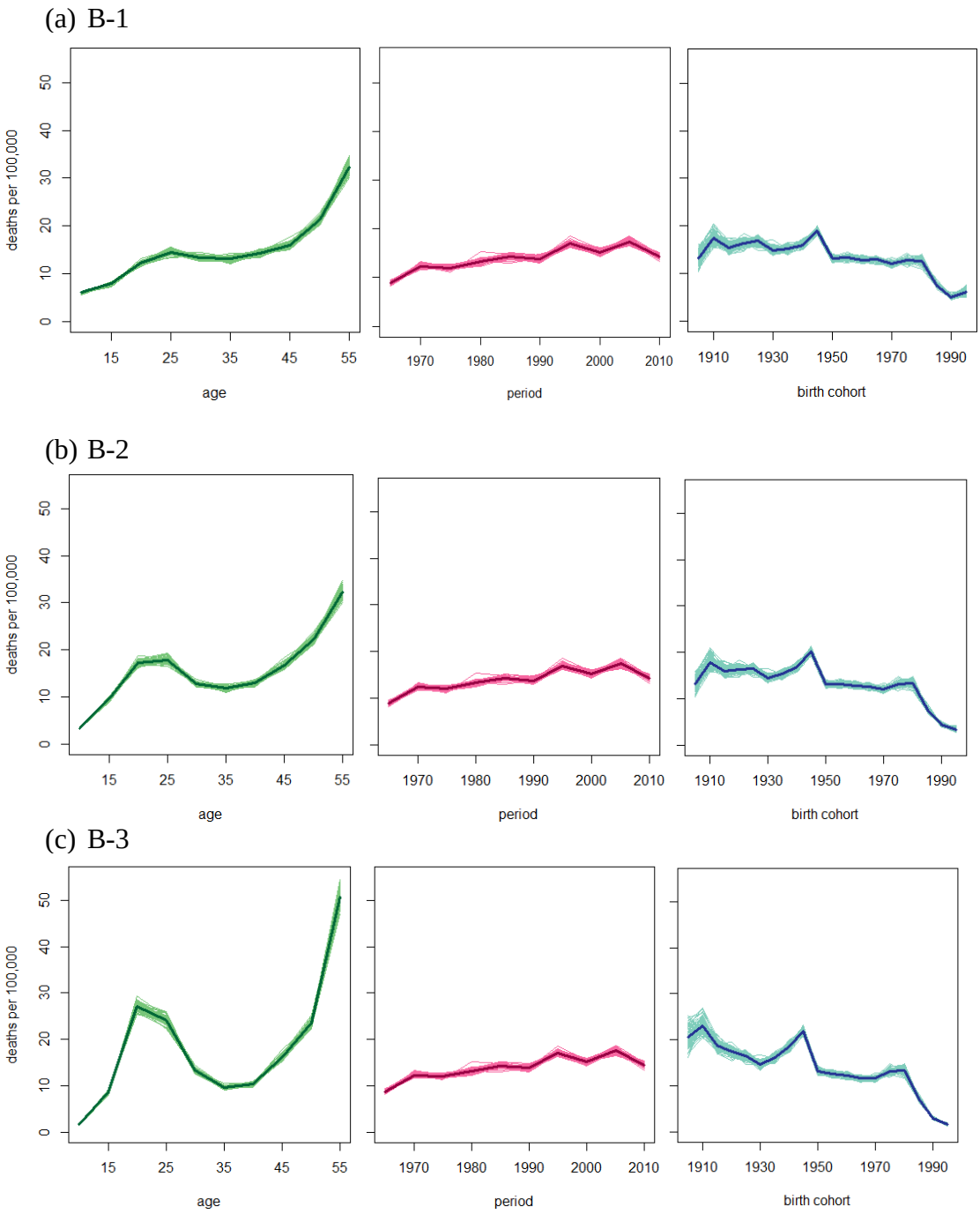



Figure 13. Visualized Descriptive Statistics of Scenarios in B-Set

Table 8. Model Fit Statistics for Nested APC Models Fitted to 100 Simulated Data B-Set

| | | A | P | C | A+C | A+P | P+C | APC |
|---|---|---|---|---|---|---|---|---|
| B-1 | AIC | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| | BIC | 0 | 0 | 0 | 71 | 0 | 0 | 29 |
| B-2 | AIC | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| | BIC | 0 | 0 | 0 | 71 | 0 | 0 | 29 |
| B-3 | AIC | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| | BIC | 0 | 0 | 0 | 71 | 0 | 0 | 29 |
| | DF | 10 | 10 | 19 | 28 | 19 | 28 | 36 |
| | N | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 |

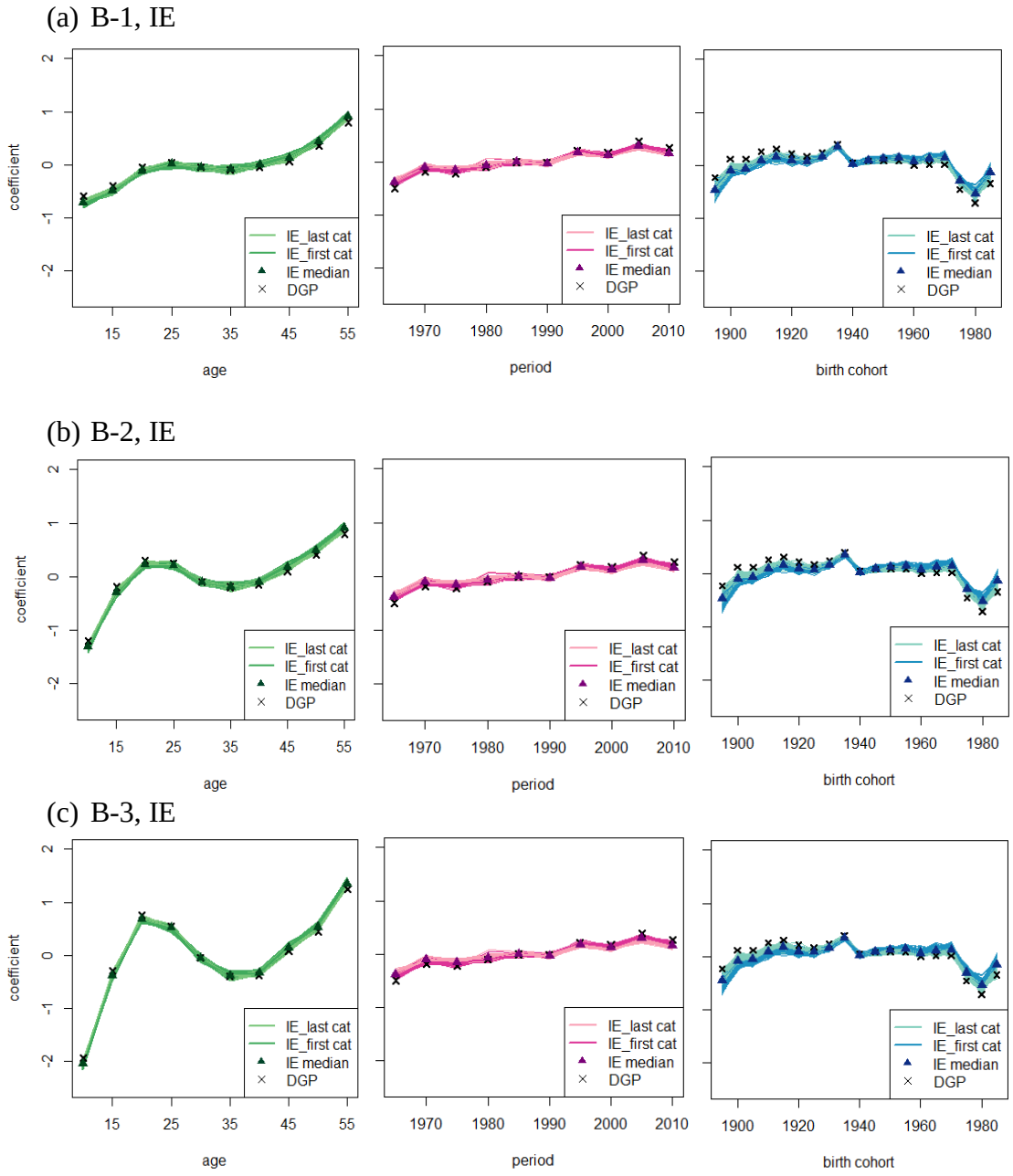



Figure 14. Estimates of the IE for 100 Simulated Data in B-Set

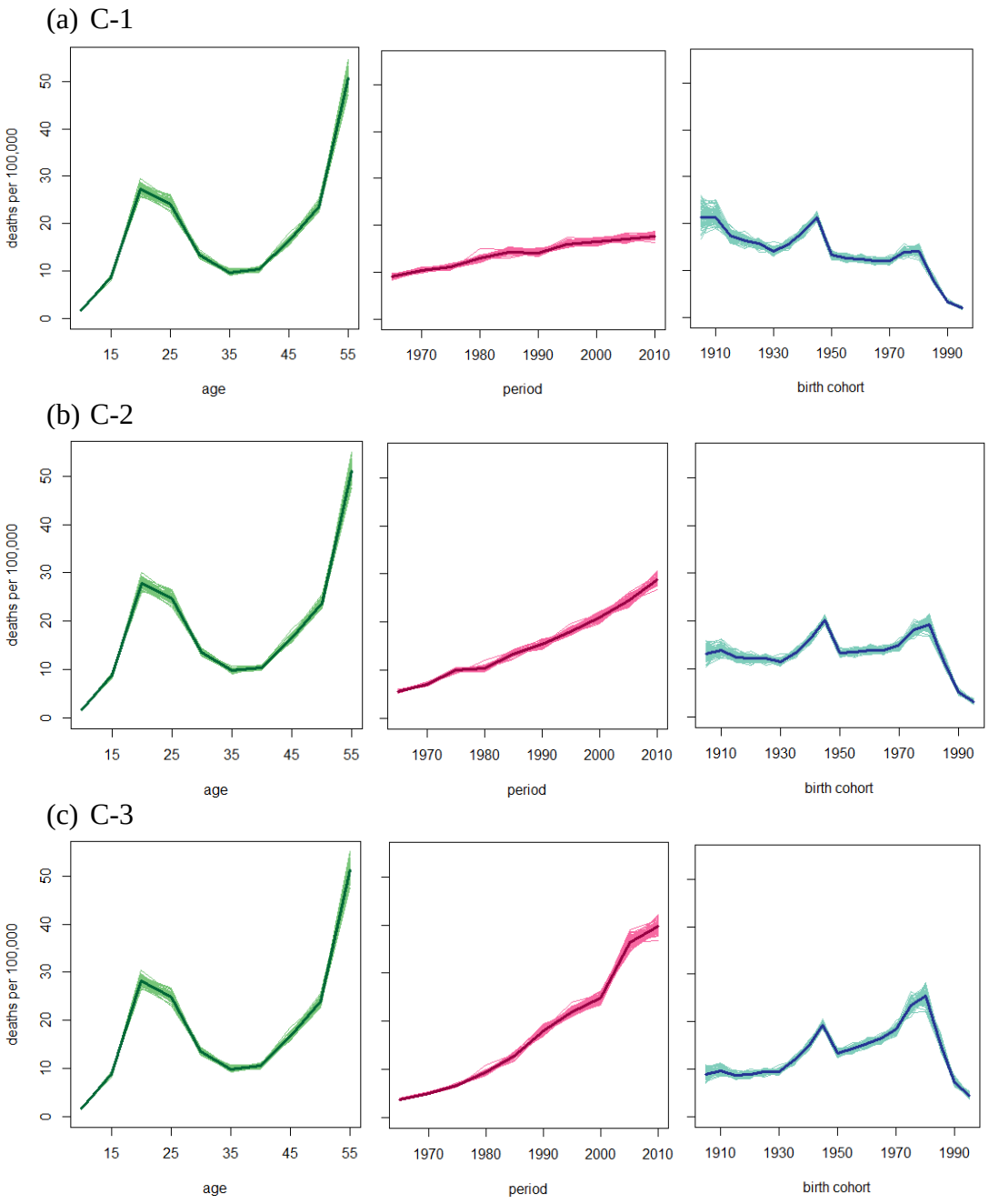
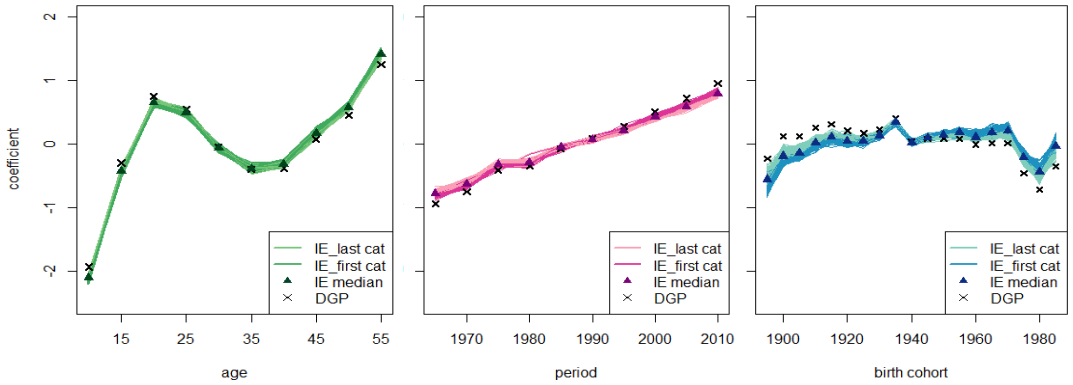



Figure 15. Visualized Descriptive Statistics of Scenarios in C-Set

Table 9. Model Fit Statistics for Nested APC Models Fitted to 100 Simulated Data in C-Set

| | | A | P | C | A+C | A+P | P+C | APC |
|---|---|---|---|---|---|---|---|---|
| C-1 | AIC | 0 | 0 | 0 | 49 | 0 | 0 | 51 |
| | BIC | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| C-2 | AIC | 0 | 0 | 0 | 12 | 0 | 0 | 88 |
| | BIC | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| C-3 | AIC | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| | BIC | 0 | 0 | 0 | 98 | 0 | 0 | 2 |
| | DF | 10 | 10 | 19 | 28 | 19 | 28 | 36 |
| | N | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 |

(a) C-2, 88 Simulated Dataset, IE




(b) C-3, 100 Simulated Dataset, IE

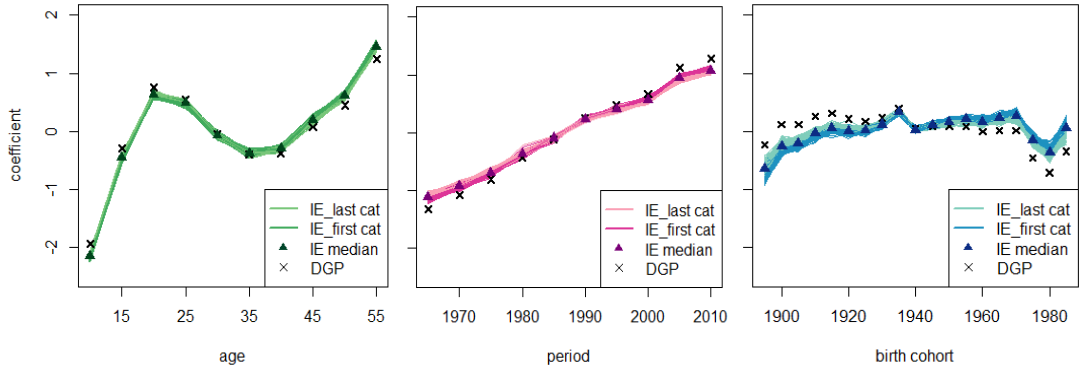



Figure 16. Estimates of the IE for Simulated Data in C-Set

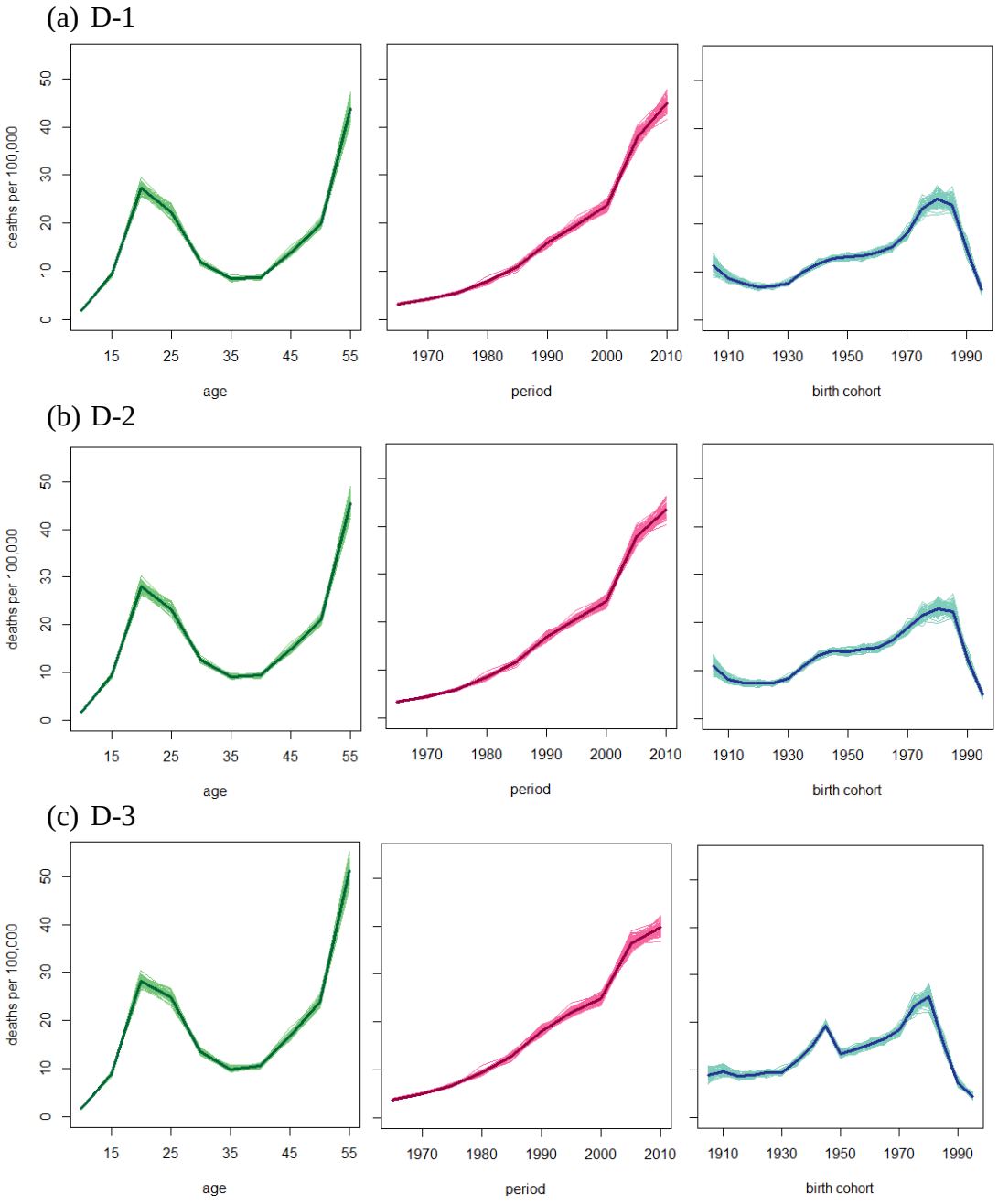



Figure 17. Visualized Descriptive Statistics of Scenarios in D-Set

Table 10. Model-Fit Statistics for Nested APC Models Fitted to 100 Simulated Data in D-Set

| | | A | P | C | A+C | A+P | P+C | APC |
|---|---|---|---|---|---|---|---|---|
| D-1 | AIC | 0 | 0 | 0 | 0 | 99 | 0 | 1 |
| | BIC | 0 | 0 | 0 | 0 | 100 | 0 | 0 |
| D-2 | AIC | 0 | 0 | 0 | 0 | 6 | 0 | 94 |
| | BIC | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| D-3 | AIC | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| | BIC | 0 | 0 | 0 | 98 | 0 | 0 | 2 |
| | DF | 10 | 10 | 19 | 28 | 19 | 28 | 36 |
| | N | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 | 50,000 |

(a) D-2, 94 Simulated Datasets, IE

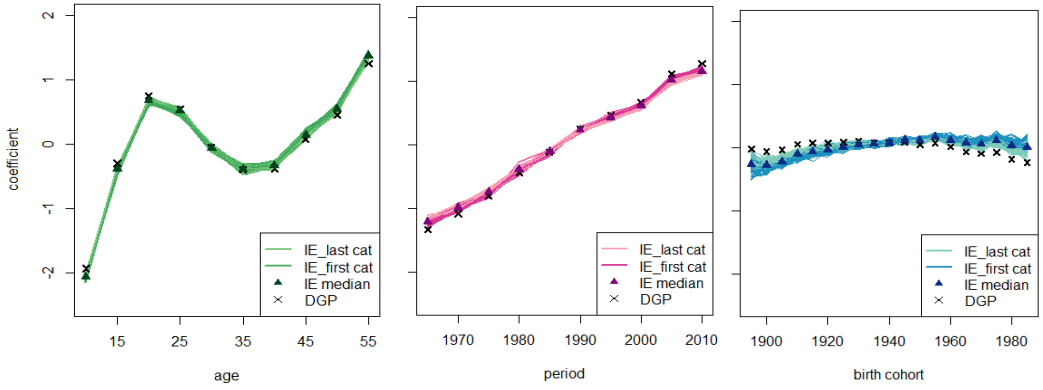



Figure 18. Estimates of the IE for Simulated Data in D-Set

# CHAPTER IV
# ESTIMATING THE EFFECTS OF AGE, PERIOD, AND COHORT ON THE EDUCATIONAL GAPS IN HEALTH USING HIERARCHICAL AGE-PERIOD-COHORT MODELS

## Introduction

All human beings experience biological aging. It accompanies health deterioration after a certain age during adulthood. The accumulation of damage at the molecular- and cell- level that occurs within an individual over time causes a loss of ability to restore homeostasis, organ failures, and eventually elevated risks of diseases and death. Gompertz demonstrated that human morbidity and mortality rates exponentially increase throughout adulthood beyond the age of 30 (Fries 1983, Fries 2002, Manton 1982). While there have been significant reductions in age-specific morbidity and mortality rates over the last century, health deterioration accompanied by individuals' aging still remains an irreversible and intractable process (Kannisto et al. 1994).

Although no one can avoid gradual health decline as a consequence of biological aging, the rate of health deterioration differs from person to person. This difference is caused by a complex combination of life events, genes, health behaviors, environments, and other unknown determinants. In particular, social epidemiologists have studied the socio-economic predictors that differentiate the rate of health deterioration. Education is one of those predictors, which significantly affects the pace of health deterioration, according to previous studies (Conti and Heckman 2010, Cutler and Lleras-Muney 2006). Less educated people tend to experience faster health deterioration than more educated people over the life course. As a result, people with higher educational attainment are

likely to show better health outcomes when measured by their self-rated health, morbidity, mortality, and mental well-being than less educated people at the same age (Feldman et al. 1989, Kaplan et al. 1987, Winkleby et al. 1992).

There are several different explanations for this inverse association between education and the rate of health deterioration. Human capital theory argues that more schooling produces more human capital, which is necessary to maintain one's health, including cognitive skills, rational and complex strategies of thinking, self-directedness and self-confidence (Becker 1994). The credential model highlights the importance of educational degrees in hiring decisions. Higher education calls for better jobs with higher income, which eventually leads to the procurement of quality health care and a healthy diet (Collins 1979).

When the rates of deterioration differ by the level of education (i.e., the deterioration rate is higher in less educated people and lower for more educated people), the consequence is a widening health gap across different educational groups as people age. Some scholars apply "Cumulative Advantage Theory (CAT)" to understand the relationship between education and health disparities over the life course (Mirowsky and Ross 2005, Ross and Wu 1996). According to this theory, effective abilities, healthy habits, and nurturing environments, which are obtained by education in early adulthood, lead to health advantages in later life stages. Conversely, early disadvantages such as economic hardship and chronic stressors cause health deterioration in the future (Mirowsky and Ross 2005). This theoretical approach attributes the widening health inequality by education over the life course to accumulated benefits and difficulties with aging.

Meanwhile, some recent studies have found evidence that another temporal dimension—birth cohort membership—is also an important factor that contributes to the diverging health gap by education (Lauderdale 2001, Lynch 2003). These studies have shown that the effects of education on health are larger among recent birth cohorts than among earlier cohorts (Lauderdale 2001, Lynch 2003). Findings of these studies may contrast with CAT. For instance, according to the CAT, the cumulative gaps of advantages and disadvantages are maximized among older people. However, the cohort studies show that older people experience the least severe inequality, because they are comprised of the earliest cohorts. These opposite trends of age and cohort highlight the importance of taking into account of the independent trajectories of the temporal terms for studying educational inequality in health (Lynch 2003). Failing to separate the independent trajectories by age and cohort may cause an appearance of converging health inequality over time, which is a false representation of what is really happening as a result of the obscured effects of age and birth cohort.

When attempting to estimate inequalities in health trajectories across an extensive range of birth cohorts and age groups, data limitations can impose some difficulties. For example, to include a wide range of birth cohorts in a statistical model, a researcher needs access to data that cover relatively long periods of observation. That is because age and cohort terms are defined in a linear relationship with period (birth year = period-age), and therefore with short time periods, only a limited number of birth cohorts can be defined. For instance, a single cross-section of data cannot differentiate age groups from birth cohorts, and it is impossible to separate the effects of age and birth cohorts on health trajectories in such data setting.

Using data from repeated cross-sectional surveys covering longer time periods provides one solution to this problem. However, there is another issue with this approach: now the data cover more periods, and another temporal dimension (i.e., secular changes over time) may confound the contributions of age and cohort to changes in health. Then, a researcher should also take into account the independent contribution of secular changes when understanding health trajectories over the life course (i.e., age) and across birth cohorts. In traditional regression-based approaches, a statistical analysis of these three temporal dimensions presents difficulties associated with the perfect linear relationship between the three terms, as discussed in Chapter 2 and Chapter 3.

The newly emerging age-period-cohort approach for repeated cross-sectional survey data can be a good solution in this context. Using a hierarchical age-period-cohort (HAPC) model, a researcher can extend the ranges of periods and cohorts by including multiple waves of data, permitting separate estimates for different ages, periods of observation, and birth cohorts. In this study, I attempt to estimate the unique contributions of the three time dimensions to health inequalities by education using the cross-classified random-effects models (CCREMs) of the HAPC approach.

As suggested by Yang and Land (2013a) and Chapter 2 of this dissertation, the HAPC models can provide valid results if they are preceded by preliminary analyses that evaluate the data structure. In this chapter, I follow the three-step procedure of the HAPC model recommended by Yang and Land (2013a), which includes a descriptive analysis using graphics (Step 1) and an examination of model selection statistics across nested models to understand the structure of given data (Step 2). Then, I apply the model that is the best match to the data structure (Step 3).

Through this process, I will achieve the following aims: First, from the perspective of social epidemiology, I will estimate how educational inequalities in health vary across age groups, periods of observation, and birth cohorts. Second, by examining the Cumulative Advantage Theory and recent findings from cohort-based studies, I will determine which theoretical perspective is upheld in a nationally-representative data source. In addition, from a methodological perspective, I provide guidelines for future studies that employ the HAPC approach. Given recent debates on the validity of the HAPC, it is important that researchers understand how to avoid misapplications through preliminary analyses (Yang and Land 2013a).

*Socioeconomic Status, Education, and Health Gradient*

The literature in social epidemiology has thoroughly documented that SES is a powerful and consistent predictor of health (Adler et al. 1994, Feinstein 1993, Winkleby et al. 1992). SES is a complex concept that can be measured by a spectrum of variables, including occupation, income, and education. Researchers have found that each of those widely used indicators represents a unique aspect of one's status in the social structure and, therefore, has a unique implication for health (Adler et al. 1994, Shavers 2007). For example, occupation is a measure for working environments (e.g., physical hazards and psychological stressors) which can predict health risks that an individual is exposed to. Income can be a proxy for access to material goods, including quality food and better housing options that can influence one's health. Education may reflect lifestyle and health behaviors that lead to variations in health risks; it may also represent self-efficacy when coping with illnesses (Shavers 2007),

Among the SES indicators, education is the most commonly used measure of SES in sociological and epidemiological studies (Adler et al. 1994, Shavers 2007). Since education is often completed during the early stages of life, it becomes a preceding condition to other aspects of the SES for most people that also provides initial placement within the social stratification system. Once an individual is located in the hierarchical social structure, it significantly affects his/her occupational opportunities and income potential. In that sense, education has not only a direct influence on health because it changes lifestyles and behaviors, but also an indirect influence on health by playing a critical role in determining other aspects of SES (Moore and Hayward 1990).

Importantly, scholars argue that education influences health throughout adulthood, even though the process of educational attainment is usually completed relatively early in life (Ross and Wu 1996, Shavers 2007). Some evidence indicates that health gaps across educational groups do not converge with age, but rather widen across the life course (Mirowsky and Ross 2005). Mirowsky and Ross (2005) emphasize that education transforms a person by putting his/her life on a particular track early in life that is difficult to alter in later years. The advantages that better education afford are therefore not temporary, but persistently embedded in one's lifestyle, social support system, health behaviors (such as diet, smoking, and drinking), and access to quality health care. In the long term, the difference that those factors produce is cumulative and causes more divergences in physiological and psychological functions.

As advantages accumulate over time for the well-educated, so do disadvantages for the less fortunate. Lower tiers of the social stratification system embody lower-status occupations and lower earnings, which may harm health by limiting access to quality

health care, exposing to dangerous and stressful work environments, and deterring from purchasing healthy food or treating illness (Dupre 2008, Ferraro and Kelley-Moore 2003). In addition, such life conditions may result in chronic stress, which can cause deteriorated immune system, promote diseases, and increase mortality in a long-term (McEwen and Seeman 1999). According to McEwen (1998), the disadvantages in health due to lower educational attainment are minor when people are young, because people are likely to be healthy during their early life stage. However, when stressors increase and coping mechanisms are lacking over the long term, the body's adaptive system may break down, which ultimately leads to a significantly elevated risk of having diseases in later life stages.

*Estimating Three Temporal Effects to Explain Health Inequality*

This explanation of the widening health gap by accumulated advantage and disadvantage is based upon Cumulative Advantage Theory (CAT). The concept of CAT was originally proposed in Robert Merton's work on differentials in academic careers (Merton 1968). It explains that a difference in early-career performance grows over time, and causes cumulative career advantage by attracting exponential resources and increased reputation. Here, the accumulation occurs through a temporal process that widens the gaps. The amount of time spent in an academic career is what determines the amount of accumulation and the consequent performance gap between individuals.

Applied to health inequalities over the life course, the time unit that determines the amount of accumulation is the time since birth, i.e., the individual's biological age. Most studies include age as the time unit (Ross and Wu 1995, Ross and Wu 1996).

However, recent studies have shown that birth cohorts—another temporal dimension—also has significant implications for health gaps over the life course (Lynch 2003). These findings suggest that it is important to study the independent contributions of both temporal terms to understand differences in health over the life course. Lynch (2003) demonstrated the importance of considering cohort effects along with age effects, showing that the patterns of age and cohort effects might suppress each other when either was not considered. Because education has undergone enormous changes over the latest few decades, different cohorts may have experienced different contents of education, which may have changed the association between education and health. Lynch highlights that ignoring either age or cohort terms in studies of health inequalities over the life course may result in completely different conclusions. He shows that the effects of education on health grow stronger with age, and that this pattern has become stronger across cohorts. This important finding could not have been observed if the author had not taken into account both age and cohort effects.

Although following individual cohorts over time is one way to control for differentiated cohort effects on health, the drawback is that longitudinal data generally cover a short follow-up period (Kitagawa and Hauser 1973) and only very few cohorts. Consequently, it is hard to track changes throughout adulthood or to compare various cohorts to each other. Using repeated cross-sectional data is an alternative strategy because they typically contain information from young adults to elderly people in each wave of data collection; this introduces the option of estimating both age and cohort trajectories while also controlling for the influence of secular changes on health over long periods of observation. If the contribution of secular changes (i.e., period effects) to

health is ignored, the model may produce biased estimates due to potentially confounded effects of age and cohort, and it will also fail to examine how health inequalities have changed over time. If there was, for instance, a public health intervention implemented at a certain point in time, the effectiveness of the intervention can be revealed by estimating period effects. Therefore, it is essential for a researcher to employ a statistical approach such as HAPC, which can disentangle the effects of the three different time dimensions for accurately evaluating the temporal changes in health inequality.

*Research Goal of This Chapter*

Once a researcher successfully estimates the independent contribution of age, period, and cohort, the results can be used to shed more light on some important temporal aspects of educational inequality in health. First, trajectories of age effects over life-course can provide a clear understanding on whether health inequality is a process of unbounded growth. There have been diverging arguments on this issue. Some scholars (Beckett 2000, House et al. 1994) argue that health advantages tend to diminish after a critical age because of social programs for the elderly and the elimination of unhealthy people by premature deaths. Other scholars argue that the accumulation lasts until the end of the life course with no convergence in the health gap (DiPrete and Eirich 2006). Aware of these different theoretical perspectives, I will test if the health gap widens without bound by estimating predicted probabilities using the HAPC model. If the CAT holds true, educational disparities in health should be relatively narrow during early life stages and subsequently widen with age, reaching a maximum at the latest life stage. If the

growth of health is a bounded process, the results on age effects will also identify when during a life course the influence of the cumulative advantage is maximized.

Second, a comparison of predicted probabilities in different educational groups by period will provide an understanding on whether the health inequality between educational levels has changed over the last two decades. Lastly, based on the results on cohort effects, it is possible to examine Lynch (2003)'s argument that the recent finding that the later cohorts are more likely to suffer from the educational inequality in health for the latest cohorts. If the finding does not hold true for the most recent cohort, the results of this study will help detect the birth cohorts who suffer the most and the least from health inequality by education.

In addition, the results of this study can help us understand what level of education most significantly reduces the risk of reporting poor health over the life course. Most prior studies have focused on the health effects of education and treated education as a continuous variable measured by the years of education attained (Mirowsky and Ross 2005, Ross and Wu 1996). Although the results of such analytic approach will show how much health can be improved by additional years of education, they do not tell us which level of educational attainment (e.g., high school or college education) is most important with regard to slowing the rate of health decline.

This study uses up-to-date data. Elo and Preston (1996) analyzed the National Longitudinal Mortality Survey (1979-1981) to investigate the relationship between health and education. In addition, Kramarow, Pastor and Gorina (2000) examined educational differences in health a decade later for older adults using 1987-1994 data. This study utilizes nationally-representative data from 1994 to 2014. Using more recent data enables

me to update the understanding on the educational inequality in health for the latest periods and also for more recent birth cohorts.

## Methods

### *Data and Measures*

I use data from the National Health Interview Survey (NHIS), an annual cross-sectional survey conducted since 1956. In this study, I include the timeframe 1997 to 2014 to update findings from previous studies. The sample in this study includes only adult individuals whose ages are between 28 and 85. The youngest age is set at 28 to accommodate the assumption that respondents have attained their highest level of education. This setting yields 18 single-year periods of observation and 58 continuous age groups. Subtracting age from period, I obtain individuals' birth year, which ranges from 1912 to 1986. Following practices recommended elsewhere (Kupper et al. 1985, Preston, Heuveline and Guillot 2000, Yang and Land 2013b), I divide individual birth years into 15 separate 5-year birth cohorts (i.e., 1912-1916, 1917-1921,..., 1982-1986).

I focus on four levels of educational attainment: (1) no high school degree, (2) high school degree, General Educational Development (GED), or equivalent, (3) some college or associate's degree, and (4) bachelor's degree or more. Individuals who answered "don't know" or refused to respond to the question about educational attainment are excluded from the analysis.

For the outcome variable to measure health status, I use self-rated health, which in NHIS is asked by the question, "Would you say your health in general is excellent, very good, good, fair, or poor?" Although individuals with different education levels may

assess their health differently in the frame of the 5-level self-rated health scale (Dowd and Zajacova 2007), "poor" self-rated health in particular is a reliable predictor of subsequent mortality across social classes (Burström and Fredlund 2001). To minimize the influence of subjective health rating across education levels and bolster the reliability of this measure, I collapse this 5-level Likert scale into two categories. Health status equal to "poor" or "fair" is coded as 1, while health status better than "fair" is coded as 0 (McGee et al. 1999). Individuals who answered "don't know" or refused to respond to the question were excluded from the analysis.

All statistical models in this chapter include gender, marital status, and race/ethnicity as covariates because these variables have critical effects on health (Annandale and Hunt 2000, Krieger et al. 2003, Umberson 1992). Missing data for independent and control variables ranges from 0% (gender) to 2.8% (education). After applying listwise deletion, the final sample includes 186,406 (18.9%) people with no high school degree, 288,328 (29.18%) people who have a high school, GED, or equivalent degree, 259,282 (26.24%) people who have some college or an associate's degree, and 254,200 (25.72%) people who have a bachelor's or higher degree. The data are stratified by education level and APC models are applied separately to each of those four stratified datasets.

*Statistical Analysis*

For each of the datasets stratified by educational level, I adopt the three-step approach to apply HAPC models. For the descriptive analysis at Step 1, I will graph the prevalence of "fair or poor" health conditions by age, period, and cohort group to acquire

an initial understanding of the data structure. At Step 2, I will fit seven nested model (i.e. A, P, C, AP, AC, PC, and APC) and compare the goodness-of-fit statistics (AIC and BIC) for gaining an understanding of the data structure and the best statistical model for the given data. Then, only for the cases where all three dimensions are operative, I will proceed with Step 3 of the analysis. At Step 3, I will employ the HAPC approach and specify cross-classified random effects models (CCREM) to estimate age, period, and cohort effects on having "fair or poor" health. If Step 2 suggests a reduced model rather than the full three-dimensional model, then I will estimate the reduced model instead. In all models, I include gender, race, and marital status as potential covariates (Centers for Disease Control 2008, Goldman, Korenman and Weinstein 1995, Jackson et al. 2006, Lillard and Panis 1996). Such a model specification may be written as follows:

$$\text{Level 1: } Y_{ijk} = \beta_{0jk} + \beta_1 \text{Age}_{ijk} + \beta_2 \text{Age}^2_{ijk} + \beta_3 \text{Female}_{ijk} + \beta_4 \text{Married}_{ijk} + \beta_5 \text{Black}_{ijk} + \beta_6 \text{Hispanic}_{ijk} + \beta_7 \text{Others}_{ijk} + e_{ijk}$$

$$\text{with } e_{ijk} \sim N(0, \sigma^2)$$

$$\text{Level 2: } \beta_{0jk} = \gamma_0 + u_{0j} + v_{0k}$$

$$\text{with } u_{0j} \sim N(0, \tau_u),\ v_{0k} \sim N(0, \tau_v),$$

where $Y_{ijk}$ represents whether the *ith* individual for $i = 1,\ldots,n_{jk}$ has "fair or poor" health status in the j-th period for $j=1,\ldots,J$ and the *kth* birth cohort for $k=1,\ldots,K$; $\beta_1$ and $\beta_2$ are fixed regression slopes for age and age-squared. Age is centered at the grand mean to reduce the association between age and age-squared terms (Reither et al. 2009); $\beta_3$, $\beta_4$, and $\beta_5$ are fixed regression slopes for gender, race, and marital status, respectively; $e_{ijk}$ is

the random individual effect that is normally distributed with mean 0 and variance $\sigma^2$. At level 2, where $\gamma_0$ is the grand mean (i.e., proportion) of the "fair or poor" health of all individuals, $u_{0j}$ is the period effect for period *j* averaged over all birth cohorts, which is assumed to follow a normal distribution with mean 0 and variance $\tau_u$, and $v_{0j}$ is the cohort effect for cohort k averaged over all periods, which is assumed to follow a normal distribution with mean 0 and variance $\tau_v$ (Yang and Land 2013).

## Results

### *Sample Characteristics*

Descriptive characteristics of the final sample are presented in Table 11. The percentage with health status equal or worse than "fair" is highest among people who have no high school degree. The percentages are lower for more educated groups, and the lowest for the most educated group. The percentage for the people who have a bachelor's or higher degree is as low as 5.56%.

Age has an inverse relationship with educational attainment. The mean age is lowest among those who have the highest educational attainment, and the highest among those who have the lowest educational attainment. Across the educational levels, the percentage of females was slightly higher in the sample than that of males. The majority of respondents in the sample are married, and the percentage of married people is higher among more educated people. In terms of race and ethnicity, whites are the majority in the total sample, while Hispanics are the majority in the lowest educated group. The combination of percentages for racial/ethnicity and educational attainment shows that more than 40% of Hispanics in this sample had no high school degree. For non-Hispanic

blacks and whites, the corresponding percentages are 21% and 12%, respectively. On the other hand, almost 60% of whites and 46.7% of non-Hispanic blacks in the sample had some college or a higher level of educational attainment.

*Descriptive Statistics*

To begin with the three-step procedure, I calculated the percentages of people who have fair or poor health status by age, period, and birth cohort for each educational group (Figure 19). First of all, it is obvious from the bar charts that the more educated people are, the healthier they are. Across all three temporal dimensions, people without a high school diploma have the highest prevalence of fair or poor health. As educational levels increase, the prevalence of fair or poor health declines across all three temporal dimensions. For people with a bachelor's or a graduate-level degree, the prevalence of fair or poor health is lower than other educational groups across all ages, periods of observation and birth cohorts.

The age-specific prevalence of fair/poor health changes at different rates across educational groups. For people who did not complete high school, the percentage reporting fair/poor health increases sharply with age, from below 10% during the late twenties to above 40% in the late fifties. Then it slightly drops during their sixties and stays around 40% thereafter. Compared to this pattern, the prevalence of fair/poor health among people with either a high school diploma or some college increases at a modest pace, reaching about 25% and 20% by the late fifties, respectively. The percentages slightly drop from their late fifties to their late sixties and increase again thereafter. The prevalence of fair/poor health among people at the highest educational level does not

show a similar bump during their late fifties. Instead, fair/poor health increases with age at a moderate pace until the late fifties and then rises more rapidly starting in the sixties. In this group, fair/poor health reaches a maximum of about 20% in the eighties. This prevalence of suboptimal health is similar to that observed among the no-high-school group in their early forties, the high-school-graduated group in their mid-fifties, and the some-college group in their sixties.

The percentages by periods are quite stable in all four educational groups. High-school-graduated and some-college groups show moderate increases across periods, although the total increase from 1997 to 2014 is only in the 3~5% range. The prevalence of fair/poor health over all periods of observation was around 30% among people who did not graduate high school, 13~18% among people who graduated high school, 10~12% among people with some college, and 5~6% among people with a bachelor's or higher degree. This indicates that the prevalence of fair/poor health is consistently 5-6 times higher across time periods for the lowest educational group compared to the highest educational group. In fact, simply graduating from high school cuts the prevalence of fair/poor health roughly in half.

The percentages across birth cohorts are essentially inverse patterns of the percentages by age. The earliest birth cohorts, which consist of elderly people in this sample, have the highest prevalence of fair/poor health, while the latest birth cohorts, which consist of the youngest people in the sample, have the lowest prevalence. The percentages among the oldest 7 cohorts in the no-high-school group stay as high as 40% and then dramatically drop for the 1947-1951 birth cohort. For the rest of the educational groups, the first couple of birth cohorts have the highest prevalence of fair/poor health,

followed by much lower percentages in subsequent birth cohorts. The decreasing pace is the most dramatic in the highest educational group.

From the descriptive plots, it is possible to make some initial educated guesses about the temporal structures of the four stratified datasets. First, for deciding if the period is an active dimension or not, the descriptive period plot is very useful because, as shown in Chapter 2, the descriptive period plot tended to reflect the actual period effects quite well. Therefore, my initial guess is that modest period effects exist for the high-school-graduated and some-college groups. For groups without a high school education or with at least a bachelor's degree, the period plots do not show any clear increasing or decreasing trend, so seems likely that period is either an inactive dimension or makes only a minor contribution to the outcome variable. However, at this point, it is premature to eliminate the possibility that period effects exist because they may be conflated with another temporal dimension, as shown in the D-set in Chapter 2. If period and another dimension have linear effects at the same time, the true period effects may be obscured in the descriptive plot. If this is the case, it may be found by model fit statistics, which, in such a situation, tends to point toward a reduced temporal structure due to confounding between those two dimensions.

Next, I surmised whether age and cohort dimensions are active by comparing their descriptive plots. As noted, the patterns of the plots for those two dimensions are close to mirror images of each other across all educational groups. This gives the impression that either age or cohort is dominant, and the other dimension is likely to have only a modest contribution to the outcome variable. Note that the weaker temporal dimension can still contribute independently to the outcome in some cases. Given the

information from the descriptive statistics, it is a reasonable guess that the high-school-graduated and some-college groups have a reduced temporal structure (i.e., AP or CP) or will be on the border between the reduced model and the three-dimensional model. The group that did not graduate from high school and the group with at least a bachelor's degree are likely to have a two-dimensional structure (i.e., AC) or even a single-dimensional structure (i.e., A) due to the very small period trend revealed in descriptive statistics. To make a clearer decision on the data structure, more information is required from the model fit statistics.

*Model Fit Statistics*

Model-fit statistics were estimated by AIC and BIC (Table 12). The AIC suggests the three-dimensional model for all four educational groups while the BIC suggests reduced models as a better fit. Because the sample sizes in all educational groups are larger than 50,000, which was the sample size used in the B, C, and D-sets in Chapter 2, the AIC's suggestion is more reliable than the BIC's. Based on the inferences obtained from the descriptive and model fit statistics, the groups having graduated high school and those having some college are likely to have structures in which either age or cohort has a dominant effect (the other making a minor but statistically significant contribution) and period effects increase modestly over this timeframe. Therefore, I move on to Step 3: application of HAPC models for these two groups.

For the groups that graduated high school or have a bachelor's degree or more, the models require careful validation. There are several temporal data structures that could yield nearly identical age and cohort plots. First, as in scenario D-6 from Chapter 2,

it is possible that period and cohort have linear effects in opposite directions and therefore offset one another. An interesting feature about such a scenario is that the model fit statistics also tend to point toward a reduced model. However, this is not the case here, because the model fit statistics suggest the three-dimensional model. Second, similar to scenarios B-3 and C-1 from Chapter 2, one or two dimensions have very little effect, pushing the data structure to the borderline of different dimensions. In this case, although those dimensions make only minor contributions, they are nevertheless detectible in some cases as active dimensions by model fit statistics. Just like the results of scenario B-3, the AIC can recognize the minor dimension as active with sufficient statistical power, and the three-dimensional HAPC model can successfully capture the true temporal effects.

Thus, in this case, it seems reasonable to estimate full HAPC models for groups without a high school degree and with at least a bachelor's degree. However, to confirm the validity of the estimates from HAPC models, I will cross-validate the results with estimates from reduced models. If the data structure is on the borderline between two and three dimensions, estimates from the reduced and three-dimensional models should not be substantially different. In this study, I estimate two additional models (i.e. A and AC) for these educational groups to cross-validate the results.

*Model Validation*

Figure 20 shows the results estimated by A, AC, and HAPC models for the least and the most educated groups. The estimates are very close. Although the A model does not have estimates for period and cohort effects and the AC model does not have

estimates for period effects, it does not strongly affect the conclusion since the cross-validation of models confirms that period and cohort have very little effect on the self-reported health of these two groups. Based on this confirmation, I will use the estimates of the HAPC for all four educational groups to understand the health gaps between them.

*HAPC Models*

Figure 20 represents the predicted probabilities of having fair/poor health by age, period, and cohort for the four educational groups, as estimated by HAPC models[2]. Overall, educational attainment has an inverse association with the predicted probabilities of having fair/poor health by age, period, and cohort. The more educated people are, the less likely their health condition is to be fair/poor. Graduating from high school makes a crucial difference for health, substantially reducing the probability of having fair/poor health across all three temporal dimensions. The probabilities of fair/poor health also drop slightly from the high-school-graduated group to the some-college group for all three temporal dimensions. From there, the probabilities decline once again for people who have bachelor's degrees or higher.

The predicted probabilities of fair/poor health increase with age (Figure 21-a), which is not surprising. However, the rate of increase differs by educational group. The least educated group experiences a very steep increase in the probability of fair/poor health relatively early in life (i.e., 30s~50s) and reaches the highest probability during their 70s. For those who graduated high school or have some college, the probabilities increase with age, but at a slower pace than the least educated group. The pace of health

[2] I also estimated the weighted HAPC models using survey weights provided by NHIS and compared the results to those of unweighted HAPC models. There are only minor differences between these two sets of estimates; therefore, in the manuscript I present only the results of unweighted models.

decline is even slower among the most educated group. The most distinctive characteristic of the most educated group is the modest increase in suboptimal health during their early years (i.e., 30s~50s) and the steeper increase at later ages (i.e., 60s~80s). People in this group do not experience a decline in poor health probabilities after their seventies, which likely means that selection by premature death is less prevalent in this group compared to the other educational groups.

Educational disparities in health clearly widen from early adulthood to retirement, before narrowing at later ages—again, likely due to mortality selection in the less-educated groups. Unfortunately, these results show that health begins to deteriorate relatively early in the life course among the least educated. As a result of these uneven rates of health deterioration, the health gap between the most and the least educated group increases steadily until the mid-60s. The health gaps between the most educated group and the high school or some college groups also reaches a maximum during the mid-60s, but these gaps are relatively narrow.

The predicted probabilities of reporting fair/poor health by period (Figure 21-b) for all four educational groups are quite stable over the survey years included in this study. None of the educational groups experienced dramatic improvement or deterioration in their health status. In other words, the health gaps between these educational groups have not changed in a meaningful way over the last 20 years.

Figure 21-c shows the trajectories for predicted probabilities of fair/poor health across birth cohorts. Health disparities between the three most highly educated groups are considerably less for cohorts born in the 1930s than cohorts born earlier in the 20$^{th}$ century. As the probabilities of fair/poor health for those who have a high school or

equivalent diploma or some college education increase after the 1940 birth cohort, the gaps between these groups and the least educated group narrow slightly. In the meantime, the gaps between the highest educated group and the other groups become wider as the probability of the highest educated group is quite stable across all birth cohorts. This result supports the argument that the upper class in the U.S. has been separating from the rest of society. Reeves (2015) points out that the advantages of education, income, and lifestyle for the upper class have become more concentrated in recent years. The widening gap between the highest educated group and the other groups shown in Figure 21-c indicates that the "pulling away" of the upper class from the other classes began to occur for cohorts born in the 1940s, and that this separation has included self-assessments of health.

**Discussion**

As discussed in the Results section, findings of this chapter support the Cumulative Advantage Theory that educational inequities in health widen over the life course (Aneshensel, Frerichs and Huba 1984, Ross and Wu 1995). The predicted probability of poor health changes much more dramatically by age than by period or birth cohort. In addition, the results show that change in educational health disparities over the life-course is a bounded process. The health gaps are greatest when people are in their mid-sixties, followed by a gradual decline. There are two plausible explanations for this partial health convergence in the latest stages of life: First, people who have poor health are disproportionately included in lower educational groups. Unfortunately, many of these individuals die relatively early in life, leaving behind a "healthy remnant" of people

in the lower educational groups at later life stages. As relatively healthy individuals who are disproportionately included in highly educated groups inevitably begin to face a biological ceiling after their seventies, the probability of poor health in these groups rapidly increases, leading to a narrowing of educational health disparities. Second, social programs such as Medicare and Social Security become operative after age 65, which may limit advantages in income and health care enjoyed by the most educated groups across most of the adult life course, thereby reducing health inequalities during the later stages of life (Willson, Shuey and Elder 2007).

When comparing the predicted probabilities of poor health across birth cohorts, the gaps between the highest educated group and the second/third highest educated groups declined for people who were born between 1917 and 1932, but then it began to widen after that. The health gap between highest and lowest educated groups did not significantly widen or narrow across birth cohorts. These findings suggest that the argument that recent cohorts tend to suffer more from educational inequality in terms of health outcomes (Lynch 2003) may be applicable to the health gaps between those with college degrees or more and those with high school degrees or some college education.

Previous studies show that the contents of education in the U.S. has changed tremendously over the last few decades, which could portend changes to the association between education and health. It is therefore plausible that high school graduates today possess more health knowledge than those with high school degrees from previous decades. Some scholars have argued that these significant renovations in the educational system are likely to be reflected in changing associations between education and health across birth cohorts (Manton, Stallard and Corder 1997). The distribution of education

has also changed over time, with the current period containing the largest number of people who have a college degree in our nation's history (Ryan and Bauman 2016).

Although recent cohorts are living in a society where the average education level is more elevated, and more knowledge about health is provided at the same level of education, the results of this study show that the health gap between highest and lowest educated groups has not significantly improved across birth cohorts. This finding provide supporting evidence for the idea of "academic inflation" which implies that elevated educational levels of a society does not necessarily mean better quality or amount of knowledge (Collins 2002). Rather, academic credentials seem to have become more critical for placing individuals in the social stratification system—and those credentials therefore play a major role in producing inequalities in occupation, income, and, furthermore, in health (Collins 2002, Hesseln and Jackson 2000).

In addition, my study findings show that educational disparities in health have not changed substantially over the period of observation spanning 1997 to 2014. Rather, the health gap between people who have bachelor's degree or more and people who have not experienced college has widened modestly. During the observational period for the data used in this study, the costs of health care have increased rapidly, which may have contributed to this slight widening in health disparities by education. However, at the same time, social programs such as the Affordable Care Act have been implemented and may have had countervailing effects on health inequalities, resulting in negligible period effects observed.

From a methodological perspective, this chapter showed how to apply HAPC models through a three-step process. To avoid the misuse of APC models, one should

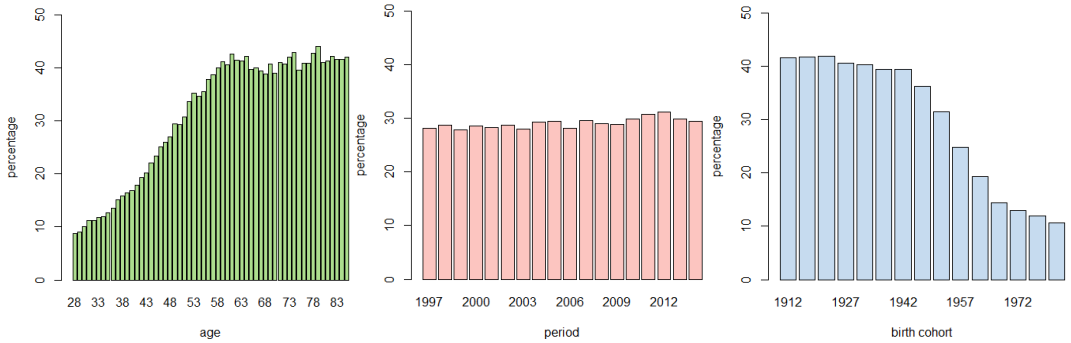
always start by understanding the relevant theories and data structures first, before mechanically applying the three dimensional model (Yang and Land 2013a). This preliminary step is essential for finding the best fitting model for the given data. It is also worth noting that the results of descriptive plots and model fit statistics should be carefully interpreted and weighed against each other. When those two procedures do not converge on a single best model, it is possible that the data structure is on the border between a reduced and the full, three-dimensional model. In this case, applying both models and cross-validating the results is another helpful method to avoid the misapplication of three-dimensional models.

## Tables and Figures

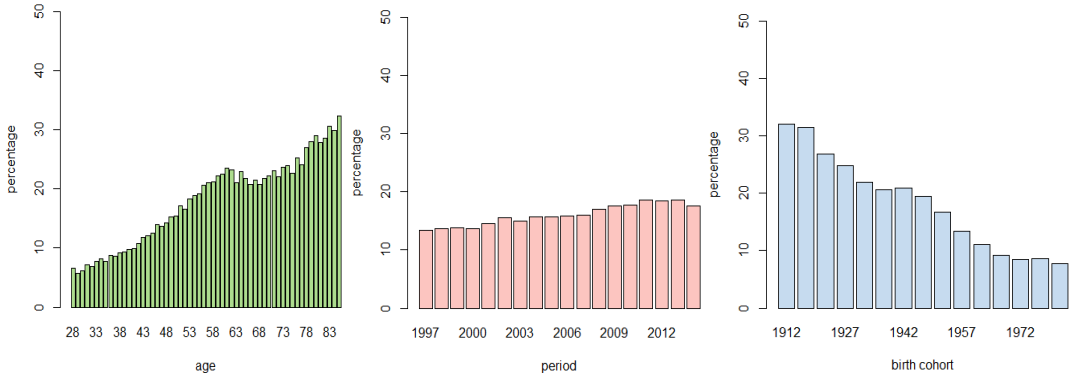
Table 11. Sample Characteristics

| | Total | NH | HS | SC | BC |
|---|---|---|---|---|---|
| **Health** | | | | | |
| Fair/Poor | 15.59 | 29.06 | 15.98 | 11.76 | 5.56 |
| Good/Very Good/Excellent | 84.41 | 70.94 | 84.02 | 88.24 | 94.44 |
| **Education** | | | | | |
| No high school | 18.86 | 100 | N/A | N/A | N/A |
| Graduated from high school | 29.18 | N/A | 100 | N/A | N/A |
| Some college | 26.24 | N/A | N/A | 100 | N/A |
| Bachelor's degree or more | 25.72 | N/A | N/A | N/A | 100 |
| **Age** | 50.61 | 53.77 | 51.45 | 48.76 | 48.45 |
| | (15.05) | (16.77) | (15.19) | (14.17) | (13.86) |
| **Gender** | | | | | |
| Male | 46.79 | 46.89 | 46.3 | 44.42 | 49.56 |
| Female | 53.21 | 53.11 | 53.7 | 55.58 | 50.44 |
| **Marital Status** | | | | | |
| Married | 63.81 | 57.09 | 62.76 | 63.49 | 71.9 |
| Unmarried | 36.19 | 42.91 | 37.24 | 36.51 | 28.1 |
| **Race/Ethnicity** | | | | | |
| White | 61.06 | 37.53 | 65.78 | 67.53 | 73.41 |
| Hispanic | 19.95 | 43.65 | 15.37 | 13.13 | 7.65 |
| NH Black | 13.17 | 14.55 | 14.53 | 14.67 | 8.92 |
| Others | 5.82 | 4.26 | 4.31 | 4.67 | 10.02 |
| N | 988,216 | 186,406 | 288,328 | 259,282 | 254,200 |

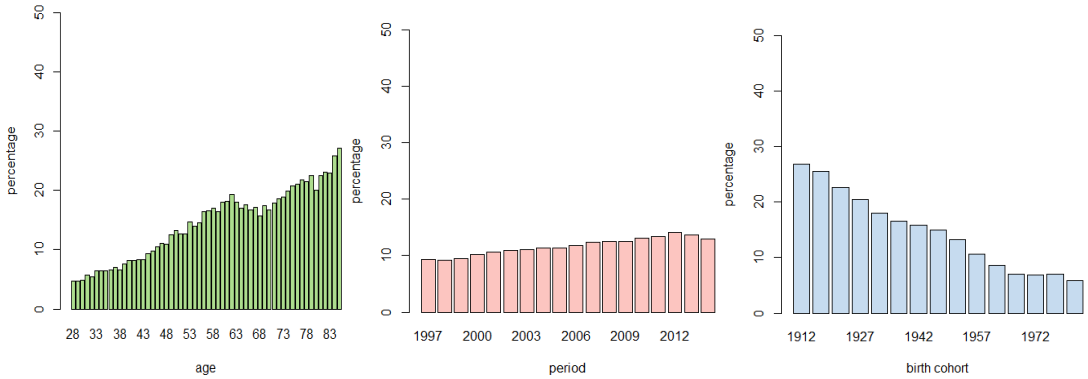
(a) No High School (NH)




(b) Graduated from High School (HS)




(c) Some College (SC)




(d) Bachelor's Degree or More (BC)

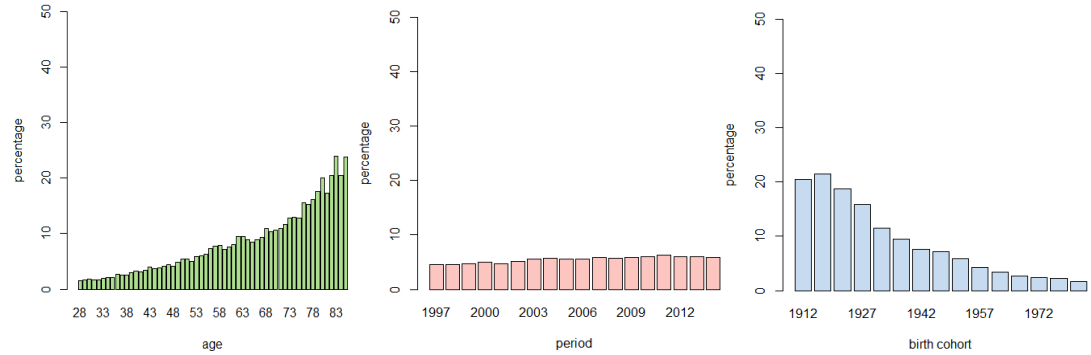
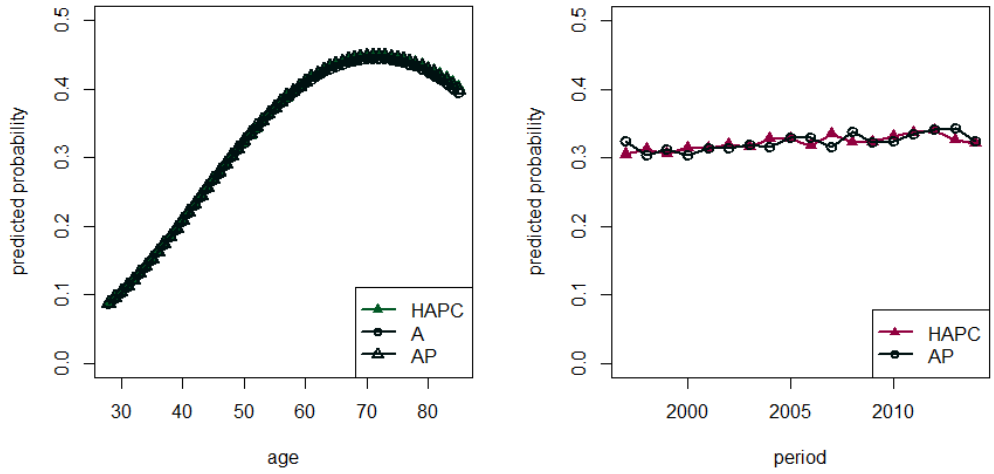



Figure 19. Descriptive Statistics – Percentage of Having Poor/Fair Health for Four Education Groups by Age, Period, and Cohort

Table 12. Goodness-of-Fit Statistics for Four Educational Groups

| | NH | HS | SC | BC |
|---|---|---|---|---|
| N | 186,406 | 288,328 | 259,282 | 254,200 |
| Data Structure designated by AIC | APC | APC | APC | APC |
| Data Structure designated by BIC | AC | AC | AC | A |

(a) No High School




(b) Bachelor's Degree or More

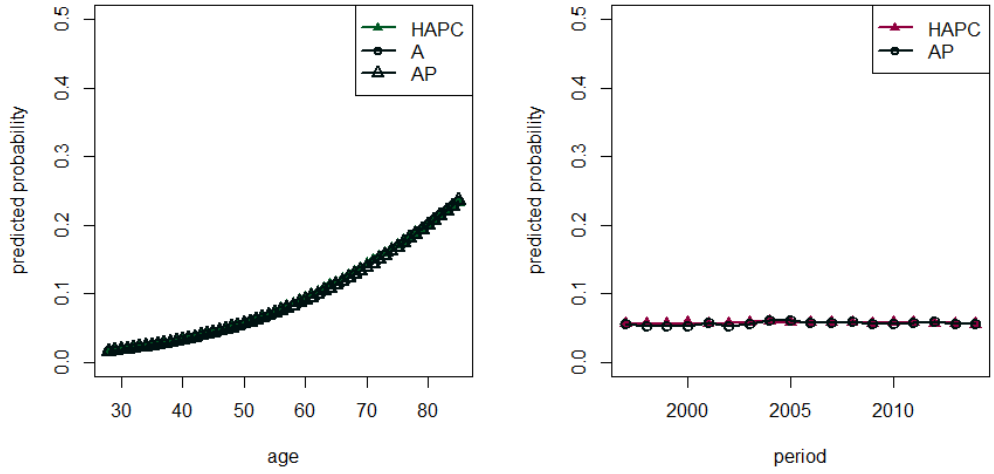
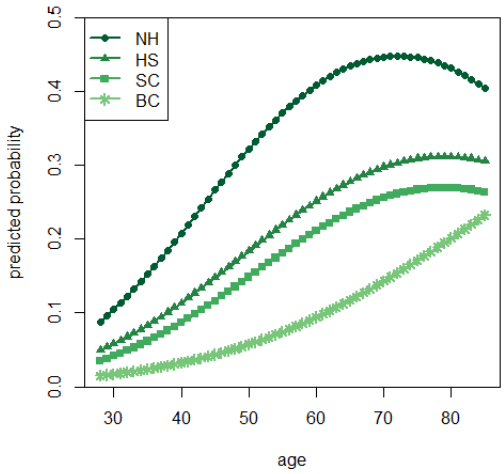
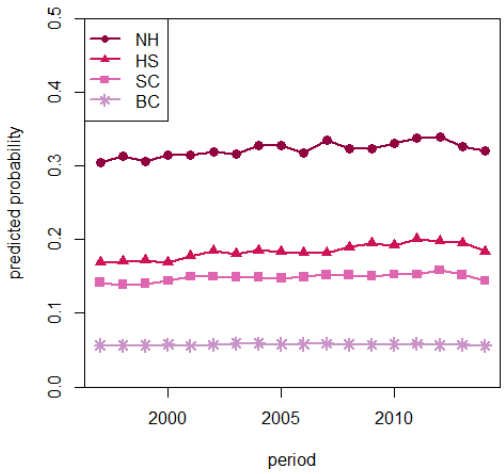



Figure 20. Comparing Estimates of A, AP, and APC Models for the Least and Most Educated Groups

(a) Age




(b) Period




(c) Cohort

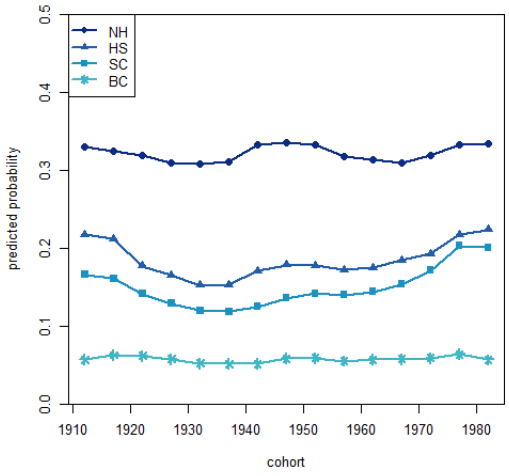



Figure 21. Predicted Probability of Having Fair/Poor Health Estimated by HAPC Models

# CHAPTER V

# CONCLUSION

To develop a statistical model, sometimes it is inevitable to make tentative assumptions first and derive the model from there. Once the model is built upon the assumptions, the results of the model are conditional upon the assumptions being satisfied. Therefore, during an empirical application of such a statistical model, complete and accurate specification of the assumptions is imperative (Poole and O'Farrell 1971). Typically, it can be done by applying some preliminary statistical analysis to better understand the given data and variables of interest. Without such procedures, a researcher should not expect that the results approximate what really happens. Furthermore, bias and inconsistency emanating from statistical models that violate key assumptions cannot be used as evidence to prove that a statistical method is invalid.

The developers of the HAPC and the IE approaches to APC modeling have emphasized that those techniques are designed for data with three active temporal dimensions. However, in their critiques on the HAPC and the IE, Bell and Jones as well as Luo did not pay much attention to the assumptions of the models that they were examining. Instead, these critics derived data structures by using their own assumptions, which often posed clear violations of APC methods. For example, they arbitrarily created data structures that have exact linear trends in two or more dimensions (Luo 2013a) or that derive the pattern of one temporal dimension from the pattern of another (Bell and Jones 2015). As a result of these clear violations, they “discovered” that HAPC and IE

models tend to fail. However, since the data structures generated in these studies do not satisfy the assumptions of HAPC and IE models, the authors cannot expect the models to accurately estimate the true age, period and cohort coefficients. Therefore, results from these studies should not be used as evidence that the models generally fail. Rather, the results should be used to highlight the importance of checking that the data satisfy the assumptions of HAPC and IE models prior to the application of these methods.

Testing APC model assumptions requires a robust set of procedures, but there has been a debate on the validity of such techniques. Yang and Land (2013) suggested the three-step application procedure, which uses descriptive (Step 1) and model fit statistics (Step 2) to understand the data structure and check model assumptions. However, critics (Bell and Jones 2014b, Bell and Jones 2015, Luo 2013a) of APC models have argued that no statistical technique could be useful for checking the assumptions of APC models because the identification problem makes statistical techniques ineffective for examining the real data structure and the model assumptions. In Chapter 2 and 3 of this dissertation, I attempted to examine the validity of descriptive and model-fit statistics when used to check the model assumptions, as no study so far has specifically focused on the preliminary steps of APC analysis. In particular, different from the arbitrary (i.e. "the cohort trends based on the period trend", Bell and Jones 2015, p.332) and unrealistic (i.e. "age, period, and cohort each have effects on the outcome variable that show an (exact) linear trend", Luo 2013, p.1953) data structures generated by critics, I attempted to simulate realistic data by borrowing patterns from the effects of empirical studies. The findings showed that graphing and model selection statistics were useful in identifying the temporal data structures prior to application of HAPC and IE models.

In Chapter 4, I used the lessons from the previous two chapters to estimate HAPC models with actual data. From the descriptive and model-fit statistics, I ascertained that the four stratified data sets had temporal structures that bordered between reduced and full three-dimensional models, and identified a handful of plausible matches. By adding cross-validation to my analysis, I demonstrated that it is possible to utilize APC methods even when the data structures and corresponding "best models" are ambiguous. The substantive results from Chapter 4 contribute to a better understanding of the disentangled effects of age, period, and cohort on educational inequalities in health. These health gaps between different educational groups increase steadily with age until people are in their 60s, at which time they begin to narrow. Neither secular changes over the past two decades nor differences across birth cohorts appear to play a major role in shaping self-rated health in the United States.

While the findings in this dissertation show that the three-step procedure is a valid approach, I also found some room for improvement when choosing the statistical tools to check model assumptions. More studies are needed to shed further light on the interpretation of descriptive and model-fit statistics, as the results are not always clear enough to adjudicate between different possible temporal data structures. For example, when interpreting the descriptive plots of age, period, and cohort, there is neither a numerical criterion of curvature of slopes that signifies an active dimension, nor is there a way to identify sufficient divergence between age and cohort plots. Although the descriptive plots did provide a helpful preliminary overview of the data structure—particularly for age and period effects—it is nevertheless challenging to determine which model to use when data structures border between two different dimensions.

In addition, we may need to develop more specific guidelines on how to use AIC and BIC when they are not consistent with each other. Previous literature (Yang and Land 2013a) on APC modeling indicates that "we can use AIC or BIC for understanding the data structure," but no guidelines are given for prioritizing AIC or BIC when different models are suggested. In this dissertation, the sample sizes were quite large, so I relied more on AIC than BIC. However, when the condition is not so obvious, the inconsistency between the fit statistics can cause uncertainty about the selection of the best model. Future studies may want to investigate which criterion is more reliable for different sample sizes, number of parameters, and other variations in the model and data.

Furthermore, studies on the AIC and BIC reveal that those two statistics are actually based on very different philosophies, and, therefore, the "best models" designated by AIC and BIC are best in different ways (Burnham and Anderson 2004, Kuha 2004). The AIC is built on Kullback-Leibler's concept of information loss, while BIC is based on Bayes factors (Burnham and Anderson 2004, Kuha 2004). When the two criteria suggest different models for any given APC data set, it may be necessary to confirm which one suggests the best model for the specific case study—i.e., one that is conceptually close to the data structure being investigated.

Despite these methodological limitations, it is worth reiterating that those two new approaches are a great methodological advancement in age-period-cohort analyses. When Norman Ryder (1965) initially came up with the concept of cohort, it was not accompanied by a methodological approach to empirically estimate cohort effects. As a result of methodologists' efforts for 10 years thereafter, Mason et al. (1973) proposed a method to estimate APC models by imposing an equality constraint, which can be

derived from external theories. Since then scholars have made even more progress in devising more robust statistical methods. After 30 years of debate and methodological innovation, it is now possible to estimate the APC effects for some data sets without any external information.

The main message that I would like to deliver in this dissertation is that while critiques of APC analyses (or any statistical method) can have value, discrepancies between our expectation or hypothesis and the results do not imply that the approach should be discarded or that the efforts to develop the idea is a "futile" quest. As Box (1976) highlights in his famous article *Science and Statistics*, "the good scientist must have the flexibility and courage to seek out, recognize, and exploit such errors. (p. 791)" I am sure that future developments in APC modeling will stem from productive exchanges between relevant critiques and rejoinders, not deliberate efforts to play with model assumptions in such a way that essentially guarantees the "discovery" that APC models fail under such conditions.

# REFERENCES

Adler, Nancy E., Thomas Boyce, Margaret A. Chesney, Sheldon Cohen, Susan Folkman, Robert L. Kahn and S. Leonard Syme. 1994. "Socioeconomic Status and Health: The Challenge of the Gradient." *American Psychologist* 49(1):15.

Aneshensel, Carol S., Ralph R. Frerichs and George J. Huba. 1984. "Depression and Physical Illness: A Multiwave, Nonrecursive Causal Model." *Journal of Health and Social Behavior*:350–71.

Annandale, Ellen and Kate Hunt. 2000. "Gender Inequalities in Health."

Becker, Gary S. 1994. "Human Capital Revisited." Pp. 15–28 in *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education (3rd Edition)*: The University of Chicago Press.

Beckett, Megan. 2000. "Converging Health Inequalities in Later Life-an Artifact of Mortality Selection?". *Journal of Health and Social Behavior*:106–19.

Bell, Andrew and Kelvyn Jones. 2013. "The Impossibility of Separating Age, Period and Cohort Effects." *Social Science & Medicine* 93:163–65.

Bell, Andrew and Kelvyn Jones. 2014a. "Another'futile Quest'? A Simulation Study of Yang and Land's Hierarchical Age-Period-Cohort Model." *Demographic Research* 30:333.

Bell, Andrew and Kelvyn Jones. 2014b. "Don't Birth Cohorts Matter? A Commentary and Simulation Exercise on Reither, Hauser, and Yang's (2009) Age–Period–Cohort Study of Obesity." *Social Science & Medicine* 101:176–80.

Bell, Andrew and Kelvyn Jones. 2015. "Should Age-Period-Cohort Analysts Accept Innovation without Scrutiny? A Response to Reither, Masters, Yang, Powers, Zheng and Land." *Social Science & Medicine* 128:331–33.

Box, George E. P. 1976. "Science and Statistics." *Journal of the American Statistical Association* 71(356):791–99.

Burnham, Kenneth P. and David R. Anderson. 2004. "Multimodel Inference Understanding Aic and Bic in Model Selection." *Sociological Methods & Research* 33(2):261–304.

Burström, Bo and Peeter Fredlund. 2001. "Self Rated Health: Is It as Good a Predictor of Subsequent Mortality among Adults in Lower as Well as in Higher Social Classes?". *Journal of Epidemiology and Community Health* 55(11):836–40.

Centers for Disease Control. 2008. "Racial/Ethnic Disparities in Self-Rated Health Status among Adults with and without Disabilities--United States, 2004-2006." *MMWR. Morbidity and mortality weekly report* 57(39):1069.

Collins, Randall. 1979. *The Credential Society: An Historical Sociology of Education and Stratification*: Academic Pr.

Collins, Randall. 2002. "Credential Inflation and the Future of Universities." *The Future of the City of Intellect: The Changing American University*:23–46.

Conti, Gabriella and James J. Heckman. 2010. "Understanding the Early Origins of the Education–Health Gradient: A Framework That Can Also Be Applied to Analyze Gene–Environment Interactions." *Perspectives on Psychological Science* 5(5):585–605.

Cutler, David M. and Adriana Lleras-Muney. 2006. "Education and Health: Evaluating Theories and Evidence." Vol.: National Bureau of Economic Research.

Dassonneville, Ruth. 2013. "Questioning Generational Replacement. An Age, Period and Cohort Analysis of Electoral Volatility in the Netherlands, 1971–2010." *Electoral Studies* 32(1):37–47.

DiPrete, Thomas A. and Gregory M. Eirich. 2006. "Cumulative Advantage as a Mechanism for Inequality: A Review of Theoretical and Empirical Developments." *Annual Review of Sociology* 32:271–97.

Dowd, Jennifer Beam and Anna Zajacova. 2007. "Does the Predictive Power of Self-Rated Health for Subsequent Mortality Risk Vary by Socioeconomic Status in the Us?". *International Journal of Epidemiology* 36(6):1214–21.

Dunteman, George H. 1989. "Principal Component Analysis. Quantitative Applications in the Social Sciences Series (Vol. 69)." Thousand Oaks, CA: Sage Publications.

Dupre, Matthew E. 2008. "Educational Differences in Health Risks and Illness over the Life Course: A Test of Cumulative Disadvantage Theory." *Social Science Research* 37(4):1253–66.

Elo, Irma T. and Samuel H. Preston. 1996. "Educational Differentials in Mortality: United States, 1979–1985." *Social Science & Medicine* 42(1):47–57.

Feinstein, Jonathan S. 1993. "The Relationship between Socioeconomic Status and Health: A Review of the Literature." *The Milbank Quarterly*:279–322.

Feldman, Jacob J., Diane M. Makuc, Joel C. Kleinman and Joan Cornoni-Huntley. 1989. "National Trends in Educational Differentials in Mortality." *American Journal of Epidemiology* 129(5):919–33.

Ferraro, Kenneth F. and Jessica A. Kelley-Moore. 2003. "Cumulative Disadvantage and Health: Long-Term Consequences of Obesity?". *American Sociological Review* 68(5):707.

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

Fries, James F. 1983. "The Compression of Morbidity." *The Milbank Memorial Fund Quarterly. Health and Society*:397–419.

Fries, James F. 2002. "Aging, Natural Death, and the Compression of Morbidity." *Bulletin of the World Health Organization* 80(3):245–50.

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):263–78.

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):453–66.

Goldman, Noreen, Sanders Korenman and Rachel Weinstein. 1995. "Marital Status and Health among the Elderly." *Social Science & Medicine* 40(12):1717–30.

Held, Leonhard and Andrea Riebler. 2013. "Comment on “Assessing Validity and Application Scope of the Intrinsic Estimator Approach to the Age-Period-Cohort (Apc) Problem”." *Demography* 50(6):1977–79.

Hesseln, Hayley and Dave Jackson. 2000. "Academic Inflation: The Devaluation of a University Degree." Pp. 112 in *Biennial Conference on University Education in Natural Resources*: Citeseer.

House, James S., James M. Lepkowski, Ann M. Kinney, Richard P. Mero, Ronald C. Kessler and A. Regula Herzog. 1994. "The Social Stratification of Aging and Health." *Journal of Health and Social Behavior*:213–34.

Jackson, Pamela Braboy, David R. Williams, A Shulz and Leith Mullings. 2006. "The Intersection of Race, Gender, and Ses: Health Paradoxes." *Gender, Race, Class and Health*.

Jeon, Sun Y., Eric N. Reither and Ryan K. Masters. 2016. "A Population-Based Analysis of Increasing Rates of Suicide Mortality in Japan and South Korea, 1985–2010." *BMC Public Health* 16(1):356.

Kannisto, Vaino, Jens Lauritsen, A. Roger Thatcher and James W. Vaupel. 1994. "Reductions in Mortality at Advanced Ages: Several Decades of Evidence from 27 Countries." *Population and Development Review*:793–810.

Kaplan, George A., Mary N. Haan, S. Leonard Syme, Meredith Minkler and Marilyn Winkleby. 1987. "Socioeconomic Status and Health." *American Journal of Preventive Medicine* 3(suppl 1):125–29.

Kitagawa, Evelyn M. and Philip M. Hauser. 1973. "Differential Mortality in the United States: A Study in Socioeconomic Epidemiology."

Kramarow, E, P Pastor and Y Gorina. 2000. "Educational Differentials in Mortality among Older Us Adults." in *Annual Meetings of the Population Association of America, March, Los Angeles*.

Krieger, Nancy, Jarvis T. Chen, Pamela D. Waterman, David H. Rehkopf and S. V. Subramanian. 2003. "Race/Ethnicity, Gender, and Monitoring Socioeconomic Gradients in Health: A Comparison of Area-Based Socioeconomic Measures—the Public Health Disparities Geocoding Project." *American Journal of Public Health* 93(10):1655–71.

Kuha, Jouni. 2004. "Aic and Bic: Comparisons of Assumptions and Performance." *Sociological Methods & Research* 33(2):188–229.

Kupper, Lawrence L., Joseph M. Janis, Ibrahim A. Salama, Carl N. Yoshizawa, Bernard G. Greenberg and H. H. Winsborough. 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):201–17.

Kupper, Lawrence L., Joseph M. Janis, Azza Karmous and Bernard G. Greenberg. 1985. "Statistical Age-Period-Cohort Analysis: A Review and Critique." *Journal of Chronic Diseases* 38(10):811–30.

Land, Kenneth C., Yang Yang and Yi Zeng. 2005. "Mathematical Demography." Pp. 659–717 in *Handbook of Population*: Springer.

Land, Kenneth C., Emma Zang, Qiang Fu, Xin Guo, Sun Y. Jeon and Eric N. Reither. 2016. "Playing with the Rules and Making Misleading Statements: A Response to Luo, Hodges, Winship, and Powers 1." *American Journal of Sociology* 122(3):962–73.

Lauderdale, Diane S. 2001. "Education and Survival: Birth Cohort, Period, and Age Effects." *Demography* 38(4):551–61.

Lillard, Lee A. and Constantijn W. A. Panis. 1996. "Marital Status and Mortality: The Role of Health." *Demography* 33(3):313–27.

Luo, Liying. 2013a. "Assessing Validity and Application Scope of the Intrinsic Estimator Approach to the Age-Period-Cohort Problem." *Demography* 50(6):1945–67.

Luo, Liying. 2013b. "Paradigm Shift in Age-Period-Cohort Analysis: A Response to Yang and Land, O'brien, Held and Riebler, and Fienberg." *Demography* 50(6):1985–88.

Luo, Liying, James S. Hodges, Christopher Winship and Daniel Powers. 2014. "The Sensitivity of the Intrinsic Estimator to Coding Schemes: A Comment on Yang, Schulhofer-Wohl, Fu, and Land." *American Journal of Sociology*:2014–1.

Lynch, Scott M. 2003. "Cohort and Life-Course Patterns in the Relationship between Education and Health: A Hierarchical Approach." *Demography* 40(2):309–31.

Manton, Kenneth G. 1982. "Changing Concepts of Morbidity and Mortality in the Elderly Population." *The Milbank Memorial Fund Quarterly. Health and Society*:183–244.

Manton, Kenneth G., Eric Stallard and Larry Corder. 1997. "Education-Specific Estimates of Life Expectancy and Age-Specific Disability in the Us Elderly Population 1982 to 1991." *Journal of Aging and Health* 9(4):419–50.

Mason, Karen Oppenheim, William M. Mason, Halliman H. Winsborough and W. Kenneth Poole. 1973. "Some Methodological Issues in Cohort Analysis of Archival Data." *American Sociological Review*:242–58.

Mason, William M. and Stephen E. Fienberg. 1985. "Introduction: Beyond the Identification Problem." Pp. 1–8 in *Cohort Analysis in Social Research*: Springer.

McEwen, Bruce S. 1998. "Stress, Adaptation, and Disease: Allostasis and Allostatic Load." *Annals of the New York Academy of Sciences* 840(1):33–44.

McEwen, Bruce S. and Teresa Seeman. 1999. "Protective and Damaging Effects of Mediators of Stress: Elaborating and Testing the Concepts of Allostasis and Allostatic Load." *Annals of the New York Academy of Sciences* 896(1):30–47.

McGee, Daniel L., Youlian Liao, Guichan Cao and Richard S. Cooper. 1999. "Self-Reported Health Status and Mortality in a Multiethnic Us Cohort." *American Journal of Epidemiology* 149(1):41–46.

Merton, Robert King. 1968. *Social Theory and Social Structure*: Simon and Schuster.

Mirowsky, John and Catherine E. Ross. 2005. "Education, Cumulative Advantage, and Health." *Ageing International* 30(1):27–62.

Moore, David E. and Mark D. Hayward. 1990. "Occupational Careers and Mortality of Elderly Men." *Demography* 27(1):31–53.

O'Brien, Robert M. 2013. "Comment of Liying Luo's Article,"Assessing Validity and Application Scope of the Intrinsic Estimator Approach to the Age-Period-Cohort Problem"." *Demography* 50(6):1973–75.

Piontek, Daniela, Ludwig Kraus, Alexander Pabst and Stéphane Legleye. 2012. "An Age–Period–Cohort Analysis of Cannabis Use Prevalence and Frequency in Germany, 1990–2009." *Journal of Epidemiology and Community Health* 66(10):908–13.

Poole, Michael A. and Patrick N. O'Farrell. 1971. "The Assumptions of the Linear Regression Model." *Transactions of the Institute of British Geographers*:145–58.

Preston, Samuel, Patrick Heuveline and Michel Guillot. 2000. "Demography: Measuring and Modeling Population Processes."

Reeves, Richard V. 2015. "The Dangerous Separation of the American Upper Middle Class." Vol. *Social Mobility Papers.* Brookings.

Reither, Eric N., Robert M. Hauser and Yang Yang. 2009. "Do Birth Cohorts Matter? Age-Period-Cohort Analyses of the Obesity Epidemic in the United States." *Social Science & Medicine* 69(10):1439–48.

Reither, Eric N., Kenneth C. Land, Sun Y. Jeon, Daniel A. Powers, Ryan K. Masters, Hui Zheng, Melissa A. Hardy, Katherine M. Keyes, Qiang Fu and Heidi A. Hanson. 2015a. "Clarifying Hierarchical Age–Period–Cohort Models: A Rejoinder to Bell and Jones." *Social Science & Medicine* 145:125–28.

Reither, Eric N., Ryan K. Masters, Yang Claire Yang, Daniel A. Powers, Hui Zheng and Kenneth C. Land. 2015b. "Should Age-Period-Cohort Studies Return to the Methodologies of the 1970s?". *Social Science & Medicine* 128:356–65.

Ross, Catherine E. and Chia-ling Wu. 1995. "The Links between Education and Health." *American Sociological Review*:719–45.

Ross, Catherine E. and Chia-Ling Wu. 1996. "Education, Age, and the Cumulative Advantage in Health." *Journal of Health and Social Behavior*:104–20.

Ryan, Camille L. and Kurt Bauman. 2016. "Educational Attainment in the United States: 2015." *Current Population Reports* 20.

Ryder, Norman B. 1965. "The Cohort as a Concept in the Study of Social Change." *American Sociological Review*:843–61.

Schwadel, Philip. 2010. "Period and Cohort Effects on Religious Nonaffiliation and Religious Disaffiliation: A Research Note." *Journal for the Scientific Study of Religion* 49(2):311–19.

Shavers, Vickie L. 2007. "Measurement of Socioeconomic Status in Health Disparities Research." *Journal of the National Medical Association* 99(9):1013.

Umberson, Debra. 1992. "Gender, Marital Status and the Social Control of Health Behavior." *Social Science & Medicine* 34(8):907–17.

Willson, Andrea E., Kim M. Shuey and Jr Elder, Glen H. 2007. "Cumulative Advantage Processes as Mechanisms of Inequality in Life Course Health 1." *American Journal of Sociology* 112(6):1886–924.

Winkleby, Marilyn A., Darius E. Jatulis, Erica Frank and Stephen P. Fortmann. 1992. "Socioeconomic Status and Health: How Education, Income, and Occupation Contribute to Risk Factors for Cardiovascular Disease." *American Journal of Public Health* 82(6):816–20.

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):75–110.

Yang, Yang and Kenneth C. Land. 2006. "A Mixed Models Approach to the Age-Period-Cohort Analysis of Repeated Cross-Section Surveys, with an Application to Data on Trends in Verbal Test Scores." *Sociological Methodology* 36(1):75–97.

Yang, Yang. 2008a. "Trends in Us Adult Chronic Disease Mortality, 1960–1999: Age, Period, and Cohort Variations." *Demography* 45(2):387–416.

Yang, Yang. 2008b. "Social Inequalities in Happiness in the United States, 1972 to 2004: An Age-Period-Cohort Analysis." *American Sociological Review* 73(2):204–26.

Yang, Yang and Kenneth C. Land. 2008. "Age–Period–Cohort Analysis of Repeated Cross-Section Surveys: Fixed or Random Effects?". *Sociological Methods & Research* 36(3):297–326.

Yang, Yang and Kenneth C. Land. 2013a. "Age-Period-Cohort Analysis." *Chapman & Hall/CRC Interdisciplinary Statistics Series. doi* 10(1201):b13902.

Yang, Yang Claire and Kenneth C. Land. 2013b. "Misunderstandings, Mischaracterizations, and the Problematic Choice of a Specific Instance in Which the Ie Should Never Be Applied." *Demography* 50(6):1969–71.

# CURRICULUM VITAE


## Sun Y. Jeon

CONTACT INFORMATION

0730 Old Main Hill
Logan, UT, 84322, USA

*Phone:* (435) 764-8819
*E-mail:* s.jeon@aggiemail.usu.edu


EDUCATION

**Ph.D., Sociology (Demography)**, December 2016, **GPA: 3.89**

**Utah State University**, Logan, UT USA

- Dissertation Title: "Do Data Structures Matter? A Simulation Study for Testing Validity of Age-Period-Cohort Models"

**M.S., Statistics**, December 2015, **GPA: 4.0**

**Utah State University**, Logan, UT USA

- Thesis Title: "A Comparison of Random Forest-based Methods for Racial/Ethnic-Specific Classification of Obesity"

**M.S., Sociology (Demography)**, May 2012

**Utah State University**, Logan, UT USA

- Thesis Title: "Demographic Evaluation of Increasing Rates of Suicide Mortality in Japan and South Korea"

**B.S., Chemical & Biomolecular Engineering**, August 2009

**Sogang University**, Seoul, South Korea

SKILLS

Statistical/Mathematical Packages: R, SAS, Stata, SPSS, Maple
Languages: SQL (mySQL, proc sql), basic python
Applications: LaTeX, MS office, and presentation software

CERTIFICATION

**SAS Base Programming for SAS 9 (A00-211)** **August, 2015 - no exp.**

(Certificate No: BP051803v9)

RESEARCH EXPERIENCE

**Presidential Doctoral Research Fellow**, Utah State University **2012-present**

- Analyzed National Health Interview Survey (NHIS) data using hierarchical age-period-cohort modeling to estimate the health gap between different educational groups throughout their life courses.
- Cleaned and merged National Longitudinal Survey of Youth (NLSY), and applied fixed-effect model to estimate the effects of work status on health outcomes.
- Applied random forest classification to National Health and Nutrition Examination Survey (NHANES) data, develop racial/ethnic-specific classification of obesity, and improved the classification accuracy compared to the current BMI-based method.
- Conducted simulation analysis using Stata and R as a lead statistician in study to prove the validity of hierarchical age-period cohort model.
- Analyzed National Latino and Asian American Study (NLAAS) data using principal component analysis to study the influence of social/familiar network on mental health among Asian and Latino immigrants.
- Applied canonical correlation and random forest classification to NHANES data to investigate the validity of biomarkers used for studies on Allostatic Load.

(*See the publication and conference presentation sections for outcomes of the projects above*)

RESEARCH EXPERIENCE (CONTD.)

**Graduate Research Assistant**, Utah State University **2010-2012**

- Analyzed vital statistics and population data to investigate increasing suicide mortality rates in Korea and Japan.
- Assisted a research project on developing a three-dimensional mortality projection method using age-period-cohort modeling, time-series analysis and mortality data from the National Vital Statistics System (NVSS).

TEACHING EXPERIENCE

**Instructor**, Utah State University

Developed curriculum in all areas including instruction, grading, preparing tests (quizzes, midterms and finals), holding office hours, and assigning final grades

- SOC 3120 "Social Statistics (using SPSS)" : Summer 2016

**Teaching Assistant**, Utah State University

Tutoring Stata, holding office hour, and grading.

- SOC 7110 "Advanced Sociological Analysis" : Spring 2015
- SOC 3120 "Social Statistics" : Spring 2015

**Teaching Training**, Utah State University

- ITA (International Teaching Assistant) Training : Summer 2015
- SOC 6800 "Strategies for Teaching" : Fall 2012

HONORS AND AWARDS

**Fellowships and Scholarships**

- Presidential Doctoral Research Fellowship, Utah State University, 2012–Current
- Young-Chul Hong Endowment, Utah State University, 2011
- Yun & Wendy Kim Scholarship, Utah State University, 2010
- Recommendation Scholarship, Sogang University, 2008
- Academic Excellence Scholarship, Sogang University, 2004, 2006

**Awards**

- Poster session winner."A Demographic Evaluation of Increasing Rates of Suicide Mortality in Japan and South Korea, Population Association of America Annual Meeting, New Orleans, LA 2013

EXTERNAL EDUCATION

**ICPSR Summer Program**

*Applied Multilevel Models for Longitudinal Data* **July, 2013**
University of Colorado at Boulder, Boulder, CO

*Categorical Data Analysis* **August, 2012**
University of Michigan, Ann Arbor, MI

*Structural Equation Models with Latent Variables* **August, 2012**
University of Michigan, Ann Arbor, MI

**Online Training**

*Mathematical Biostatistics Boot Camp 1* **May, 2016**
Coursera
(*Completion certificate license number: JUWK3PCTWSQE*)

*Programming for Everybody (Python)* **August, 2015**
Coursera
(*Completion certificate license number: 9L4LAV8Lv3*)

*SQL for Newbs: Beginner Data Analysis* **July, 2015**
Udemy
(*Completion certificate: https://www.udemy.com/certificate/UC-SJJLLN9F/*)

## Publications

Kenneth C. Land, Qiang Fu, Xin Guo, **Sun Y. Jeon**, Eric N. Reither, and Emma Zang. *Forthcoming.* "Playing with the Rules and Making Misleading Statements: A Reponse to Luo, Hodges, Winship, and Powers", *American Journal of Sociology*

**Sun Y. Jeon**, Eric N. Reither, and Ryan K. Masters. 2016. "A Population-based Evaluation of Increasing Rates of Suicide Mortality in Japan and South Korea", *BMC Public Health*

Eric N. Reither, Kenneth C. Land, **Sun Y. Jeon**, Daniel A. Powers, Ryan K. Masters, Hui Zheng, Mellisa A. Hardy, Katherine M. Keyes, Qiang Fu, Heidi A. Hanson, Ken R. Smith, Rebecca L. Utz, and Y. Clair Yang. 2015. "Clarifying Hierarchical Age-Period-Cohort Models: A Rejoinder to Bell and Jones.", *Social Science and Medicine*

Andrew E. Burger, Eric N. Reither, David Ramos, and **Sun Y. Jeon**. 2011. "Racial and Ethnic Disparities in Seasonal Influenza Vaccination among Utah Adults, 2000-2008" *Utah's Health: An Annual Review*

## Conference Presentations

So-Jung Lim, Joongbaeck Kim, Hyeyoung Woo, and **Sun Y. Jeon**. "Relationship Between Non-standard Employment and Health in South Korea: The Role of Gender and Marital Status", Oral Presentation, Society for the Study of Social Problems Annual Meeting, Seattle WA, 2016

**Sun Y. Jeon**. "Developing Racial/Ethnic Specific Classification of Obesity for White, Hispanic, and Black Males in the U.S.", Oral Presentation, Population Association of America Annual Meeting, Washington DC, 2016

So-Jung Lim, Joongbaeck Kim, Hyeyoung Woo, and **Sun Y. Jeon**. "Relationship Between Non-standard Employment and Health in South Korea: The Role of Gender and Marital Status", Poster Presentation, Population Association of America Annual Meeting, Washington DC, 2016

Eric N. Reither, Patrick M. Krueger, Paul E. Peppard, **Sun Y. Jeon**, and Lauren Hale."Sleep, Obesity and the Physical and Psychosocial Wellbeing of Young Adults in the U.S.", Poster Presentation, SLEEP, Seattle, WA 2015

**Sun Y. Jeon** and Eric N. Reither. "Application of a Classification Method for Studies of Allostatic Load", Poster Presentation, Population Association of America Annual Meeting, San Diego, CA 2015

**Sun Y. Jeon**. "Effects of Family Networks on Mental Health among Latino Immigrants in the United States, Presentation, Utah State University Graduate Research Symposium, Logan, UT 2014

**Sun Y. Jeon**. "Effects of Family Networks on Mental Health among Latino Immigrants in the United States, Poster Presentation, Population Association of America Annual Meeting, Boston, MA 2014

**Sun Y. Jeon** and Eric N. Reither. "A Demographic Evaluation of Increasing Rates of Suicide Mortality in Japan and South Korea, Poster session, International Union for the Scientific Study of Population, Busan, South Korea 2013

**Sun Y. Jeon** and Eric N. Reither. "A Demographic Evaluation of Increasing Rates of Suicide Mortality in Japan and South Korea, Poster session, Population Association of America Annual Meeting, New Orleans, LA 2013

Andrew E. Burger, Eric N. Reither, David Ramos, and **Sun Y. Jeon**. "Racial and Ethnic Disparities in Seasonal Influenza Vaccination among Utah Adults, 2000-2008" Western Social Science Association Annual Conference, Salt Lake City, UT 2011

Mortality in Japan and South Korea, Poster session, Population Association of America Annual Meeting, New Orleans, LA 2013

Andrew E. Burger, Eric N. Reither, David Ramos, and **Sun Y. Jeon**. "Racial and Ethnic Disparities in Seasonal Influenza Vaccination among Utah Adults, 2000-2008" Western Social Science Association Annual Conference, Salt Lake City, UT 2011

## Working Papers

- **Sun Y. Jeon**. "Developing Racial/Ethnic Specific Classification of Obesity for White, Hispanic, and Black Males in the U.S." in progress
- **Sun Y. Jeon** "Application of Random Forest Classification for Studies of Allostatic Load," in progress
- So-jung Lim, Joongbaeck Kim, Hyeyoung Woo, and **Sun Y. Jeon**, "Relationship between Non-standard Employment and Health in South Korea: The Role of Gender and Marital Status," in progress
- Hyeyoung Woo, So-jung Lim, and **Sun Y. Jeon**, "Does Marriage Still Matter? Parental Marital Status and Children's Health in Korea," in progress

## Professional Affiliation

**Professional Membership**

Population Association of America
American Sociological Association

## Voluntary Service

**Reviewer for Journals**

Journal of Personality and Social Psychology
Epidemiology and Psychiatric Sciences

**Vice President**, Sociology Graduate Student Association (2013-2014)