The Report Committee for Minle Xu
Certifies that this is the approved version of the following report:

Bayesian Ridge Estimation of Age-Period-Cohort Models

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Daniel A. Powers

Carlos M. Carvalho

Bayesian Ridge Estimation of Age-Period-Cohort Models

by

Minle Xu, L.L.B.; M.S.

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Statistics

The University of Texas at Austin

August 2014

Dedication

This report is dedicated to my supportive parents and husband *Kai Yin*.

Acknowledgements

I want to express my sincere gratitude to my supervisor Dr. Daniel Powers for his excellent guidance, patience, and providing me with an excellent atmosphere for completing my report. He has offered great support to me both as a Master student in Statistics and a Ph.D. student in Sociology. He is a model sociologist and statistician for me because of his broad and in-depth knowledge and devotion to his students. I would never have been able to finish my report without his support and help.

I would like to give thanks to my committee member Dr. Carlos Carvalho for his insightful suggestions, time, and attention.

I want to thank Dr. Mary Parker for her kindness, encouragement, and patience.

I would like to acknowledge the academic, financial, and administrative support from the Department of Statistics and Data Sciences.

I would also like to thank my parents and my brother. They are always supporting and encouraging me with their best wishes.

Finally, I would like to thank my husband, Kai Yin. He is always there cheering me up and stands by me through the good times and bad.

Abstract

Bayesian Ridge Estimation of Age-Period-Cohort Models

Minle Xu, M.S. Stat

The University of Texas at Austin, 2014

Supervisor: Daniel A. Powers

Age-Period-Cohort models offer a useful framework to study trends of time-specific phenomena in various areas. Yet the perfect linear relationship among age, period, and cohort induces a singular design matrix and brings about the identification issue of age, period, and cohort model due to the identity Cohort = Period – Age. Over the last few decades, multiple methods have been proposed to cope with the identification issue, e.g., the intrinsic estimator (IE), which may be viewed as a limiting form of ridge regression. This study views the ridge estimator from a Bayesian perspective by introducing a prior distribution(s) for the ridge parameter(s). Data used in this study describe the incidence rate of cervical cancer among Ontario women from 1960 to 1994. Results indicate that a Bayesian ridge model with a common prior for the ridge parameter yields estimates of age, period, and cohort effects similar to those based on the intrinsic estimator and to those based on a ridge estimator. The performance of Bayesian models with distinctive priors for the ridge parameters of age, period, and cohort effects is affected more by the choice of prior distributions. In sum, a Bayesian ridge model is an alternative way to deal with the identification problem of age, period, and cohort model. Future studies

should further investigate the influences of different prior choices on Bayesian ridge models.

Table of Contents

List of Tables

List of Figures

INTRODUCTION

Over the last few decades the age-period-cohort (APC) model has become one of the core approaches in demography and sociology to study the trends of a multitude of social phenomena. The application and impact of APC models has spread beyond areas in social sciences to epidemiology and biostatistics. Discussions about using APC models to separate cohort effects from age and period effects on time-specific phenomena originated eighty years ago among social scientists (Mason & Wolfinger, 2002).

The first temporal component of the APC model, age, specifies variation in the outcome of interest pertaining to different age groups due to biological process of aging, cumulated social experience, and changes in social roles and statuses. The period component represents influences associated with time periods that affect people of all age groups at the same time because of significant social, cultural, economic, political changes. Cohort refers to variations related to groups of people who experience an initial event, typically birth or marriage at the same year or years, and undergo subsequent social and historical events at the same ages (Yang & Land, 2013). For instance, age, period, and cohort are all related to the behavior of consumers. Therefore, age, period, and cohort make distinct contributions to account for time-specific social phenomena. Eliminating one of the three variables will leave results subject to spurious effects (Mason, Winsborough, Mason, & Poole, 1973).

Despite the sound theoretical and conceptual rationale for incorporating age, period, and cohort simultaneously in one model to study time-specific social phenomena, there is no consensus in terms of how to solve the fundamental identification problem of APC models. This methodological challenge results from the exact linear relationship between age, period, and (birth) cohort: cohort = period - age. Consequently, it is impossible to obtain valid estimations of the distinct effects of age, period, and cohort from standard regression-type models.

A variety of methods have been proposed to solve the identification problem of APC models in recent decades, for instance, constrained generalized linear models (CGLM), the ridge estimator, the intrinsic estimator, and hierarchical APC-cross-classified fixed effects and random effects models (Fienberg & Mason, 1978; Fu, 2000; Yang, Fu, & Land 2004; Yang & Land 2008). In the following two sections, this study reviews the identification problem of APC model, current solutions to the identification problem in detail, and then introduces the Bayesian ridge model as an alternative to solving the identification problem of APC model by using data on the incidence rate of cervical cancer among Ontario women from 1960 to 1994.

THE IDENTIFICATION PROBLEM

Prior to discussing the existing methods that address the identification problem of APC model, we first review the classical identification problem. As early as the 1970's, Mason and colleagues (1973) specified the APC multiple classification model for cross-classified data. In the age by period two-way table, the rows and columns represent the main effects of age and period respectively, with the diagonals representing the interaction between age and period—the cohort effects. The APC multiple classification model is specified as

$$g(Y_{ij}) = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ij}, \tag{2-1}$$

where $i = 1, 2, \ldots, \alpha$ for the ith age group; $j = 1, 2, \ldots, p$ for jth period; and $k = 1, 2, \ldots, \alpha + p - 1$for the kth cohort. We can interpret the distinctive effects of age, period, and cohort through an analysis of variance (ANOVA) framework by imposing centered effects normalization, with $\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{p} \beta_j = \sum_{k=1}^{a+p-1} \gamma_k = 0$. Y_{ij} denotes the outcome of interest for those from the ith age group at the jth period, $g(.)$ is the link function for a generalized linear model, and μ is the grand mean of the dependent variable. The APC parameters are normalized so that α_i denotes the difference between the grand mean μ and the mean of the ith age group, β_j denotes the difference between the grand mean μ and the mean of the jth period group, and γ_k denotes the difference between the grand mean μ and the mean of the kth cohort. In a linear model specification, ε_{ij} would denote a random error with mean 0 and variance σ^2. When Y_{ij} is normally distributed, model (2-1) can also be written in matrix form for a linear model as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2-2}$$

where $\mathbf{Y}$ is a column vector of outcomes, **X** is the design matrix of dummy variable column vector, and **β** is a model parameter vector,

$$\boldsymbol{\beta} = (\mu, \alpha_1, \cdots \alpha_{\alpha-1}, \beta_1, \cdots \beta_{p-1}, \gamma_1, \cdots \gamma_{\alpha+p-2})^T, \tag{2-3}$$

where $\boldsymbol{\varepsilon}$ is a vector of random errors with mean 0 and variance σ^2. The ordinary least squares method can be used to obtain the estimates of the model parameter vector **β**:

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}. \tag{2-4}$$

However, a unique estimator **b** does not exist due to the perfect linear relationship among age, period, and cohort. In this case, the design matrix **X** is one less than full rank. As the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is singular and has an eigenvalue of 0, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is not invertible unless special numerical methods such as a Moore-Penrose generalized inverse is used. In other words, there are infinite solutions of **b** that fit the data equally well as a result of the perfect linear relationship among age, period, and cohort. This is the fundamental identification issue of the unconstrained APC model.

CURRENT SOLUTIONS TO THE IDENTIFICATION PROBLEMS

Several decades ago, scholars started to address the identification problems of APC models. One early method proposed by Mason and colleagues (1973) was to impose at least one constraint on the parameter vector $\boldsymbol{\beta}$. For instance, the effects of two age groups, two periods, or two cohorts can be constrained to be the same with a priori reasoning. With such a constraint, APC models become just-identified and unique estimators of model parameters exist. Even though different choices of equality will not affect model fit, the coefficients and significance of age, period, and cohort vary considerably and the results can be difficult to interpret with arbitrary choices. Thus, in order to use the constrained generalized linear model (CGLM), it is crucial to justify the assumption of equality based on theoretical reasons. However, such theoretical information is not always available and differs in every situation.

Another method commonly used to deal with the identification problem caused by perfect multicollinearity is the ridge estimator. Fu (2000) first introduced the ridge estimator to the APC multiple classification model whose design matrix has one less than full rank. The ridge estimator overcomes the identification issue by adding a ridge penalty to the diagonal of matrix $\mathbf{X}^T\mathbf{X}$. Let $\mathbf{X}$ be as the $m \times n$ ($n < m$) design matrix and $\mathbf{I}$ the $m \times m$ identity matrix. Letting λ be the shrinkage or ridge parameter ($\lambda \geq 0$), the ridge estimator is defined as

$$\mathbf{b}_R = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{3-1}$$

This equation shows that ridge parameter induces bias except when λ is equal to 0. Typically, the values of λ lie in the range of $(10^{-4}, 1)$. As λ increases, the bias increases but variance decreases. The optimal value of λ that produces a little bias but substantially lowers the variance is the λ that minimizes the generalized cross-validation (GCV).

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\hat{y}_i}{1-\mathrm{tr}(\mathbf{H})/n}\right)^2. \tag{3-2}$$

The hat matrix $\mathbf{H}$ is $\mathbf{X}\,(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}$ here and the trace of $\mathbf{H}$ is the sum of its diagonal elements. Ultimately, the ridge estimator yields better estimation with smaller mean square error.

Yang et al. (2004) employed another approach—the intrinsic estimator— to cope with the identification problem of APC model. Given that the design matrix $\mathbf{X}$ is one less than full column rank, the parameter space $\mathbf{b}$ of APC model can be decomposed into the sum of two linear subspaces:

$$\mathbf{b} = \mathbf{B} + t\mathbf{B}_0, \tag{3-3}$$

where t is a real value for a specific solution, $\mathbf{B}_0$ refers to the null subspace corresponding to the zero eigenvalue of matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and only relies on the design matrix $\mathbf{X}$ (the number of age, period, and cohorts), and $\mathbf{B}$ represents the complement non-null subspace orthogonal to the null space and is the intrinsic estimator. One way to compute intrinsic estimator is to use the Moore-Penrose generalized inverse of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ denoted by $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{+}$ (Fu & Hall, 2006):

$$\mathbf{b}_{\mathrm{IE}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{+}\,\mathbf{X}^{\mathrm{T}}\mathbf{Y}. \tag{3-4}$$

Here we focus on describing another approach to calculate intrinsic estimator—the principle component regression method:

$$\mathbf{b}_{\mathrm{IE}} = (\mathbf{Q}\mathrm{L}_0^{-1}\mathbf{Q}^{\mathrm{T}})\,\mathbf{X}^{\mathrm{T}}\mathbf{Y}, \tag{3-5}$$

where $\mathbf{Q}$ is the $n \times n$ orthogonal matrix of eigenvectors of matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, $\mathbf{L}$ is an $n \times n$ diagonal matrix composed of the eigenvalues of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$: $\ell_1, \cdots, \ell_n$, and $\mathbf{Q}\mathbf{L}_0^{-1}\mathbf{Q}^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$. Accordingly, ℓ_0^{-1} is an $n \times n$ diagonal matrix with values $\ell_1^{-1}, \cdots, \ell_{n-1}^{-1}, 0$ on the diagonal. Thus, the intrinsic estimator is obtained by eliminating eigenvalue 0 via principle components. Moreover, intrinsic estimator can be generalized as a limiting form of ridge estimator (Fu, 2000). Specifically, intrinsic estimator is the limit of a coefficient vector from a ridge regression with a vanishingly small shrinkage penalty $\lambda \to 0^+$. When $\lambda > 0$, the variance of the ridge estimator is smaller than that of the intrinsic estimator. If λ is set to be a very small positive number, the ridge estimator will be almost equal to intrinsic estimator. Therefore, we might choose to use the ridge estimator rather than intrinsic estimator in practice. However, one difficulty of using ridge estimator lies in determining the optimal value of λ for a given dataset.

When the range of the outcome variable for an APC model is bounded in the population (e.g. binary outcome), Browning et al. (2012) proposed a generic approach using maximum entropy estimator to address the deification issue of APC models. If $Y \in [Y_{\min}, Y_{\max}]$, $\beta_k \in [Y_{\min} - Y_{\max}, Y_{\max} - Y_{\min}]$ for all k. That is, the bounded Y leads to the set (partial) identification of $\boldsymbol{\beta}$. The parameter vector $\boldsymbol{\beta}$ falls into a closed, convex parameter space. The central idea is to reparameterize the parameter vector $\boldsymbol{\beta}$ to a probability distribution over the set of possible solutions that has maximum entropy and choose the most uninformative (flat) probability distribution based on available information in the data. Suppose the identified set for $\boldsymbol{\beta}$ is given by

$$\mathcal{B} = \{\boldsymbol{\beta} | \boldsymbol{\beta} = \mathbf{SP}\}, \tag{3-6}$$

Where $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \cdots, \boldsymbol{s}_J]$, the vector $\mathbf{s}_j$ represents the vertices of $\mathcal{B}$; **P** is a vector of non-negative weights that sum to 1 and are used to form all of the convex combinations of the vertices. Hence $\mathbf{X\beta}$ can be expressed as **XSP** and **P** is treated as a discrete probability distribution over the *J* multivariate outcomes represented by the columns of the matrix **S**. We need to choose a distribution that does not overly favor one outcome over another and the probabilities are nonnegative and sum to one. The entropy function

$$\mathrm{H}(\mathbf{P}) = -\mathbf{P}^{\mathrm{T}} \log\mathbf{P} \tag{3-7}$$

is an objective function that is maximized when the probabilities are uniform. The problem becomes a maximum entropy problem with a unique solution. If $\mathbf{P}^*$ is the vector that solves the problem, then $\boldsymbol{\beta}^* = \mathbf{SP}^*$, which can be interpreted as the expected value of a discrete multidimensional random variables consistent with the maximum entropy probability distribution.

The mixed effects model for APC analysis developed by Yang and Land (2008) can be conducted when repeated cross-section sample surveys are available. The availability of individual level observations enable us to group age and period into one-year length and cohort into meaningful multiple year intervals (e.g. five-year birth cohort). The classification of age, period, and cohort into unequal intervals eliminates the identification issue and finite solutions to equation (2-4) can be obtained. Because individuals from the same birth cohorts or survey years may share some similarities unique to their cohorts or survey periods, hierarchical regression models should be employed to account for the nonindependence of error terms and estimate error variances.

Specifically, Yang and Land (2008) proposed a cross-classified random-effects model (CCREM) and a cross-classified fixed-effects model (CCFEM). A basic level 1 equation for both models can be specified as:

$$Y_{ijk}= \beta_{0jk} + \beta_1 Age_{ijk} + \varepsilon_{ijk} \sim N\,(0, \sigma^2). \tag{3-8}$$

The level 2 equation of CCREM is:

$$\beta_{0jk} = \gamma_0 + u_{0j} + v_{0k},\ u_{0j} \sim N\,(0,\ \tau_u),\ v_{0k} \sim N\,(0,\ \tau_v). \tag{3-9}$$

u_{0j} and v_{0k} are the residual random effects of period j (averaged over all cohorts) and cohort k (averaged over all periods) on β_{0jk}. The level 2 equation of CCFEM is:

$$\beta_{0jk} = \gamma_0 + \sum_{j=2}^{J} \gamma_{1j}\, \mathrm{Period}_j + \sum_{k=2}^{K} \gamma_{2k}\, \mathrm{Cohort}_k, \tag{3-10}$$

where the effects of periods and cohorts are respectively estimated by J-1 dummy variables for periods and K-1 dummy variables for cohorts. Conventionally, the choice between CCREM and CCFEM depends on two conditions: 1) The CCREM requires that the level 2 effects are independent of level 1 predictor variables; 2) The relatively small total number of birth cohorts and periods suggests modeling these contextual effects as fixed. However, empirical results done by Yang and Land (2008) favor CCREM due to the unbalanced data design of repeated cross-section surveys.

Although the ridge estimator is an accessible approach to deal with the identification problem of the APC model, a suitable method to find the optimal λ for a given dataset may require further investigation. Fu (2000) suggested using a GCV approach to select an optimal value of λ. An alternative way to determine the optimal λ involves in Bayesian analysis. A general Bayesian interpretation of the ridge estimator has been noted in 1970s (Hsiang, 1975; Marquardt, 1970). Congdon (2006) explicated the use of Bayesian ridge priors as one possible solution to multicollinearity. As far as I know, no one has applied a Bayesian ridge approach to APC multiple classification models which are subject to perfect linear relationship between age, period and cohort. In this paper, I will utilize Bayesian ridge priors to solve the identification problem of APC model using data on cervical cancer incidence rates among Ontario women from 1960 to 1994. I will then compare the results to those obtained by using the intrinsic estimator and using a conventional ridge estimator.

Table 1 Cervical cancer Incidence rates in Ontario women 1960-1994 (per 10^5 person-years)

Age/Year	60-64	65-69	70-74	75-79	80-84	85-89	90-94
20-24	3.89	3.24	2.90	2.05	2.19	1.76	1.73
25-29	16.01	11.18	8.92	9.74	8.48	7.43	7.54
30-34	26.02	21.14	16.23	15.84	14.54	13.67	12.71
35-39	38.84	25.09	21.07	18.74	18.80	18.04	18.18
40-44	47.65	32.50	22.71	20.01	18.78	16.19	18.12
45-49	51.48	36.69	22.15	19.20	17.74	17.29	18.31
50-54	49.12	37.26	25.51	18.41	16.66	15.41	14.07
55-59	51.48	40.87	34.70	21.83	16.97	17.69	13.73
60-64	47.68	42.80	29.76	22.71	20.16	17.69	16.94
65-69	40.44	39.17	31.44	28.79	23.35	19.26	19.16
70-74	42.4	35.32	27.78	24.31	20.27	20.19	14.95
75-79	42.44	36.68	28.75	25.22	21.17	21.08	19.43
80-84	41.50	29.74	31.54	22.31	20.04	15.25	21.28
85+	30.79	32.43	37.10	19.81	16.42	14.87	12.06

METHODS

Before introducing the Bayesian ridge approach, a brief review of the Bayesian statistical method is presented here. Unlike the frequentist statistical paradigm that treats a parameter θ as an unknown fixed parameter, Bayesian statistical method views θ as a random quantity and uses a prior probability distribution to describe its variation. This prior distribution of θ is updated by taking account of information from the data to obtain the posterior distribution of θ. According to Bayes theorem, the posterior distribution of θ is summarized as:

$$p(\theta|y) = p(y|\theta)\, p(\theta) \,/\, p(y), \tag{4-1}$$

where $p(y|\theta)$ is the likelihood function, $p(\theta)$ is the prior distribution of θ before seeing the data. $p(y)$ is the marginal distribution of the data defined as $p(y) = \int p(y|\theta)p(\theta)d(\theta)$ and this integral can be complicated and hard to compute. However, since θ is integrated out, $p(y)$ is a normalizing constant that guarantees $p(\theta|y)$ is a proper density. Bayes theorem is usually expressed as $p(\theta|y) \propto p(y|\theta)\, p(\theta)$. One commonly used Bayes estimator is the mean of the posterior distribution $p(\theta|y)$ given by

$$\widehat{\theta} = \int \theta p(\theta|y)d(\theta). \tag{4-2}$$

Other summary statistics include posterior median, mode, variance, credible interval, and interquartile range. When the posterior distribution $p(\theta|y)$ is from a known density function, such summary statistics can be easily calculated. However, this is usually not the case especially when dealing with high-dimensional models. Under such circumstances, Bayesian statisticians have resorted to sampling-based estimation methods—Markov chain

Monte Carlo (MCMC) to draw inferences about θ. Sample summary statistics calculated based on relatively large samples from the posterior distribution using iterative MCMC methods tend to equate posterior summary statistics. One useful Markov chain algorithm is the Gibbs sampler, which samples iteratively from the full conditional posterior distribution of each parameter obtained from the joint density distribution. Each parameter is updated sequentially and conditional on all the other parameters. When models involve standard distributions, the conditional posterior distributions of the parameters are also likely to be standard densities and sampling from such conditional posterior distributions is straightforward.

BAEYSIAN RIDGE MODEL

The ridge estimator proposed to solve the identification problem can be viewed from a Bayesian perspective (Congdon, 2006). For the standard regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon}$ distributed normally with mean 0 and variance σ^2, the prior on $\boldsymbol{\beta}$ can be assumed to be from a common normal density with mean zero and variance equal to σ^2/λ. Then the mean of the posterior distribution of $\boldsymbol{\beta}$ has the form $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$, which is identical to the ridge estimator. Different ridge priors for age, period, and cohort coefficients can also be specified. The inclusion of different ridge priors extends the model to the form of generalized ridge estimates and the posterior mean of $\boldsymbol{\beta}$ then becomes $(\mathbf{X}^T\mathbf{X} + \boldsymbol{\Lambda}\mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y}$, where $\boldsymbol{\Lambda}$ represents a vector of λ's.

Data used to demonstrate and compare the Bayesian ridge prior model with models estimated by the intrinsic estimator and the ridge estimator was originally presented in Fu's study (2000). The data documented the cervical cancer incidence rates of Ontario women aged 20 and above from 1960 to 1994. As shown in Table 1, there are 98 observations (or data cells), with 14 age groups, 7 period groups, and 20 diagonals of birth cohorts. A log transformation was applied to the incidence rates of cervical cancer. The APC is then specified as

$$\log(Y_{ij}) = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ij}, \quad (4\text{-}3)$$

where Y_{ij} is the cervical cancer rate for age group i and period j, i = 1, 2,…,14 , j = 1, 2,…,7 , and k =1, 2,…, 20 . ANOVA normalization was used to center the parameters in model (4-3). And the last age, period, and cohort category was used as the reference category respectively. Therefore,

$$\log(Y_{ij}) = \mu^* + \alpha_i^* + \beta_j^* + \gamma_k^* + \varepsilon_{ij}, \quad (4\text{-}4)$$

where $\mu^* = \mu + \bar{\alpha} + \bar{\beta} + \bar{\gamma}$, $\alpha_i^* = \alpha_i - \bar{\alpha}$, $\beta_j^* = \beta_j - \bar{\beta}$; $\gamma_k^* = \gamma_k - \bar{\gamma}$, and α_i^* = 1, 2,…,13, j = 1, 2,…,6 , and k =1, 2,…,19. The Bayesian model with one ridge prior for age, period, and cohort coefficients therefore can be summarized as follows:

Likelihood function for the model: $f(Y|\mu^*, \beta, \sigma^{-2}, \lambda)$[1]

Prior distributions: $p(\mu^*, \beta, \sigma^{-2}, \lambda) = p(\mu^*)\, p(\beta)\, p(\sigma^{-2})\, p(\lambda)$

The joint posterior distribution:

$$p(\mu^*, \beta, \sigma^{-2}, \lambda|Y) \propto f(Y|\mu^*, \beta, \sigma^{-2}, \lambda)\, p(\mu^*, \beta, \sigma^{-2}, \lambda)$$

As sampling directly from the joint posterior distribution is not feasible here, a Gibbs sampler that works with conditional distributions for each parameter is used. The Gibbs sampler then works as follows:

---

[1] For the simplicity of expression, β represents all the parameters for age, period, and cohort effects.

1: Start with a vector of starting values for all the parameters: $(\mu_0^*, \beta_0, \sigma_0^{-2}, \lambda_0)$,

2: Sample μ_1^* from $p(\mu_1^*|\beta_0, \sigma_0^{-2}, \lambda_0)$,

3: Sample β_1 from $p(\beta_1|\mu_1^*, \sigma_0^{-2}, \lambda_0)$,

4: Sample σ_1^{-2} from $p(\sigma_1^{-2}|\mu_1^*, \beta_1, \lambda_0)$,

5. Sample λ_1 from $p(\lambda_1|\mu_1^*, \beta_1, \sigma_1^{-2})$,

6. Then repeat step 2 to step 5: e.g., sample μ_2^* from $p(\mu_2^*|\beta_1, \sigma_1^{-2}, \lambda_1)$.

Conditionally conjugate priors were used for all the parameters in the APC model. First, a normal density with $N(0, \sigma^2/\lambda)$ was used as the common prior distribution for all the age, period, and cohort coefficients. The noninformative prior distribution of μ^* is distributed as $N(0, 10^4)$ and a vague gamma prior was used for the precision of the error term (Gelman, et al., 2013):

$$\sigma^{-2} \sim \text{gamma } (.001, .001). \tag{4-5}$$

λ is the Bayesian ridge penalty, and a noninformative prior distribution such as gamma (.001, .001) can be assigned to λ. Here the prior distribution of λ is specified as

$$\lambda \sim \text{gamma } (1,1), \tag{4-6}$$

since the posterior estimation of age, period, and cohort effects are very similar to those using the noninformative gamma prior. The Bayesian estimation of any model parameters can be gained once the Markov chain has been run for a large number of iterations. For instance, the posterior mean and standard error of β based on M draws of $(\mu^*, \beta, \sigma^{-2}, \lambda)$ can be calculated as follows:

$$\hat{\beta} = \frac{1}{M} \sum_{M=1}^{M} \beta^M, \tag{4-7}$$

$$\mathrm{SE}(\beta) = \sqrt{\frac{1}{M-1}\sum_{M=1}^{M}(\beta^{M} - \hat{\beta})^{2}} \quad . \tag{4-8}$$

Different Ridge Priors

An idea that fits substantively better with the APC theory is to define three different priors for age, period, and cohort effects rather than using a common Bayesian ridge prior. Supposing that for λ_A, λ_P, and λ_C correspond to the ratio of the error variance and variance of age, period, cohort coefficients, for instance, $\lambda_A = \sigma^2/\sigma_A^2$. In other words, the age, period, cohort coefficients have distinctive variances σ_A^2, σ_P^2, and σ_C^2 in this case. The exchangeable ridge priors for the age, period, cohort coefficients are specified as

$$\alpha_i^* \sim N(0, \sigma^2/\lambda_A), \tag{4-9}$$

$$\beta_j^* \sim N(0, \sigma^2/\lambda_P), \tag{4-10}$$

$$\gamma_k^* \sim N(0, \sigma^2/\lambda_C). \tag{4-11}$$

And the priors used for λ_A, λ_P, and λ_C in the Bayesian model (a) were:

$$\lambda_A \sim \text{gamma}(1,1), \tag{4-11}$$

$$\lambda_P \sim \text{gamma}(1,1), \tag{4-12}$$

$$\lambda_C \sim \text{gamma}(1,100). \tag{4-13}$$

The prior distributions of μ^* and the precision of the error term remain unchanged. To test the influence of priors on model performance, the priors used for λ_A, λ_P, and λ_C in the Bayesian model (b) are:

$$\lambda_A \sim \text{gamma}(1,1), \tag{4-14}$$

$$\lambda_P \sim \text{gamma}(1,1), \tag{4-15}$$

$$\lambda_C \sim \text{gamma}(1,1). \tag{4-16}$$

In the present study, all analyses were conducted using the statistical software R (R Core Team, 2013) and Bayesian inferences using Gibbs sampler were conducted using Jags (Plummer, 2003) via the R package "rjags" (Plummer, 2014). The first 10,000 iterations were used as burn-in and all parameter estimation was based on 50,000 posterior draws.

Table 2 APC Model Estimation from Intrinsic Estimator, Ridge Estimator, and Bayesian Model with a Ridge Prior for APC Effects.

	Intrinsic Estimator	Ridge Estimator	Bayesian Posterior Mean	95% Credible Interval
Intercept	2.945(0.014)	2.939(0.014)	2.941(0.014)	(2.913, 2.968)
Age 20-24	-1.879(0.042)	-1.858(0.116)	-1.850(0.101)	(-2.045, -1.660)
Age 25-29	-0.509(0.039)	-0.503(0.099)	-0.501(0.087)	(-0.665, -0.337)
Age 30-34	0.047(0.039)	0.047(0.084)	0.046(0.075)	(-0.096, 0.187)
Age 35-39	0.316(0.039)	0.312(0.070)	0.310(0.063)	(0.189, 0.431)
Age 40-44	0.368(0.039)	0.362(0.057)	0.360(0.053)	(0.257, 0.462)
Age 45-49	0.354(0.040)	0.347(0.047)	0.345(0.045)	(0.256, 0.433)
Age 50-54	0.244(0.040)	0.237(0.041)	0.236(0.041)	(0.155, 0.316)
Age 55-59	0.298(0.040)	0.292(0.041)	0.290(0.041)	(0.209, 0.371)
Age 60-64	0.273(0.040)	0.268(0.047)	0.267(0.046)	(0.178, 0.355)
Age 65-69	0.278(0.039)	0.274(0.057)	0.273(0.053)	(0.170, 0.375)
Age 70-74	0.122(0.039)	0.120(0.070)	0.121(0.063)	(0.001, 0.241)
Age 75-79	0.138(0.039)	0.138(0.084)	0.139(0.075)	(-0.003, 0.281)
Age 80-84	0.036(0.039)	0.040(0.099)	0.042(0.087)	(-0.121, 0.207)
Period 60-64	0.476(0.026)	0.476(0.056)	0.475(0.050)	(0.381, 0.570)
Period 65-69	0.270(0.026)	0.269(0.042)	0.269(0.039)	(0.195, 0.344)
Period 70-74	0.081(0.026)	0.080(0.031)	0.081(0.030)	(0.022, 0.139)
Period 75-79	-0.103(0.026)	-0.104(0.026)	-0.103(0.026)	(-0.155, -0.052)
Period 80-84	-0.190(0.026)	-0.190(0.031)	-0.190(0.030)	(-0.248, -0.132)
Period 85-89	-0.263(0.026)	-0.262(0.042)	-0.262(0.039)	(-0.336, -0.188)
Cohort -1879	0.090(0.098)	0.079(0.184)	0.082(0.164)	(-0.236, 0.398)

Table 2, Cont.

Cohort 1876-1884	0.309(0.070)	0.298(0.157)	0.296(0.139)	(0.031, 0.560)
Cohort 1881-1889	0.334(0.058)	0.329(0.137)	0.326(0.121)	(0.094, 0.554)
Cohort 1886-1894	0.268(0.052)	0.266(0.119)	0.264(0.105)	(0.064, 0.463)
Cohort 1891-1899	0.156(0.047)	0.158(0.103)	0.156(0.091)	(-0.017, 0.327)
Cohort 1896-1904	0.180(0.044)	0.183(0.086)	0.182(0.077)	(0.035, 0.328)
Cohort 1901-1909	0.133(0.041)	0.137(0.071)	0.136(0.064)	(0.013, 0.259)
Cohort 1906-1914	0.210(0.042)	0.216(0.059)	0.215(0.055)	(0.109, 0.321)
Cohort 1911-1919	0.148(0.043)	0.155(0.049)	0.155(0.048)	(0.061, 0.249)
Cohort 1916-1924	-0.013(0.043)	-0.004(0.044)	-0.003(0.044)	(-0.089, 0.086)
Cohort 1921-1929	-0.133(0.043)	-0.123(0.044)	-0.121(0.044)	(-0.208, -0.034)
Cohort 1926-1934	-0.205(0.042)	-0.195(0.049)	-0.193(0.048)	(-0.286, -0.099)
Cohort 1931-1939	-0.233(0.041)	-0.224(0.058)	-0.222(0.055)	(-0.327, -0.116)
Cohort 1936-1944	-0.234(0.040)	-0.228(0.070)	-0.228(0.063)	(-0.350, -0.105)
Cohort 1941-1949	-0.189(0.042)	-0.186(0.086)	-0.185(0.076)	(-0.330, -0.039)
Cohort 1946-1954	-0.102(0.045)	-0.101(0.102)	-0.102(0.090)	(-0.273, 0.070)
Cohort 1951-1959	-0.138(0.050)	-0.140(0.119)	-0.140(0.104)	(-0.340, 0.059)
Cohort 1956-1964	-0.145(0.057)	-0.150(0.137)	-0.150(0.120)	(-0.379, 0.079)
Cohort 1961-1969	-0.190(0.069)	-0.199(0.157)	-0.198(0.138)	(-0.460, 0.067)
λ	—	0.050	0.078(0.023)	(0.041, 0.132)
Posterior variance of error	—	—	0.011(0.002)	(0.008, 0.018)
Posterior variance of APC coefficients	—	—	0.150(0.036)	(0.095, 0.235)

RESULTS

Table 2 presents estimates of the APC model parameters using the intrinsic estimator, the ridge estimator, and the Bayesian model with a common prior for age, period, and cohort effects. The three approaches generated very similar patterns for the age, period, and cohort trends as shown by the estimates and levels of significance. The 95% credible interval indicates that the significance of age, period, and cohort effects from the Bayesian ridge prior model is consistent with results from the intrinsic and ridge estimators. For instance, the 95% credible interval for the age effect of the group aged 30 to 34 is (-0.096, 0.187). The inclusion of zero in this interval implies that the age effect of the group aged 30 to 34 is 0. The results from the intrinsic or ridge estimator also indicate that the group aged 30 to 34 has no significant effects on women's cervical cancer incidence rate since the ratio of the age coefficient to its standard error is less than 1.96. Generalized cross-validation (GCV) was used for selection of the optimal λ for ridge estimator and the plot of GCV(λ) is shown in Figure 1 which illustrates that the minimum value of GCV is about 0.017 in this case and the corresponding value for λ is 0.050. The posterior mean of λ is 0.078 and the 95% credible interval indicates the true mean of λ is within the interval (0.041, 0.132) with 95% probability. The ridge parameter (λ = 0.050) is within the 95% credible interval.

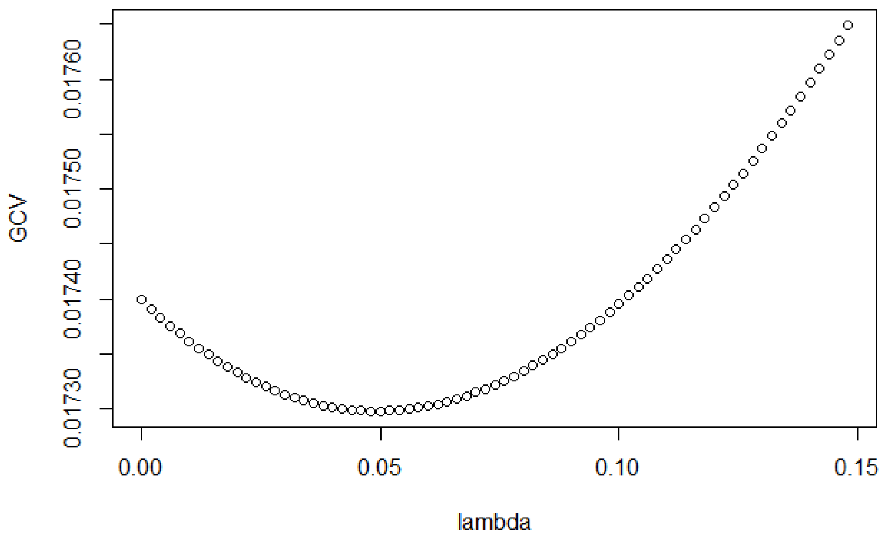


Figure 1 Selection of Lambda for Ridge Estimator via GCV

Figure 2 presents the graphical convergence diagnosis of the MCMC algorithms of selected parameters due to the limited space here. For each selected parameter, the trace plot shows the posterior sample values of a parameter during the runtime of the chain and the marginal density plot is the smoothened histogram of the parameter values from the trace plot. The first three parameters represent the effects of the first age group (20-24), the first period (1960-1964), the first cohort group (-1879). The trace plots provide evidence of satisfactory convergence of the MCMC algorithms for these three parameters. The last three parameters represent the error variance, ridge parameter, and the variance of the APC effects. The trace plots indicate each chain is mixing well here. The Gelman-Rubin (GR) convergence diagnostic is used as a formal test for convergence that assesses whether parallel chains with dispersed initial values converge to the same target distribution. The GR diagnostic shows that the scale reduction factor (SRF) for each parameter is equal to one indicating no difference between the chains for a particular parameter. The multivariate potential SRF is also one, suggesting the joint convergence of the chains over all the parameters. Figure 3 shows the GR diagnostic plots for selected parameters. For each parameter, the GR plot shows the development of Gelman and Rubin's shrink factor as the number of iterations increases and the shrink factor of each parameter eventually stabilized around one.

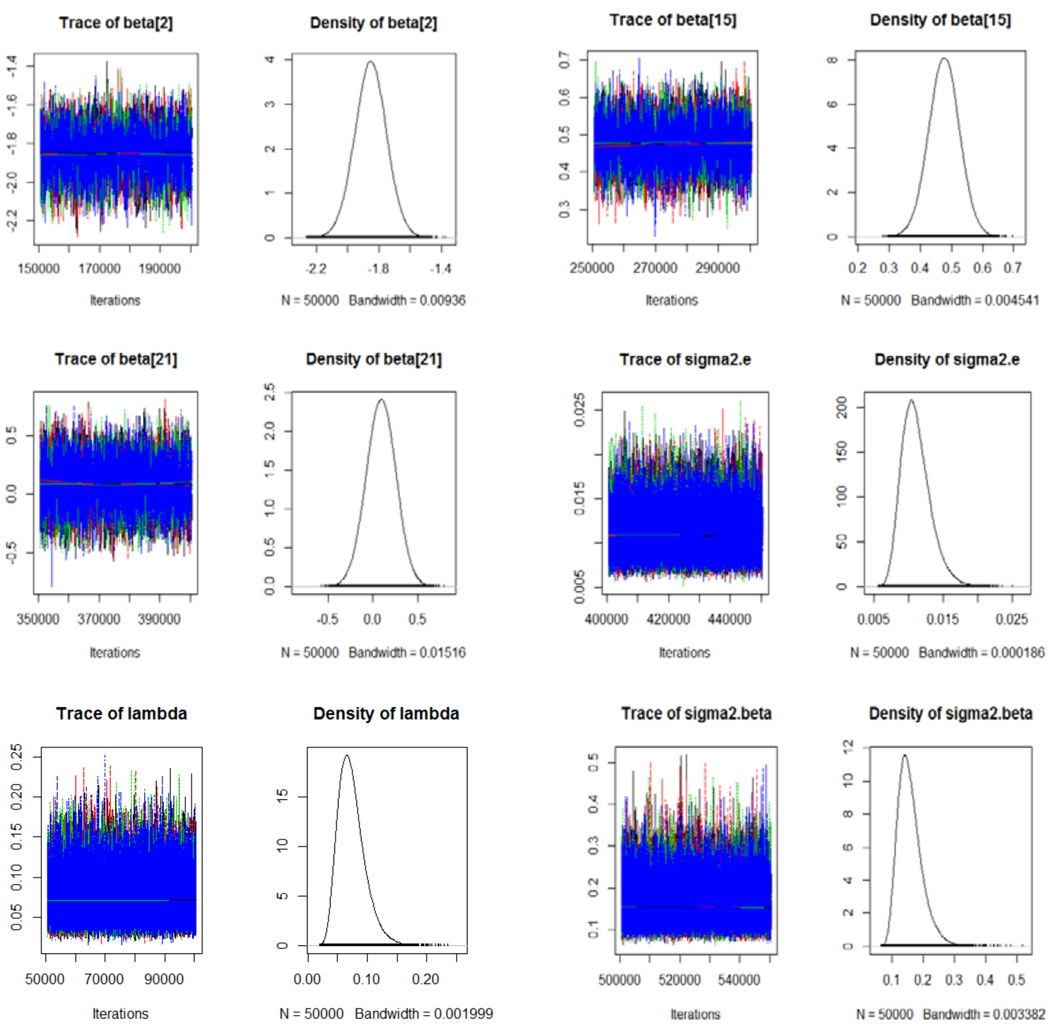


Figure 2 Trace Plots and Density Plots for the Posterior Samples of Selected Parameters.

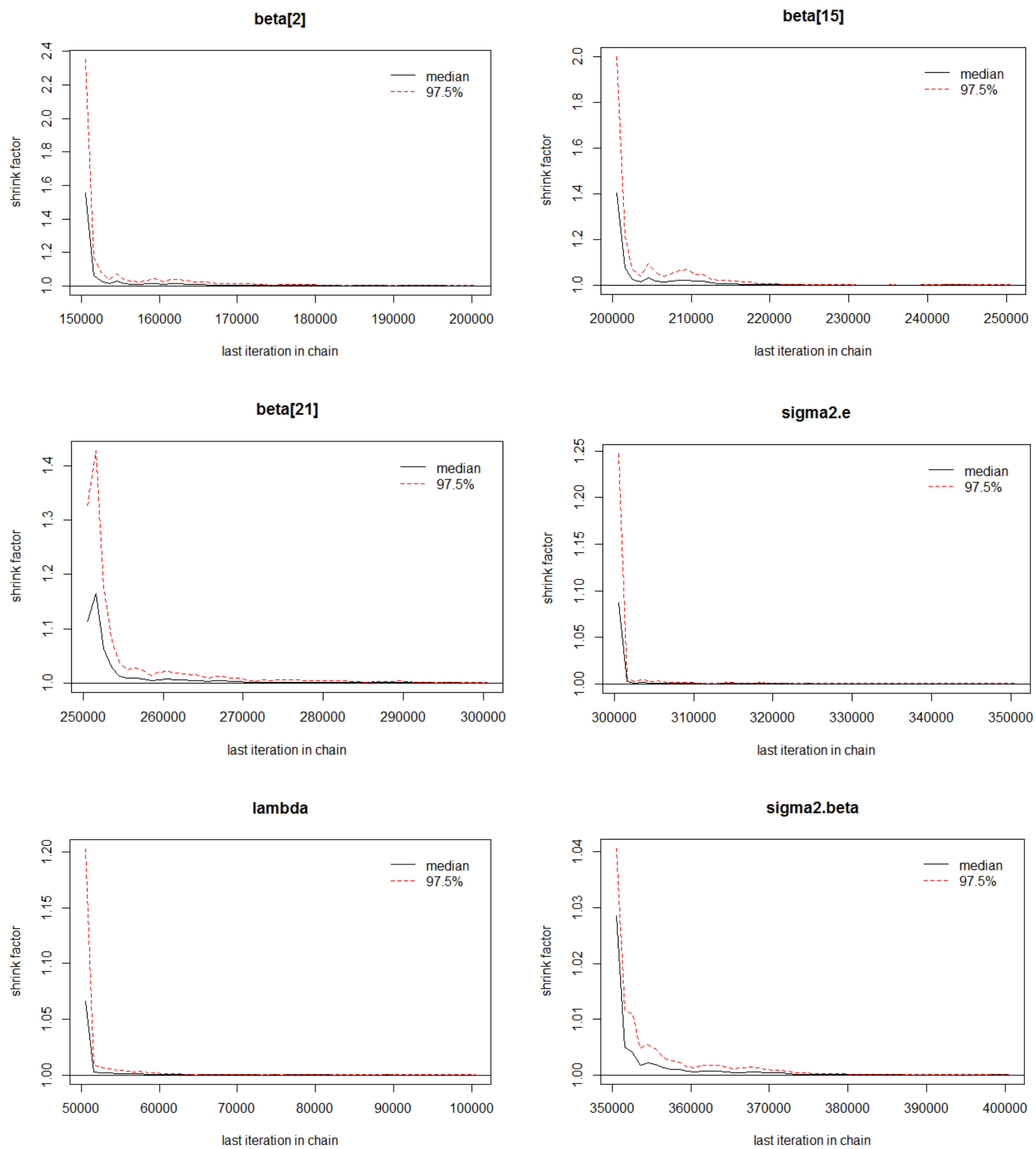


Figure 3 Plots of Gelman-Rubin's Diagnostic of Selected Parameters.

Table 3 APC Model Estimation from Bayesian Model (a) with Different Ridge Priors for APC Effects.

	Bayesian Posterior Mean	2.50% Credible Interval	97.50% Credible Interval
Intercept	2.943	2.918	2.968
Age 20-24	-1.912	-2.131	-1.613
Age 25-29	-0.544	-0.730	-0.291
Age 30-34	0.016	-0.142	0.225
Age 35-39	0.289	0.158	0.457
Age 40-44	0.347	0.241	0.476
Age 45-49	0.339	0.253	0.434
Age 50-54	0.237	0.163	0.312
Age 55-59	0.298	0.223	0.372
Age 60-64	0.281	0.185	0.368
Age 65-69	0.293	0.164	0.401
Age 70-74	0.146	-0.024	0.277
Age 75-79	0.170	-0.040	0.328
Age 80-84	0.077	-0.176	0.263
Period 60-64	0.492	0.349	0.598
Period 65-69	0.281	0.182	0.361
Period 70-74	0.088	0.024	0.145
Period 75-79	-0.102	-0.149	-0.055
Period 80-84	-0.194	-0.252	-0.130
Period 85-89	-0.273	-0.352	-0.173
Cohort -1879	0.030	-0.306	0.470

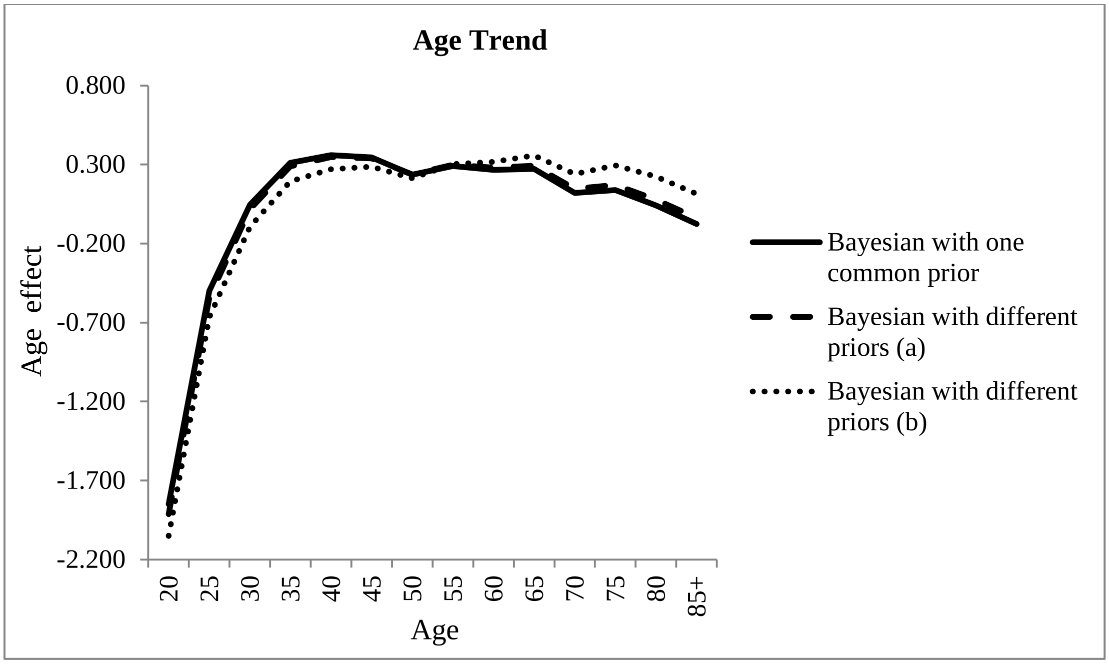
Table 3, Cont.

Cohort 1876-1884	0.245	-0.048	0.639
Cohort 1881-1889	0.278	0.022	0.627
Cohort 1886-1894	0.221	-0.002	0.522
Cohort 1891-1899	0.118	-0.073	0.375
Cohort 1896-1904	0.149	-0.013	0.361
Cohort 1901-1909	0.110	-0.024	0.280
Cohort 1906-1914	0.194	0.083	0.327
Cohort 1911-1919	0.139	0.047	0.240
Cohort 1916-1924	-0.012	-0.092	0.069
Cohort 1921-1929	-0.124	-0.204	-0.045
Cohort 1926-1934	-0.189	-0.288	-0.098
Cohort 1931-1939	-0.210	-0.340	-0.102
Cohort 1936-1944	-0.206	-0.375	-0.075
Cohort 1941-1949	-0.157	-0.368	0.003
Cohort 1946-1954	-0.066	-0.320	0.124
Cohort 1951-1959	-0.096	-0.398	0.125
Cohort 1956-1964	-0.098	-0.442	0.158
Cohort 1961-1969	-0.136	-0.527	0.155
λ_A	0.029	0.011	0.062
λ_P	0.164	0.037	0.443
λ_C	0.078	0.032	0.142
Posterior variance of age coefficients	0.365	0.162	0.778
Posterior variance of period coefficients	0.080	0.022	0.231

Table 3, Cont.

Posterior variance of cohort coefficients	0.135	0.058	0.309
Posterior variance of error	0.009	0.007	0.013

Results from Bayesian model (a) with different ridge priors for age, period, and cohort effects are shown in Table 3. The Bayesian posterior estimation of age, period, cohort effects is similar to that from the Bayesian model with one common prior for the APC effects. To better illustrate the APC trends, Figure 4 shows the age, period, and cohort trends from Bayesian models with distinct specifications for the ridge priors. The solid line represents the Bayesian model with one common prior for the ridge parameter λ which is distributed as gamma(1, 1). The dashed line represents the Bayesian model (a) specifying different priors for λ_A, λ_P, and λ_C with λ_A and λ_P distributed as gamma(1, 1) while λ_C is distributed as gamma(1,100). The dotted line represents the Bayesian model (b) using a gamma(1, 1) prior for λ_A, λ_P, and λ_C respectively. Figure 4 clearly shows that the patterns of age, period, and cohort trends from model (a) resemble those from the Bayesian model with a common prior. For Bayesian model (b), the age and period patterns are akin to that from model (a); whereas the pattern of cohort differs from those from model (a). For instance, there are significant differences in incidence rates of cervical cancer between the early cohorts (born in the late 19th century) and latter cohorts (born in late 20th century) from model (a). However, the incidence rates of cervical cancer for the early cohorts do not significantly differ from those of the latter cohorts from model (b).



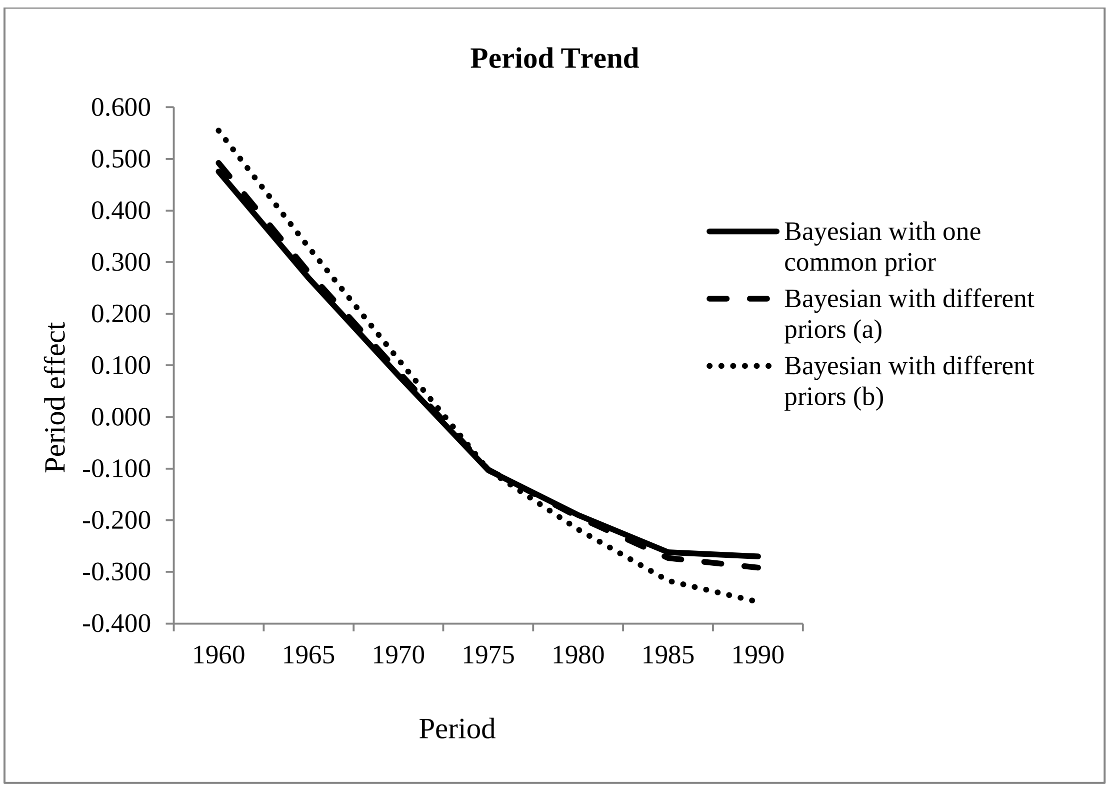
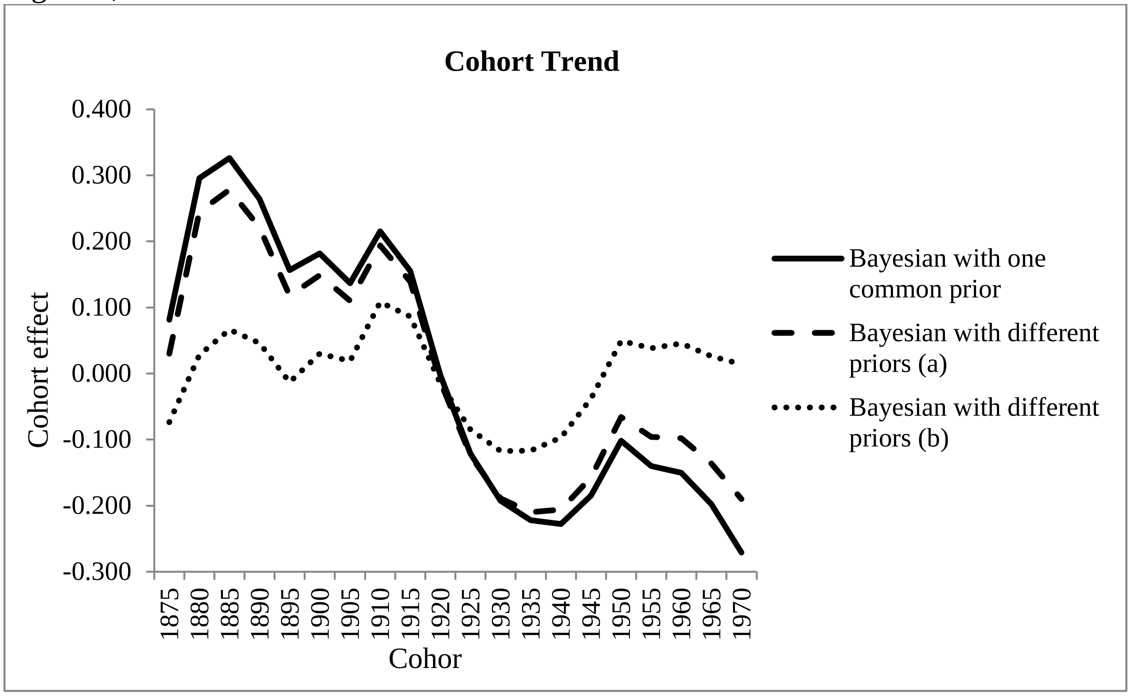


Figure 4 Bayesian Models for Age, Period, and Cohort Trends on Cervical Cancer Incidence Rates in Ontario Women.

Figure 4, Cont.



DISCUSSION

The age, period, cohort accounting model serves as a critical framework to study time-specific phenomena, such as mortality, fertility, and disease rates. The importance of separating age, period, and cohort effects for time-specific phenomena poses a challenge to obtain unique estimates of age, period, and cohort effects simultaneously due to the perfect linear relationship between age, period and cohort. The last few decades have witnessed a proliferation of methods proposed to deal with the identification problem caused by this particular form of multicollinearity, e.g., the intrinsic estimator and the ridge estimator. This paper builds upon the traditional ridge estimator but approaches the identification problem from the Bayesian interpretation of ridge estimation.

In this paper, a Bayesian ridge prior model was used to estimate the age, period, and cohort effects. Results from the Bayesian model with one common ridge prior for age, period, and cohort effects are almost identical to those from a traditional ridge estimator and the intrinsic estimator, suggesting that Bayesian ridge prior model is a useful alternative method to solve the identification problem in APC models. The downside of using the conventional ridge estimator is that one has to specify an optimal value for the ridge parameter in advance based on some criteria. For the Bayesian ridge model, there is no need to assign a single value to the ridge parameter because it is considered as a random variable. We can obtain a series of summary statistics from the posterior samples of the ridge parameter. Further, the random property of the ridge parameter λ in the Bayesian model makes the interpretation of the 95% credible interval more straightforward than the 95% confidence interval from traditional statistics.

A natural extension of the Bayesian model with a common prior for the ridge parameter is to define disparate priors for the corresponding ridge parameters for age, period, and cohort effects. This approach accords with the theory of APC modeling in essence and is of great advantage if prior information on the age, period, and cohort effects is available from meta-analysis based on previous findings. Under this circumstance information from the relevant literature can be incorporated into model estimation by specifying informative priors for age, period, and cohort ridge parameters and the posterior estimation of age, period, cohort effects will be more accurate and close to the true values. The current study demonstrates that the choice of appropriate prior distributions for the ridge parameters is very important as it will affect the posterior means of the age, period, and cohort effects, especially the pattern of the cohort trend in this case.

Although this study touches upon the sensitivity issue associated with choices of prior distributions, it is beyond the scope of this study to thoroughly examine the influences of different prior distributions on the APC model performance. However, one should be cautious when choosing prior distributions for the ridge parameters as the choices of informative priors will impose large influence on the posterior estimation, especially when sample size is small. If no prior information is available, the use of noninformative or diffuse prior distributions is recommended because noninformative priors are more objective compared to subjective elicited priors and leads to Bayesian posterior means close to the maximum likelihood estimates (Congdon, 2006).

To conclude, the Bayesian ridge model provides an alternative way to cope with the identification problem inherent in the APC model due to the perfect linear relationship between age, period, and cohort. Even though noninformative priors can be used to obtain Bayesian estimates of age, period, and cohort effects, informative priors based on the APC theory or previous empirical findings will make the posterior estimation more meaningful.

References

Browning, M., Crawford, I., & Knoef, M. (2012).The age-period cohort problem: Set identification and point identification (CEMMAP working paper CWP02/12). Retrieved from http://dx.doi.org/10.1920/wp.cem.2012.0212

Congdon, P. (2006). Bayesian Statistical Modelling. Wiley Series in Probability and Statistics. doi:10.1002/9780470035948

Fienberg, S. E., & Mason, W. M. (1979). Identification and estimation of Age-Period-Cohort models in the analysis of discrete archival data. *Sociological Methodology*, 10, 1-67. doi:10.2307 /270764

Fu, W. J. (2000). Ridge estimator in singular design with application to age-period-cohort analysis of disease rates. *Communications in Statistics - Theory and Methods*, 29, 263–278. doi:10.1080/03610920008832483

Fu, W. J., & Hall, P. (2006). Asymptotic properties of estimators in age-period-cohort analysis. *Statistics & Probability Letters*, 76, 1925–1929. doi:10.1016/j.spl.2006. 04.051

Gelman, A., Carlin, J. B., Stern, H. S., Runson, D. B., Vehtari A., & Rubin, D. B. (2013). Bayesian data analysis (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Hsiang, T. C. (1975). A Bayesian view on ridge regression. *The Statistician*, 24, 267-268. doi:10.2307/2987923

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12, 591-612. doi:10.2307/12 67205

Mason, K. O., Mason, W. M., Winsborough, H. H., & Poole, W. K. (1973). Some methodological issues in cohort analysis of archival data. *American Sociological Review*, 38, 242-258. doi:10.2307/2094398

Mason, W. M., & Wolfinger, N. H. (2001). Cohort analysis. *International Encyclopedia of the Social & Behavioral Sciences*, 2189–2194. doi:10.1016/b0-08-043076-7/0 0401-0

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing.* Retrieved from http://www.ci.tuwien.ac.at/Conferences/D SC-2003/

Plummer, M. (2014). rjags: Bayesian graphical models using MCMC. R package version 3-13. http://CRAN.R-project.org/package=rjags

Yang, Y., Fu, W. J., & Land, K. C. (2004). A methodological comparison of Age-Period-Cohort models: The intrinsic estimator and conventional generalized linear models. *Sociological Methodology*, 34, 75–110. doi:10.1111/j.0081-1750.20 04.00148.x

Yang, Y., & Land, K. (2013). Age-Period-Cohort analysis. Chapman & Hall/CRC Interdisciplinary Statistics Series. doi:10.1201/b13902

Yang, Y., & Land, K. C. (2008). Age-Period-Cohort analysis of repeated cross-section surveys: fixed or random effects? *Sociological Methods & Research*, 36, 297–326. doi:10.1177/004912 4106292360