Bayesian Ridge Estimation of Age-Period-Cohort Models

Minle Xu

Daniel A. Powers

The University of Texas at Austin

Abstract

Age-Period-Cohort (APC) models offer a framework to study trends of the three temporal dimensions underlying age by period tables. However, the perfect linear relationship among age, period, and cohort leads to a well-known identification issue due perfect colinearity from the identity Cohort = Period – Age. A number of methods have been proposed to deal with this identification issue, e.g., the intrinsic estimator (IE), which may be viewed as a limiting form of ridge regression. Bayesian regression offers an alternative approach to modeling tabular age, period, cohort data. This study views the ridge estimator from a Bayesian perspective by introducing prior distributions for the ridge parameters, which permits these parameters to be estimated from the data rather than being assigned (and fixed) a-priori. Results show 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 with a shrinkage penalty obtained from cross-validation. The performance of Bayesian models with distinctive priors for the ridge parameters of age, period, and cohort effects is, however, affected by the choice of prior distributions. Further investigation of the influence of the choice of prior distributions is therefore warranted.

Introduction

Over the past few decades the age-period-cohort (APC) model has become a core approach for the investigation of trends in numerous social phenomena in demography and sociology. The application and impact of APC models has spread beyond areas in social sciences to epidemiology and biostatistics. Discussions about the use and applicability of 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 age-period-cohort accounting model for age by period tabular data arrays involves three temporal components. The first component, 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 some existing methods that address the identification problem in APC models, 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

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

where $i = 1,\ldots,a$ for the ith age group; $j = 1,\ldots,p$ for jth period; and $k = 1,\ldots,a+p-1$ for the kth cohort. We can interpret the distinctive effects of age, period, and cohort through an analysis of variance (ANOVA) framework by imposing a centered-effects normalization in which

$$\sum_{i=1}^{a}\alpha_i = \sum_{j=1}^{p}\beta_j = \sum_{k=1}^{a+p-1}\gamma_k = 0\,, \tag{2}$$

where 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 (or a suitable transformation of the Y_{ij}), and μ is the grand mean of the dependent variable. The APC parameters are normalized so that each APC effect, α_i, β_j, and γ_k, represent deviations from the grand mean. In a linear model specification, ε_{ij} would denote a random error with mean 0 and variance σ^2. Generalized linear models would not necessarily include an error term or the accompanying residual variance parameter.

When Y_{ij} is continuous, model (1) can be written in matrix form as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

where $\mathbf{Y}$ is a column vector of outcomes, $\mathbf{X}$ is the design matrix with composed of a unit vector and an ANOVA-coded design matrix normalized using the last category level of each APC factor as reference, $\boldsymbol{\beta}$ is the parameter vector,

$$\beta = (\mu, \alpha_1, \ldots, \alpha_{a-1}, \beta_1, \ldots, \beta_{p-1}, \gamma_1, \ldots, \gamma_{a+p-2})^T, \tag{4}$$

and $\boldsymbol{\varepsilon}$ is a vector of random errors with mean 0 and variance σ^2. In an identified model, ordinary least squares can be used to obtain estimates of the model parameter vector $\boldsymbol{\beta}$ as

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{5}$$

However, a unique estimator $\mathbf{b}$ does not exist due to the perfect linear dependence among age, period, and cohort. In this case, the design matrix $\mathbf{X}$ is one less than full rank, and the $\mathbf{X}^T\mathbf{X}$ matrix is singular and is not invertible without special numerical methods such as a Moore-Penrose generalized inverse or singular-value decomposition. In the case of the unconstrained APC model, there are infinitely many solutions of $\mathbf{b}$ that fit the data equally well as a result of this linear dependency. This is the fundamental identification issue pertaining to the unconstrained APC model.

Current Solutions to the Identification Problem

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 estimates of model parameters exist. Even though different choices of equality constraints 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 (Glenn 1976). However, such theoretical information is not always available and differs in every situation.

Ridge regression is another method commonly used to deal with the identification problem caused by perfect multicollinearity. Ridge regression was proposed over 50 years ago as an estimator to accommodate models with highly-correlated predictors (Hoerl 1962; Hoerl and Kennard 1970; Marquart 1970). Modern variants of ridge regression methods exist today in the form of the lasso and lars, estimators (Tibshirani 1996; Efron 2004), which are known collectively as regularization methods. These methods are commonly applied to high-dimensional problems where the goal is to select an optimal subset of predictors having coefficients with minimum variance. Kupper and Janis (1980) were perhaps the first to suggest that ridge regression might be applied to APC models. Fu (2000) applied the ridge estimator to the APC multiple classification model.

The ridge estimator overcomes the identification issue by adding a ridge penalty to the diagonal of $\mathbf{X}^T\mathbf{X}$. Increasing this penalty shrinks the parameter vector toward 0. Let $\mathbf{X}$ be the $n \times m$ $(m < n)$ 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

$$\mathbf{b}_R = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \tag{6}$$

Equation (6) shows that ridge parameter induces bias except when λ is equal to 0. Typically, the values of λ lie in the range of (1.0^{-8}, 1). Like many shrinkage estimators, the ridge estimator yields biased estimates. A tradeoff is a smaller mean square error. In particular, increasing λ results in estimates that are more biased relative to OLS, but with a smaller mean square error. The choice of the shrinkage parameter is critical. In the unconstraind APC model, any choice of λ will produce the same model fit when gauged by criteria such as the residual sum of squares. A ridge trace plot is typically examined to show the behavior of the coefficient vector under varying values of λ. Alternatively, cross-validation measures can be constructed to find the optimal value of λ that produces a little bias but substantially lowers the variance is the λ that minimizes a generalized cross-validation (GCV) measure (Fu 2000).

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - \mathrm{tr}(\mathbf{H})/n}\right)^2 , \tag{7}$$

where $\mathbf{H}$ is the "hat" matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T$, and $\mathrm{tr}(\mathbf{H})$ is the sum of diagonal elements of $\mathbf{H}$. This approach requires repeated model fitting over a range of values of λ in search of the minimum GCV value.

Yang et al. (2004) popularized the intrinsic estimator (IE)—which has been demonstrated by Fu (2000) to be a limiting form of ridge regression-—to cope with the identification problem of the APC model. Given that the design matrix $\mathbf{X}$ is one less than full column rank, the parameter space $\mathbf{b}$ of the APC model can be decomposed into the sum of two linear subspaces:

$$\mathbf{b} = \mathbf{B} + t\mathbf{B}_0, \tag{8}$$

where t is a real value for a specific solution, $\mathbf{B}$ refers to the null subspace corresponding to the zero eigenvalue of $\mathbf{X}^T\mathbf{X}$ and only relies on the design matrix (i.e., 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 $\mathbf{X}^T\mathbf{X}$ denoted by $(\mathbf{X}^T\mathbf{T})^+$ (Fu & Hall, 2006):

$$\mathbf{b}_{\mathrm{IE}} = (\mathbf{X}^T\mathbf{X})^+\mathbf{X}^T\mathbf{Y}\,. \tag{9}$$

This approach is equivalent to a principle component regression

$$\mathbf{b}_{\mathrm{IE}} = (\mathbf{Q}\mathbf{L}_0^{-1}\mathbf{Q}^T)\mathbf{X}^T\mathbf{Y}\,, \tag{10}$$

where $\mathbf{Q}$ is the $m \times m$ orthogonal matrix of eigenvectors of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{L}$ is an $m \times m$ diagonal matrix containing the eigenvalues of $\mathbf{X}^T\mathbf{X}$, $\ell_1,\ldots,\ell_m$ and $\mathbf{Q}\mathbf{L}\mathbf{Q}^T = \mathbf{X}^T\mathbf{X}$. To accommodate the singular design, $\mathbf{L}_0$ in Eq. (9) is defined as the $m \times m$ diagonal matrix with values $\ell_1,\ldots,\ell_{m-1,}0$ on the diagonal. Therefore, in this specification, the intrinsic estimator is obtained by eliminating eigenvalue 0 via principle components, yielding a

principal components regression model. The intrinsic estimator has been shown to be a limiting form of the ridge estimator (Fu, 2000), with a vanishingly small shrinkage penalty $\lambda \to 0^+$, where 0^+ is a value close to, but not equal to, 0. When $\lambda > 0$, the variance of the ridge estimator is smaller than that of the intrinsic estimator. Thus, if λ is set to be a very small positive number, the ridge estimator will produce results nearly equal to those of the intrinsic estimator. Therefore, as noted by Knupper and Janis (1980) and Fu (2000), researchers might choose to use the ridge estimator rather than the intrinsic estimator for APC analysis. However, a difficulty of the ridge estimator lies in determining the optimal value of λ for a given dataset, i.e., a value that produces optimal shrinkage of the APC coefficient vector for that data.

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 presents an added step in modeling. Fu (2000) suggested using a generalized cross-validation (GCV) approach to select an optimal ridge penalty. As noted earlier, this approach requires a series of ridge regressions carried out over a grid of λ values in search of the value yielding the smallest GCV. While this is a straightforward procedure, an alternative approach to determine the optimal shrinkage parameter would be to determine it jointly along with other APC parameters using Bayesian methods. A general Bayesian interpretation of the ridge estimator has been recognized since the 1970s (Hsiang, 1975; Marquardt, 1970). Congdon (2006) explicated the use of Bayesian ridge priors as a possible solution to multicollinearity. However, as

far as we know, a Bayesian ridge approach has not been applied to the APC multiple classification model, which is subject to perfect linear dependence. In this paper, we utilize Bayesian ridge priors to deal with the identification problem of APC model using data on cervical cancer incidence rates among Ontario women from 1960 to 1994.[1] We then compare the results to those obtained using the intrinsic estimator and using a conventional ridge estimator.

Methods

Before introducing the Bayesian ridge approach, we will briefly review Bayesian statistical methods. 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

$$p(\theta \mid y) = \frac{p(y \mid \theta) p(\theta)}{p(y)}, \tag{11}$$

where $p(y \mid \theta)$ is the likelihood function, $p(\theta)$ is the prior distribution of θ before seeing the data, and $p(y)$ is the marginal distribution of the data defined as

[1] Identification may be less an issue using a Bayesian approach where inference is carried out using simulation, as opposed to the traditional numerical methods using least squares.

$p(y) = \int p(y \mid \theta) p(\theta) d\theta$. This integral can be complicated and is often analytically intractable. However, since θ is integrated out, $p(y)$ is a normalizing constant that guarantees that $p(\theta \mid y)$ is a proper density. Therefore, Bayes' theorem is usually expressed as $p(\theta \mid y) \propto p(y \mid \theta) p(\theta)$. One commonly used Bayes estimator is the mean of the posterior distribution of $p(\theta \mid y)$ given by

$$\hat{\theta} = \int p(\theta \mid y) d\theta \tag{12}$$

Other summary statistics include the posterior median, mode, variance, credible interval, and interquartile range. When the posterior distribution $p(\theta \mid y)$ 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 to 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 conditionally 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.

A Bayesian Ridge Model Specification

The ridge estimator proposed for the APC 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 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. Noninformative priors are usually adopted so that the inferences are predominantly based on information from the data. However, a Bayesian approach has advantages over the other frequentist methods (e.g., the conventional ridge estimator) because the specification of priors can draw upon information from previous research as well as take into account uncertainties associated with estimating the parameters of the present study. Moreover, priors based on past research facilitate more meaningful interpretation of inferences.

The data used here to demonstrate and compare the Bayesian ridge prior model with models estimated by the intrinsic estimator and the ridge estimator were originally presented by Fu's (2000). The data document 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 is applied to the incidence rates of cervical cancer, yielding the following APC model specification

$$\log(Y_{ij}) = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ij}, \tag{13}$$

where Y_{ij} is the cervical cancer rate for age group i and period j, $i = 1,\ldots,14$, $j = 1,\ldots,7$, and $k = 1,\ldots, 20$. ANOVA normalization is used to center the parameters in model Eq. (13). And the last age, period, and cohort category is used as reference. Therefore,

$$\log(Y_{ij}) = \mu^* + \alpha_i^* + \beta_j^* + \gamma_k^* + \varepsilon_{ij}, \tag{14}$$

where $\mu^* = \bar{\alpha} + \bar{\beta} + \bar{\gamma}$, $\alpha_i^* = \alpha_i - \bar{\alpha}$, $\beta_j^* = \beta_j - \bar{\beta}$; $\gamma_k^* = \gamma_i - \bar{\gamma}$, and $i = 1,\ldots,13$, $j = 1,\ldots,6$, and $k = 1,\ldots,19$. For purposes of exposition, let β denote the complete APC parameter vector (i.e., excluding the grand mean). The Bayesian model with a single ridge prior for age, period, and cohort coefficients therefore can be summarized as follows:

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

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

The joint posterior distribution: $p(\mu^*, \beta, \sigma^{-2}, \lambda) \mid Y) \propto f(Y \mid \mu^*, \beta, \sigma^{-2}, \lambda)p(\mu^*, \beta, \sigma^{-2}, \lambda)$

As sampling directly from the joint posterior distribution is not feasible in this case, a Gibbs sampler that works with conditional distributions for each parameter is used. The

[2] Variance components in Bayesian models are typically parameterized in terms of precision, i.e., σ^{-2} rather than variance.

Gibbs sampler sequentially samples each parameter from the conditional distribution as follows:

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

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

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

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

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

6. Repeat steps 2 through 5: e.g., sample μ_2^* from $p(\mu_2^* \mid \beta_1, \sigma_1^{-2}, \lambda_1)$

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

$$\sigma^{-2} \sim gamma(0.001, 0.001). \tag{15}$$

The Bayesian ridge penalty may be assigned a noninformative prior

$$\lambda \sim gamma(1,1) \tag{16}$$

since the posterior means of the age, period, and cohort effects are very similar to those using the noninformative gamma prior. Key aspects of the distribution of 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 obtained as summary statistics:

$$\hat{\beta} = \frac{1}{M}\sum_{i=1}^{M}\beta_i \tag{17}$$

$$\mathrm{SE}(\hat{\beta}) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\beta_i - \hat{\beta})^2} \tag{18}$$

APC-Specific 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. Suppose that λ_A, λ_P, and λ_C correspond to the ratio of the error variance to the variances of the age, period, cohort coefficients. For example, let $\lambda_A = \sigma^2 / \sigma_A^2$, with a similar expression applying to the period and cohort effect shrinkage parameters. In other words, the age, period, cohort coefficients are permitted to have distinct variances σ_A^2, σ_P^2, and σ_C^2. The exchangeable ridge priors for the age, period, cohort coefficients are then specified as

$$\alpha_i^* \sim N(0, \sigma^2 / \lambda_A), \tag{19}$$

$$\beta_j^* \sim N(0, \sigma^2 / \lambda_P), \tag{20}$$

$$\gamma_k^* \sim N(0, \sigma^2 / \lambda_C), \tag{21}$$

and the priors used for λ_A, λ_P, and λ_C are

$$\lambda_j \sim gamma(1,1), \qquad j \in \{A, P\} \tag{22}$$

and

$$\lambda_C \sim gamma(1,100). \tag{23}$$

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 APC-specific shrinkage parameters in model (b) are defined as:

$$\lambda_j \sim gamma(1,1), \qquad j \in \{A,P,C\} \tag{24}$$

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 rjags package (Plummer, 2014). The first 10,000 iterations were used as burn-in and all parameter estimation was based on 50,000 posterior draws.

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 generate 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 risk of cervical cancer among women aged 30 to 34 is insignificant given that 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 the conventional ridge estimator and the GCV plot is shown in Figure 1, which illustrates that the minimum value of GCV is about 0.017 corresponding to $\lambda = 0.050$. The posterior mean of λ from the Bayesian implementation is similar in magnitude, with $\lambda = 0.078$. The 95% credible interval indicates the true mean of λ is within the interval (0.041, 0.132) with 95% probability. In this case, the conventional ridge parameter ($\lambda = 0.050$) is within the 95% credible interval.

Figure 2 presents the graphical convergence diagnosis of the MCMC algorithm for selected parameters. For each selected parameter, the trace plot shows the posterior sample values of that parameter during the runtime of the chain. The marginal density plot is the smooth 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 that each chain is mixing well. 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.

Results from Bayesian model (a) with different ridge priors for age, period, and cohort effects are shown in Table 3. The estimated posterior means of the age, period, cohort effects are similar to those from the model with a common prior for the APC effects. However, we see that each coefficient's vector is now subject to differential shrinkage toward zero, with the period effects being most affected. To better illustrate the APC trends, Figure 4 shows the age, period, and cohort trends from the Bayesian models with distinct specifications for the ridge priors. The solid line represents the model with a common prior for the ridge parameter, which is distributed as gamma (1, 1). The dashed line represents Bayesian model (a) specifying different priors for the ridge penalties with λ_A and λ_P distributed as gamma(1,1) while λ_C is distributed as gamma(1,100). The dotted line represents Bayesian model (b) using a gamma(1,1) prior for λ_A, λ_P *and* λ_C. 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 similar to those from model

(a), whereas the cohort trend differs from that of model (a). In particular, 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) based on results 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) due to greater shrinkage of the cohort parameter vector towards 0.

Discussion

The age-period-cohort accounting model serves as a critical framework to study temporal change in phenomena such as mortality, fertility, and disease rates. The importance of separating age, period, and cohort effects for time-specific phenomena poses a challenge in estimating 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, the ridge estimator, the partial least squares approach of Tu et al. (2011;2013), the maximum-entropy approach of Browning et al. (2013). These approaches tend to agree on solutions more often than not. This paper builds upon the traditional ridge estimator but approaches the identification problem from the Bayesian interpretation of ridge estimation. In so doing, it avoids the inherent limitations related to solving systems of linear equations in favor of iterated conditional sampling.

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 a-priori criteria based on cross-validation. Although the optimal ridge estimator enables the model to be identified, it does not have any meaningful interpretation. For the Bayesian ridge model, there is no need to assign a single value to the ridge parameter because it is considered a random variable and is able to incorporate uncertainties about the age, period, and cohort effects for a specific study and simultaneously take advantage of information from existing research. We can obtain a series of summary statistics from the posterior samples of the ridge parameter. Further, the random property of the ridge parameter 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 considerable 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. Given the relative small sample size, 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 with respect to 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 often leads to Bayesian posterior means close to the maximum likelihood estimates (Congdon, 2006).

To conclude, the Bayesian ridge model provides an effective way to cope with the identification problem inherent in the classical age-period-cohort accounting model. Although 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 may render 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.

Hoerl, A. E. (1962). Application of ridge analysis to regression problems, *Chemical Engineering Progress*, 58: 54-59.

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: biased estimation for non-orthoganal problems," *Technometrics,* 12: 55-67

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

Kupper, J. J., and Janis J. M. (1980). The multiple classification model in age, period, and cohort analysis: theoretical considerations. Institute of Statistics Mimeo No. 1311 1980; Department of Biostatistics University of North Carolina, Chapel Hill.

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12, 591-612. doi:10.2307/1267205

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/b0-08-043076-7/00401-0

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/DSC-2003/

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

Tu, Y. K., Smith, G. D. and Gilthorpe, M. S. (2011). A new approach to age-period-cohort analysis using partial least squares: the trend in blood pressure in Glasgow Alumni Cohort. *Plos One*. doi:1371/journal.pone. 001901

Tu, Y. K., Kramer N, and Lee, W. (2013). Addressing the identification problem in age-period-cohort analysis: a tutorial on the use of partial least squares and principle components analysis. *Epidemiology*, 23:583-593.

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.0081-1750.2004.00148.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

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

Table 2: Alternative Estimates of Age, Period, and Cohort 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)
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)

(*Table 2 Continued*)

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	-	0.05	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)

Table 3: Estimates from Bayesian Model (a) with Different Ridge Priors for the 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
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

(*Table 3 Continued*)

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

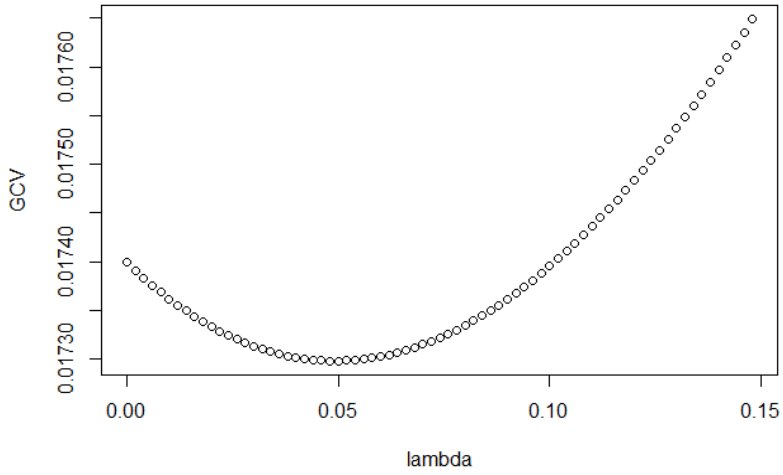


Figure 1: Selection of Lambda for Ridge Estimator via GCV

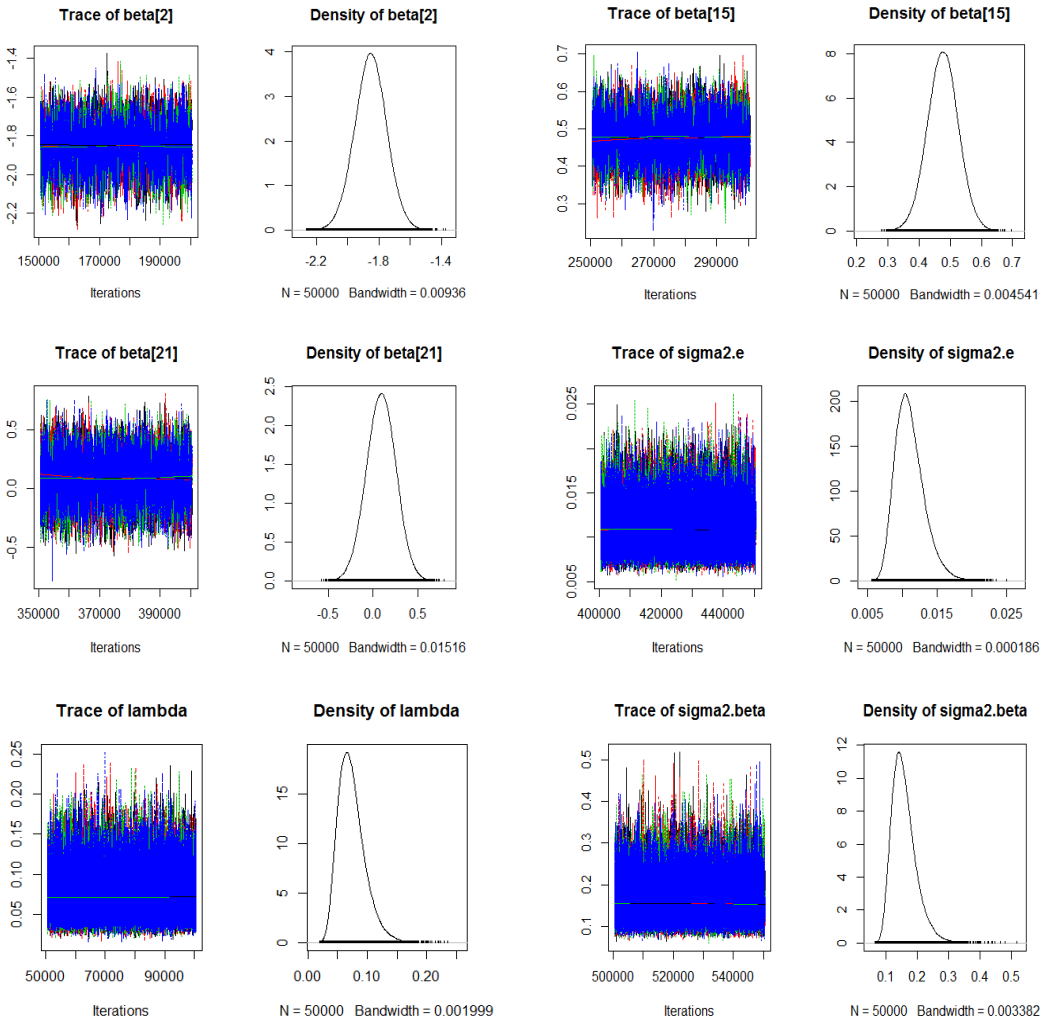


Figure 2: Trace Plots and Density Plots for the Posterior Samples for Selected Parameters.

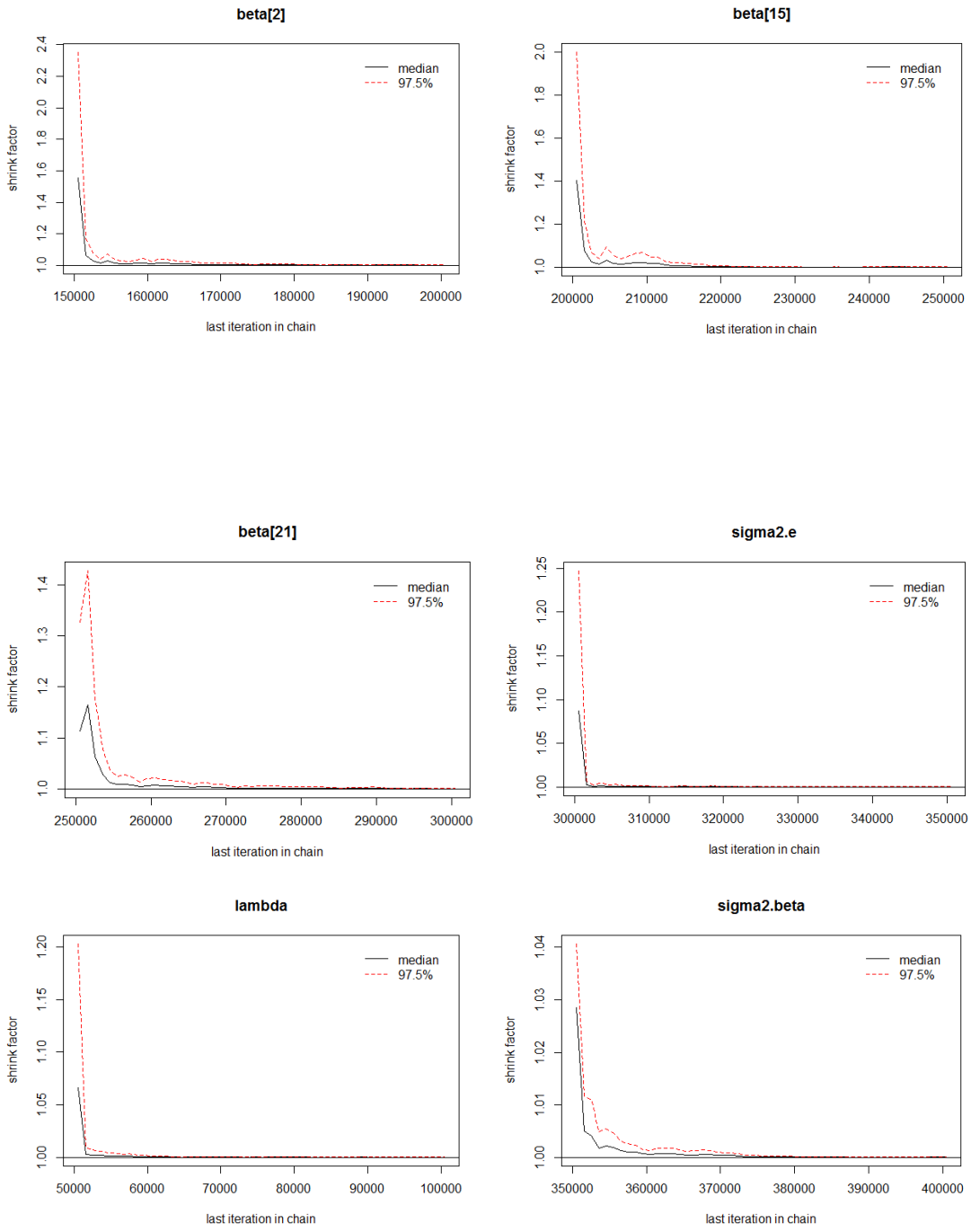
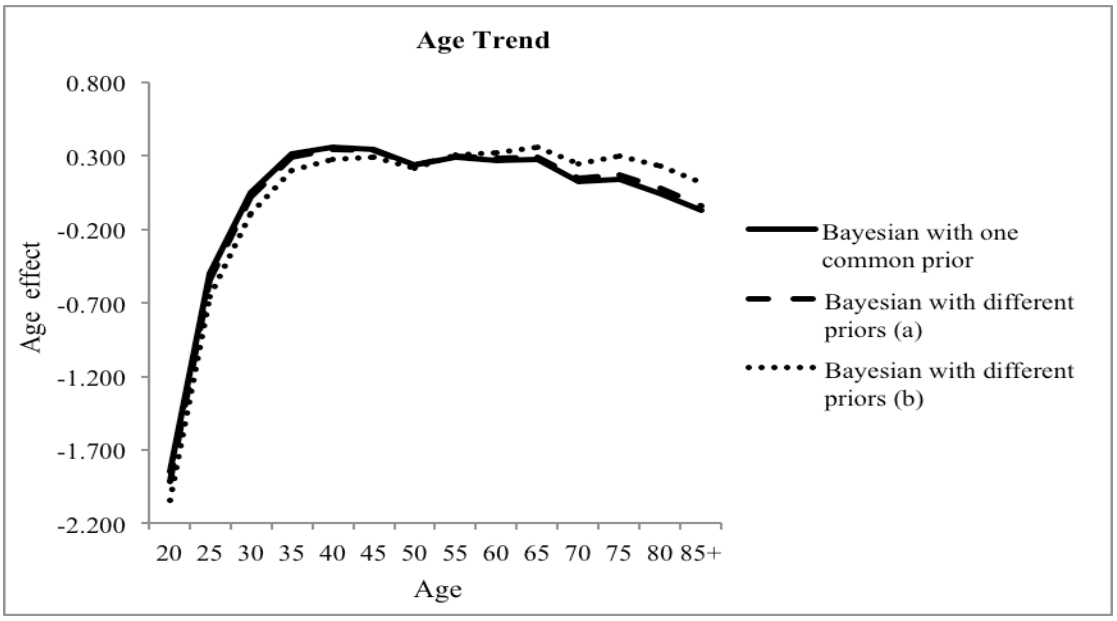


Figure 3: Plots of Gelman-Rubin's Diagnostic for Selected Parameters.



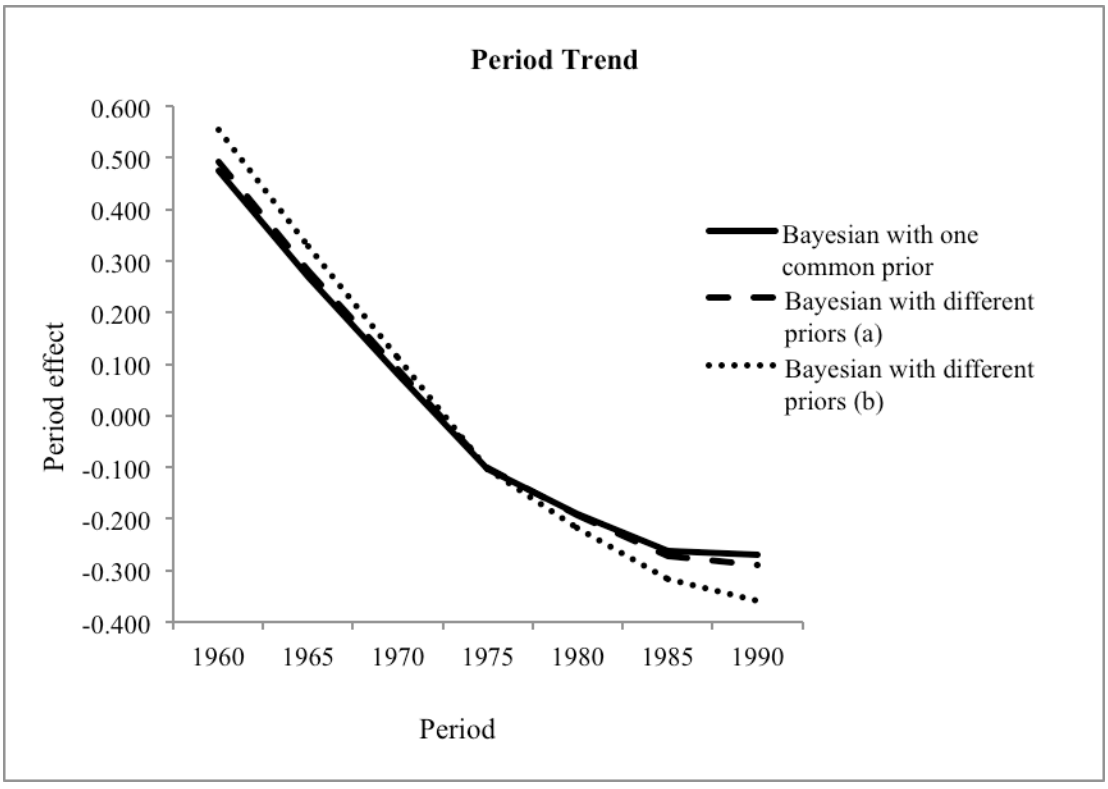
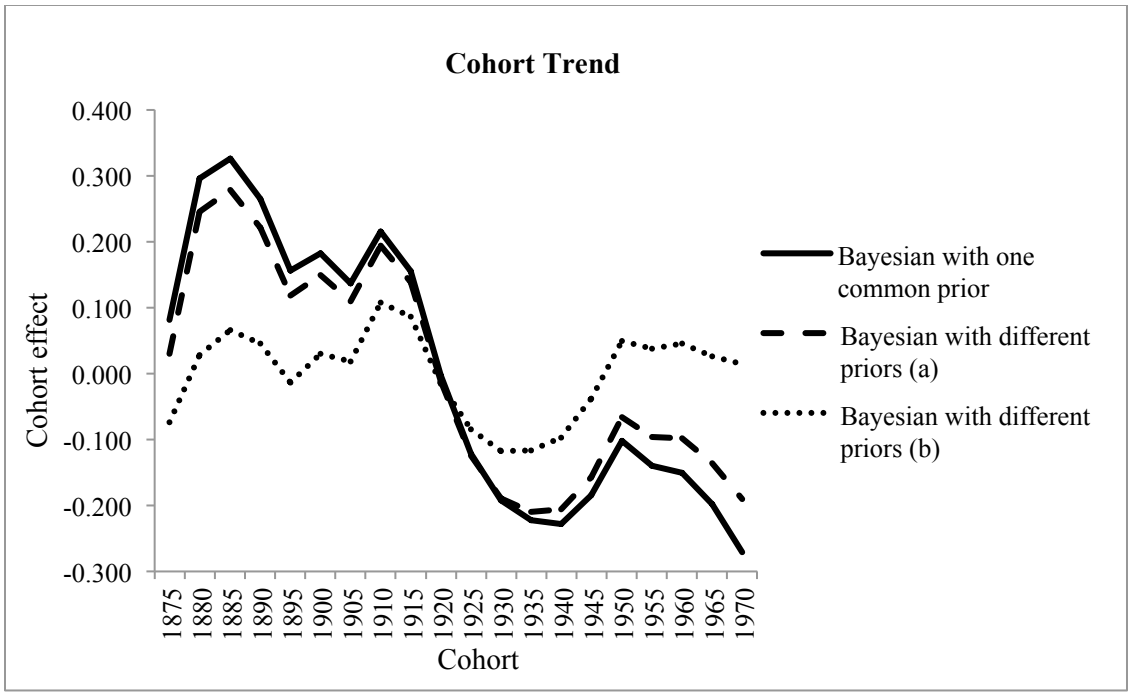


Figure 4 Bayesian Models for Age, Period, and Cohort Trends on Cervical Cancer Incidence Rates in Ontario Women.



(*Figure 4 Continued*)