Solving the Identifiability Problem with the Lasso Regularization in Age-Period-Cohort Analysis

Beverly Fu

Okemos High School, Okemos, Michigan 48864

Email: fubeverly99@gmail.com

Abstract

The age-period-cohort analysis (APC) has been a popular tool to analyze data in demography, economics, marketing research, public health and social studies. However, the multiple estimators of the APC multiple classification model result in indetermination of the model parameters, leading to the difficult identifiability problem. In this paper, I apply the Lasso regularization method to the APC models and demonstrate that it leads to a resolution of the identifiability problem and yields sensible trend estimation in two studies, a study of US female breast cancer mortality rates, and a study of homicide arrest rates.

1. Introduction

The age-period-cohort (APC) analysis has been a popular tool to analyze data displayed in a rectangular table with certain numbers of rows and columns, where each cell of the table contains one data point, often a disease rate or a social event rate, such as breast cancer mortality rate or homicide arrest rate. The rows of the table are often consecutive age groups, and the columns are consecutive periods (calendar years). If the age groups and periods are of the same time span, the diagonals of the table represent birth cohorts or generations. Data in such a table are often analyzed with the analysis of variance (ANOVA) models, such as the two-way ANOVA models with fixed age and period effects. However, if the two-way ANOVA model cannot achieve a good fit, the birth cohort effects may also be considered in the model. Table 1 displays the US female breast cancer mortality rates during 1980 – 2009 in 13 age groups and 6 periods with 5 year span in each age group and period. Since the age, period and cohort satisfy a linear relationship **Period – Age = Cohort**, the APC multiple classification model with fixed effects of the age, period and cohort (1) suffers from an identifiability problem [Kupper et al 1985], where multiple estimators fit the model equally well.

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

where Y_{ij} is the log-transformed rate in the *i*-th age group and *j*-th period with $i = 1, \ldots, a$, and $j = 1, \ldots, p$. μ is the model intercept, and α_i β_j, and γ_k are the effects of the *i*-th age group, *j*-th period and *k*-th cohort on the diagonal, respectively. The cohort index $k = 1, \ldots, a + p - 1$. ε_{ij} is the random error term with mean 0 and common variance σ^2. Model (1) needs side conditions on the parameters α_i, β_j and γ_k as a special fixed effect ANOVA,

either by the parameter centralization $\sum_{i=1}^{a}\alpha_i = \sum_{j=1}^{p}\beta_j = \sum_{k=1}^{a+p-1}\gamma_k = 0$, or by specifying reference levels, such as $\alpha_1 = \beta_1 = \gamma_1 = 0$.

2. The Identifiability Problem

The identifiability problem can be explained with ease by rewriting model (1) in a matrix form

$$\boldsymbol{Y} = X\boldsymbol{b} + \boldsymbol{\varepsilon}\,, \qquad (2)$$

where $\boldsymbol{Y}, \boldsymbol{b}, \boldsymbol{\varepsilon}$ are the vectors of the responses, the model parameters and the random errors, respectively. The matrix X is a singular design matrix of the ANOVA model with 1-less than its full rank [Kupper et al 1985], even after specifying the above side conditions. Hence, the usual least-squares method yields multiple estimators because the matrix X^TX has multiple generalized inverse matrices, leading to multiple sets of parameter estimates and the indetermination of the parameters. Figure 1 illustrates the identifiability problem through multiple sets of parameter estimates of the US female breast cancer mortality data in Table 1. The curves in the same color present the parameter estimates by the same estimator, and curves in different colors are estimates with different estimators. It is shown that it would be difficult to determine which set of estimates provides accurate estimation of the parameters and thus the trends in the age, period and cohort.

Given the multiple sets of the parameter estimates, interpretation of the varying trends in the age, period and cohort is impossible. Hence in practice, an extra constraint on the parameters is often specified based on the investigator's prior knowledge about the disease or event under investigation. For example, the first two age effects are often assumed to be identical based on the assumption that in the early ages, disease mortality does not vary largely and thus it is hoped that such an assumption yields reasonable mortality trends. However, as pointed out in [Kupper et al 1985], often seemingly reasonable assumptions lead to insensible trend estimation, because the extra constraint may not be satisfied by the true parameters, resulting in biased estimation. Another practical approach is to identify the estimable functions that do not vary with the extra constraint, thus leading to invariant estimation of certain linear combinations of the parameters – the estimable functions. However, it was observed that the nonlinear components of the trends, such as the curvature or higher order characteristics are invariant, but the overall slope varies with the constraint and thus may not be estimable [Kupper et al 1985, Holford, 1991]. Hence the identifiability problem was deemed unsolvable during the past 40 years. Consequently, many studies with data in the APC format cannot be analyzed with accurate estimation of the parameters and the trends in the age, period and cohort.

In recent years, a number of methods have been studied aiming to address the identifiability problem. A smoothing approach was studied by applying smoothing to the parameter estimates [Heuer 1997, Fu 2008], leading to full determination of the trends. Another promising approach is the intrinsic estimator [Fu 2000], which is often computed via the principal component analysis by taking the eigenvectors of all nonzero eigenvalues of the singular matrix X^TX as principal components, estimating the effects of the principal components and further transforming back to the original scale of the age, period and

cohort. It has been shown that this method yields robust estimation for finite samples and consistent estimation for diverging samples with many desirable properties [Fu 2016].

The Lasso regularization has been studied extensively [Tibshirani 1996, Knight and Fu 2000, Zhao and Yu 2006, Hastie et al 2008], not only because it provides an automatic approach to variable selection, which is computationally efficient for models with a moderate or large number of covariates [Fu 1998, Efron et al 2004], and also possesses the oracle properties for variable selection [Fan and Li 2001, Zou 2006], which ensures the correctness of variable selection. Motivated by the desirable properties of the Lasso, I apply the Lasso regularization to the APC models.

The Lasso regularization takes the following approach to the parameter estimation in linear models. It minimizes the penalized residual sum of squares with an L_1 penalty on the L parameters $\beta_1, \dots, \beta_L$.

$$\min_{\beta_1,\dots,\beta_L} \sum_{i=1}^{n} [y_i - (\mu + x_{i1}\beta_1 + \cdots + x_{iL}\beta_L)]^2 \; + \; \lambda \sum_{l=1}^{L} |\beta_l| \,, \quad \text{for } \lambda \geq 0 \,.$$

For large enough $\lambda > 0$, it yields some parameter estimates $\hat{\beta}_l = 0$, thus achieving variable selection of the remaining nonzero ones. Often the optimal value of λ is selected with the Bayesian information criterion (BIC) [Schwarz 1978] below.

$$BIC(\lambda) = \sum_{i=1}^{n} [y_i - \left(\mu + x_{i1}\hat{\beta}_1 + \cdots + x_{iL}\hat{\beta}_L\right)]^2 \; + \; L \, log(n) \,.$$

3. Lasso regularization approach to the identifiability problem

Since the aim of the APC models is to determine the relative scale of the parameter estimates and to achieve further trend estimation, but not variable selection to determine which effects are zero, the naïve approach of direct application of the Lasso regularization to the APC models may not serve the purpose. Instead, I apply the Lasso regularization to the principal components [Jolliffe 1986] and study which principal components the Lasso selects and which ones it deselects. Let x_{ijl} be the loading of the l-th principal component on the observation y_{ij}, and let ϑ_l, $l = 1, \dots, L$ be the effects of the principal components. One may write the residual sum of squares of the PCA model and its Lasso regularization as

$$\min_{\vartheta_1,\dots,\vartheta_L} \textstyle\sum_{i,j}^{a,p} [y_{ij} - (\mu + x_{ij1}\vartheta_1 + \cdots + x_{ijL}\vartheta_L)]^2 + \;\; \lambda \; \sum_{l=1}^{L} |\vartheta_l| \,. \qquad (3)$$

I demonstrate that this approach of the Lasso regularization to the APC models yields accurate estimation using the US female breast cancer mortality data and the homicide arrest rate data in [O'Brien 2000] by comparing the trend estimates in the age, period and cohort between the Lasso regularization method and the intrinsic estimator method. Figures 2 and 3 present the trends in the age, period and cohort for the cancer mortality data and the homicide arrest rate data, respectively. It is shown that the Lasso regularization yields trend estimates within the 95% confidence interval of the intrinsic estimator, indicating the two methods yield very close estimates. It is shown in Figure 2 that the breast cancer mortality sharply increases in the early ages from 20 to 40, then gradually slows down and plateaus around age 80-84, but never decreases with age. The mortality also increases in period, at a much faster pace from 1980 to mid 1990s than later

from 1995 to the mid 2000s. The cohort presents a constant decreasing trend from 1900 to the 1980, followed by a flat trend afterwards. Notice that although the period presents an increasing trend, its scale is much smaller than the cohort trend and the age trend, indicating that one needs to emphasize the age effect and the cohort effect in order to more efficiently lower breast cancer mortality. The increasing age trend can be well explained by cancer epidemiology, while the decreasing cohort trend may be explained by the improved education about breast cancer risk. The slightly increasing period trend from 1995 to 2000 indicates more effort is needed in fighting breast cancer mortality. In Figure 3, it is shown that the homicide arrest rate increased sharply from late teens to early twenties, then goes down sharply till age 45, which reflects that seniors are less aggressive in committing violent crime, such as homicide. The slowly decreasing cohort trend from 1910 to 1955 and the sharply increasing trend from 1960 to 1980 indicates younger generations born after 1960 are more aggressive than the older generations. The period trend seems to show an inverse relationship between the homicide arrest rate and the economy. The increasing trend of homicide arrest rate between 1960 to 1970 and the relative high level between 1970 to 1985 seem to be associated with poor economy during those periods of time, while the recent downtrend of homicide arrest rate from 1990 to 2010 may be explained by the improved economy.

Overall, the Lasso regularization method yields accurate estimation of the trends in the age, period and cohort, leading to sensible interpretation for the above two studies.

4. Conclusion

The age-period-cohort models have broad applications in demography, economics, marketing research, public health studies and social sciences. The difficult identifiability problem has been studied in the literature during the past 40 years. Novel approaches are needed to resolve the identifiability problem. In this paper, I apply the powerful Lasso regularization method to the principal components of the singular design matrix of the APC models. It yields accurate estimation of the age, period and cohort effects, and leads to sensible trend estimation. The parameter estimates are shown to be within the 95% confidence interval of the intrinsic estimator – a proven approach for APC analysis. It is hoped that the Lasso regularization method provides another alternative method to address the identifiability problem.

References:

Efron, B. Hastie, T. Johnstone, I. and Tibshirani, R. (2004) Least angle regression, *Annals of Statistics*, 32 (2): 407-499.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of American Statistical Association*, 96; 456: 1348-1360.

Fu WJ. (1998) Penalized regressions: the Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, 7 (3):397-416.

Fu, WJ. (2000) Ridge estimator in singular design with application to age-period-cohort analysis of disease rates, *Communications in Statistics—Theory and Method*, 29, 263–278.

Fu, WJ (2008) A smoothing cohort model in age–period–cohort analysis with applications to homicide arrest rates and lung cancer mortality rates, *Sociological Methods and Research*, 36 (3): 327-361.

Fu, WJ. (2016) Constrained estimator and consistency of a regression model on a Lexis diagram, *Journal of the American Statistical Association*, 111 (513):180-199.

Hastie, T. Tibshirani, R. and Friedman, J. (2008) *The Elements of Statistical Learning, Data Mining, Inference and Prediction*, Springer, New York.

Heuer, C. (1997) Modeling of time trends and interactions in vital rates using restricted regression splines, *Biometrics*, 53, 161-177.

Holford, TR. (1991) Understanding the effects of age, period, and cohort on incidence and mortality rates. *Annual Review of Public Health* 12:425-457.

Jolliffe, IT. (1986) *Principal Component Analysis*, Springer, New York.

Knight, K. and Fu, W. (2000) Asymptotics for lasso-type estimators, *Annals of Statistics*, 28, 5: 1356-1378.

Kupper, LL. Janis, JM. Karmous, A. and Greenberg, BG. (1985) Statistical age-period-cohort analysis: a review and critique, *Journal of Chronic Disease*, 38, 811-830.

O'Brien, R. (2000) Age, period, cohort characteristic model, *Social Science Research*, 29,1:123-139.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics.* 6(2), 461-464.

Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society*, Series B, 58, 267–288.

Zhao, P. and Yu, B. (2006) On model selection consistency of Lasso, *Journal of Machine Learning Research* 7: 2541-2563.

Zou, H. (2006) The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418-1429.

Table 1. Breast Cancer Mortality Rate (10-5 person-year) among US Females

Age	1980-84	1985-89	1990-94	1995-99	2000-04	2005-09
20-24	0.1	0.1	0.1	0.1	0.1	0.1
25-29	1.3	1.2	1.1	1.0	0.8	0.7
30-34	5.5	5.0	4.3	3.9	3.5	2.9
35-39	12.8	13.0	11.1	9.6	8.4	7.2
40-44	23.3	23.5	21.8	18.0	15.6	13.8
45-49	37.1	37.8	35.7	30.2	25.1	22.1
50-54	56.8	55.1	51.8	44.8	38.1	32.4
55-59	75.1	73.0	66.2	58.1	51.4	44.6
60-64	87.4	90.9	82.8	70.8	63.8	58.7
65-69	99.7	102.3	100.6	85.4	76.0	70.0
70-74	110.2	118.4	116.5	105.6	93.1	83.2
75-79	122.8	130.4	135.6	122.3	113.5	103.9
80-84	141.1	150.9	156.2	148.0	139.3	131.7

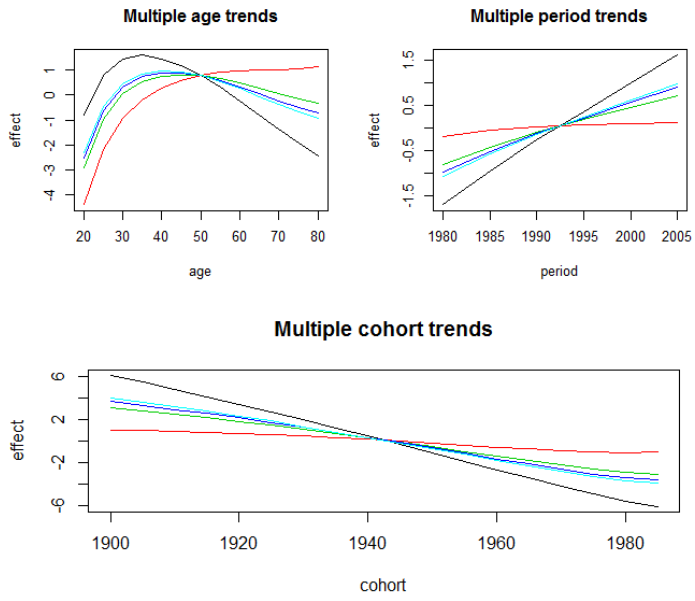

Figure 1. Illustration of the identifiability problem using US female breast cancer mortality data. The curves in the same color are estimates of age, period and cohort effects in the same set by one estimator. Curves in different colors present estimates with different estimators.

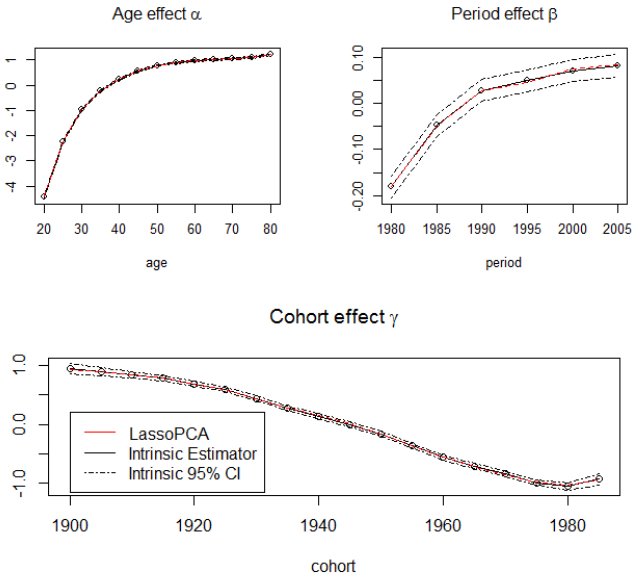

Figure 2. Trend comparison by the Lasso estimator on the principal components (LassoPCA) and the intrinsic estimator in analyzing the US female breast cancer mortality rate data. The Lasso estimates lie in the 95% confidence interval of the intrinsic estimator.

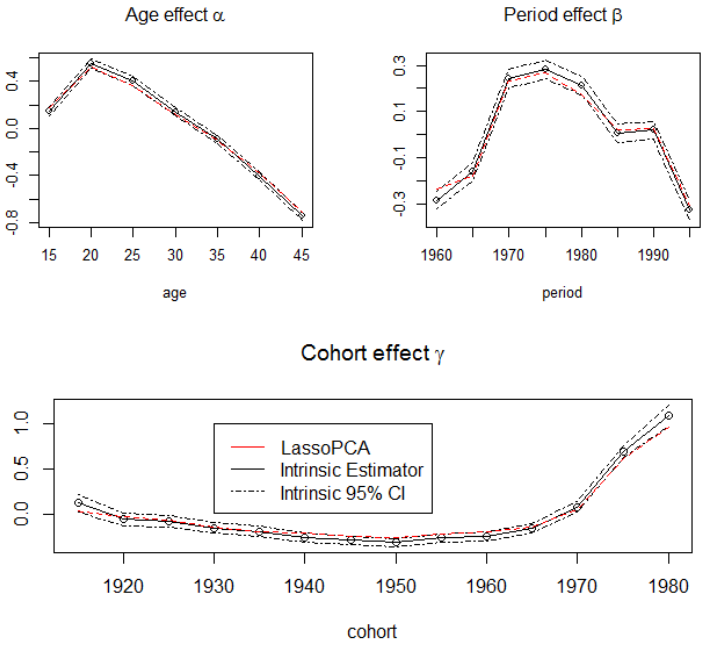

Figure 3. Trend comparison by the Lasso estimator on the principal components (LassoPCA) and the intrinsic estimator in analyzing the homicide arrest rate data. The Lasso estimates lie in the 95% confidence interval of the intrinsic estimator.