Title	Age-period-cohort analysis for trends in body mass index in Ireland
Authors	Jiang, Tao;Gilthorpe, Mark S.;Shiely, Frances;Harrington, Janas M.;Perry, Ivan J.;Kelleher, Cecily C.;Tu, Yu-Kang
Publication date	2013-09-25
Original Citation	JIANG, T., GILTHORPE, M. S., SHIELY, F., HARRINGTON, J. M., PERRY, I. J., KELLEHER, C. C. & TU, Y.-K. 2013. Age-period-cohort analysis for trends in body mass index in Ireland. BMC Public Health, 13:889, 1-7. http://dx.doi.org/10.1186/1471-2458-13-889
Type of publication	Article (peer-reviewed)
Link to publisher's version	10.1186/1471-2458-13-889
Rights	© Jiang et al.; licensee BioMed Central Ltd. 2013. This article is published under license to BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. - creativecommons.org/licenses/by/2.0
Download date	2024-03-28 13:24:08
Item downloaded from	https://hdl.handle.net/10468/2265

ADDITIONAL FILES

Introduction to PLS

The fundamental problem with APC analysis is perfect collinearity amongst the variables age, period and cohort. The direct consequence of this is that the data matrix is not of full rank, i.e. it is singular and therefore not invertible. This means that more commonly used regression models such as ordinary least squares regression, does not work. Partial least squares (PLS) regression is not affected by the singularity of the data matrix hence it can be applied to APC analysis.

PLS attempts to extract weighted components **t** of the explanatory variables, maximising the covariance between the response variable and **t**. For example, for the variables Age, Period and Cohort: $\mathbf{t}_i = w_{i1}\text{Age} + w_{i2}\text{Period} + w_{i3}\text{Cohort}$, where the weights w_{ij} are subject to the constraints that $\sum_{j=1}^{3} w_{ij}^2 = 1$ and the components are mutually orthogonal (i.e. the pair-wise correlations are zero).

This differs from principal components analysis (PCA), which extracts components based on the amount of variance each component explains within the data matrix. Therefore, given a set of explanatory variables, principle components analysis will always extract the same components regardless of the response variable.

The maximum number of PLS (and PCA) components that can be extracted is equal to the rank of the data matrix. In our case, although we have three variables (age, period and cohort), we only have two degrees of freedom; hence our data matrix is rank two. This is a direct result of perfect collinearity amongst the variables. Therefore, only two components can be extracted.

Having obtained the coefficients for the PLS components, it is necessary to recover the coefficients for the original predictor variables. For example, suppose we take the full 2-component model in our analysis of BMI, we would obtain the following:

$$BMI = \beta_1 \mathbf{t}_1 + \beta_2 \mathbf{t}_2$$

$$= \beta_1 (w_{11} Age + w_{12} Period + w_{13} Cohort) + \beta_2 (w_{21} Age + w_{22} Period + w_{23} Cohort)$$
$$= (\beta_1 w_{11} + \beta_2 w_{21}) Age + (\beta_1 w_{12} + \beta_2 w_{22}) Period + (\beta_1 w_{13} + \beta_2 w_{23}) Cohort$$

with β_1 and β_2 being the estimated coefficients for the 1st and 2nd PLS components respectively. We can therefore, extract the separate regression coefficients for Age, for instance, to be $\beta_1 w_{11} + \beta_2 w_{21}$ via simple algebraic manipulation.

Further Properties

Since we have three variables but only two degrees of freedom, the two extracted components for both principal components analysis and PLS span the same space which is why the 2-component model for PLS gives the same output as that for principal components analysis. Although the extracted components from both methods are not the same, since they span the same space, the resulting coefficients for the original covariates will be identical.

Given we are trying to estimate three separate coefficients but we only have two degrees of freedom, a constraint is necessary. PCA (and PLS) implicitly applies the constraint that the sum of two coefficient estimates is equal to the third.

Theorem 1: In a classic APC model where the only covariates are age, cohort and period, PCA will obtain parameter estimates λ_1, λ_2 and λ_3 for age, cohort and period respectively such that $\lambda_1 + \lambda_2 = \lambda_3$.

Proof: Let the PCA algorithm produce two components $\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ and $\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$ with coefficients β_1 and β_2 respectively. Then we have $\begin{aligned} \lambda_1 &= \beta_1 x_1 + \beta_2 y_1 \\ \lambda_2 &= \beta_1 x_2 + \beta_2 y_2 \\ \lambda_3 &= \beta_1 x_3 + \beta_2 y_3 \end{aligned}$ hence it is sufficient to prove $x_1 + x_2 = x_3$ and $y_1 + y_2 = y_3$.

Let A be the variance in Age, C be the variance in Cohort and B be the covariance between Age and Cohort, the covariance matrix for our data will be $\begin{pmatrix} A & B & A+B \\ B & C & B+C \\ A+B & B+C & A+2B+C \end{pmatrix}$. The components extracted by PCA are the eigenvectors ordered in order of size of their corresponding eigenvalues.

Consider arbitrary eigenvector $\begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}$ with corresponding eigenvalue λ

$$\begin{pmatrix} A & B & A+B \\ B & C & B+C \\ A+B & B+C & A+2B+C \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \lambda \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}$$

Hence $\begin{cases} Az_1 + Bz_2 + (A+B)z_3 = \lambda z_1 \\ Bz_1 + Cz_2 + (B+C)z_3 = \lambda z_2 \\ (A+B)z_1 + (B+C)z_2 + (A+2B+C)z_3 = \lambda z_3 \end{cases}$

Summing the first two equations will give you the left hand side of the third, hence we have

$\lambda(z_1 + z_2) = \lambda z_3$

Therefore #

$z_1 + z_2 = z_3$ or $\lambda = 0$

Since PCA will not extract the component with the zero eigenvalue, all extracted components have the property that $z_1 + z_2 = z_3$ hence we have $x_1 + x_2 = x_3$ and $y_1 + y_2 = y_3$. [*QED*]

PLS is similar to PCA, only that it considers the covariance between the covariates and the outcome. As a result, it inherits the same constraint on the coefficients as does PCA. This constraint is directly inherited from the mathematical relationship amongst the variables; hence it is a reasonable constraint to impose.