



*The World's Largest Open Access Agricultural & Applied Economics Digital Library*

**This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.**

**Help ensure our sustainability.**

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

[aesearch@umn.edu](mailto:aesearch@umn.edu)

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

*No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.*

## Age–period–cohort modeling

Mark J. Rutherford  
Department of Health Sciences  
University of Leicester, UK  
mjr40@le.ac.uk

Paul C. Lambert  
Department of Health Sciences  
University of Leicester, UK  
paul.lambert@le.ac.uk

John R. Thompson  
Department of Health Sciences  
University of Leicester, UK  
john.thompson@le.ac.uk

**Abstract.** Age–period–cohort models provide a useful method for modeling incidence and mortality rates. It is well known that age–period–cohort models suffer from an identifiability problem due to the exact relationship between the variables (cohort = period – age). In 2007, Carstensen published an article advocating the use of an analysis that models age, period, and cohort as continuous variables through the use of spline functions (Carstensen, 2007, *Statistics in Medicine* 26: 3018–3045). Carstensen implemented his method for age–period–cohort models in the Epi package for R. In this article, a new command is introduced, `apcfit`, that performs the methods in Stata. The identifiability problem is overcome by forcing constraints on either the period or cohort effects. The use of the command is illustrated through an example relating to the incidence of colon cancer in Finland. The example shows how to include covariates in the analysis.

**Keywords:** st0211, `apcfit`, `poprisktime`, age–period–cohort models, incidence rates, mortality rates, Lexis diagrams

## 1 Introduction

An age–period–cohort (APC) model provides a modeling tool that can be used to summarize the information that is routinely collected by cancer registries and registries for other diseases. Classically, APC models fit the effects of age, period, and cohort as factors. It has become common practice to report the age and period effects in 5-year intervals, resulting in ten-year overlapping intervals for the relevant cohorts. However, through the increase in computer power and the use of restricted cubic (natural) splines (Durrleman and Simon 1989), it has been shown that it is possible to analyze the effects as continuous variables (Carstensen 2007). This article builds on the work of Carstensen and explains how the method and the extensions have been made available in Stata.

APC models suffer from an identifiability problem. The date of birth can be calculated directly from the age at diagnosis and the date of diagnosis. If fitted directly in a generalized linear model (GLM) this leads to overparameterization and, consequently, the exclusion of one of the terms. It is therefore necessary to fit constraints to the model to extract identifiable answers for each of the parameters. This step is

required because each of the components of the model provides different insights into the trends of the disease over time. The age effect provides information on the rates of disease in terms of different age groups. The period effect can highlight changes in treatment that could affect all ages simultaneously. The cohort effects are associated with long-term exposures, with different generations being exposed to different risks (Robertson, Gandini, and Boyle 1999).

Other Stata commands are available that apply constraints to overcome the identifiability issue for APC models. The `apc_ie` and `apc_cglim` commands are available to download from the `apc` package, which can be found via the Statistical Software Components (SSC) archive. The `apc_cglim` command uses a single equality constraint. The age, period, and cohort terms are fitted as factors, and a constraint that sets two of the categories from different components equal to one another is applied to overcome the lack of identifiability (Yang, Fu, and Land 2004). The `apc_ie` command uses the intrinsic estimator, which employs a principal components regression to arrive at the constrained estimates for the age, period, and cohort effects. The two approaches are described in detail and compared by Yang, Fu, and Land (2004). A good overview of techniques available to carry out APC models is given by Land (2008).

The `apcfit` command differs from the existing approaches by using restricted cubic splines to model the three variables. `apcfit` gives estimates for the three effects (age, period, and cohort) that can then be combined to give the predicted rates. The estimates for the three components are also interpretable individually and can be plotted to show incidence and mortality trends over the different time scales. The graphs that can be produced provide a clear and simple depiction of the data. A further benefit of `apcfit` is the potential for further modeling to investigate the effect of covariates.

## 2 Methods

The general APC model can be described using the following equation:

$$\ln \{\lambda(\mathbf{a}, \mathbf{p})\} = f(\mathbf{a}) + g(\mathbf{p}) + h(\mathbf{c})$$

where  $f()$ ,  $g()$ , and  $h()$  are functions,  $\lambda$  refers to the rate,  $\mathbf{a}$  refers to the age variable,  $\mathbf{p}$  refers to the period variable, and  $\mathbf{c}$  refers to the cohort variable. This model can be used to predict the incidence or mortality rate for any given combination of age and period. However, because of the direct relationship between the terms,  $\mathbf{c} = \mathbf{p} - \mathbf{a}$ , the components of this model cannot be uniquely determined. The model needs to be constrained in some way to ensure that three functions showing the age, period, and cohort effects can be extracted. Carstensen (2007) details a method that allows this to be achieved.

In essence, the method proposed by Carstensen uses restricted cubic (natural) splines for the age, period, and cohort terms within a GLM framework with a Poisson family error structure, a log link function, and an offset of  $\log(\text{person risk-time})$ . However, to overcome the identifiability problem, transformations are made to the spline basis vectors for the period and cohort effects using matrix transformations.

Having performed the matrix transformations, a GLM is fitted within Stata using the adjusted spline basis vectors. Using this GLM as a foundation, it is possible to extend the analysis to include covariates. The data required to do this have observations for each unique age–period combination for every level of the covariate of interest. It is then possible to adjust for the effect of the covariate by including the term in the GLM. It is also possible to include interaction terms between the covariate and age, period, and cohort.

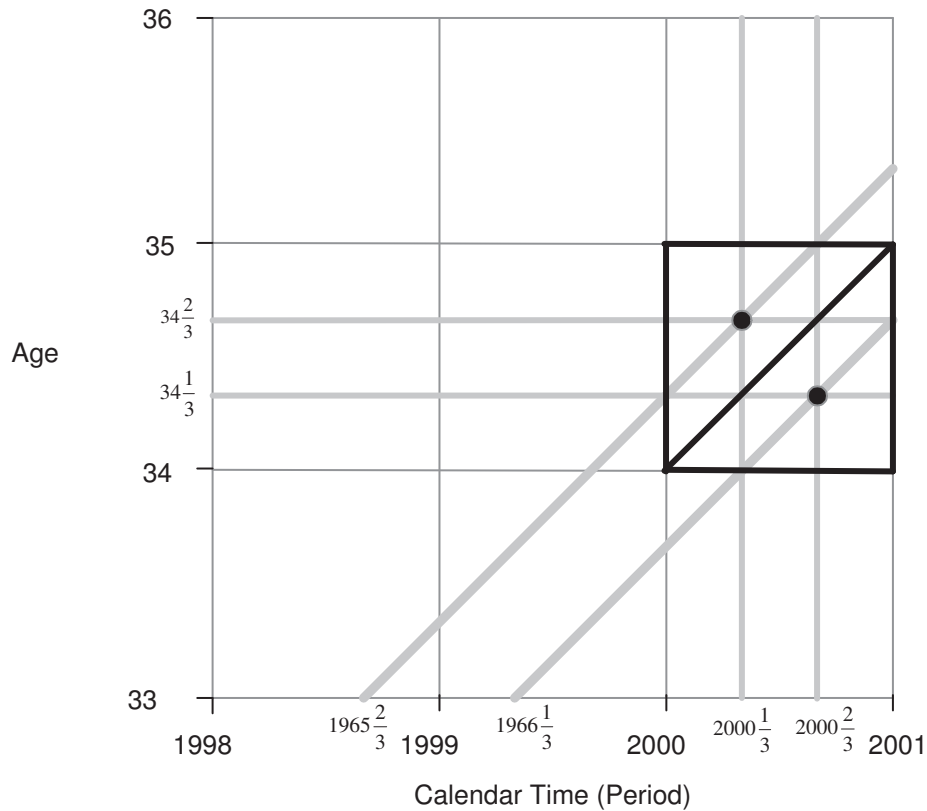


Figure 1. Snapshot of a Lexis diagram indicating the reasoning behind the use of the average values that are offset by  $1/6$  for the triangular subsets (compared with the average values for the squares of a Lexis diagram)

## 2.1 Form of the data

Cancer registries and other disease registries typically collect data that could be summarized in a Lexis diagram. A Lexis diagram summarizes a population's disease status

over calendar time against age. For example, Lexis diagrams can be used to depict the number of new cases of a disease by category for age and period of diagnosis (see figure 1). A Lexis diagram is usually split into five-year intervals for period and age; we suggest that yearly intervals should be used. The cells of a Lexis diagram can also be further subdivided by cohort by using information on patients' dates of birth.

To appropriately fit the models allowed by the `apcfit` command, it is necessary to ensure that the data are in the right form. In practice, the dataset will have one observation for each of the subsets of the Lexis diagram. The width of the intervals for the age and period terms will be dictated by the availability of population figures for the intervals. Each observation will consist of these explanatory variables: number of events (cases), population risk time (person-years), mean age, period, and cohort.

To analyze data that are set out in the form of a Lexis diagram, the appropriate averages for age and period need to be used for the triangular subsections of the diagram (see figure 1; the dots in the center of the two triangles give the average values that should be used). Carstensen (2007) highlights that when the diagram is split into yearly age and period categories, the necessary values differ from conventional averages by one sixth. The conventional averages that would be used are at the center of the square; that is, at age  $34\frac{1}{2}$  and period  $2000\frac{1}{2}$ . These values are different from those at the center of the two triangles by one sixth. Making this distinction provides data that can then model the full extent of the Lexis diagram, taking into account both the upper and lower triangular subsets. The reasoning behind the averages used for the values of age, period, and cohort is illustrated in figure 1. The set of three lines that pass through the center of each triangle indicates the values that should be used for age, period, and cohort. The distinction between the upper and lower triangular subsets is defined by the patients' years of birth. This can again be seen from the partial Lexis diagram given as figure 1; the upper triangular subset relates to patients who were born in 1965, whereas the lower triangular subset relates to patients who were born in 1966.

Once the data are set up with the appropriate averages for age and period, it is necessary to ensure that the population risk-time is calculated for each triangular section of the Lexis diagram.

### Person-years (person risk-time)

For the majority of countries, it is possible to obtain population figures in one-year age classes for each calendar year. For example, the population data used in the example in section 5 were obtained from Statistics Finland (Statistics Finland). It is then possible to use these figures in the calculation of risk-time for each triangular subset of the Lexis diagram.

The command `poprisktime` can be used to calculate the population risk-time from population data using the formula suggested by Sverdrup (1967) as given in Carstensen (2007). The syntax of the command is detailed in the following section. The `poprisktime` command adjusts the averages for the age and period variables provided that the dataset is split into yearly intervals. The population risk-time should be cal-

culated for every possible combination of age and period. The dataset containing the values for the population risk-time ( $Y$ ) is merged with the dataset containing the number of cases as part of the command. An example of the form of data that are required to carry out the `poprisktime` command is given at the beginning of the example in section 5.

## 2.2 Matrix transformations

The main function of the `apcfit` command is to make transformations to the spline basis vectors so that the resulting output has a clear and sensible interpretation in spite of the identifiability issue. The matrix transformations are performed using Mata and they remove the trend from the cohort and period terms. The so-called drift term is then added to either the cohort or period terms, depending on the selected parameterization. This is why the cohort term, the period term, or both terms are constrained to have 0 slope (see `param()` in section 3.3).

The appropriate spline basis vectors are combined into matrices relating to each of the components (age, period, and cohort) of the model. Let these three design matrices be  $\mathbf{M}_A$ ,  $\mathbf{M}_P$ , and  $\mathbf{M}_C$ . The method requires that the period and cohort matrices be detrended. This is achieved by projecting the columns of the matrices onto the orthogonal complement of a two-column matrix,  $\mathbf{X}$ . In the case of the detrending of the period matrix:  $\mathbf{X} = [1 \mid P]$ , where  $P$  is the column of all the values of period.

The form of a general inner product that allows weighting is

$$\langle \mathbf{x} \mid \mathbf{y} \rangle = \sum_i x_i w_i y_i = \mathbf{x}' \mathbf{W} \mathbf{y}$$

where  $\mathbf{W} = \text{diag}(w_i)$ . The projection matrix on the column space of  $\mathbf{X}$  with respect to a general inner product is

$$\mathbf{P}_W = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$$

and the projection of  $\mathbf{M}$  on the orthogonal complement is

$$(\mathbf{I} - \mathbf{P}_W)\mathbf{M}$$

Once the period and cohort matrices have been projected in this way, the next stage is to reduce the number of columns of the matrices to ensure that they are of full rank. The columns of the matrices are also pivoted during this process. The rank of the matrices and the pivoting vector to be used are obtained by using Mata's `hqrnd()` function. The columns required to ensure that the matrices are of full rank are selected using the `select()` function in Mata. The matrices are then centered around the relevant reference points by subtracting a row corresponding to the reference point from each of the rows of  $\mathbf{M}$ . A column of 1s is then attached at the beginning of the  $\mathbf{M}_A$  matrix to ensure that the intercept is part of the age effects. The intercept term is contained within the age effects so that the age term carries the rate dimension. Then according to the parameterization, the drift column is added to the front of either the

period or cohort matrix. For full details of the matrix operations, see the appendix of Carstensen's article (2007).

## 2.3 Weights

The weighting matrix can take on any form; however, three logical choices for the weights are:  $w_i = 1$ ,  $w_i = D_i$  (where  $D_i$  is the number of cases for an observation), and  $w_i = Y_i$  (where  $Y_i$  is the population risk-time for an observation). Carstensen (2007) suggests using a weighting that is based upon the number of cases (D). Using equal weights (of 1) during the process of the detrending is a method that is attributed to Holford (1983). Using different values for the weights produces different estimates for the drift term.

# 3 The apcfit command

## 3.1 Syntax

```
apcfit [if] [in], age(varname) cases(varname) poprisktime(varname)
      [period(varname) cohort(varname) agefitted(newvar) perfitted(newvar)
      cohfitfitted(newvar) refper(#) refcoh(#) dextr(weighted|holford)
      param(ACP|APC|AdCP|AdPC|AP|AC) level(#) dfa(#) dfp(#) dfc(#)
      nper(#) bknotsa(numlist) bknotsp(numlist) bknotsc(numlist)
      knotsa(numlist) knotsp(numlist) knotsc(numlist)
      knotplacement(equal|weighted) adjust replace]
```

## 3.2 Note

`rcsgen` must be installed to run `apcfit`. `rcsgen` can be installed from the SSC archive. `rcsgen` generates basis functions for restricted cubic splines.

## 3.3 Options

`age(varname)` specifies the variable that refers to the age values. `age()` is required.

`cases(varname)` specifies the variable that refers to the number of cases or deaths for a given age and period. `cases()` is required.

`poprisktime(varname)` specifies the variable that refers to the population risk-time (person-years) for a given age and period. The population risk-time can be calculated using the `poprisktime` command. `poprisktime()` is required.

`period(varname)` specifies the variable that refers to the period values. This variable must be specified in all cases except when the age-cohort parameterization option is specified.

`cohort(varname)` specifies the variable that refers to the cohort values. If this variable is not given, the cohort values are calculated from the period and age variables according to the equality  $\text{cohort} = \text{period} - \text{age}$ .

`agefitted(newvar)` can specify the name of the fitted rate values given for age in the output. The default is `agefitted(agefitted)`. This can be useful when the user wants to compare the results of more than one parameterization.

`perfitted(newvar)` can specify the name of the fitted relative risk (RR) values given for period in the output. The default is `perfitted(perfitted)`. This can be useful when the user wants to compare the results of more than one parameterization.

`cohfitted(newvar)` can specify the name of the fitted RR values given for cohort in the output. The default is `cohfitted(cohfitted)`. This can be useful when the user wants to compare the results of more than one parameterization.

`refper(#)` can specify the reference period for the model. The default is to take the reference period to be the median date of diagnosis among the cases.

`refcoh(#)` can specify the reference cohort for the model. The default is to take the reference cohort to be the median date of birth among the cases.

`drextr(weighted|holford)` specifies the method of drift extraction for the model.

`drextr(weighted)` lets the drift extraction depend on the weighted average (by number of cases) of the period and cohort effects. The default is `drextr(weighted)`.

`drextr(holford)` uses a naïve average over all the values of the estimated effects, disregarding the number of cases.

`param(ACP|APC|AdCP|AdPC|AP|AC)` specifies the parameterization of the APC model.

`param(ACP)` dictates that the age effects should be rates for the reference cohort, the cohort effects should be RR relative to the reference cohort, and the period effects should be RR constrained to be 0 on average (on the log scale) with 0 slope. The default is `param(ACP)`.

`param(APC)` dictates that the age effects should be rates relative to the reference period, the period effects should be RR relative to the reference period, and the cohort effects should be RR constrained to be 0 on average (on the log scale) with 0 slope.

`param(AdCP)` dictates that the age effects should be rates for the reference cohort, and the cohort and period effects should be RR constrained to be 0 on average (on the log scale) with 0 slope. The drift term is missing from this model, and so the fitted values do not multiply to the fitted rates.



**param(AdPC)** dictates that the age effects should be rates for the reference period, and the cohort and period effects should be RR constrained to be 0 on average (on the log scale) with 0 slope. The drift term is missing from this model, and so the fitted values do not multiply to the fitted rates.

**param(AP)** dictates that the age effects should be rates for the reference period, and the period effects should be RR relative to the reference period. The cohort effects are not included in this model. Therefore, there is no identifiability issue.

**param(AC)** dictates that the age effects should be rates for the reference cohort, and the cohort effects should be RR relative to the reference cohort. The period effects are not included in this model. Therefore, there is no identifiability issue.

**level(#)** specifies the confidence level, as a percentage, for confidence intervals. The default is **level(95)**.

**dfa(#)** specifies the degrees of freedom used for the natural (restricted) cubic spline relating to the age variable. The default is **dfa(5)** (unless **knotsa()** is specified). The (df-1) internal knots are placed at the centiles of the data, depending on the value specified.

**dfp(#)** specifies the degrees of freedom used for the natural (restricted) cubic spline relating to the period variable. The default is **dfp(5)** (unless **knotsp()** is specified). The (df-1) internal knots are placed at the centiles of the data, depending on the value specified.

**dfc(#)** specifies the degrees of freedom used for the natural (restricted) cubic spline relating to the cohort variable. The default is **dfc(5)** (unless **knotsc()** is specified). The (df-1) internal knots are placed at the centiles of the data, depending on the value specified.

**nper(#)** specifies the units to be used in reported rates. For example, if the analysis time is in years, specifying **nper(1000)** results in rates per 1000 person-years. The default is **nper(1)**.

**bknotsa(numlist)** specifies the lower and upper boundary knots for the age variable; the default is the upper and lower values of the variable.

**bknotsp(numlist)** specifies the lower and upper boundary knots for the period variable; the default is the upper and lower values of the variable.

**bknotsc(numlist)** specifies the lower and upper boundary knots for the cohort variable; the default is the upper and lower values of the variable.

**knotsa(numlist)** specifies the knots for the age variable if the **dfa()** option is not used; the default is to use **dfa(5)** if neither option is specified.

**knotsp(numlist)** specifies the knots for the period variable if the **dfp()** option is not used; the default is to use **dfp(5)** if neither option is specified.

**knotsc(numlist)** specifies the knots for the cohort variable if the **dfc()** option is not used; the default is to use **dfc(5)** if neither option is specified.

`knotplacement(equal | weighted)` specifies the method of knot placement for the spline terms.

`knotplacement(equal)` means that the knots are placed at equally spaced centiles of the respective variables, depending on the number of knots that are used. This is the default.

`knotplacement(weighted)` means that the knots are placed at centiles of the variables that are dependent on the number of cases. For example, if there are more cases in the higher ages, the knots would be concentrated at the higher ages.

`adjust` specifies that the constrained variables be given relative to a reference point rather than averaging to zero on the log scale. This option cannot be applied to the age-period and age-cohort parameterizations. This option alters the variable that is constrained to be 0 on average (on the log scale) with 0 slope to still have 0 slope but to make the RRs relative to the reference point that is specified (or the median, if not specified). Adjusting the third variable to be relative to a reference point alters the interpretation of the age effects.

`replace` specifies that the default fitted value variables for age, period, and cohort should be replaced by the new run of the command. This will work only if the default names are used for the original model and if all the variables are still in the dataset.

## 4 The poprisktime command

### 4.1 Syntax

```
poprisktime using filename, age(varname) period(varname) cohort(varname)
      cases(varname) agemin(#) agemax(#) permin(#) permax(#) [pop(string)
      poprisktime(newvar) covariates(varlist) missingreplace]
```

### 4.2 Datasets

The using dataset refers to the dataset that contains population data split into yearly intervals of age and period. The master dataset should contain information on the number of cases split by age, period, and cohort. The intervals for the age and period variables should, again, be of length 1. Examples of the form of the data that is required will be detailed in the next section.

### 4.3 Options

`age(varname)` specifies the age variable that must have the same name in both datasets. In the population dataset (the using dataset), age values one less and one greater than those specified by `agemax()` and `agemin()` are required to avoid missing values.

**period**(*varname*) specifies the period variable that must have the same name in both datasets. In the population dataset (the using dataset), period values one greater than that specified by **permax**() are required. Population data are also necessary for at least as low as the **permin**() value. Missing values will be generated for the population risk-time variable if the population data do not at least satisfy these requirements.

**cohort**(*varname*) specifies the cohort variable in the master dataset.

**cases**(*varname*) specifies the variable in the master dataset that contains the number of cases.

**agemin**(#) specifies the minimum age in the output dataset. In the population dataset (the using dataset), an age that is less than this is required.

**agemax**(#) specifies the maximum age in the output dataset. In the population dataset (the using dataset), an age that is greater than this is required.

**permin**(#) specifies the minimum period in the output dataset. In the population dataset (the using dataset), a period that is equal to or less than this is required.

**permax**(#) specifies the maximum period in the output dataset. In the population dataset (the using dataset), a period that is greater than this is required.

**pop**(*string*) specifies the name of the variable in the using dataset that refers to the population figures. If this option is not specified, it is assumed that the variable is called **pop**.

**poprisktime**(*newvar*) specifies the name of the new variable that is added to the file to specify the population risk-time. The default is to name the variable **Y**.

**covariates**(*varlist*) can specify any covariates, such as a sex variable, by which the two datasets are split. If this option is specified, the covariates are included in the **merge** statement. If the dataset is split by covariates, this option must be specified so that the variables by which the data are merged uniquely identify the observations in the datasets. Both datasets must be split by the same covariates.

**missingreplace** specifies whether the missing values for the cases variable should be replaced with a zero in the merging process. This should be an appropriate assumption, unless missing data were present in the data beforehand. If **missingreplace** is not specified, the **cases**() variable is likely to contain missing values. A warning is given to indicate the presence of missing values that most probably should be replaced with values of 0.

## 5 Example

Colon cancer data from Finland will be used to illustrate the use of the **poprisktime** and the **apcfit** commands. The data cover diagnoses between 1980 and 2003 for all regions of Finland. It was decided to restrict the age range of the dataset to people who

were, at time of diagnosis, equal to or greater than 20 but less than 80 years of age. To highlight the possibility of including covariates into the analysis, the gender of patients was included when collapsing the dataset into unique records of age, period, and cohort. The data were collapsed into yearly intervals for age and period, leading to an upper and lower cohort value for each unique combination of age and period (according to their date of birth).

This collapse led to  $(80 - 20) \times (2004 - 1980)$  different age–period categories, each of which was further subdivided by date of birth into two categories. This gave a total of 2,880 observations for (D,Y)—one for each triangular subset. However, because the Finnish dataset contains sufficient information to include a sex term in the dataset, the dataset actually contains 5760 ( $= 2880 \times 2$ ) observations. To increase the dataset to include the sex term, the calculations for population risk-time were done for each gender.

## 5.1 Creating the dataset

Getting the data into the appropriate form has been facilitated by the creation of the `poprisktime` command. The `poprisktime` command uses the `merge` command to ensure that the data are matched in terms of the unique values of age, period, and cohort. It therefore requires the definition of a using dataset and needs the master dataset to be in the appropriate form. The data that are required for the master dataset for `poprisktime` take the form:

```
. list A P D C if A==30 & P<1990, noobs
```

A	P	D	C
30	1980	1	1950
30	1982	2	1952
30	1983	1	1952
30	1983	1	1953
30	1984	1	1953
30	1985	1	1954
30	1985	2	1955
30	1986	1	1956
30	1987	1	1956
30	1987	3	1957
30	1988	1	1958
30	1989	1	1959

It should be noted that these values refer to the left endpoints of the relevant classes of age (A), period (P), and cohort (C). This form would result in the two triangular subsets highlighted in figure 1 being listed as having the same age and period values (A = 34 and P = 2000), but different cohort values (C = 1965 and C = 1966). The penultimate column (D) refers to the number of cases.

The data that are required for the using dataset for the `poprisktime` command should take the form

```
. list A P pop if A==30 & P<1990, noobs
```

A	P	pop
30	1980	84828
30	1981	81245
30	1982	83554
30	1983	80213
30	1984	80514
30	1985	79804
30	1986	80150
30	1987	77427
30	1988	73556
30	1989	75669

The dataset (produced by `poprisktime`) that can be used by `apcfit` takes the form

```
. poprisktime using popdatanosex, age(A) period(P) cohort(C) cases(D) agemax(80)
> agemin(20) permin(1980) permax(2004) missingreplace
. list A P C D Y if A>30 & A<31 & P<1990, noobs
```

A	P	C	D	Y
30.333	1980.667	1950.334	1	40605.83
30.333	1981.667	1951.334	0	41758
30.333	1982.667	1952.334	2	40086.33
30.333	1983.667	1953.334	1	40245
30.333	1984.667	1954.334	0	39909.17
30.333	1985.667	1955.334	2	40069.83
30.333	1986.667	1956.334	1	38721.5
30.333	1987.667	1957.334	3	36779.33
30.333	1988.667	1958.334	1	37818.17
30.333	1989.667	1959.334	1	37791
30.667	1980.333	1949.666	0	42433.5
30.667	1981.333	1950.666	0	40654.5
30.667	1982.333	1951.666	0	41799.67
30.667	1983.333	1952.666	1	40115.83
30.667	1984.333	1953.666	1	40257.33
30.667	1985.333	1954.666	1	39909.33
30.667	1986.333	1955.666	0	40066.67
30.667	1987.333	1956.666	1	38710.33
30.667	1988.333	1957.666	0	36794
30.667	1989.333	1958.666	0	37872

It should be noted that to create the data used during this example, all the above datasets were also split by a covariate for gender.

## 5.2 Basic model

Having set up the data into the correct form, as detailed in the previous section, the `apcfit` command can be applied. The simplest form of the `apcfit` command uses the defaults for the options and only requires the specification of the data, as shown below:

```
. apcfit, age(A) period(P) cases(D) poprisktime(Y) nper(100000)
Iteration 0:  log likelihood = -10229.164
Iteration 1:  log likelihood = -9772.3502
Iteration 2:  log likelihood = -9757.1651
Iteration 3:  log likelihood = -9757.0919
Iteration 4:  log likelihood = -9757.0919

Generalized linear models               No. of obs      =       5760
Optimization      : ML                  Residual df      =       5745
                                          Scale parameter =         1
Deviance          = 6592.292062          (1/df) Deviance = 1.147483
Pearson           = 6554.902862          (1/df) Pearson  = 1.140975
Variance function: V(u) = u             [Poisson]
Link function     : g(u) = ln(u)         [Log]
                                          AIC              = 3.393087
Log likelihood    = -9757.091911        BIC              = -43151.9
```

D	OIM			z	P> z	[95% Conf. Interval]	
	Coef.	Std. Err.					
_spA1_intct	-9.142635	.0348165	-262.60	0.000	-9.210874	-9.074396	
_spA2	1.702715	.0358606	47.48	0.000	1.632429	1.773	
_spA3	-.0312765	.0262975	-1.19	0.234	-.0828187	.0202658	
_spA4	.0775714	.0206082	3.76	0.000	.0371802	.1179627	
_spA5	.0135517	.0117284	1.16	0.248	-.0094356	.036539	
_spA6	.0332615	.0065958	5.04	0.000	.020334	.046189	
_spP1	.0201192	.007845	2.56	0.010	.0047432	.0354951	
_spP2	.0025498	.0067633	0.38	0.706	-.010706	.0158056	
_spP3	.0103832	.0071587	1.45	0.147	-.0036476	.024414	
_spP4	.0029901	.0075344	0.40	0.691	-.0117772	.0177573	
_spC1_ldrft	.0107694	.0011545	9.33	0.000	.0085067	.0130321	
_spC2	.0099424	.0224079	0.44	0.657	-.0339763	.0538611	
_spC3	-.0080999	.0155304	-0.52	0.602	-.0385389	.022339	
_spC4	-.0415647	.0163449	-2.54	0.011	-.0736	-.0095293	
_spC5	-.0198339	.0153494	-1.29	0.196	-.0499182	.0102504	
Y	(exposure)						

The `apcfit` command saves the adjusted spline basis as `_spA*` for the age variable, `_spP*` for the period variable, and `_spC*` for the cohort variable, which allows other models to be fit using the `glm` command (providing that the appropriate family, link, and offset are used). As a result, (providing that the dataset was appropriately split for any given covariate), further models can be fit that can account for interactions.

The following set of commands can be used to list the estimated fitted values and their confidence intervals by age. The `apcfit` command creates new variables in the dataset containing the fitted values for the first occurrence of each unique value. In the interest of saving space, the output displayed below is limited to ages less than twenty-five years. The rates given are per 100,000 person-years because the `nper(100000)` option is used.

```
. sort A
. list A agefitted agefitted_lci agefitted_uci if agefitted!=. & A<=25, noobs
> abbreviate(13) divider separator(10)
```

A	agefitted	agefitted_lci	agefitted_uci
20.333	.7674873	.5526323	1.065875
20.667	.7808697	.5675128	1.074438
21.333	.8083171	.5981881	1.09226
21.667	.8225069	.6140866	1.101665
22.333	.8517627	.6468446	1.121598
22.667	.8669665	.6638144	1.132291
23.333	.8984786	.6987644	1.155273
23.667	.9149407	.7168638	1.167748
24.333	.9492388	.7541356	1.194817
24.667	.9672476	.7734399	1.209619

Figure 2 displays the fitted incidence of Finnish colon data for males and females combined. The default for `apcfit` is to make the reference points at the median value (with respect to the number of cases) for the period and cohort variable, respectively. The median cohort for the Finnish colon data is 1926.33, and the median period is 1993.67. To correctly interpret results, it is vital that the values of the reference points are known. The `apcfit` command returns the values of the reference points as r-class values `refcoh` and `refper`. The default parameterization was used in the model because the `param()` option was not specified. This means that the age values are relative to the reference cohort, having been adjusted for the effect of period. Under the default parameterization, the period effect is constrained to have 0 slope and to be 0 on average (on the log scale). This is due to the fact that the period effects are detrended and the drift term is then added to the cohort effects. The cohort effect is relative to the reference cohort and is allowed a slope through the inclusion of the drift term in the cohort effect.

(Continued on next page)

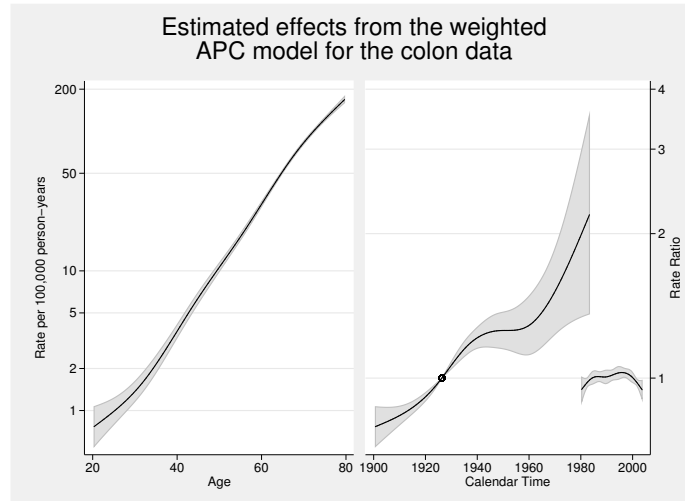


Figure 2. Graph for APC model for Finnish colon data. The leftmost solid line refers to the estimated age effect, the longest of the solid lines on the RR half of the graph refers to the estimated cohort effect, and the shortest line in the RR half of the graph refers to the estimated period effect. The respective regions surrounding the lines provide the 95% confidence intervals. The circle indicates the reference point.

The degrees of freedom were taken to be five for each of the spline bases for the three variables (age, period, and cohort). It is interesting to alter the degrees of freedom for any one of the variables, particularly the cohort variable, although this might lead to overfitting if the number is increased too much. The decision on the number of degrees of freedom can be aided through the use of the Akaike's information criterion (AIC) values. AIC values can be obtained via the use of the `estat ic` command. A lower AIC value suggests a better fitting model.

### 5.3 Including covariates

Including covariates in the analysis is a relatively simple process provided that the data are in the correct form. The only difficulty lies in appropriately splitting the dataset and the consequent calculation of the population risk-time. Population figures that are split by gender are usually available, making it feasible to calculate the population risk time for males and females separately.

Refitting the same model as in section 5.2, our code is

```
. apcfit, age(A) period(P) cases(D) poprisktime(Y) agefitted(ageAll)
> perfitted(perAll) cohfitted(cohAll) nper(100000)
. glm D _spA* _spP* _spC*, lnoffset(Y) family(poisson) nocons
(output omitted)
```



Using this structure, it is possible to add terms to the GLM to take into account the effect of gender. The simplest method for the inclusion of the sex term into the GLM is to assume a proportional effect for gender. The method for including this term is shown below. The covariate for gender is coded as 0 for female and 1 for male. The `eform` option is used to report the term for gender as an incidence rate ratio (IRR) (males relative to females).

```
. glm D _spA* _spP* _spC* sex, family(poisson) lnoffset(Y) nocons eform
Iteration 0:   log likelihood = -10183.733
Iteration 1:   log likelihood = -9719.7909
Iteration 2:   log likelihood = -9704.9493
Iteration 3:   log likelihood = -9704.8737
Iteration 4:   log likelihood = -9704.8737

Generalized linear models               No. of obs   =       5760
Optimization      : ML                  Residual df   =       5744
                                          Scale parameter =        1
Deviance          =  6487.855702         (1/df) Deviance =  1.129501
Pearson           =  6473.673607         (1/df) Pearson  =  1.127032
Variance function: V(u) = u             [Poisson]
Link function     : g(u) = ln(u)         [Log]
                                          AIC           =  3.375303
Log likelihood    = -9704.873731         BIC           = -43247.68
```

D	OIM		z	P> z	[95% Conf. Interval]	
	IRR	Std. Err.				
_spA1_intct	.0000998	3.55e-06	-259.28	0.000	.0000931	.000107
_spA2	5.51232	.1977228	47.59	0.000	5.138099	5.913797
_spA3	.9671249	.0254402	-1.27	0.204	.9185265	1.018295
_spA4	1.079698	.0222534	3.72	0.000	1.036952	1.124207
_spA5	1.01307	.0118833	1.11	0.268	.9900445	1.03663
_spA6	1.033405	.0068161	4.98	0.000	1.020132	1.046851
_spP1	1.020693	.0080073	2.61	0.009	1.005119	1.036508
_spP2	1.002559	.0067806	0.38	0.706	.9893567	1.015937
_spP3	1.01047	.0072336	1.45	0.146	.9963913	1.024748
_spP4	1.003077	.0075577	0.41	0.683	.9883726	1.017999
_spC1_ldrft	1.010528	.0011671	9.07	0.000	1.008243	1.012818
_spC2	1.007732	.0225837	0.34	0.731	.9644266	1.052982
_spC3	.9915087	.0153987	-0.55	0.583	.9617826	1.022154
_spC4	.9601781	.0156951	-2.49	0.013	.9299038	.991438
_spC5	.9799076	.0150411	-1.32	0.186	.9508666	1.009835
sex	1.158693	.0166619	10.24	0.000	1.126492	1.191814
Y (exposure)						

The output given above shows that, in Finland, males have about a 16% greater incidence of colon cancer than females across the entire dataset when adjusting for the other effects. The  $p$ -value for the sex term highlights that the effect for gender is significant at the 5%, and even the 0.1%, level. This measure of significance, however, assumes that the effect of gender is proportional over both time scales and date of birth (cohort).

### Full interaction using the adjusted spline variables

The other extreme is to fit a full interaction for gender with all the spline terms in the model. In theory, this approach should give results equivalent to fitting two separate models for each gender (which can be achieved by using an `if` qualifier as part of the `apcfit` command). However, the results given in this case are not exactly equivalent if the default drift extraction is used because the default drift extraction uses a weighting that is based upon the number of cases. Therefore, the weighting takes into account the cases for both males and females when fitting the model with all patients included but only takes into account, for example, the number of cases for males in the model fitted exclusively for males. This means that when the full interaction is fitted, the fitted values are slightly different from the values obtained by fitting two separate models. This difference would be magnified if the number of cases for males and females were markedly different—for example, if the dataset related to cases of breast cancer. However, in most cases, the difference will be negligible. If the drift extraction suggested by Holford is used (which gives equal weight to all the observations), then the model with the full interaction term for sex is entirely equivalent to the two separate models for each of the genders.

### Using reduced splines to model the interaction

Fitting the full gender interaction with all the components of the model may result in overfitting. Technically, it involves using the same number of knots for the original components and for the difference between the genders when fewer knots would suffice. However, it is possible to perform an analysis that uses fewer knots for the differences between the genders whilst still maintaining the greater number of knots for the baseline shape. Spline bases with a reduced number of degrees of freedom can be created to model the effect of the interaction with gender.

It is possible that some of the components of the model, such as the period effect, may not vary significantly with gender. The likelihood-ratio test can be used to compare nested models. It should be noted that the identifiability problem is reintroduced when modeling the effect of gender using interaction terms. Therefore, it is only possible to fit an interaction with at most two of the age, period, and cohort components (unless the full adjusted spline terms are used, as described above).

Fitting the effect of gender using just a single model rather than the two separate models allows the calculation of the time-dependent IRR for the genders. This ratio can be plotted against the relevant time scale to show the differences between the genders in terms of the incidence rate. This approach can also be implemented with a reduced set of spline bases to model the difference between the genders.

The code given below generates spline terms to model the interaction with sex for the age and cohort effects. The spline terms have a reduced number of degrees of freedom compared with those used to calculate the original fitted values using `apcfit`. It was decided to illustrate the method using reduced splines with degrees of freedom equal to three (compared with eight degrees of freedom for each of the components of the original

APC model). It is important to center the generated reduced set of splines for the cohort effect around the reference cohort. Centering can be achieved using the fact that the original cohort splines will take a value of zero when the line of the spline design matrix for the cohort effects corresponds to the reference cohort. This approach assumes that the reference cohort selected is a part of the original dataset of cohort values. If that is not the case, further work is required to center the reduced set of cohort splines.

```
. apcfit, age(A) period(P) cases(D) poprisktime(Y) dfa(8) dfp(8) dfc(8)
> nper(100000)

. *** Generate the reduced set of splines for age ***
. rcsgen A, gen(newA) df(3) orthog
Variables newC1 to newC3 were created
. local dfreda=wordcount(r(knots))-1
. *** Obtain the relevant interaction terms ***
. forvalues i = 1/`dfreda' {
.     generate sexnewA`i'=sex*newA_`i'
. }

. *** Generate the reduced set of splines for cohort ***
. rcsgen C, gen(newC) df(3) orthog
. local dfredc=wordcount(r(knots))-1
. *** Obtain the relevant interaction terms ***
. forvalues i = 1/`dfredc' {
.     quietly generate sexnewC`i'=sex*newC_`i'
. }

. *** Center the reduced Cohort splines on the reference cohort ***
. forvalues i = 1/`dfredc' {
.     quietly summarize sexnewC`i' if _spC1==0 & sex==1
.     quietly generate sexnewCref`i'=r(mean)
.     quietly replace sexnewC`i'=sexnewC`i'-sexnewCref`i' if sex==1
. }

. drop sexnewCref*

. *** Fit the reduced splines as part of the GLM ***
. glm D _sp* sex sexnewA* sexnewC*, lnoffset(Y) f(p) nocons
```

The required estimates for each of the genders can be obtained from this model by using the `partpred` command, which allows for the calculation of partial predictions from the last fitted model. The `partpred` package is available to download from the SSC archive. The `partpred` command can also be used to calculate the relevant time-dependent IRRs for the difference between the genders in terms of the age and cohort effects. The required code for the prediction of the age effects, and their confidence intervals, for both genders is

```
. partpred agefem if sex==0, for(_spA*) eq(D) eform ci(agefemlci agefemuci)
. partpred agemal if sex==1, for(_spA* sex sexnewA*) eq(D) eform
> ci(agemallci agemaluci)
```

(Continued on next page)

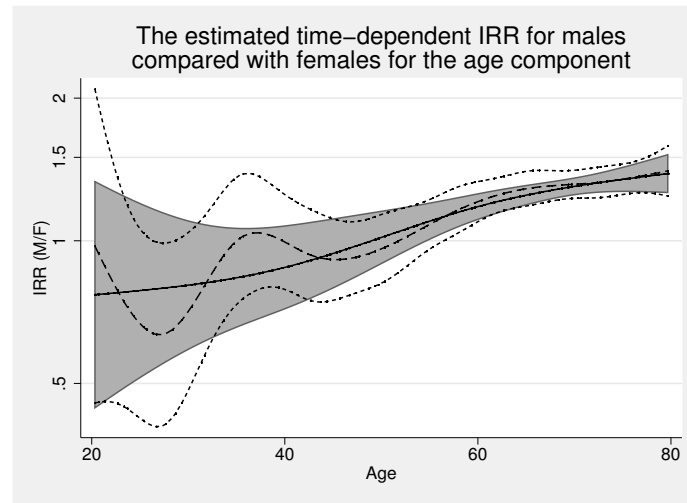


Figure 3. Graph of time-dependent IRR for the age-by-sex interaction. The solid line is the IRR (males compared with females) for the age-by-sex interaction using a reduced set of splines with three degrees of freedom (from the model using reduced splines for age and cohort, having fitted a model with eight degrees of freedom for each of the components). The shaded region around this line gives the relevant 95% confidence interval. The dashed line gives the IRR for sex in terms of the age component from the model with the full interaction for sex (from the previous section) using 8 degrees of freedom for each of the components. The dotted lines form the appropriate 95% confidence interval.

The results of the analysis using a reduced set of spline bases can be compared with the analysis that uses a full interaction for each of the components. The major differences that will be observed between these two analyses relate to the two fundamental differences in assumptions between the models. First, the model that fits a full interaction with the sex term also allows for a difference in terms of the period effect. The model fitted with the reduced sets of splines for age and cohort assumes that the period effect is the same for both genders. Second, the model fitted with the reduced sets of splines for age and cohort assumes that the effect of gender can be modeled using fewer knots than the model that fits the full interaction with gender. Figure 3 shows a comparison of the two analyses in terms of the effect of gender on the IRR for the age component. Using an increased number of degrees of freedom could lead to an overfitting of the IRR for any of the components, which is apparent in figure 3. It can be seen that using the results of the reduced sets of splines gives a more believable and interpretable effect of gender over the age time scale. In fact, it could be argued that it may only be necessary to model the interaction using a linear function, considering how straight the line is for the reduced splines model. However, there is a danger that a reduction in the degrees of freedom may make the spline function too inflexible to follow a more complex pattern.

### Continuous covariates

The example given above is simplistic in that it includes a single binary variable as the only covariate. Carrying out more complex scenarios is also feasible using the output given by `apcfit`. Providing that the appropriately split population data are available for a continuous covariate, it is possible to carry out more complex interaction models. It is possible to also use splines to describe the effect of the continuous covariate and create spline-by-spline interactions for any of the components of the APC model. Even though presentation of the results for these models can become difficult, they do have clear advantages in terms of flexibility.

## 5.4 Future incidence

Having obtained the models for incidence via the `apcfit` command, it is possible to extrapolate to make predictions about future incidence. The fact that restricted cubic (natural) splines are linear beyond the boundary knot can be used to make linear extensions to the predicted values. Using a linear regression analysis of the final few estimates of an APC analysis to predict future trends has been common practice for some time (Osmond 1985; Bray and Møller 2006). Utilizing the fact that restricted cubic splines are linear beyond the boundary knot to make predictions is conceptually similar to using linear regression. However, for accurate predictions to be given using this method, the boundary knots must be placed within the range of the data. It is also possible to simply project the drift term using the `apcfit` models to make predictions for future incidence (Møller et al. 2003, 2007). Further work is planned for investigating the effectiveness of these methods for giving predictions of incidence.

## 6 Conclusion

Carstensen suggested that APC models should be fit using continuous variables for the age, period, and cohort effects (Carstensen 2007). This procedure can now be accomplished in Stata by using the `apcfit` command. The command provides the possibility of producing a clear and attractive graphical display of the data. It also provides the foundations required to predict the future rates of incidence and mortality.

## 7 Acknowledgments

We would like to acknowledge that this method was suggested and implemented in R by Bendix Carstensen. We would also like to thank Bendix for his useful comments during the drafting of this article. This work was carried out as part of a PhD thesis that is funded by the Medical Research Council. Finally, we would like to thank the Finnish Cancer Registry for access to the data.

## 8 References

- Bray, F., and B. Møller. 2006. Predicting the future burden of cancer. *Nature Reviews Cancer* 6: 63–74.
- Carstensen, B. 2007. Age–period–cohort models for the Lexis diagram. *Statistics in Medicine* 26: 3018–3045.
- Durrleman, S., and R. Simon. 1989. Flexible regression models with cubic splines. *Statistics in Medicine* 8: 551–561.
- Holford, T. R. 1983. The estimation of age, period and cohort effects for vital rates. *Biometrics* 39: 311–324.
- Land, K. C. 2008. Disentangling Age-Period-Cohort Effects: New Models, Methods, and Empirical Applications.  
<http://help.pop.psu.edu/help-by-statistical-method/apc/Land-Presentation.ppt/view>.
- Møller, B., H. Fekjær, T. Hakulinen, H. Sigvaldason, H. H. Storm, M. Talbäck, and T. Haldorsen. 2003. Prediction of cancer incidence in the Nordic countries: Empirical comparison of different approaches. *Statistics in Medicine* 22: 2751–2766.
- Møller, H., L. Fairley, V. Coupland, C. Okello, M. Green, D. Forman, B. Møller, and F. Bray. 2007. The future burden of cancer in England: Incidence and numbers of new patients in 2020. *British Journal of Cancer* 96: 1484–1488.
- Osmond, C. 1985. Using age, period and cohort models to estimate future mortality rates. *International Journal of Epidemiology* 14: 124–129.
- Robertson, C., S. Gandini, and P. Boyle. 1999. Age-period-cohort models: A comparative study of available methodologies. *Journal of Clinical Epidemiology* 52: 569–583.
- Statistics Finland. [http://www.stat.fi/index\\_en.html](http://www.stat.fi/index_en.html).
- Sverdrup, E. 1967. Statistiske metoder ved dødelighetsundersøkelser. Statistical Memoirs. (in Norwegian). Institute of Mathematics, University of Oslo.
- Yang, Y., W. J. Fu, and K. C. Land. 2004. A methodological comparison of age-period-cohort models: The intrinsic estimator and conventional generalized linear models. *Sociological Methodology* 34: 75–110.

### About the authors

Mark Rutherford is a PhD student at the University of Leicester, UK. He is currently working on a PhD thesis relating to the prediction of future cancer burden on society. The PhD is funded by a Medical Research Council grant.

Paul Lambert is a reader in medical statistics at the University of Leicester, UK. His main interest is in the development and application of methods in population based cancer research.

John Thompson is a professor in the department of Health Sciences at the University of Leicester with a research interest in Genetic Epidemiology. He teaches on the department's MSc in Medical Statistics and is a longtime Stata user and enthusiast.