Projecting cancer incidence using restricted cubic splines.

Mark J. Rutherford ¹  Paul C. Lambert ¹,²  
John R. Thompson ¹

¹Department of Health Sciences, University of Leicester, UK.
²MEB, Karolinska Institutet, Stockholm.

Nordic and Baltic Stata UGM.  
11th November, 2011.
Why do we need to project into the future?

- Health and planning officials need to plan treatment and care.
  - Need to know how many new cases of cancer there will be.
  - Need to know the type/severity of cancer to cost out treatment/care.
  - Need to know how many patients have cancer at a given moment in time (prevalence).

- Assuming that the current rates will remain the same is often inadequate.

- However, making predictions can be dangerous and difficult.
Firstly, we feel that using cubic splines will have a benefit over the traditional factor method approach.
Firstly, we feel that using cubic splines will have a benefit over the traditional factor method approach.

- Splines are a collection of polynomials that are joined at a pre-defined number of points; known as knots.
- The number and location of knots can affect the fit.
- Restricted cubic splines use cubic polynomials between knots.
- The further restriction is linearity beyond the boundary knots.
Firstly, we feel that using cubic splines will have a benefit over the traditional factor method approach.

- Splines are a collection of polynomials that are joined at a pre-defined number of points; known as knots.
- The number and location of knots can affect the fit.
- Restricted cubic splines use cubic polynomials between knots.
- The further restriction is linearity beyond the boundary knots.

Secondly, the traditional approach to incidence projections uses data over a long range to base the projections upon. Using splines we can use more “up-to-date” trends to make the projections.
Restricted Cubic Splines: Boundary Knots

- Upper Boundary Knot 80th
- Upper Boundary Knot 100th
Incidence Models

- Age-Period-Cohort (APC) models used to model incidence data.

For Incidence Data:

- **Age** would be the age of the subject at diagnosis.
- **Period** would refer to the calendar time at which the diagnosis was made.
- **Cohort** would refer to the patient’s date (or cohort) of birth.

- Trend of the disease in terms of all 3 of these variables.
- However, Age-Period-Cohort models suffer from an identifiability issue making appropriate modelling of all 3 terms difficult.
In the age-period-cohort setting we fit spline functions to each of the three components; age, period, and cohort.

Constraints need to be made because of the lack of identifiability of the model.

The identifiability issue stems from the fact that there is an exact relationship between the variables:

- Age at Diagnosis (A) ↔ Date of Birth (C) → Date of Diagnosis (P)
## Data Format

```
Data
```

```
Data
.list * in 1/10

<table>
<thead>
<tr>
<th>A</th>
<th>P</th>
<th>C</th>
<th>sex</th>
<th>D</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.333</td>
<td>1980.667</td>
<td>1960.334</td>
<td>0</td>
<td>0</td>
<td>18528.5</td>
</tr>
<tr>
<td>20.333</td>
<td>1981.667</td>
<td>1961.334</td>
<td>0</td>
<td>0</td>
<td>18524.33</td>
</tr>
<tr>
<td>20.333</td>
<td>1982.667</td>
<td>1962.334</td>
<td>0</td>
<td>0</td>
<td>18810.67</td>
</tr>
<tr>
<td>20.333</td>
<td>1983.667</td>
<td>1963.334</td>
<td>0</td>
<td>0</td>
<td>18569.5</td>
</tr>
<tr>
<td>20.333</td>
<td>1984.667</td>
<td>1964.334</td>
<td>0</td>
<td>1</td>
<td>18116</td>
</tr>
<tr>
<td>20.333</td>
<td>1985.667</td>
<td>1965.334</td>
<td>0</td>
<td>0</td>
<td>18053.83</td>
</tr>
<tr>
<td>20.333</td>
<td>1986.667</td>
<td>1966.334</td>
<td>0</td>
<td>0</td>
<td>17783.5</td>
</tr>
<tr>
<td>20.333</td>
<td>1987.667</td>
<td>1967.334</td>
<td>0</td>
<td>0</td>
<td>17144.83</td>
</tr>
<tr>
<td>20.333</td>
<td>1988.667</td>
<td>1968.334</td>
<td>0</td>
<td>0</td>
<td>15806.83</td>
</tr>
<tr>
<td>20.333</td>
<td>1989.667</td>
<td>1969.334</td>
<td>0</td>
<td>0</td>
<td>15487.17</td>
</tr>
</tbody>
</table>
```
apcfit was described in a Stata Journal article available in Issue 4 of 2010. apcfit is used for fitting age-period-cohort models when not making projections.

```
apcfit, age(A) period(P) cases(D) poprisktime(Y)
```

net sj 10-4 st0211

The extension to making the projections involves a little care in setting up the data and making the knot selection.

A future update of the command/an associated command will hopefully make projections simpler from apcfit.
### Stata Output

```
. quietly apcfit, age(A) period(P) cases(D) poprisktime(Y) nper(100000)
. glm, noheader

|         | Coef.  | Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|---------|--------|-----------|-------|------|---------------------|
| _spA1_intct | -9.142635 | 0.0348165 | -262.60 | 0.000 | -9.210874 to -9.074396 |
| _spA2    | 1.702715  | 0.0358606 | 47.48  | 0.000 | 1.632429 to 1.773036  |
| _spA3    | -0.0312765 | 0.0262975 | -1.19  | 0.234 | -0.0828187 to 0.0202658 |
| _spA4    | 0.0775714  | 0.0206082 | 3.76   | 0.000 | 0.0371802 to 0.1179627 |
| _spA5    | 0.0135517  | 0.0264987 | 0.50   | 0.617 | -0.0094356 to 0.0363539 |
| _spA6    | 0.0332615  | 0.0065958 | 5.04   | 0.000 | 0.020334 to 0.046189  |
| _spP1    | 0.0201192  | 0.007845  | 2.56   | 0.010 | 0.0047432 to 0.0354961 |
| _spP2    | 0.0026498  | 0.0067633 | 0.38   | 0.706 | -0.010706 to 0.0158056 |
| _spP3    | 0.0103832  | 0.0071587 | 1.45   | 0.147 | -0.003647 to 0.024414 |
| _spP4    | 0.0029901  | 0.0075344 | 0.40   | 0.691 | -0.0117772 to 0.0177573 |
| _spC1_ldrft | 0.0107694  | 0.0011545 | 9.33   | 0.000 | 0.0085067 to 0.0130321 |
| _spC2    | 0.0099424  | 0.0224079 | 0.44   | 0.657 | -0.0339763 to 0.0538611 |
| _spC3    | -0.0080999 | 0.0155304 | -0.52  | 0.599 | -0.0385389 to 0.022339 |
| _spC4    | -0.0415647 | 0.0163494 | -2.54  | 0.011 | -0.0736 to -0.0095293 |
| _spC5    | -0.0198339 | 0.0153494 | -1.29  | 0.196 | -0.0499182 to 0.0102504 |
| ln(Y)    | 1       |           |       |      | 1 (exposure)          |
```
Restricted Cubic Splines

![Graph showing rate ratios over age and calendar time with labels for age, cohort, and period.]
Comparison of Factor Model to the Model using Splines for the Age Effects

Rate per 100,000 person-years

Age

Comparison of Factor Model to the Model using Splines for the Age Effects

Mark Rutherford

Stata UGM. 11th November, 2011.
We want to project the incidence rates into the future.

When projecting it is “safer” to make simple assumptions (i.e. Linearity).

There is no information (data) to allow projections of a complicated shape.
Simple Description

- We want to project the incidence rates into the future.
- When projecting it is “safer” to make simple assumptions (i.e. Linearity).
- There is no information (data) to allow projections of a complicated shape.

**Question:** How do we draw the straight lines at the end of our observed data?

- One way is to take the trend over the entire period of observed data and project that. (OLD)
- Or, we could use the restriction of the cubic splines being linear beyond the boundary knot. (NEW)
New Projection Method - Spline Restriction
Old Projection Method - Spline Drift

In the graph, we see the comparison between old and new projection methods using splines. The incidence rate is plotted against age at diagnosis, showing a clear upward trend with age.

The graphs below illustrate the rate ratio over different years and periods. The new method (solid line) exhibits a more consistent and realistic trend compared to the old method (dashed line).

Key observations:
- The solid line represents the improved projection method using splines.
- The dashed line indicates the old projection method.
- The graphs highlight the drift correction achieved with the new method.

This visual representation facilitates a better understanding of the incidence trends and the effectiveness of the new projection technique.
Results - Colon Cancer (Males)
Results - Pancreatic Cancer (Females)

Total Number of Cases


Calendar Time

True Rate  Spline Drift  Spline Restriction
Results - Lung Cancer (Males)
Sensitivity to Boundary Knot - Lung (Males)
Making projections is a dangerous game.

Steps can be taken to use more recent information to “project” into the future.

Using splines and more finely split data as opposed to the factor models with coarsely split data seems better.

A considered approach to making the projections needs to be made.

A “one method, fits all” approach is inadequate.

Need to use external information - e.g. screening program introductions, and expected peaks/troughs.
References

B. Carstensen.
Age-period-cohort models for the Lexis diagram.

Age-period-cohort modelling.

Bjørn Møller, Harald Fekjær, Timo Hakulinen *et al.*

H. Møller, L. Fairley, V. Coupland *et al.*

Projecting cancer incidence using age-period-cohort models incorporating restricted cubic splines.