Projecting cancer incidence using restricted cubic splines.

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

[1]Department of Health Sciences,
University of Leicester, UK.

[2]MEB, 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.

What can be improved?

- Firstly, we feel that using cubic splines will have a benefit over the traditional factor method approach.

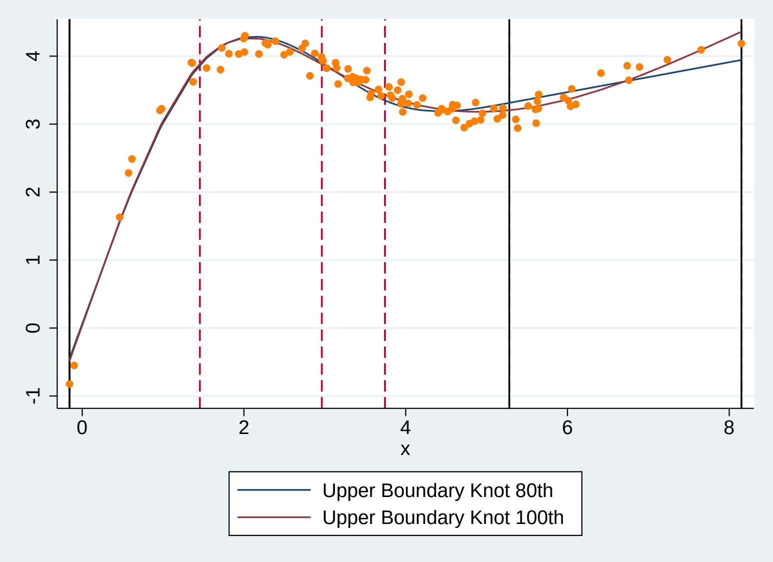
What can be improved?

- 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 effect the fit.
  - Restricted cubic splines use cubic polynomials between knots.
  - The further restriction is linearity beyond the boundary knots.

What can be improved?

- 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 effect 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



Incidence Models

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

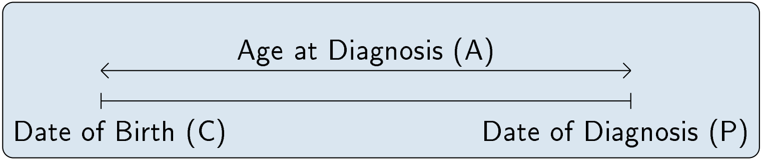
For Incidence Data:

- *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.

Splines for Age-Period-Cohort (APC) models

- 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:



Data Format

Data

```
. list * in 1/10
```

	A	P	C	sex	D	Y
1.	20.333	1980.667	1960.334	0	0	18528.5
2.	20.333	1981.667	1961.334	0	0	18524.33
3.	20.333	1982.667	1962.334	0	0	18810.67
4.	20.333	1983.667	1963.334	0	0	18569.5
5.	20.333	1984.667	1964.334	0	1	18116
6.	20.333	1985.667	1965.334	0	0	18053.83
7.	20.333	1986.667	1966.334	0	0	17783.5
8.	20.333	1987.667	1967.334	0	0	17144.83
9.	20.333	1988.667	1968.334	0	0	15806.83
10.	20.333	1989.667	1969.334	0	0	15487.17

How to do this in Stata?

- `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`.

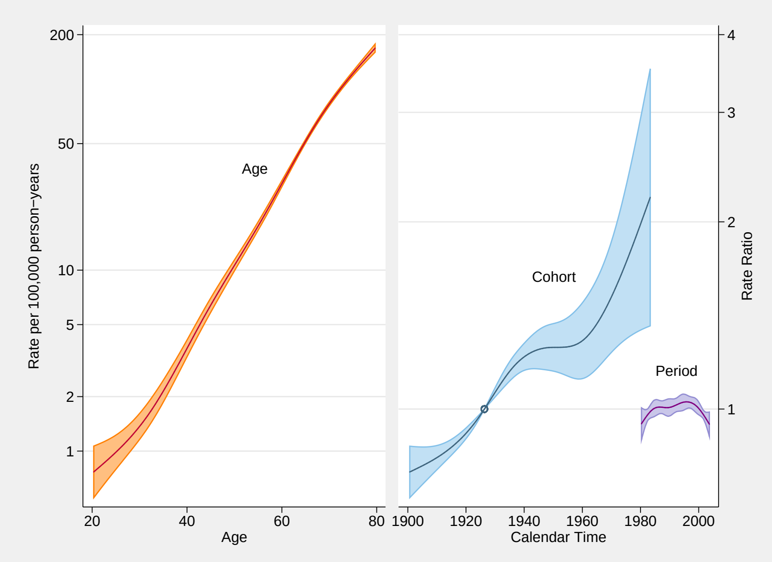
Stata Output

apcfit

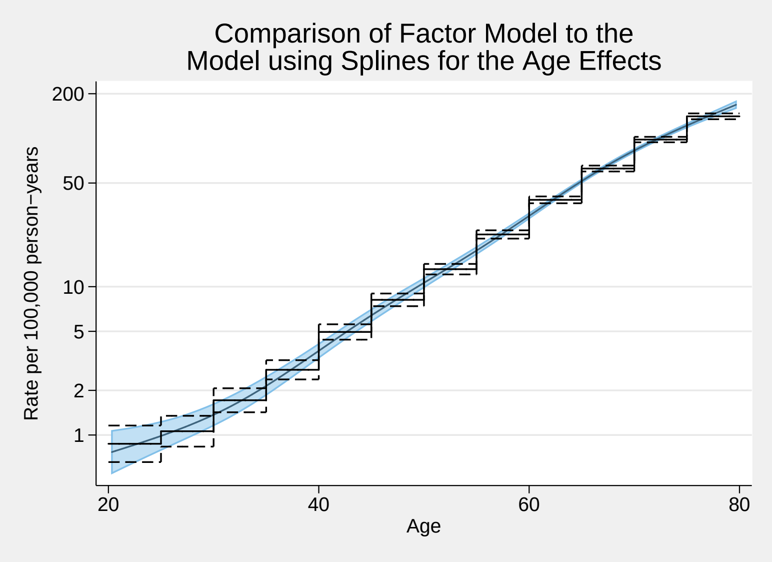
```
. quietly apcfit, age(A) period(P) cases(D) poprisktime(Y) nper(100000)
. glm, noheader
```

D	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
_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
ln(Y)	1	(exposure)				

Restricted Cubic Splines



Restricted Cubic Splines vs Factor Method



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.

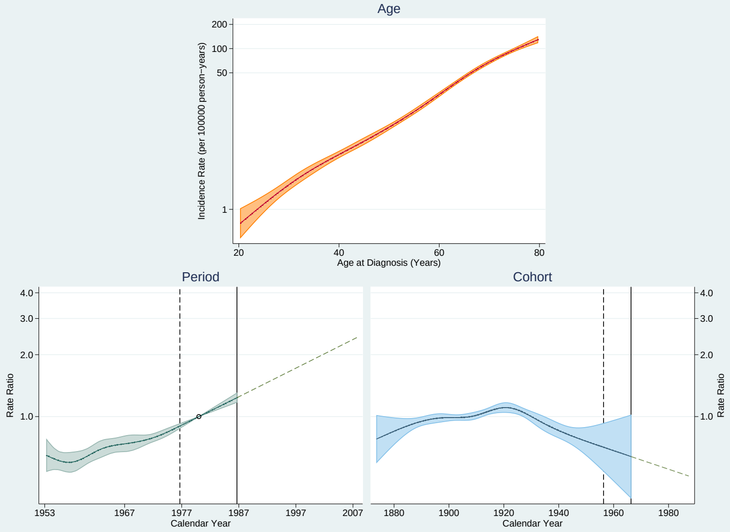
Simple Description

- 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.

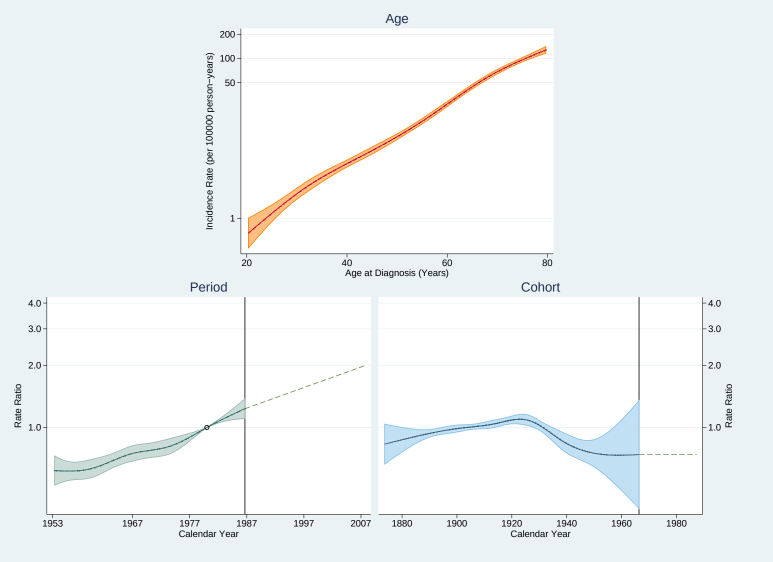
Question: How do we draw the straight lines at the end of our observed data?

- 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)

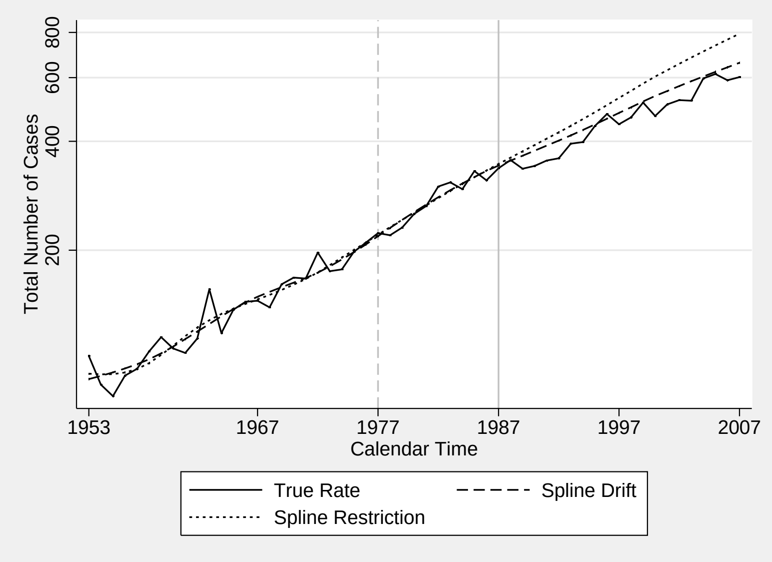
New Projection Method - Spline Restriction



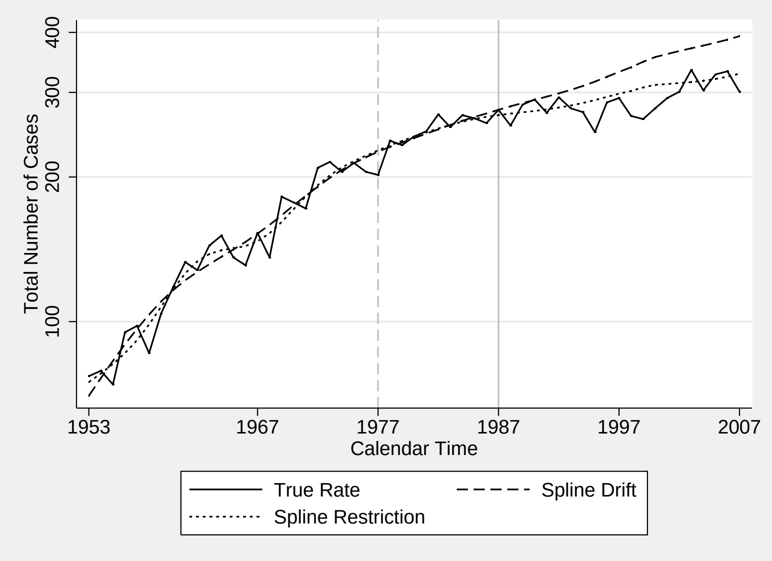
Old Projection Method - Spline Drift



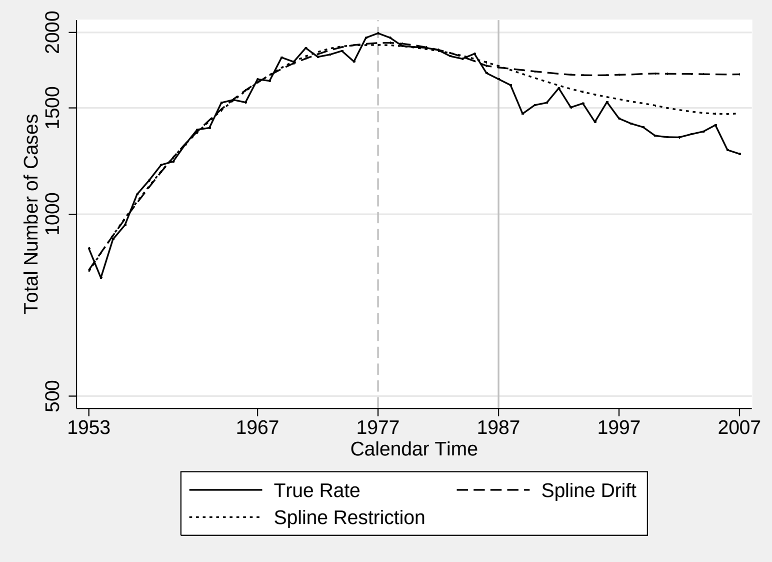
Results - Colon Cancer (Males)



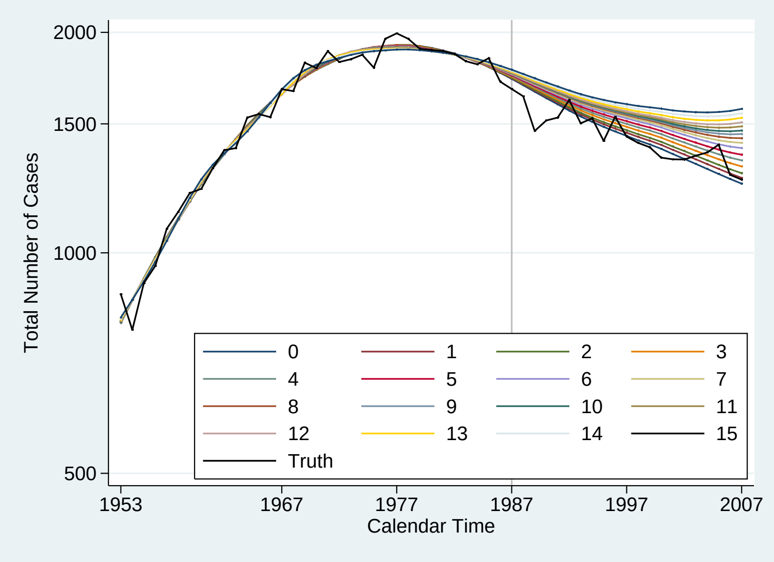
Results - Pancreatic Cancer (Females)



Results - Lung Cancer (Males)



Sensitivity to Boundary Knot - Lung (Males)



Discussion

- 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.
Statistics in Medicine, **26** :3018–3045, 2007.

M.J. Rutherford, P.C. Lambert, and J.R. Thompson.
Age-period-cohort modelling.
Stata Journal, **10** :606–627, 2010.

Bjørn Møller, Harald Fekjær, Timo Hakulinen *et al.*
Prediction of cancer incidence in the Nordic countries: empirical comparisons of different approaches.
Statistics in Medicine, **22** :2751–2766, 2003.

H. Møller, L. Fairley, V. Coupland et al.
The future burden of cancer in England: Incidence and numbers of new patients in 2020.
Br J Cancer, **96** :1484–1488, 2007.

M.J. Rutherford, P.C. Lambert, and J.R. Thompson.
Projecting cancer incidence using age-period-cohort models incorporating restricted cubic splines.
Statistics in Medicine, (submitted) 2011.