Fitting joinpoint models for cancer trends in Stata

Paul C Lambert[1,2]

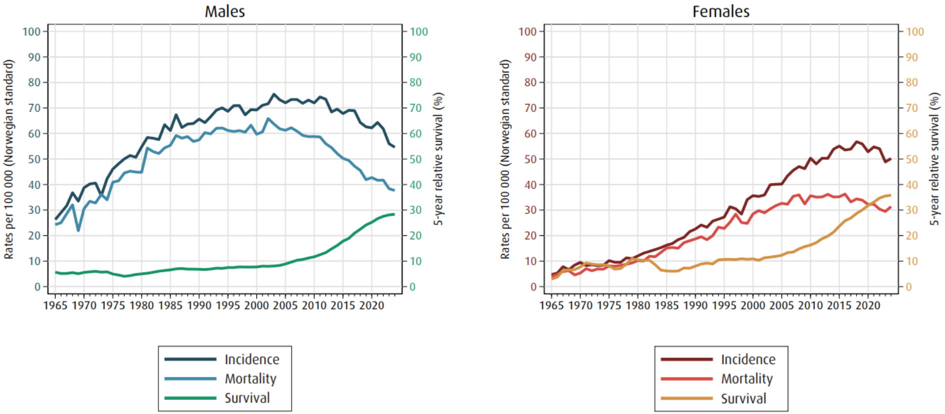
[1] Cancer Registry of Norway, FHI, Norway
[2] Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Northern European Stata Conference, Stockholm
29 August 2025

Introduction

- We are interested in monitoring cancer trends[1].

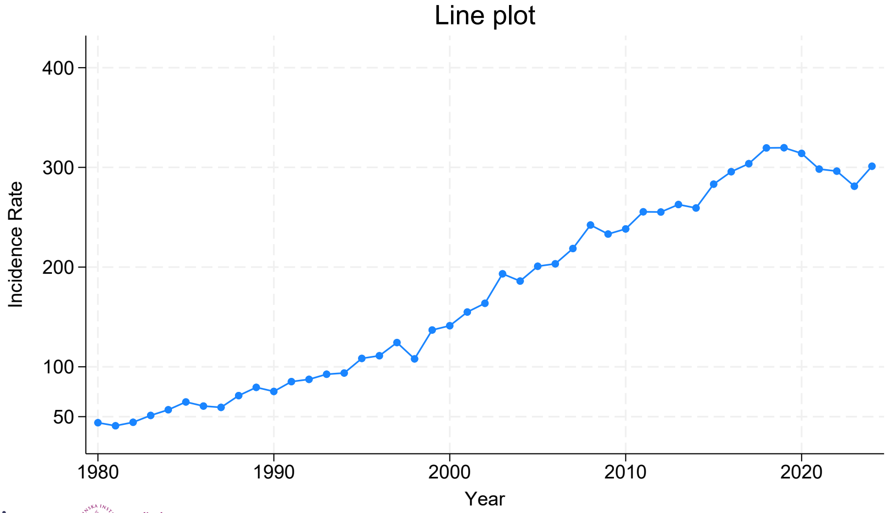
Figure 9.1–L: Lung, trachea (ICD-10 C33–34)



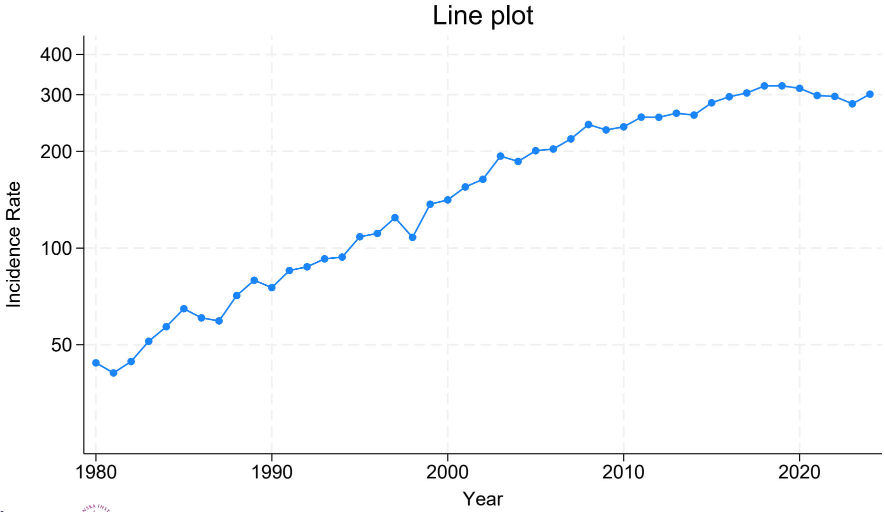
Summarzing Trends

- We want to summarise trends over time.
- We can explore graphically, but it is also useful to summarise trends numerically
- This talk will shows an implementation of joinpoint models in Stata.
- Joinpoint modes use linear splines to look at changes in trends over time. They include selecting the number of knots and their locations[2].
- They are descriptive models.

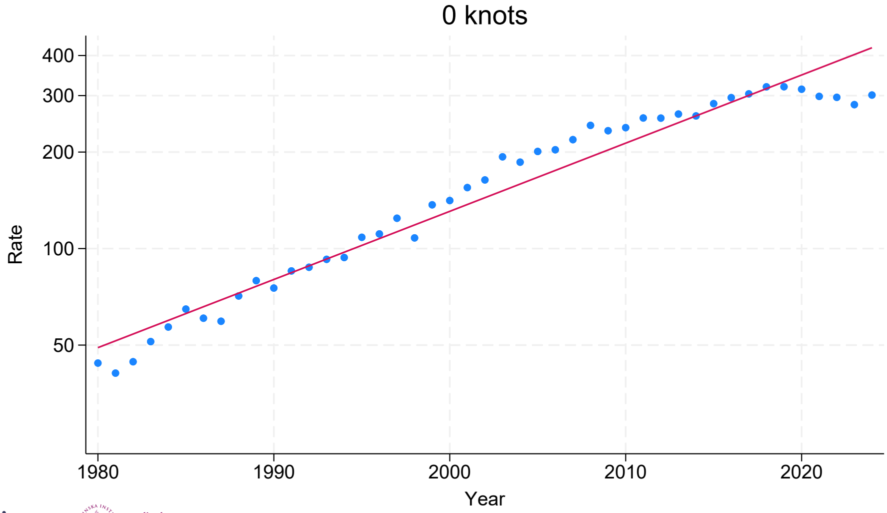
Incidence of lung cancer in Norway (females 70-79)



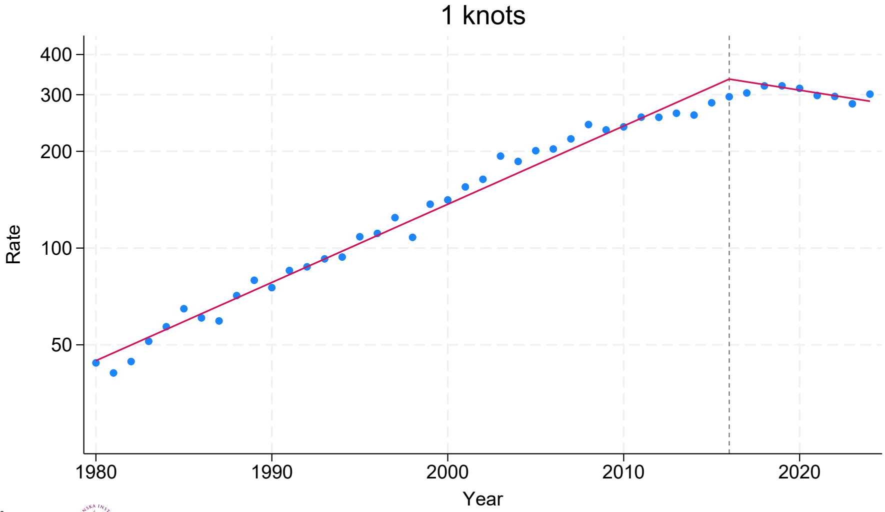
Incidence of lung cancer in Norway (females 70-79)



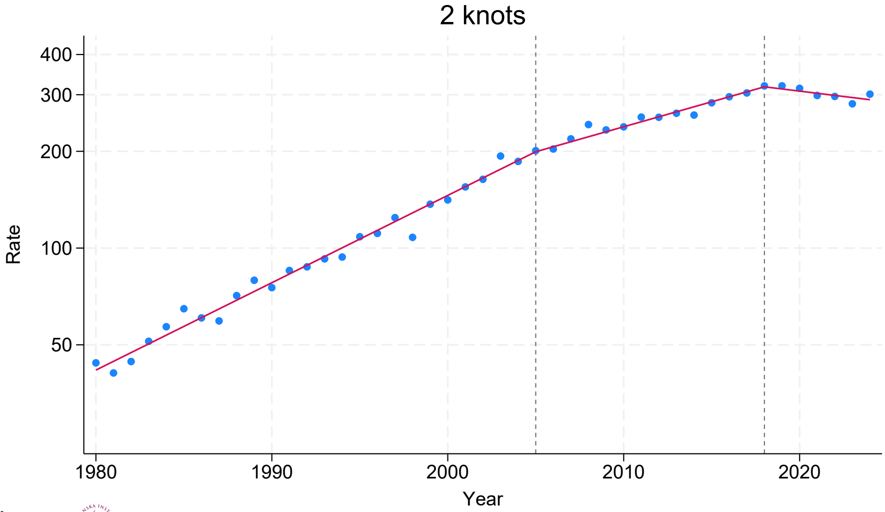
Incidence of lung cancer in Norway (females 70-79)



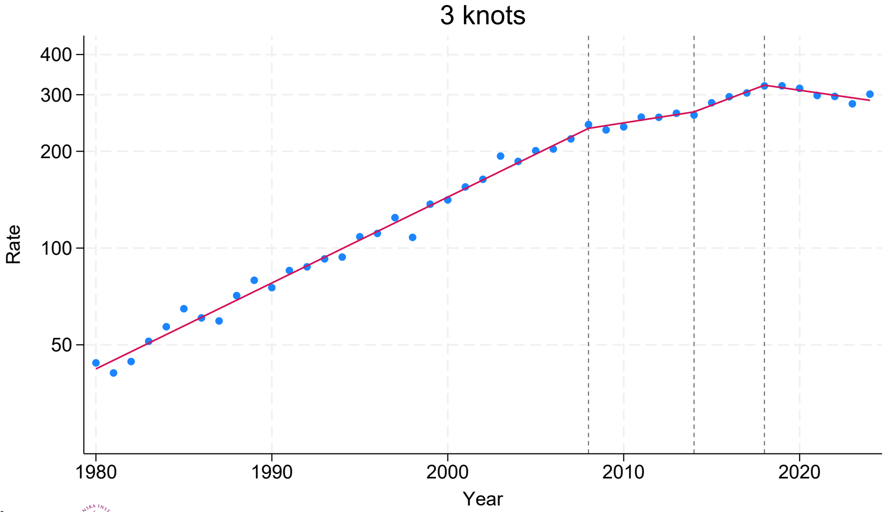
Incidence of lung cancer in Norway (females 70-79)



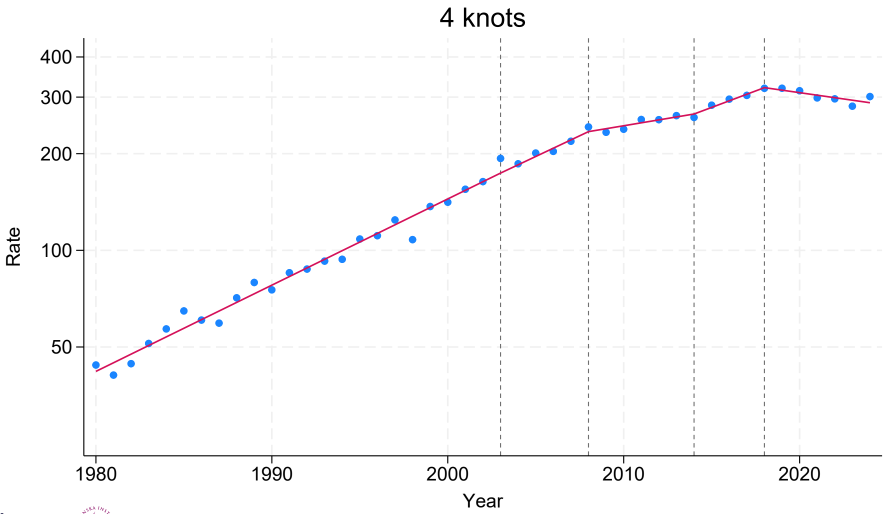
Incidence of lung cancer in Norway (females 70-79)



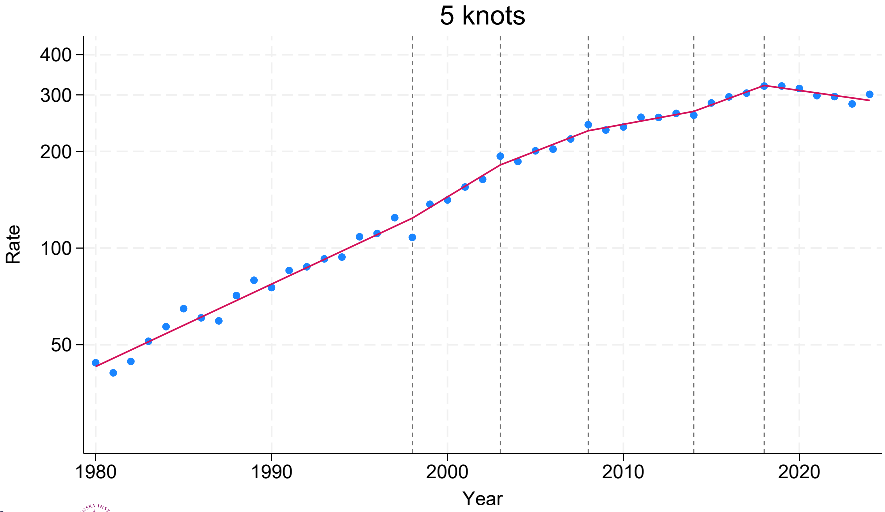
Incidence of lung cancer in Norway (females 70-79)



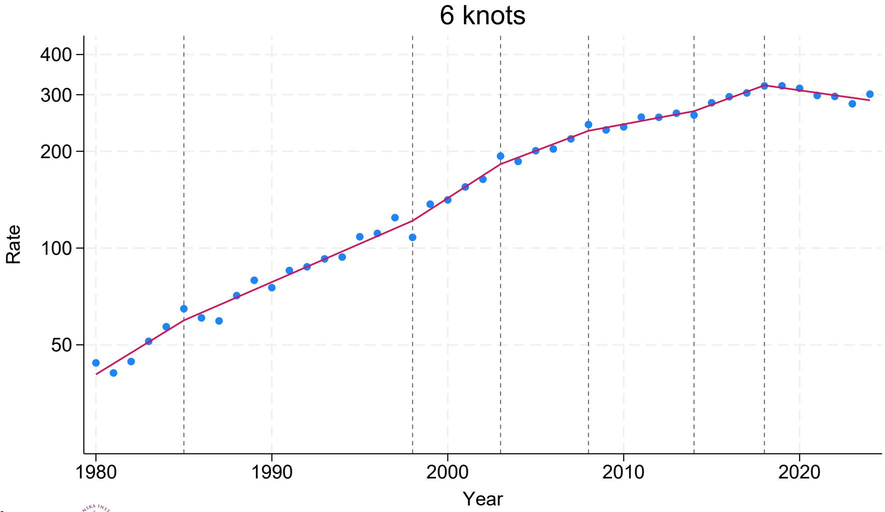
Incidence of lung cancer in Norway (females 70-79)



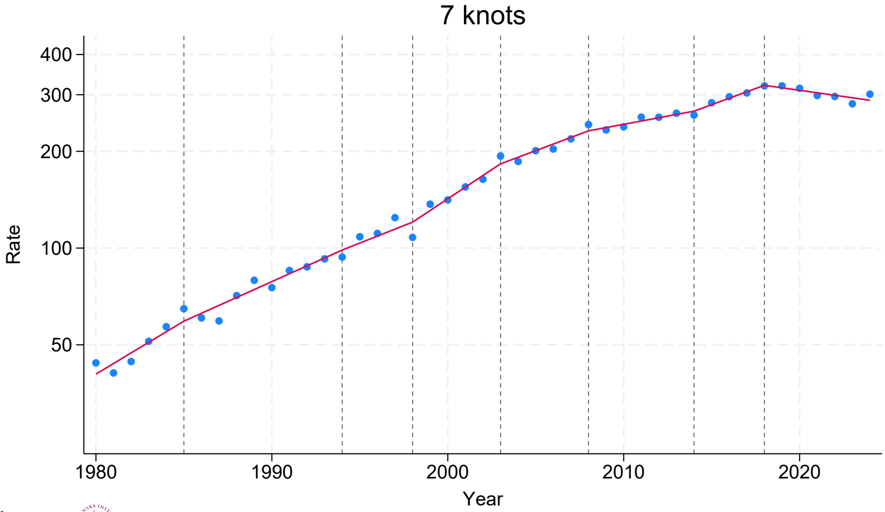
Incidence of lung cancer in Norway (females 70-79)



Incidence of lung cancer in Norway (females 70-79)



Incidence of lung cancer in Norway (females 70-79)



Joinpoint model

- For K knots $\tau_1, \ldots, \tau_K$, the joinpoint model is

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} \delta_k (x_i - \tau_k)^+ + \epsilon_i$$

  where $u^+ = \max(0, u)$
- We want to estimate
  - The number of knots, K,
  - The location of the knots, τ_k, $k = 1, \ldots, K$,
  - The intercept and gradient before the first knot, β_0, β_1,
  - The change in gradient at each knot, δ_k, $k = 1, \ldots, K$,
- For cancer trends the outcome, y is usually a log(rate).
- Rates are estimated with uncertainty: incoporate weights, $w_i = 1/SE(y_i)^2$.

Annual percent change (APC)

- We are assuming linearity between knots (joinpoints).
- The gradient for the increase/decrease in log(rate) per year can be obtained.
- If γ_j is the gradient is the j^{th} segment, then the annual percent change (APC) is

$$APC_j = 100 \times (e^{\gamma_j} - 1)$$

- with K knots (joinpoints) there are $K + 1$ segments.
- Common way to summarise trends.

NIH Joinpoint Software

- Developed by the Surveillance Research Program at the National Cancer Institute (NIH).
- Available as a standalone, point and click application, and as command line versions.
- Good software with detailed documentation and papers describing methodology[2–5].
- https://surveillance.cancer.gov/joinpoint/

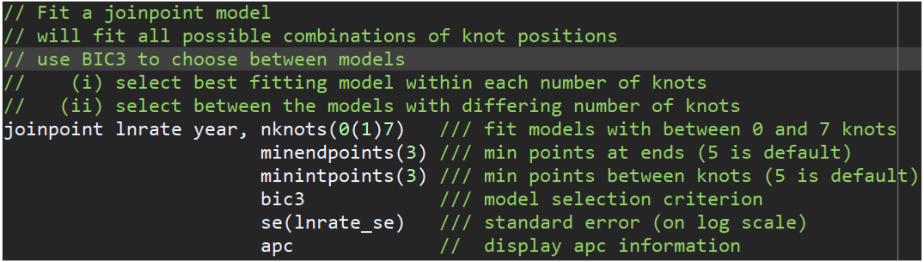
NIH Joinpoint Software

- Developed by the Surveillance Research Program at the National Cancer Institute (NIH).
- Available as a standalone, point and click application, and as command line versions.
- Good software with detailed documentation and papers describing methodology[2–5].
- https://surveillance.cancer.gov/joinpoint/
- Much more convenient for us to fit a model from within Stata.
  - Loop over many cancers / age groups /sex etc.
- The NIH joinpoint software very useful for checking for consistency of results.

Joinpoint models

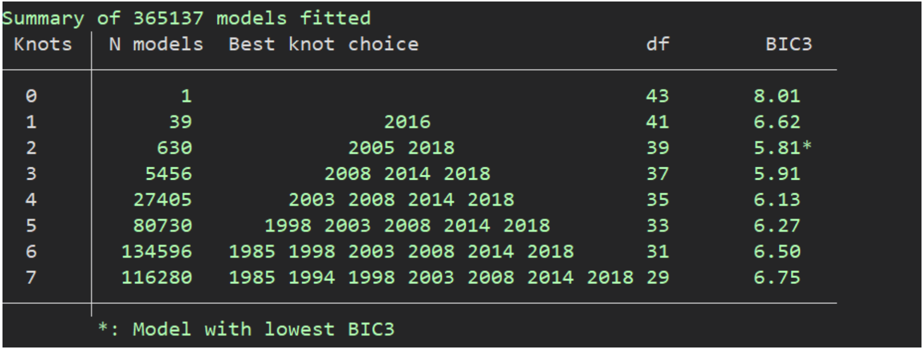
- To fit a joinpoint model, we need to initial specify,
  - The minimum and maximum number of knots.
  - The minimum number of data points between knots.
  - The minimum number of data points before the first knot and after the last knot.
  - How to choose between different models.
- I will first show an example of using the `joinpoint` command, and then discuss some details of the implementation.

Example of joinpoint



```
// Fit a joinpoint model
// will fit all possible combinations of knot positions
// use BIC3 to choose between models
//    (i) select best fitting model within each number of knots
//   (ii) select between the models with differing number of knots
joinpoint lnrate year, nknots(0(1)7)   /// fit models with between 0 and 7 knots
                       minendpoints(3) /// min points at ends (5 is default)
                       minintpoints(3) /// min points between knots (5 is default)
                       bic3            /// model selection criterion
                       se(lnrate_se)   /// standard error (on log scale)
                       apc             //  display apc information
```

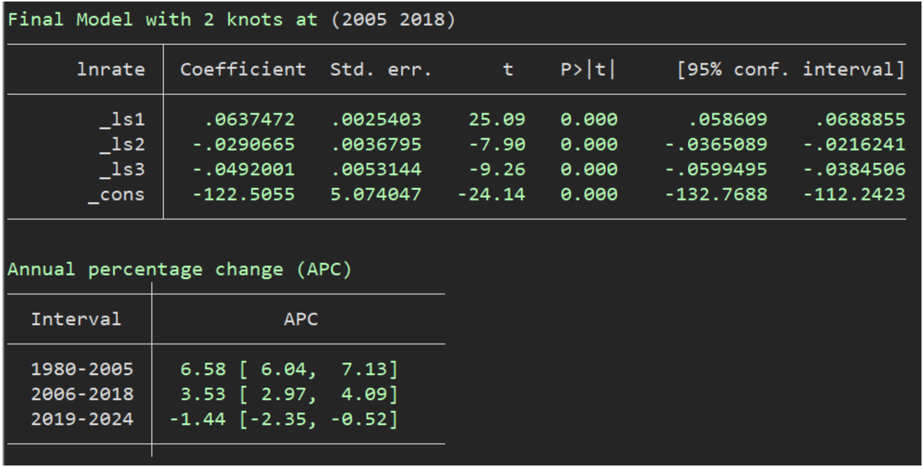
Example of `joinpoint`: Output 1



```
Summary of 365137 models fitted
 Knots  │ N models  Best knot choice                   df         BIC3
────────┼───────────────────────────────────────────────────────────────────
        │
  0     │       1                                      43        8.01
  1     │      39                2016                  41        6.62
  2     │     630             2005 2018                39        5.81*
  3     │    5456           2008 2014 2018             37        5.91
  4     │   27405        2003 2008 2014 2018           35        6.13
  5     │   80730      1998 2003 2008 2014 2018        33        6.27
  6     │  134596   1985 1998 2003 2008 2014 2018      31        6.50
  7     │  116280   1985 1994 1998 2003 2008 2014 2018 29        6.75
        │
────────┼───────────────────────────────────────────────────────────────────
         *: Model with lowest BIC3
```

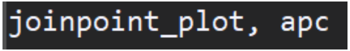
- You can save details of all fitted models using the `savemodelfit` option.

Example of `joinpoint`: Output 2

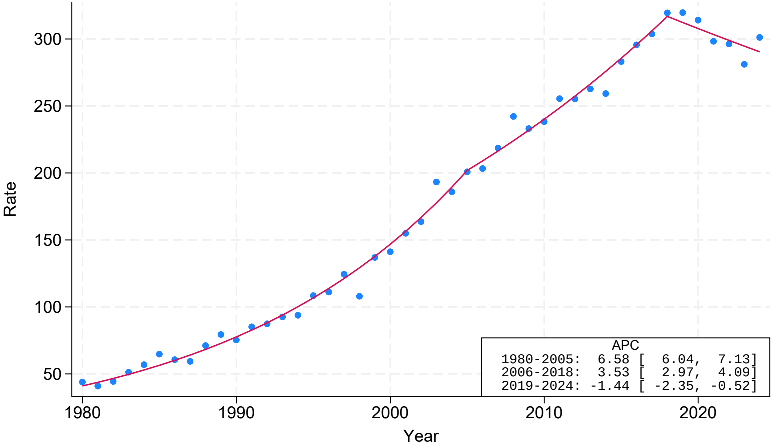


```
Final Model with 2 knots at (2005 2018)

------------------------------------------------------------------------------
      lnrate | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        _ls1 |   .0637472   .0025403    25.09   0.000      .058609    .0688855
        _ls2 |  -.0290665   .0036795    -7.90   0.000    -.0365089   -.0216241
        _ls3 |  -.0492001   .0053144    -9.26   0.000    -.0599495   -.0384506
       _cons |  -122.5055   5.074047   -24.14   0.000    -132.7688   -112.2423
------------------------------------------------------------------------------


Annual percentage change (APC)

-----------------------------------
  Interval |           APC
-----------+-----------------------
 1980-2005 |  6.58 [ 6.04,   7.13]
 2006-2018 |  3.53 [ 2.97,   4.09]
 2019-2024 | -1.44 [-2.35, -0.52]
-----------------------------------
```

Example of joinpoint: Plotting



```
joinpoint_plot, apc
```



```
            APC
1980-2005:  6.58 [  6.04,  7.13]
2006-2018:  3.53 [  2.97,  4.09]
2019-2024: -1.44 [ -2.35, -0.52]
```

Knot selection

- `joinpoint` will fit all possible knot combinations
- Mata function first works out all possible combinations of knots, subject to:
  - minimum and maximum number of knots
  - minimum data points between knots
  - minimum data points before first and after last knot.
- Loops over all combinations, fit model, store results.
- Finally, select the best model using a specified criterion.

Computationally Intensive

- We end up having to fit many models.
- For example, with 80 data points with a maximum of 5 knots and a minimum of 5 data points between knots there are **2,496,185** different models.
- These are regression models, so we could use `regress`.
- However, we can also use `mm_ls()` from Ben Jann's `moremata` set of functions.
- What is the speed gain?
  - I will compare fitting 100,000 models.

Using regress

```
clear
set obs 80
gen x = _n
gen y = runiform()
gensplines x, df(3) type(bs) gen(bs) degree(1)

timer clear
timer on 1
forvalues i = 1/100000 {
  qui regress y bs*
}
timer off 1
```

Using mm_ls()

```
mata:
  bs = st_data(.,("bs1","bs2","bs3"),.)
  y  = st_data(.,"y",.)

  timer_on(2)
  for(i=1;i<=100000;i++) {
    S = mm_ls(y, bs,1,1)
  }
  timer_off(2)
end
```

Compare times

Method	Time (seconds)	Relative Speed
`regress`	t_1	
`mm_ls()`	t_2	$100 \times t_2 / t_1$

- Have a guess at what $100 \times t_2 / t_1$ is?

Compare times

Method	Time (seconds)	Relative Speed
`regress`	453.35	
`mm_ls()`	1.43	0.31

- Dramatic increase in speed.
- In `joinpoint` models are fitted using `mm_ls()`, with the final model fitted using `regress`.
- Important when we loop over many cancer sites, age groups etc.

Adjusting SEs for model selection.

- We are selecting the number of knots and their locations. This model selection should be accounted for when quantifying uncertainty.
- 1 df for each knot in the model and an additional df to account for selection of a knot position.
  - Use `dof()` option when using `regress`.
- Similar to selecting powers in fractional polynomials[6].
- Additionally, Kim and Kim[7] and others showed that improved coverage with fitting a linear spline model without continuity constraints and using the submatrix that corresponds to the slope parameters.
  - Use `ereturn repost ...` to modify `e(V)`.

Model Selection

- Different ways to select between models.

$$BIC = \frac{\ln(SSE)}{N} + \frac{2(k+2)\ln(N)}{N}$$

$$BIC3 = \frac{\ln(SSE)}{N} + \frac{3(k+2)\ln(N)}{N}$$

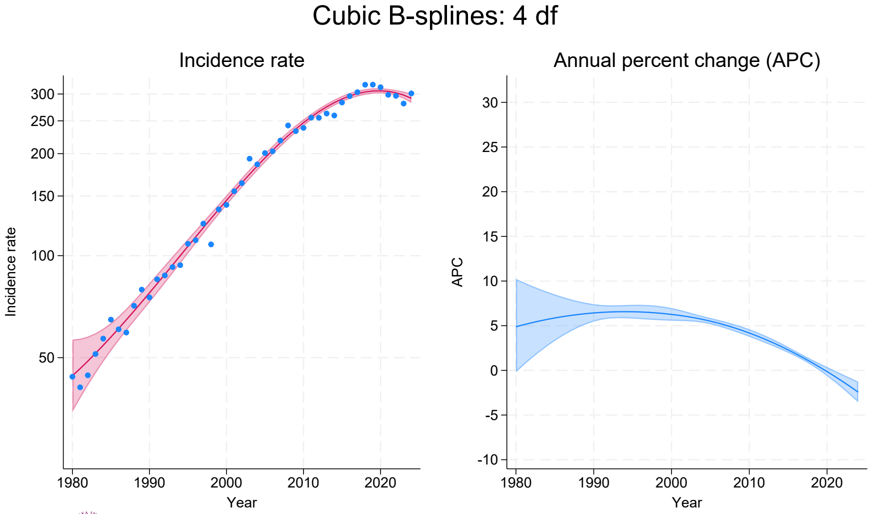
$$WBIC = (1-w)BIC + wBIC3$$

w increases with increasing change in gradient between knots

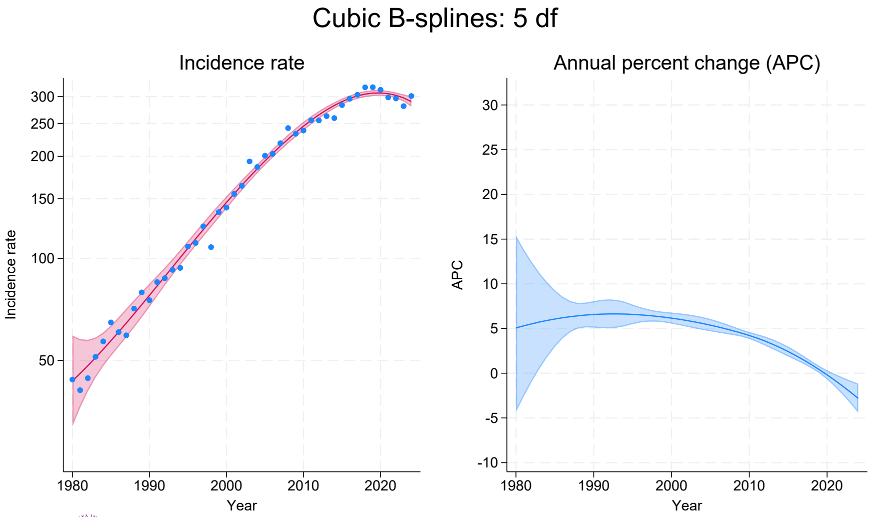
Why not cubic splines?

- Initially I was not keen on these models.
- I am used to using more flexible spline functions, e.g. natural splines, B-splines, I-splines, M-splines as part of the `gensplines` command.
- The world is usually more complex that a series of straight lines.
- However, joinpoint models provide a quick, easy to interpret summary of trends over time.
- I do not see joinpoint models as selecting an exact time of change, but as a useful tool in descriptive analysis.
- Particularly useful when analysing many cancer sites with a varity of subgroups.

Incidence of lung cancer in Norway (females 70-79)



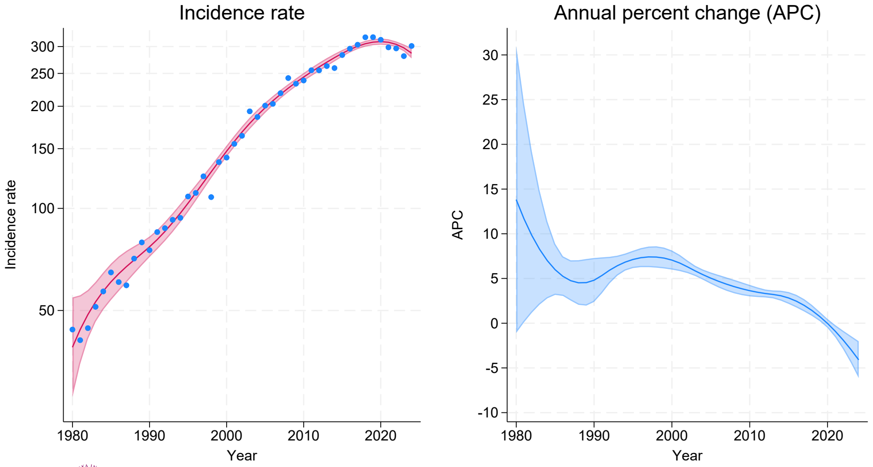
Incidence of lung cancer in Norway (females 70-79)



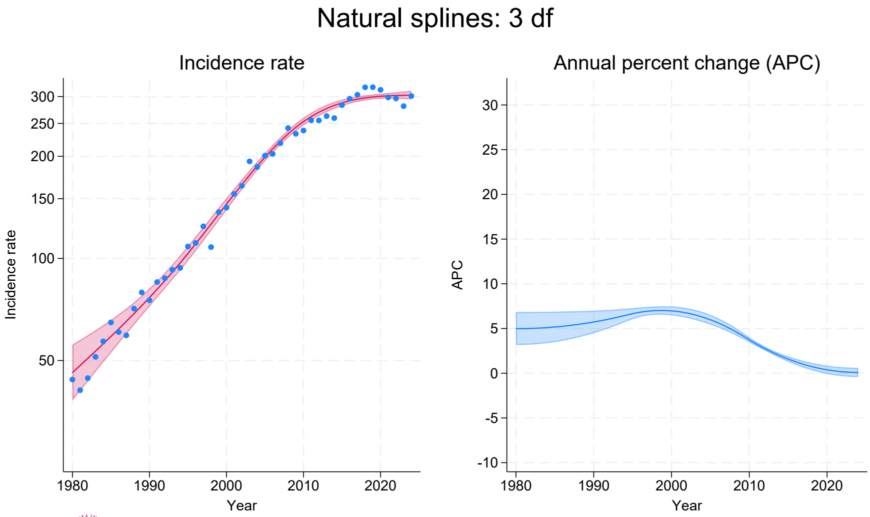
Cancer Registry of Norway FHI Karolinska Institutet

Incidence of lung cancer in Norway (females 70-79)

Cubic B-splines: 6 df



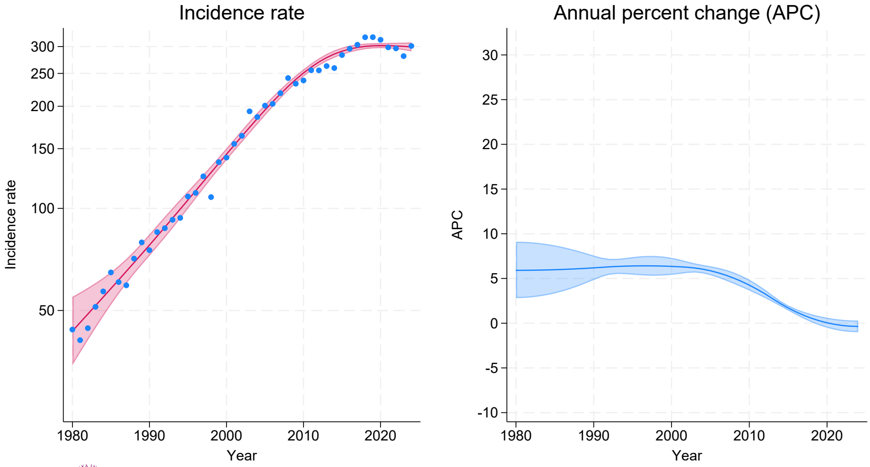
Incidence of lung cancer in Norway (females 70-79)



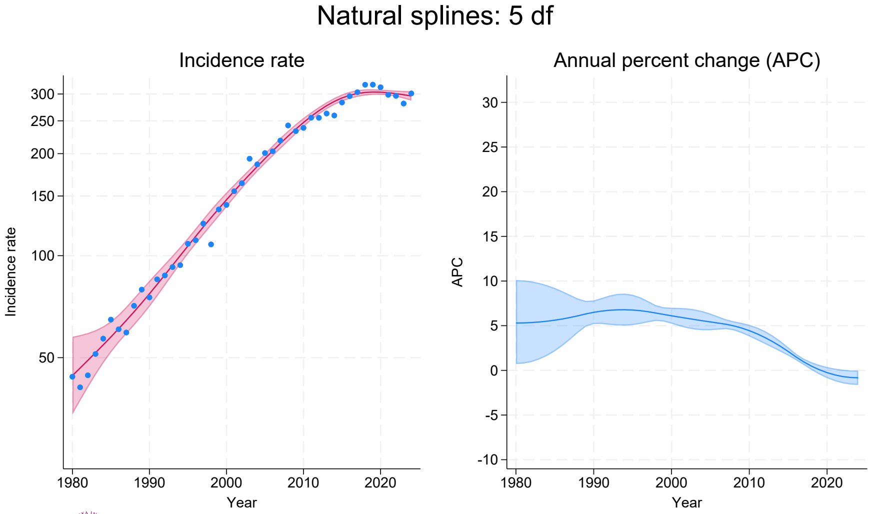
Cancer Registry of Norway FHI Karolinska Institutet

Incidence of lung cancer in Norway (females 70-79)

Natural splines: 4 df



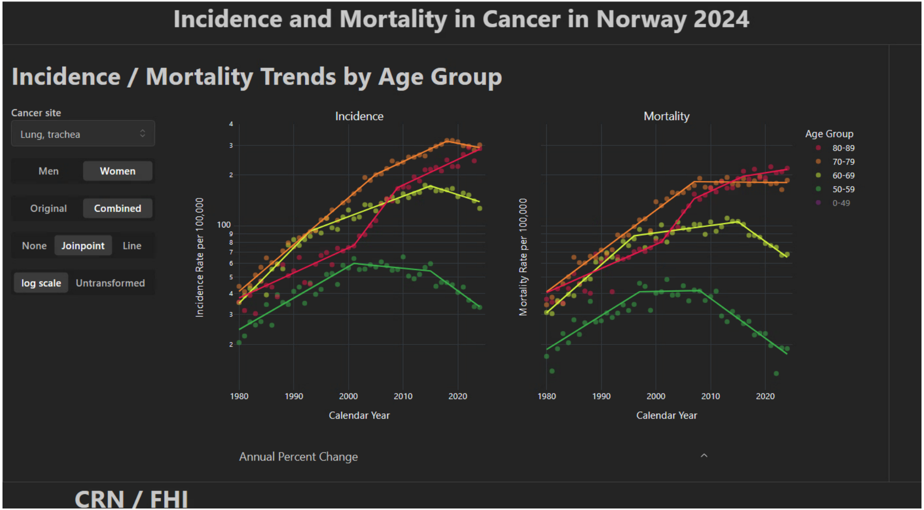
Incidence of lung cancer in Norway (females 70-79)



Interative App

- At the Cancer Registry of Norway we have used these models, initially for internal use.
- We are interested in a range of cancer, often stratified by age/sex.
- We export model estimates etc from Stata to Python and use in an interative plotly Dash app.

Interative App Example



Incidence and Mortality in Cancer in Norway 2024

Incidence / Mortality Trends by Age Group

Cancer site

Lung, trachea

Men Women

Original Combined

None Joinpoint Line

log scale Untransformed

Annual Percent Change

CRN / FHI

Availability

- Still some final checking to do before release.
- Test version will be available on my website this weekend!.
- SSC version coming soon.

Summary

- Joinpoint models useful descriptive tool for cancer trends.
- Good stand alone software exists, but useful to have a Stata implementation.
- Still a couple of things to add.
  - Average annual percent change (AAPC)
  - Confidence intervals for knot locations
- Will be released on SSC soon.

References

1. *Cancer in Norway 2023 - Cancer Incidence, Mortality, Survival and Prevalence in Norway.* (Cancer Registry of Norway, Oslo, 2024).
2. Kim, H., Chen, H., Byrne, J., Wheeler, B. & Feuer, E. J. Twenty Years since Joinpoint 1.0: Two Major Enhancements, Their Justification, and Impact. *Statistics in Medicine* **41,** 3102–3130 (2022).
3. Kim, H.-J., Fay, M. P., Feuer, E. J. & Midthune, D. N. Permutation Tests for Joinpoint Regression with Applications to Cancer Rates. *Statistics in Medicine* **19,** 335–351 (2000).
4. Clegg, L. X., Hankey, B. F., Tiwari, R., Feuer, E. J. & Edwards, B. K. Estimating Average Annual per Cent Change in Trend Analysis. *Statistics in Medicine* **28,** 3670–3682 (2009).
5. Kim, H. *et al.* Improved Confidence Interval for Average Annual Percent Change in Trend Analysis. *Statistics in Medicine* **36,** 3059–3074 (2017).
6. Royston, P. & Altman, D. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. *Applied Statistics* **43,** 429–467 (1994).
7. Kim, J. & Kim, H.-J. Applications of Asymptotic Inference in Segmented Line Regression. *Communications in Statistics - Theory and Methods* **50,** 5585–5606 (2021).