Epidemiology with R

Bendix Carstensen Steno Diabetes Center
Gentofte, Denmark
http://BendixCarstensen.com

Université Bordeaux
23 January 2015

http://BendixCarstensen.com/Epi

Study types and data types

Epidemiology with R
23 January 2015
Université Bordeaux
http://BendixCarstensen.com/Epi

study-types

Epidemiological study types

- **Cross-sectional** studies:
  What is disease status at a particular date
- **Follow-up** studies:
  What is the rate of disease occurrence
  - Fixed cohorts, population based surveys
  - Dynamic cohorts
  - An entire population followed through registers
    Medical demography
- **Case-control** studies:
  Compare cases with non-cases.
  - Sampling based on disease status
  - Partial measures of disease occurrence/presence

Epidemiological data types

- **Continuous** (metric) responses can emerge from any observational design.
- **Categorical** response data essentially always **derived** from follow-up data:
  - Tables of counts from a cross-sectional study.
  - Tables of counts and follow-up time.
  - Tables of case-control status and exposure.
- Continuous and categorical explanatory variables occur in any design.

Cross-sectional studies

- What **fraction of the population** has a certain characteristic (such as a diagnosis of diabetes or other disease).
- **Observations:** the entire population (or a sample of it) classified by disease status
- The **likelihood** is a binomial likelihood for

$$p = \mathrm{P}\left\{\text{presence of disease}\right\}$$

- . . . that is, how p depends on explanatory variables.

Follow-up studies

- **Medical demography** — describing the entire population w.r.t. disease status over time
  - An entire population is followed for a particular event of interest (CVD, death, . . .)
- **Epidemiological** (observational) study
  Part of the population (a cohort) is followed for a limited period of time
  - May not necessarily be generalizable.
  - — but can elucidate the size of exposure effects on disease occurrence.
  - Neither exposures nor outcomes need be representative — only their relationship.

Follow-up studies

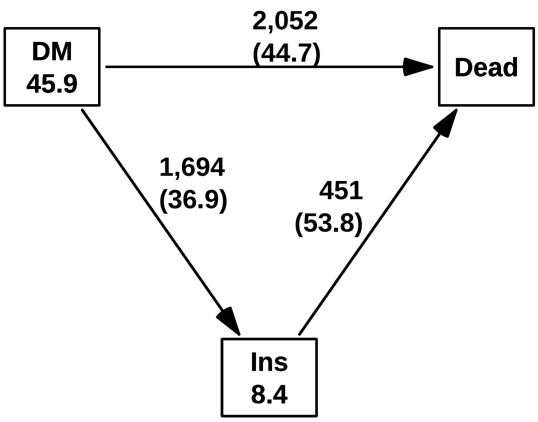
- **Observations** are (empirical) rates:
  (d, y): d events during y follow-up time
  (risk time, exposure time, person-years)
- **Models** for occurrence rates:

$$\lambda(t) = \mathrm{P}\left\{\text{event in } (t, t+h) \mid \text{ no event till } t\right\}/h$$

- The **likelihood** for this is proportional to a Poisson likelihood (if λ is constant):

$$\text{log-lik} = \ell(\lambda|d, y) = d\log(\lambda) + \lambda y$$

How a follow-up study looks



Follow-up studies

- Each transition can be considered separately
- Rates modelled seperately (or jointly)
- Probabilities can be derived from estimated rates
- Simplest probability is:

$$S(t) = \mathrm{P}\left\{\text{survive till time } t\right\}$$

- Other probabilities of interest, *e.g.*:

$$P_c(t) = \mathrm{P}\left\{\text{die from cause } c \text{ before } t\right\}$$

— depend on more than one rate.

Case-control studies

- Events (cases) are sampled.
- But risk time is not...
  — it is replaced by a carefully chosen sample of the non-event persons.
- The likelihood is a binomial likelihood for

$$p = \mathrm{P}\{\text{case} \mid \text{included in the study}\}$$

  which contains the parameters of interest (and some not of any interest) *e.g.* rate-ratios.

Simple analyses of relationships

Epidemiology with R
23 January 2015
Université Bordeaux
http://BendixCarstensen.com/Epi

simp-ana

Variables in models

Response variables must be numeric.

- Metric (a measurement with units)
- Binary (two values coded 0/1)
- Failure (does the subject fail at end of follow-up) and FU-time
- Count (aggregated failure data)

Explanatory variables can be:

- Numeric
- Factor (classes)

Simple analyses

- Response as a function of exposure:

$$y = \mu + \beta x$$

- — **controlled** for levels of a confounder:

$$y = \mu + \beta x + \delta_c$$

confounding refers to the **change** in β

- — **stratified** by levels of an effect modifier:

$$y = \mu + \beta_e x + \gamma_e$$

effect modification is called **interaction** in statistics

Simple analyses made simple: `effx`

```
effx( response, type = "metric",
                 fup = NULL,
            exposure,
              strata = NULL,
             control = NULL,
             weights = NULL,
               alpha = 0.05,
                base = 1,
              digits = 3,
                data = NULL )
```

— so let's see how it works; exercise

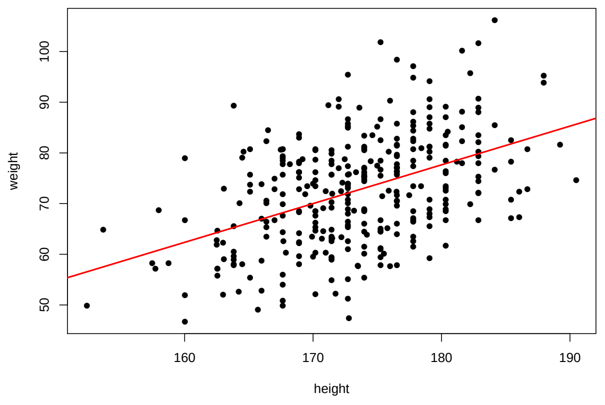
Linear relationships

Epidemiology with R
23 January 2015
Université Bordeaux
http://BendixCarstensen.com/Epi

lin-mod

Linear models

```
> options( show.signif.stars=FALSE, width=60 )
> library( Epi )
> data( diet )
> names( diet )

 [1] "id"         "doe"        "dox"        "dob"
 [5] "y"          "fail"       "job"        "month"
 [9] "energy"     "height"     "weight"     "fat"
[13] "fibre"      "energy.grp" "chd"

> with( diet, plot(weight~height,pch=16) )
> abline( lm(weight~height,data=diet), col="red", lwd=2 )
```

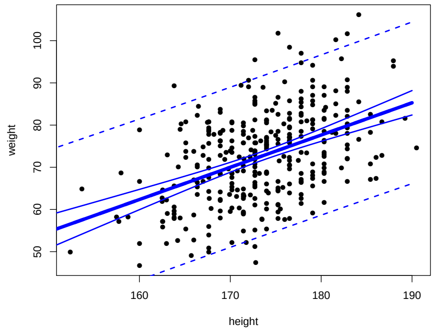




```
> with( diet, plot(weight~height,pch=16) )
> abline( lm(weight~height,data=diet), col="red", lwd=2 )
```


Linear models, extracting estimates

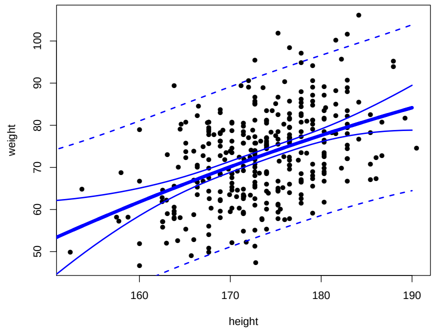
```
> ml <- lm( weight ~ height, data=diet )

> summary( ml )

Call:
lm(formula = weight ~ height, data = diet)

Residuals:
     Min       1Q   Median       3Q      Max
-24.7361  -7.4553   0.1608   6.9384  27.8130

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -59.91601   14.31557  -4.185 3.66e-05
height        0.76421    0.08252   9.261  < 2e-16

Residual standard error: 9.625 on 330 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.2063, Adjusted R-squared:  0.2039
F-statistic: 85.76 on 1 and 330 DF,  p-value: < 2.2e-16

> round( ci.lin( ml ), 4 )
```

Linear models, prediction




```
> ml <- lm( weight ~ height, data=diet )
> nd <- data.frame( height = 150:190 )
> pr.co <- predict( ml, newdata=nd, interval="conf" )
> pr.pr <- predict( ml, newdata=nd, interval="pred" )
> with( diet, plot( weight ~ height, pch=16 )  )
> matlines( nd$height, pr.co, lty=1, lwd=c(5,2,2), col="blue" )
> matlines( nd$height, pr.pr, lty=2, lwd=c(5,2,2), col="blue" )
```


non-Linear models, prediction




```
> mq <- lm( weight ~ height + I(height^2), data=diet )
> nd <- data.frame( height = 150:190 )
> pr.co <- predict( mq, newdata=nd, interval="conf" )
> pr.pr <- predict( mq, newdata=nd, interval="pred" )
> with( diet, plot( weight ~ height, pch=16 )  )
> matlines( nd$height, pr.co, lty=1, lwd=c(5,2,2), col="blue" )
> matlines( nd$height, pr.pr, lty=2, lwd=c(5,2,2), col="blue" )
```


Curved relationships

Epidemiology with R
23 January 2015
Université Bordeaux
http://BendixCarstensen.com/Epi

crv-mod

Testis cancer

Testis cancer in Denmark:

```
> options( show.signif.stars=FALSE )
> library( Epi )
> data( testisDK )
> str( testisDK )

  'data.frame': 4860 obs. of  4 variables:
   $ A: num  0 1 2 3 4 5 6 7 8 9 ...
   $ P: num  1943 1943 1943 1943 1943 ...
   $ D: num  1 1 0 1 0 0 0 0 0 0 ...
   $ Y: num  39650 36943 34588 33267 32614 ...

> head( testisDK )

    A    P D        Y
  1 0 1943 1 39649.50
  2 1 1943 1 36942.83
  3 2 1943 0 34588.33
  4 3 1943 1 33267.00
  5 4 1943 0 32614.00
  6 5 1943 0 32020.33
```


Cases, PY and rates

```
> stat.table( list(A=floor(A/10)*10,
+                  P=floor(P/10)*10),
+             list( D=sum(D),
+                   Y=sum(Y/1000),
+                rate=ratio(D,Y,10^5) ),
+             margins=TRUE, data=testisDK )

   -------------------------------------------------------------
          -------------------------------P----------------------
   A           1940      1950      1960      1970      1980      1990
   -------------------------------------------------------------
   0          10.00      7.00     16.00     18.00      9.00     10.00
            2604.66   4037.31   3884.97   3820.88   3070.87   2165.54
               0.38      0.17      0.41      0.47      0.29      0.46

   10         13.00     27.00     37.00     72.00     97.00     75.00
            2135.73   3505.19   4004.13   3906.08   3847.40   2260.97
               0.61      0.77      0.92      1.84      2.52      3.32

   20        124.00    221.00    280.00    535.00    724.00    557.00
            2225.55   2923.22   3401.65   4028.57   3941.18   2824.58
               5.57      7.56      8.23     13.28     18.37     19.72
```


Linear effects in glm

How do rates depend on age?

```
> ml <- glm( D ~ A, offset=log(Y), family=poisson, data=testisDK
> round( ci.lin( ml ), 4 )

              Estimate StdErr         z P    2.5%   97.5%
  (Intercept)  -9.7755 0.0207 -472.3164 0 -9.8160 -9.7349
  A             0.0055 0.0005   11.3926 0  0.0045  0.0064

> round( ci.exp( ml ), 4 )

              exp(Est.)   2.5%  97.5%
  (Intercept)    0.0001 0.0001 0.0001
  A              1.0055 1.0046 1.0064
```

Linear increase of log-rates by age

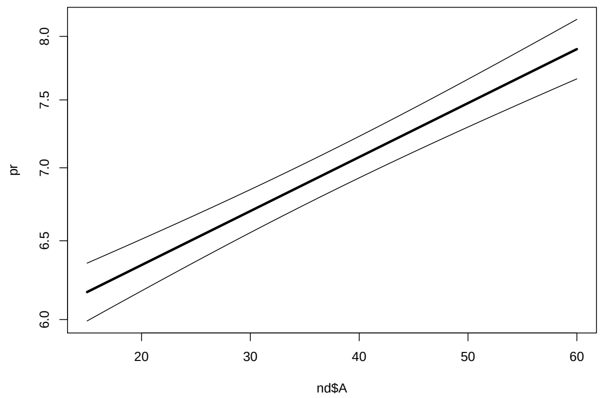
Linear effects in glm

```
> nd <- data.frame( A=15:60, Y=10^5 )
> pr <- ci.pred( m1, newdata=nd )
> head( pr )

  Estimate     2.5%    97.5%
1 6.170105 5.991630 6.353896
2 6.204034 6.028525 6.384652
3 6.238149 6.065547 6.415662
4 6.272452 6.102689 6.446937
5 6.306943 6.139944 6.478485
6 6.341624 6.177301 6.510319

> matplot( nd$A, pr,
+          type="l", lty=1, lwd=c(3,1,1), col="black", log="y" )
```


Linear effects in glm

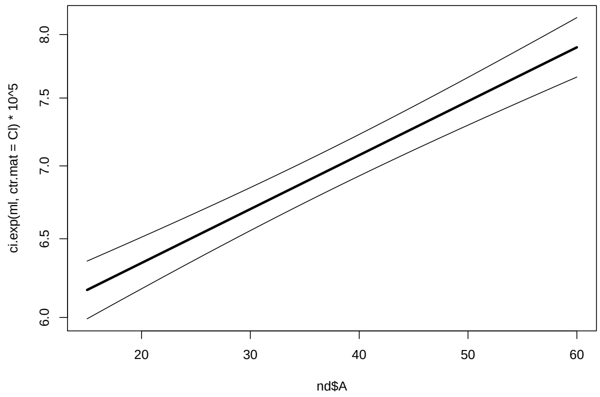
```
> round( ci.lin( m1 ), 4 )

            Estimate StdErr         z P    2.5%   97.5%
(Intercept)  -9.7755 0.0207 -472.3164 0 -9.8160 -9.7349
A             0.0055 0.0005   11.3926 0  0.0045  0.0064

> C1 <- cbind( 1, nd$A )
> head( C1 )

     [,1] [,2]
[1,]    1   15
[2,]    1   16
[3,]    1   17
[4,]    1   18
[5,]    1   19
[6,]    1   20

> matplot( nd$A, ci.exp( m1, ctr.mat=C1 ),
+          type="l", lty=1, lwd=c(3,1,1), col="black", log="y" )
```


Linear effects in glm




```
> matplot( nd$A, pr,
+          type="l", lty=1, lwd=c(3,1,1), col="black", log="y" )
```

Linear effects in glm




```
> matplot( nd$A, ci.exp( ml, ctr.mat=Cl )*10^5,
+          type="l", lty=1, lwd=c(3,1,1), col="black", log="y" )
```


Quadratic effects in glm

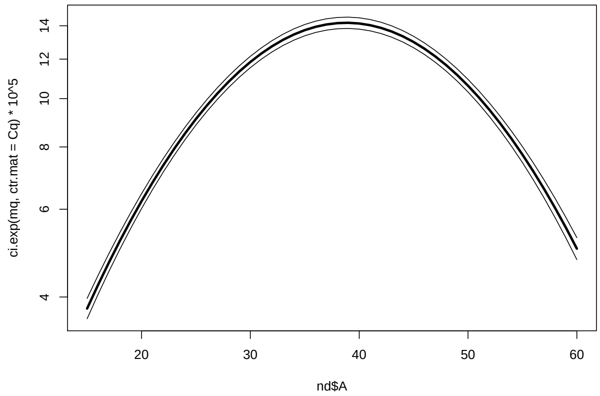
How do rates depend on age?

```
> mq <- glm( D ~ A + I(A^2),
+            offset=log(Y), family=poisson, data=testisDK )
> round( ci.lin( mq ), 4 )

              Estimate StdErr         z P     2.5%    97.5%
(Intercept) -12.3656 0.0596 -207.3611 0 -12.4825 -12.2487
A              0.1806 0.0033   54.8290 0   0.1741   0.1871
I(A^2)        -0.0023 0.0000  -53.7006 0  -0.0024  -0.0022

> round( ci.exp( mq ), 4 )

            exp(Est.)  2.5%  97.5%
(Intercept)    0.0000 0.0000 0.0000
A              1.1979 1.1902 1.2057
I(A^2)         0.9977 0.9976 0.9978
```


Quadratic effect in glm

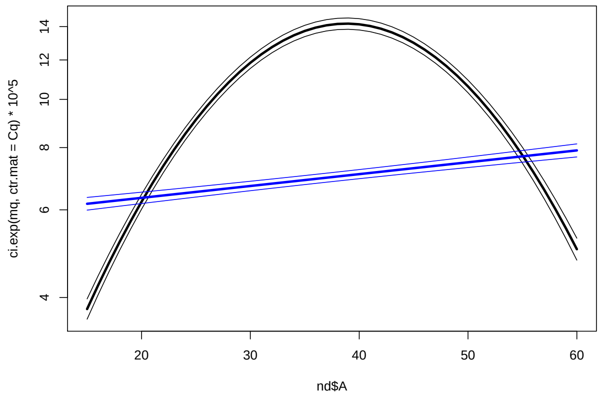
```
> round( ci.lin( mq ), 4 )

              Estimate StdErr         z P     2.5%    97.5%
(Intercept) -12.3656 0.0596 -207.3611 0 -12.4825 -12.2487
A              0.1806 0.0033   54.8290 0   0.1741   0.1871
I(A^2)        -0.0023 0.0000  -53.7006 0  -0.0024  -0.0022

> Cq <- cbind( 1, 15:60, (15:60)^2 )
> head( Cq, 4 )

     [,1] [,2] [,3]
[1,]    1   15  225
[2,]    1   16  256
[3,]    1   17  289
[4,]    1   18  324

> matplot( nd$A, ci.exp( mq, ctr.mat=Cq )*10^5,
+          type="l", lty=1, lwd=c(3,1,1), col="black", log="y" )
```

Quadratic effect in glm




```
> matplot( nd$A, ci.exp( mq, ctr.mat=Cq )*10^5,
+           type="l", lty=1, lwd=c(3,1,1), col="black", log="y" )
```


Quadratic effect in glm

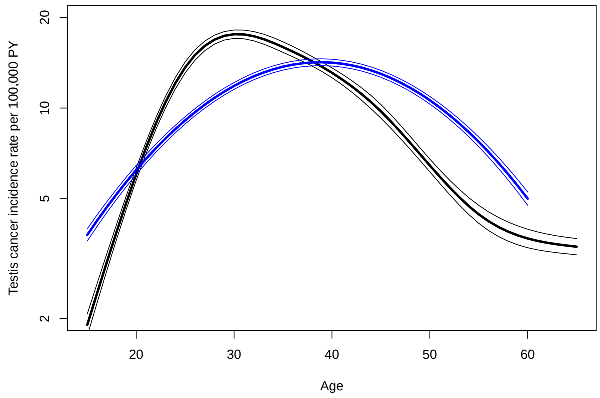



```
> matplot( nd$A, ci.exp( mq, ctr.mat=Cq )*10^5,
+           type="l", lty=1, lwd=c(3,1,1), col="black", log="y" )
> matlines( nd$A, ci.exp( ml, ctr.mat=Cl )*10^5,
+            type="l", lty=1, lwd=c(3,1,1), col="blue" )
```


Spline effects in glm

```
> library( splines )
> ms <- glm( D ~ Ns(A,knots=seq(15,65,10)),
+              offset=log(Y), family=poisson, data=testisDK )
> round( ci.exp( ms ), 3 )

                                exp(Est.)   2.5%  97.5%
(Intercept)                         0.000  0.000  0.000
Ns(A, knots = seq(15, 65, 10))1     8.548  7.650  9.551
Ns(A, knots = seq(15, 65, 10))2     5.706  4.998  6.514
Ns(A, knots = seq(15, 65, 10))3     1.002  0.890  1.128
Ns(A, knots = seq(15, 65, 10))4    14.402 11.896 17.436
Ns(A, knots = seq(15, 65, 10))5     0.466  0.429  0.505

> aa <- 15:65
> As <- Ns( aa, knots=seq(15,65,10) )
> head( As )

                1 2           3          4           5
[1,] 0.0000000000 0  0.00000000 0.00000000  0.00000000
[2,] 0.0001666667 0 -0.02527011 0.07581034 -0.05054022
[3,] 0.0013333333 0 -0.05003313 0.15009940 -0.10006626
[4,] 0.0045000000 0 -0.07378197 0.22134590 -0.14756393
```

Spline effects in glm

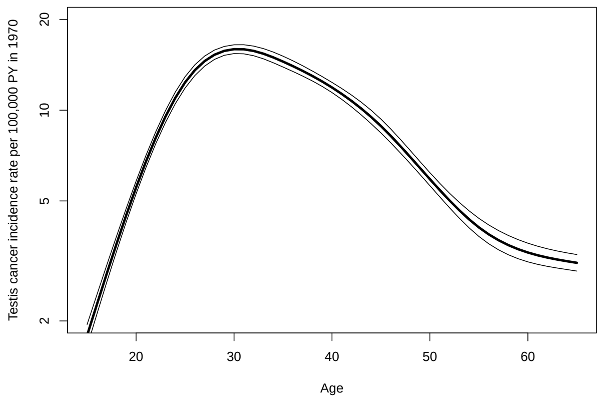



```
> matplot( aa, ci.exp( ms, ctr.mat=cbind(1,As) )*10^5,
+          log="y", xlab="Age", ylab="Testis cancer incidence ra
+          type="l", lty=1, lwd=c(3,1,1), col="black", ylim=c(2,
> matlines( nd$A, ci.exp( mq, ctr.mat=Cq )*10^5,
+           type="l", lty=1, lwd=c(3,1,1), col="blue" )
```


Adding a linear period effect

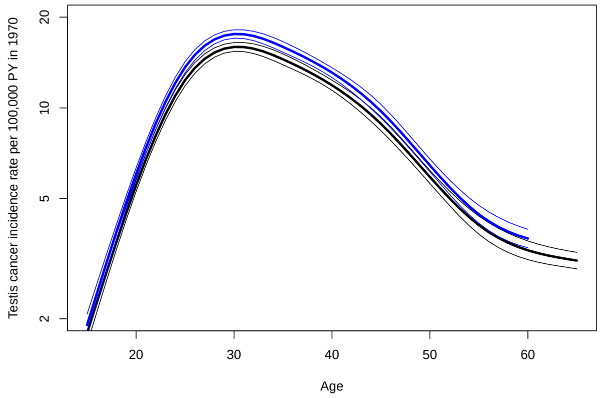
```
> msp <- glm( D ~ Ns(A,knots=seq(15,65,10)) + P,
+                offset=log(Y), family=poisson, data=testisDK )
> round( ci.lin( msp ), 3 )

                             Estimate StdErr       z     P
(Intercept)                   -58.105  1.444 -40.229 0.000
Ns(A, knots = seq(15, 65, 10))1   2.120  0.057  37.444 0.000
Ns(A, knots = seq(15, 65, 10))2   1.700  0.068  25.157 0.000
Ns(A, knots = seq(15, 65, 10))3   0.007  0.060   0.110 0.913
Ns(A, knots = seq(15, 65, 10))4   2.596  0.097  26.631 0.000
Ns(A, knots = seq(15, 65, 10))5  -0.780  0.042 -18.748 0.000
P                               0.024  0.001  32.761 0.000

> Ca <- cbind( 1, Ns( aa, knots=seq(15,65,10) ), 1970 )
> head( Ca )

           1 2           3          4           5
[1,] 1 0.0000000000 0  0.00000000 0.00000000  0.00000000 1970
[2,] 1 0.0001666667 0 -0.02527011 0.07581034 -0.05054022 1970
[3,] 1 0.0013333333 0 -0.05003313 0.15009940 -0.10006626 1970
[4,] 1 0.0045000000 0 -0.07378197 0.22134590 -0.14756393 1970
```


Adding a linear period effect




```
> matplot( aa, ci.exp( msp, ctr.mat=Ca )*10^5,
+          log="y", xlab="Age",
+          ylab="Testis cancer incidence rate per 100,000 PY in
+          type="l", lty=1, lwd=c(3,1,1), col="black", ylim=c(2,
```

Adding a linear period effect

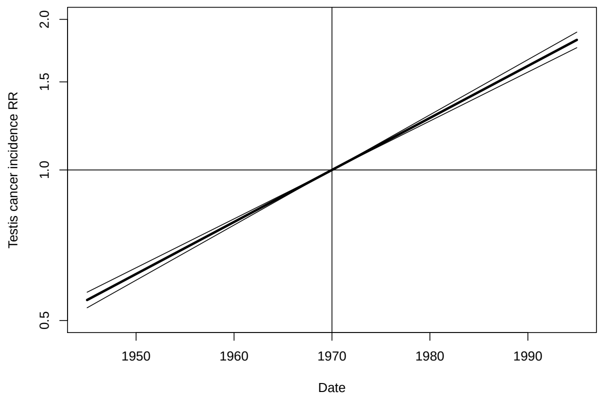



```
> matplot( aa, ci.exp( msp, ctr.mat=Ca )*10^5,
+          log="y", xlab="Age",
+          ylab="Testis cancer incidence rate per 100,000 PY in
+          type="l", lty=1, lwd=c(3,1,1), col="black", ylim=c(2,
> matlines( nd$A, ci.pred( ms, newdata=nd ),
+           type="l", lty=1, lwd=c(3,1,1), col="blue" )
```


The period effect

```
> round( ci.lin( msp ), 3 )

                                Estimate StdErr       z     P
(Intercept)                      -58.105  1.444 -40.229 0.000
Ns(A, knots = seq(15, 65, 10))1    2.120  0.057  37.444 0.000
Ns(A, knots = seq(15, 65, 10))2    1.700  0.068  25.157 0.000
Ns(A, knots = seq(15, 65, 10))3    0.007  0.060   0.110 0.913
Ns(A, knots = seq(15, 65, 10))4    2.596  0.097  26.631 0.000
Ns(A, knots = seq(15, 65, 10))5   -0.780  0.042 -18.748 0.000
P                                  0.024  0.001  32.761 0.000

> pp <- seq(1945,1995,0.2)
> Cp <- cbind( pp ) - 1970
> head( Cp )

        pp
[1,] -25.0
[2,] -24.8
[3,] -24.6
[4,] -24.4
[5,] -24.2
[6,] -24.0
```


Period effect

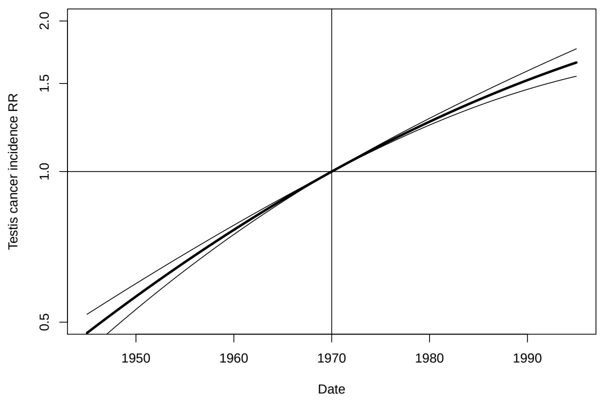



```
> matplot( pp, ci.exp( msp, subset="P", ctr.mat=Cp ),
+          log="y", ylim=c(0.5,2), xlab="Date",
+          ylab="Testis cancer incidence RR",
+          type="l", lty=1, lwd=c(3,1,1), col="black" )
> abline( h=1, v=1970 )
```

A quadratic period effect

```
> mspq <- glm( D ~ Ns(A,knots=seq(15,65,10)) + P + I(P^2),
+              offset=log(Y), family=poisson, data=testisDK
> round( ci.exp( mspq ), 3 )

                                 exp(Est.)   2.5%  97.5%
(Intercept)                          0.000  0.000  0.000
Ns(A, knots = seq(15, 65, 10))1      8.356  7.478  9.337
Ns(A, knots = seq(15, 65, 10))2      5.513  4.829  6.295
Ns(A, knots = seq(15, 65, 10))3      1.006  0.894  1.133
Ns(A, knots = seq(15, 65, 10))4     13.439 11.101 16.269
Ns(A, knots = seq(15, 65, 10))5      0.458  0.422  0.497
P                                    2.189  1.457  3.291
I(P^2)                               1.000  1.000  1.000

> Cq <- cbind( pp-1970, pp^2-1970^2 )
> head( Cq )

      [,1]      [,2]
[1,] -25.0 -97875.00
[2,] -24.8 -97096.96
[3,] -24.6 -96318.84
[4,] -24.4 -95540.64
```


A quadratic period effect




```
> matplot( pp, ci.exp( mspq, subset="P", ctr.mat=Cq ),
+          log="y", ylim=c(0.5,2), xlab="Date",
+          ylab="Testis cancer incidence RR",
+          type="l", lty=1, lwd=c(3,1,1), col="black" )
> abline( h=1, v=1970 )
```

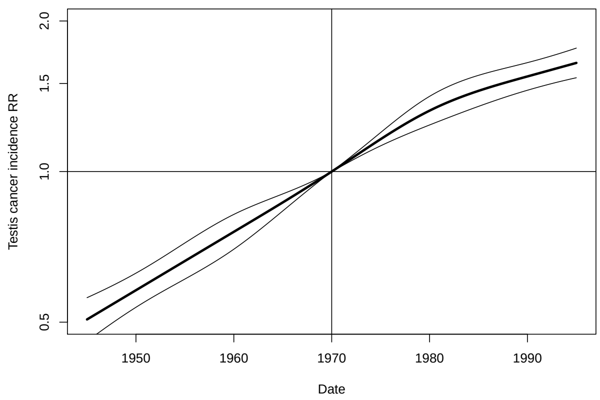

A spline period effect

```
> msps <- glm( D ~ Ns(A,knots=seq(15,65,10)) +
+              Ns(P,knots=seq(1950,1990,10)),
+              offset=log(Y), family=poisson, data=testisDK
> round( ci.exp( msps ), 3 )

                                     exp(Est.)   2.5%  97.5%
(Intercept)                              0.000  0.000  0.000
Ns(A, knots = seq(15, 65, 10))1          8.327  7.452  9.305
Ns(A, knots = seq(15, 65, 10))2          5.528  4.842  6.312
Ns(A, knots = seq(15, 65, 10))3          1.007  0.894  1.133
Ns(A, knots = seq(15, 65, 10))4         13.447 11.107 16.279
Ns(A, knots = seq(15, 65, 10))5          0.458  0.422  0.497
Ns(P, knots = seq(1950, 1990, 10))1      1.711  1.526  1.918
Ns(P, knots = seq(1950, 1990, 10))2      2.190  2.028  2.364
Ns(P, knots = seq(1950, 1990, 10))3      3.222  2.835  3.661
Ns(P, knots = seq(1950, 1990, 10))4      2.299  2.149  2.459
```

A spline period effect

```
> Cs <- Ns(                pp   ,knots=seq(1950,1990,10))
> Cr <- Ns(rep(1970,length(pp)),knots=seq(1950,1990,10))
> head( Cs, 4 )

       1         2          3         4
  [1,] 0 0.1267731 -0.3803194 0.2535463
  [2,] 0 0.1217022 -0.3651066 0.2434044
  [3,] 0 0.1166313 -0.3498939 0.2332626
  [4,] 0 0.1115604 -0.3346811 0.2231207

> head( Cr, 4 )

               1         2         3          4
  [1,] 0.6666667 0.1125042 0.1624874 -0.1083249
  [2,] 0.6666667 0.1125042 0.1624874 -0.1083249
  [3,] 0.6666667 0.1125042 0.1624874 -0.1083249
  [4,] 0.6666667 0.1125042 0.1624874 -0.1083249

> ci.exp( msps, subset="P" )

                                      exp(Est.)     2.5%    97.5
  Ns(P, knots = seq(1950, 1990, 10))1  1.710808 1.525946 1.91806
```


Period effect

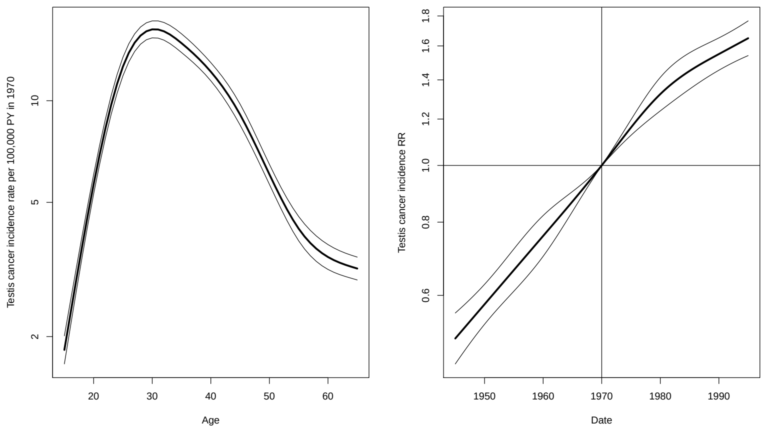



```
> matplot( pp, ci.exp( msps, subset="P", ctr.mat=Cs-Cr ),
+          log="y", ylim=c(0.5,2), xlab="Date",
+          ylab="Testis cancer incidence RR",
+          type="l", lty=1, lwd=c(3,1,1), col="black" )
> abline( h=1, v=1970 )
```


Period effect

```
> par( mfrow=c(1,2) )
> Cap <- cbind( 1, Ns(                aa   ,knots=seq(15,65,10)),
+              Ns(rep(1970,length(aa)),knots=seq(1950,1990,1
> matplot( aa, ci.exp( msps, ctr.mat=Cap )*10^5,
+          log="y", xlab="Age",
+          ylab="Testis cancer incidence rate per 100,000 PY in
+          type="l", lty=1, lwd=c(3,1,1), col="black" )
> matplot( pp, ci.exp( msps, subset="P", ctr.mat=Cs-Cr ),
+          log="y", xlab="Date", ylab="Testis cancer incidence R
+          type="l", lty=1, lwd=c(3,1,1), col="black" )
> abline( h=1, v=1970 )
```

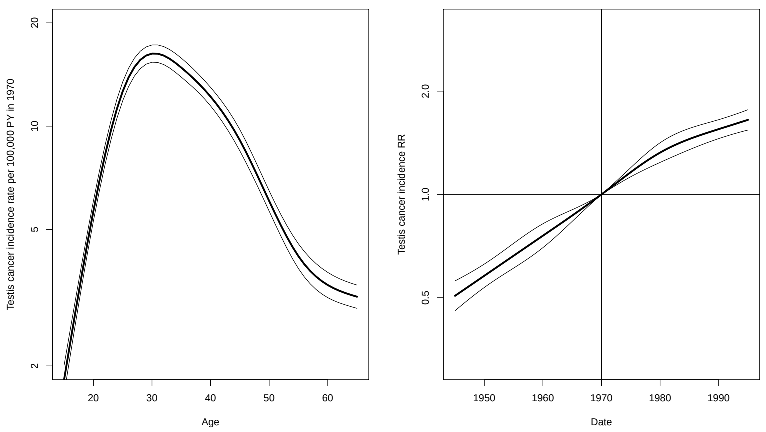
Age and period effect



Period effect

```
> par( mfrow=c(1,2) )
> Cap <- cbind( 1, Ns(                    aa   ,knots=seq(15,65,10)),
+                  Ns(rep(1970,length(aa)),knots=seq(1950,1990,1
> matplot( aa, ci.exp( msps, ctr.mat=Cap )*10^5,
+           log="y", xlab="Age",
+           ylim=c(2,20), xlim=c(15,65),
+           ylab="Testis cancer incidence rate per 100,000 PY in
+           type="l", lty=1, lwd=c(3,1,1), col="black" )
> matplot( pp, ci.exp( msps, subset="P", ctr.mat=Cs-Cr ),
+           log="y", xlab="Date",
+           ylim=c(2,20)/sqrt(2*20), xlim=c(15,65)+1930,
+           ylab="Testis cancer incidence RR",
+           type="l", lty=1, lwd=c(3,1,1), col="black" )
> abline( h=1, v=1970 )
```


Age and period effect



Age and period effect with `ci.exp`

- In rate models there is always one term with the **rate** dimension.
  Usually **age**
- But it must refer to a specific **reference** value for all **other** variables (`P`).
- **All** parameters must be used in computing rates, at reference value.
- For the "other" variables, report the RR **relative** to the reference point.
- Only parameters relevant for the variable (`P`) used.
- Contrast matrix is a **difference** between prediction points and the reference point.