A hazards approach to the biometric analysis of infant mortality

Göran Broström and Tommy Bengtsson

2023-03-06 23:22:33

Abstract

A variation of the Bourgeois-Pichat biometric analysis of infant mortality is suggested. In the original model, cumulative mortality in the last eleven months of infancy is assumed to follow a uniform distribution given a log-cube transformation of age. Instead, we assume an exponential distribution. The difference is that while the denominator is constant in the Bourgeois-Pichat model, equal to the number of births, in our model, the denominator is the current population at risk. We argue that our assumption is more satisfactory from a theoretical point of view, since it focus on the conditional probability of dying. Our model gives different estimates of endogenous and exogenous mortality and, in addition, the model fit is slightly better, especially in cases with higher levels of infant mortality.

Contents

1 Introduction

In this version the cumulative hazard functions are depicted as-is, but on the log-cube time scale. See the new **R** function *drawBB2*.

[Figure 1 about here.]

Lois Bourgeois-Pichat saw the first year of life not only as the period when mortality is highest, but also the period when improvements should be possible (Bourgeois-Pichat, 1951a,b; Bourgeouis-Pichat, 1952). He argued that infant deaths should be divided into two categories, endogenous and exogenous deaths. Endogenous deaths are, by definition, due to inherited factors, or acquired during gestation or delivery. They typically occur at the initial period of life, though sometimes much later and include congenital debility, prematurity, malformations, and disease in early life. Exogenous infant mortality, which Bourgeois-Pichat regards as accidental, are deaths for which the society must hold itself responsible and to reduce endogenous mortality, medical intervention is essential (Bourgeouis-Pichat, 1952). Since it is, based on causes of deaths, difficult to distinguish between endogenous and exogenous factors, either because they are inaccurate or difficult to make due to the diseases themself, or the information do not exist, Bourgeois-Pichat offered a solution to the problem.

Bourgeois-Pichat's biometric model to differentiate between endogenous and exogenous mortality is based on an idea of a universal law governing the distribution of deaths in the first year of life (Bourgeouis-Pichat, 1952). He assumes that all deaths taking place in the last eleven months of the first year are exogenous. He argues that, although there are also some endogenous mortalities during this period, they do not affect the development of mortality much. Supported by examples from mid-twentieth century Western and Southern Europe, the US and Canada, New Zeeland, and other countries, Bourgeois-Pichat argued that the cumulative infant deaths after the first month follows a linear development given a log-cube transformation of age (Bourgeouis-Pichat, 1952). Mortality in the first month of life is then divided into an endogenous and an exogenous assuming that the same linearity also exists in this period. It leads to the conclusion that exogenous deaths in the first month of life represent 25 percent of the deaths from the second to twelfth months (Bourgeouis-Pichat, 1952).

Applying this method, Bourgeois-Pichat also finds deviations from the linear death pattern after the first month. To take one example, the curve for Sardinia in 1948 bend strongly upward after four months (Bourgeouis-Pichat, 1952, Figure 12). To take another example, in the case of Quebec 1944–1947, the curve bends down quite strongly after the sixth month (Bourgeouis-Pichat, 1952, Figure 8). A downward bend is also what has been found using historical Swedish parish data (Bengtsson, 1999, Figure 6b; Bengtsson and Lundh, 1999, Figures 6 and 7; Lynch et al., 1998, Figures 1 and 2; Sundin and Tedebrand, 1981, Figure 6d), though less pronounced than in Quebec. The curve for Sweden 1910–1946 show, however, no such downward bend (Bourgeois-Pichat, 1951b, Figure 9; Bourgeouis-Pichat, 1952).

Bourgeois-Pichat's biometric method has also been used to identify problems with data recording (Wrigley, 1977; Wrigley and Schofield, 1981). While a high level of endogenous mortality cannot be used as a criterion for high data quality, the opposite holds true. A very low level of endogenous mortality indeed indicates data problems, just like large deviations from the normal sex ratio at birth do. These two criteria are, in fact, often used in historical studies to evaluate data recording (see Bengtsson et al 2004). They become even more precise if they are applied to different social strata, since under-recording of early deaths often have a social gradient, possibly due to the costs involved in a burial (Bengtsson, 1999; Bengtsson and Lundh, 1999).

We suggest a variation of the Bourgeois-Pichat biometric analysis of infant mortality. Instead of assuming that the cumulative mortality in the last eleven months of infancy follows a uniform distribution, given a log-cube transformation of age, we assume an exponential distribution. The difference is that while the denominator is constant in the Bourgeois-Pichat model, equal to the number of births, in our model, the denominator is the current population at risk. This assumption is more satisfactory from a theoretical point of view, since it models the conditional probability of dying. The growth in birth weight also follows this distribution (Bourgeois-Pichat, 1951a,b). In addition, the model fit is slightly better for our model, especially in cases with high level of infant mortality, common in pre-modern societies. The advantage is that not only is our assumption more attractive from a theoretical point of view, and therefore easier to interpret, but also that it is makes it easier to estimate exogenous and endogenous infant mortality with standard survival analysis programs.1 In addition, we give examples from eighteenth Sweden, where the curve do not follow the uniform distribution during the last eleven month of infancy, despite high levels of mortality in the first month of life.

2 The Bourgeois-Pichat procedure

Central in the procedure suggested by Bourgeois-Pichat (1951a) is *the log-cube transform*, see Equation (1).

$$g(t) = C \log^3(t+1),\ 0 \leq t \leq 365, \tag{1}$$

where t is age measured in days and C is a normalizing constant,

$$C = \frac{365}{\log^3(366)}.$$

The constant C is chosen so that $g(365) = 365$, see Figure 1. Note that C is *not* part of the original definition of the log-cube transform, but provided here only to make graphical comparisons easier to interpret. It makes no difference otherwise.

Assume that a cohort of infants is followed over time from birth to age one. There are no drop-outs (no right censoring except at age 365 days). The exact age in days at each observed death is noted and transformed by g, and due to imperfect time measuring there may be tied death ages. As an illustrative example, we use a data set from northern Sweden, covering the years 1861–1950.

For the first and last 30 years in this data set, we have the results shown in Figure 2.

Table 1: Life table, 1861-1890.

Day	g(Day)	Population	Deaths
0.25	0.0	50034	564
1.00	0.6	49470	138
2.00	2.4	49332	81
3.00	4.7	49251	83
4.00	7.4	49168	64
5.00	10.2	49102	73
6.00	13.1	49028	57

[Figure 2 about here.]

The cumulative numbers of death are plotted against the corresponding death ages on the g time scale, and as is seen, the fit to linearity after 28 days (68 on the g scale) is excellent for both time periods.

3 The hazard-based procedure

Instead of using the total number of births throughout in the denominator as in the Bourgeois-Pichat procedure, we suggest using the present risk-set size, that is, the the total number of infants still alive immediately prior to the death age under investigation. It is easily achieved by the use of the function `risksets` in the package `eha`. See Table 1 for an illustration.

It starts with 50034 live births, of which 564 dies on the day of birth, on average at the age of six hours (a quarter of a day), and so on.

The plot corresponding to Figure 2 is shown in Figure 3.

```
## rate =  0.0002201588 g(28) =  67.76404 y0 =  0.01491885
```

[Figure 3 about here.]

```
## rate =  0.0001000851 g(28) =  67.76404 y0 =  0.006782171
```

See Figure 4 for a comparison of the two curves.

[Figure 4 about here.]

The difference between the two curves increases as infant mortality increases, but both methods show an excellent fit to linearity.

4 Theoretical considerations

We note that the Bourgeois-Pichat method requires that no right censorings (infants lost to follow-up before one year of age) are present, and no left truncations (infants under observation only from an age later than birth). The hazards approach, on the other hand, allows left truncation and right censoring as long as they are *non-informative* in the usual sense. Often, though, this not very important, since new-born babies tend to be geographically stationary for their first year of life.

Throughout the rest of the paper, everything happens on the g time scale.

4.1 Post-neonatal mortality

We concentrate on the *postneonatal* period, since that is the period where Bourgeois-Pichat claims that the *cumulative distribution function* (CDF) is *Uniform*, and we suggest that the CDF is *exponential*, that is, the *cumulative hazard function* is "uniform" (linear).

In order to see this, the data set from above is *left truncated* at age 28 days (at 68 on the "g" scale), that is, we are considering the *conditional* survival distribution, given survival to age 30.

[Figure 5 about here.]

The *Exponential* fit is excellent, with a slight edge for the early period data, where post-neonatal mortality is high. However, one question pops up: If the exponential assumption is in fact true, will it also hold for subgroups of the data?

The theoretical ansver is "No", because a mixture of two exponential populations with different mortality rates cannot itself be exponential. A heuristic argument for that: Assume two equally-sized groups at "birth", but with different levels of mortality. As time passes, there will be more deaths in the high-mortality group, and that means that the proportion survivors in the low-level group will increase. Therefore, the population-level mortality will seem to decrease over time, and not being constant, as the exponentiality prescribes.

We can see what happens when the two periods in Figure 5 are joined, see Figure 6.

[Figure 6 about here.]

Not what we expected: The fit seems to be extremely good in both figures! However, the rate is close to zero, and so differences are small and hard to notice. And for practical use, they are not interesting.

What about the original Bourgeois-Pichat model, is it sensitive to taking subsets? Let us see.

[Figure 7 about here.]

In Figure 7 we notice a weak tendency towards convexity for boys. Not a big deal.

4.2 Exogeneous mortality

We now consider the whole infant age span (on the g scale), and compare the uniform and exponential densities, especially of interest is the fraction of deaths that occur in the neonatal phase. Manfredini (2004) argues that a less satisfactory property of the *B-P* model is that this fraction is constant, 19.7 percent, independent of the overall level of infant mortality. This is not the case for the hazards model, see Figure 8. For instance, for $p = 0.2$ the fraction in question is 21.5 percent.

[Figure 8 about here.]

4.3 Endogeneous mortality

In order to calculate the *endogeneous* infant mortality, a simple two-step procedure leads to the goal.

1. Estimate the post-neonatal mortality following the exponential route. It reduces to a simple occurrence/exposure calculation: The total number of post-neonatal deaths D is divided by the total postneonatal exposure E on the g scale. So

   $$\hat{\lambda} = \frac{D}{E}$$

   is the estimated hazard function (constant), and the cumulative hazard function is

   $$\hat{H}(t) = \hat{\lambda}t,\ t > 0$$

2. On the full infant interval, estimate the total cumulative hazard rate $A(t)$ with the usual Nelson-Aalen estimator $\hat{A}(t)$ (Nelson, 1972; Aalen, 1978). Then, subtract $H(t)$ from $\hat{A}(t)$ to get $\hat{E}(t)$.

   $$\hat{E}(t) = \hat{A}(t) - \hat{\lambda}t, 0 < t < g(28). \quad (2)$$

   Note that subtraction and addition of competing risks are okay on the hazards scale, but not with probabilities, which is yet another argument in favor of the hazards approach.

Let us do it with the given data, and the first period 1861–1890.

4.3.1 Step 1: Estimate post-neonatal (exogenous) mortality

Post-neonatal mortality is the same as exogenous mortality in the sense that the hazard functions are the same on the post-neonatal age interval. So the first step involves data left truncated at age 28 days (or at 68) on the g scale. We exemplify the numerical procedure with the aid of the **eha** (Broström, 2022, 2021). The first lines of the data set are (Table 2)

and the number of rows is 221773, and the number of deaths is 17675, resulting in a crude death probability of 80 per thousand live births.

So, our first step is to truncate and rescale. We show how it is done in **R**:

Table 2: First five rows of data frame.

birthdate	sex	enter	exit	event	period	icd.chapter
1861-01-01	girl	0	163	1	1861-1890	NA
1861-01-01	boy	0	365	0	1861-1890	J
1861-01-01	boy	0	365	0	1861-1890	R
1861-01-01	boy	0	365	0	1861-1890	NA
1861-01-01	boy	0	57	1	1861-1890	NA

```
postneo <- eha::age.window(infant, c(28, 365))
gpostneo <- eha::age.window(ginfant, c(g(28), g(365)))
print(postneo[1:5, ], row.names = FALSE) # print first five rows
```

```
##   birthdate  sex enter exit event    period icd.chapter
##  1861-01-01 girl    28  163     1 1861-1890        <NA>
##  1861-01-01  boy    28  365     0 1861-1890           J
##  1861-01-01  boy    28  365     0 1861-1890           R
##  1861-01-01  boy    28  365     0 1861-1890        <NA>
##  1861-01-01  boy    28   57     1 1861-1890        <NA>
```

```
print(gpostneo[1:5, ], row.names = FALSE) # first five rows after...
```

```
##   birthdate  sex    enter     exit event    period icd.chapter
##  1861-01-01 girl 67.76404 235.4138     1 1861-1890        <NA>
##  1861-01-01  boy 67.76404 365.0000     0 1861-1890           J
##  1861-01-01  boy 67.76404 365.0000     0 1861-1890           R
##  1861-01-01  boy 67.76404 365.0000     0 1861-1890        <NA>
##  1861-01-01  boy 67.76404 118.8162     1 1861-1890        <NA>
```

Now, the post-neonatal hazard function is *constant*, and its ML estimator is simply the *occurrence/exposure* rate, or in **R**,

```
D <- sum(gpostneo$event) # No. of deaths
E <- with(gpostneo, sum(exit - enter)) # exposure
(rate <- D/E)
```

```
## [1] 0.0001648328
```

The rate is a very small number as a consequence of the very small time unit (implying large total exposure time), originally *day*.

So we are done with the simple post-neonatal period.

4.3.2 Step 2: Estimate total and endogenous mortality

Here we focus on neonatal period.

The estimation of the cumulative hazard function for the total neonatal mortality is easily achieved by calling the function `hazards` in the *eha* package, see also Figure 9.

```
neo <- age.window(ginfant, c(0, g(28)))
par(las = 1)
fit <- coxreg(Surv(enter, exit, event) ~ 1, data = neo)
hneo <- hazards(fit, cum = TRUE)
x <- c(0, hneo[[1]][, 1])
y <- c(0, 1000 * cumsum(hneo[[1]][, 2]))
plot(x, y, type = "l", xlab = "Day", ylab = "by 1000")
lines(x, 1000 * rate * x, col = "blue", lty = 2)
abline(h = 0, v = 0)
text(8, 15, "Total", col = "black")
text(8, 5, "Exogenous", col = "blue")
```

[Figure 9 about here.]

Next, take the difference between "Total" and "Exogenous" to get "Endogenous", Figure 10.

[Figure 10 about here.]

5 Biometric analysis in practice

The original procedure of Bourgeos-Pichat is compared to the hazard based procedure for some typical cases from the real world.

5.1 Västerbotten 1921–1950

The crude IMR in Västerbotten 1921–1950 was around 56 per thousand, a rather low figure in context. Let us perform the biometric analysis with these data, see Figure 11.

```
## rate =  0.0001000541 g(28) =  67.76404 y0 =  0.006780067
```

[Figure 11 about here.]

There is a good hazards model fit, and we see that the endogenous mortality clearly dominates the early days of neonatal mortality, and almost vanishes towards the start of the post-neonatal period.

The B-P model fit is slightly worse, but not much to bother about.

5.2 Västerbotten 1890–1920

The crude IMR in Umeå 1891–1920 was around 95 per thousand, clearly higher than the later time period. Let us perform the biometric analysis with these data, see Figure 12.

```
## rate =  0.0001993882 g(28) =  67.76404 y0 =  0.01351135
```

[Figure 12 about here.]

The dominance of endogenous mortality in early life is still clear. The hazards fit is still slightly better.

5.3 Västerbotten 1861–1890

The crude IMR in Umeå 1861–1890 was around 105 per thousand, highest of the three time periods. The biometric analysis is shown in Figure 13.

```
## rate =  0.0002201398 g(28) =  67.76404 y0 =  0.01491756
```

[Figure 13 about here.]

The conclusion here is almost the same as for the later time periods, good model fit and endogenous dominance in the very early days of life. However, the B-P method seems to have a slight upper hand.

5.4 Scania 1710–1800.

```
## rate =  0.0005921974 g(28) =  67.76404 y0 =  0.04012968
```

[Figure 14 about here.]

Very bad fits in both cases.

5.5 Skellefteå 1820–1835

An example of severe under-registration during the neonatal period, see Figure 15. There is no room for endogeneous deaths at all, but otherwise a reasonably good fit with both methods.

```
## rate =  0.000327796 g(28) =  67.76404 y0 =  0.02221278
```

[Figure 15 about here.]

6 Conclusion

If anything, the hazards method never performs worse than the B-P method. But the real strength of the hazards method is that it fits naturally into general modern survival analysis with censored and truncated data, and also proportional hazards models with covariates.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6:701–726.

Bengtsson, T. (1999). The vulnerable child. Economic insecurity and child mortality in pre-industrial Sweden: A case study of Västanfors, 1757–1850. *European Journal of Population*, 15:117–151.

Bengtsson, T. and Lundh, C. (1999). Child and infant mortality in the nordic countries prior to 1900. Technical Report 66, Department of Economic History, Lund University, Lund. Lund Papers in Economic History.

Bourgeois-Pichat, J. (1951a). La mesure de la mortalité infantile. I. Principes et métodes. *Population (French Edition)*, 6:233–248.

Bourgeois-Pichat, J. (1951b). La mesure de la mortalité infantile. II. Les causes de décès. *Population (French Edition)*, 6:459–480.

Bourgeouis-Pichat, J. (1952). An analysis of infant mortality. Population Bulletin 2.

Broström, G. (2021). *Event History Analysis with R, Second Edition*. Chapman & Hall/CRC, Boca Raton.

Broström, G. (2022). *eha: Event History Analysis*. R package version 2.10.1. https://CRAN.R-project.org/package=eha.

Lynch, K. A., Greenhouse, J. B., and Brändström, A. (1998). Biometric modeling in the study of infant mortality. *Historical Methods*, 31(2):53–64.

Manfredini, M. (2004). The bourgeois-pichat's method and the influence of climate: New evidence from late 19th-century Italy. *Social Biology*, 51:24–36.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14:945—-965.

Sundin, J. and Tedebrand, L.-G. (1981). Mortaility and morbidity in Swedish iron foundries, 1750-1875. In Brändström, A. and Sundin, J., editors, *Tradition and Transition: Studies in Microdemography and Social Change*. Demographic Data Base, Umeå University, Umeå.

Wrigley, E. (1977). Birth and baptisms: The use of anglican baptism registers as a source of information about the numbers of births in England before the beginning of civil registration. *Population Studies*, 31:281–312.

Wrigley, E. and Schofield, R. (1981). *The population history of England 1541–1871: A reconstruction*. Edward Arnold, London.

List of Figures

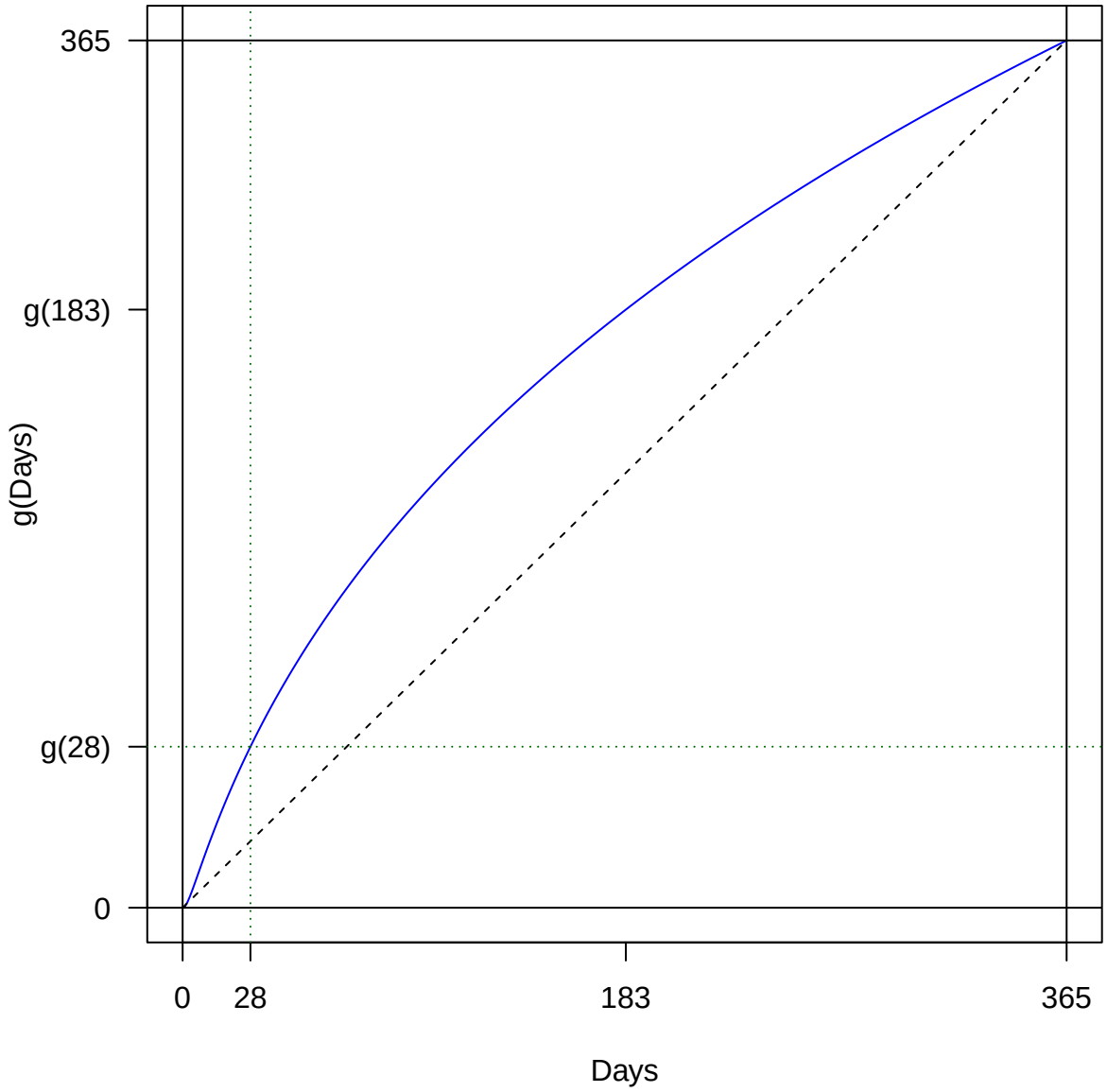

Figure 1: The log-cube transform of time in days versus the identity transform (dashed). Note that g(365) = 365.

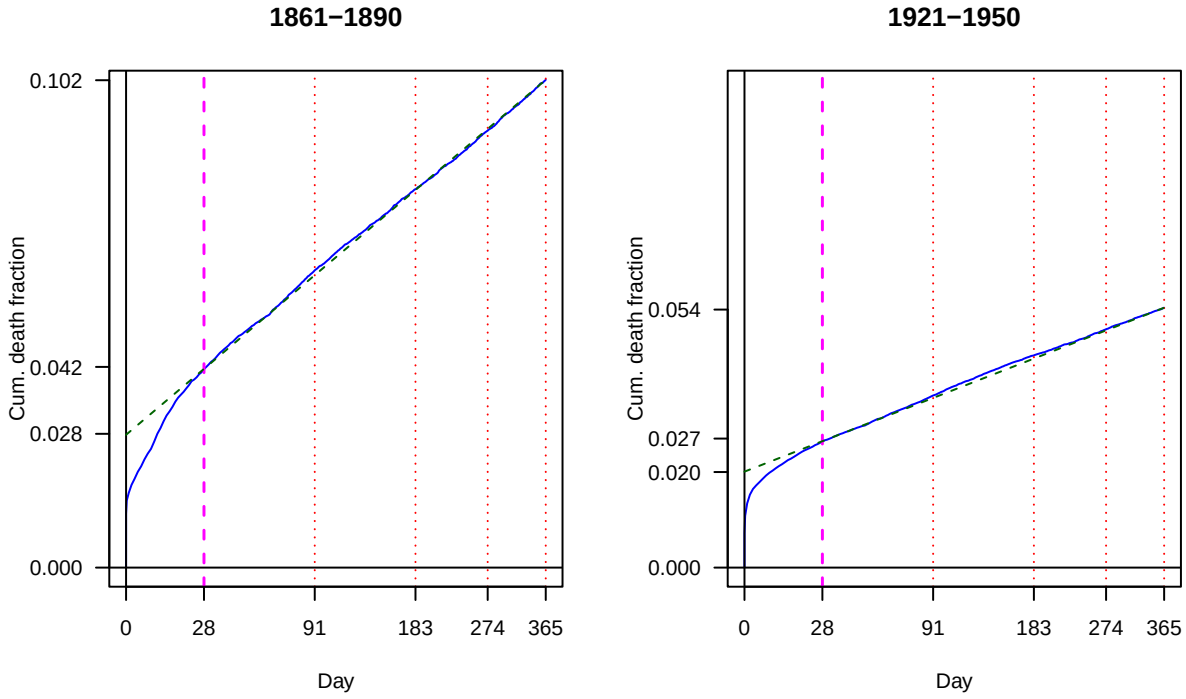

Figure 2: The periods 1861–1890 and 1921–1950 in Västerbotten, Bourgeois-Pichat method.

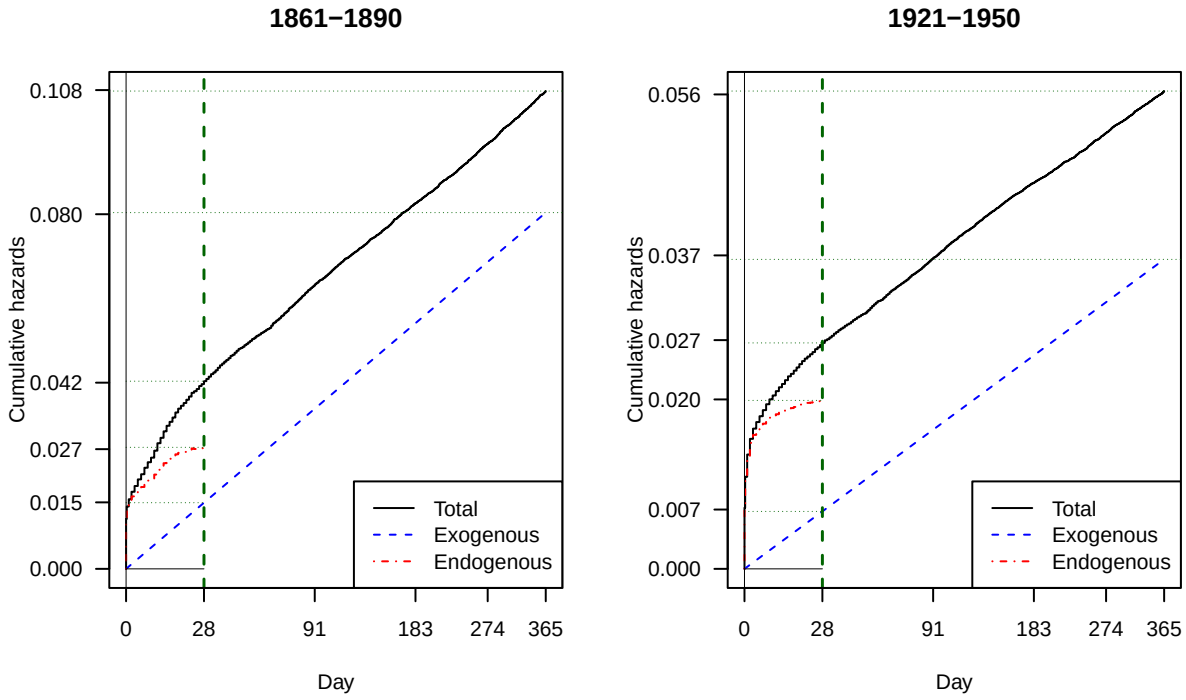

Figure 3: The periods 1861–1890 and 1921–1950 in Västerbotten, hazards method.

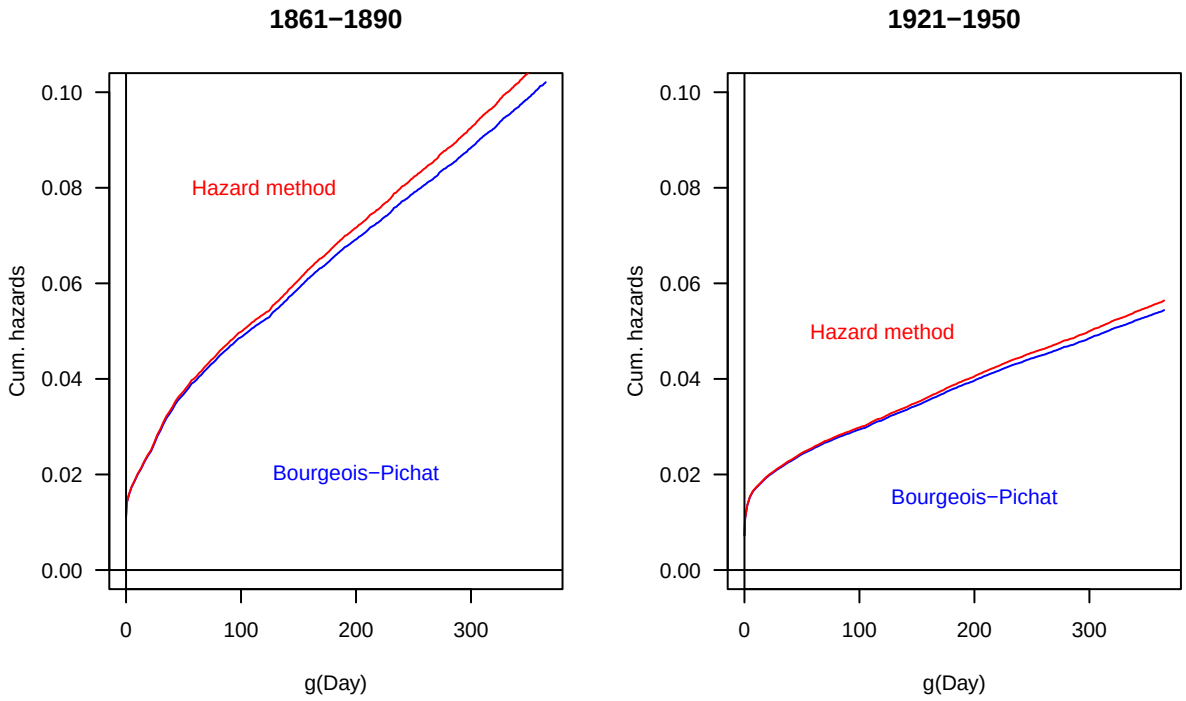

Figure 4: Bourgeois-Pichat and 'hazards method' plots.

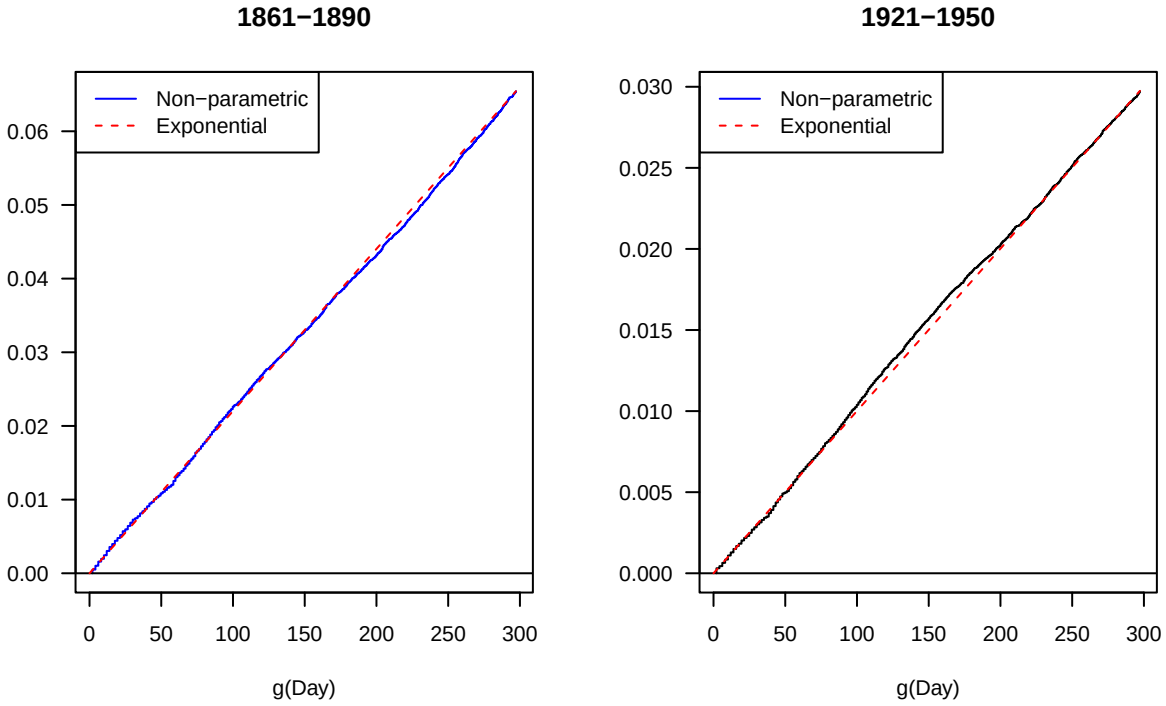

Figure 5: Exponential fits to postneonatal data, Västerbotten.

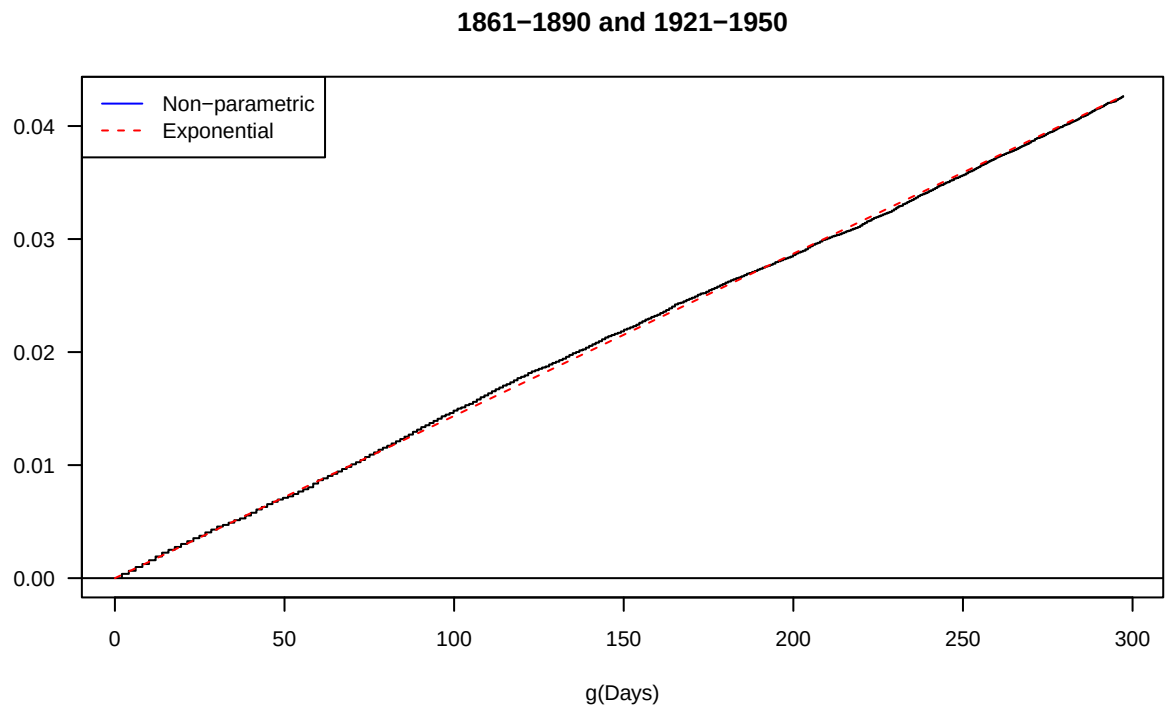

Figure 6: Exponential fit for two periods, Västerbotten.

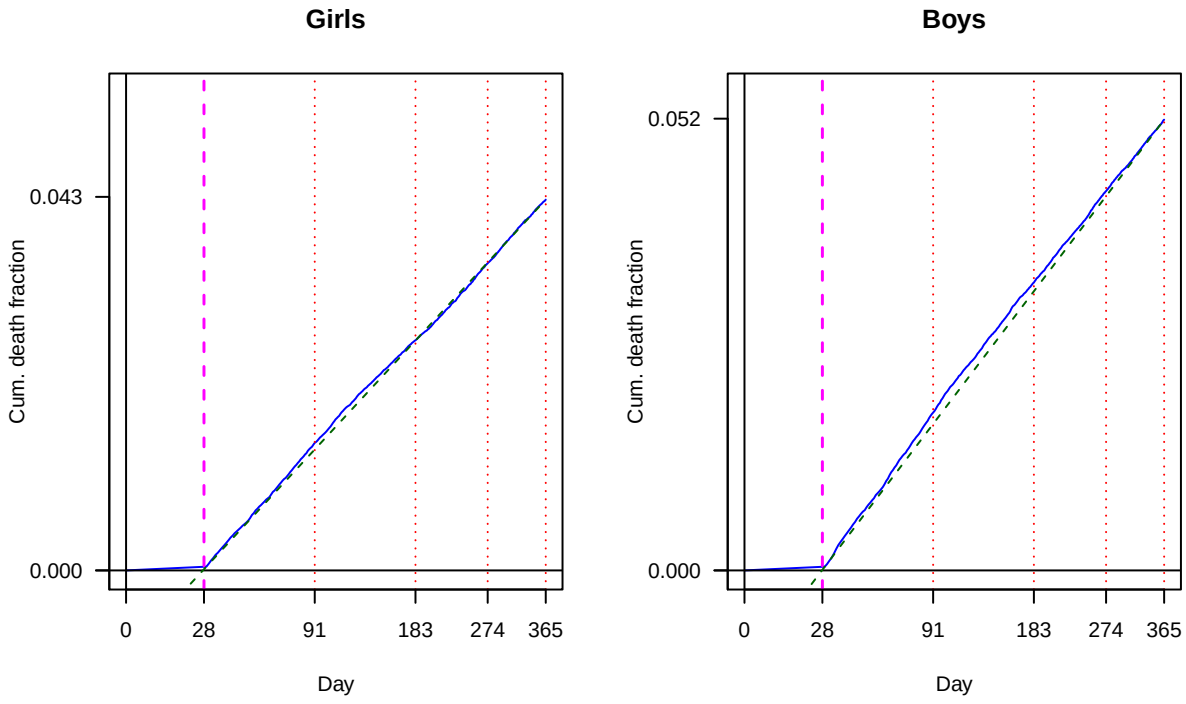

Figure 7: The period 1861-1890, Bourgeouis-Pichat method by sex.

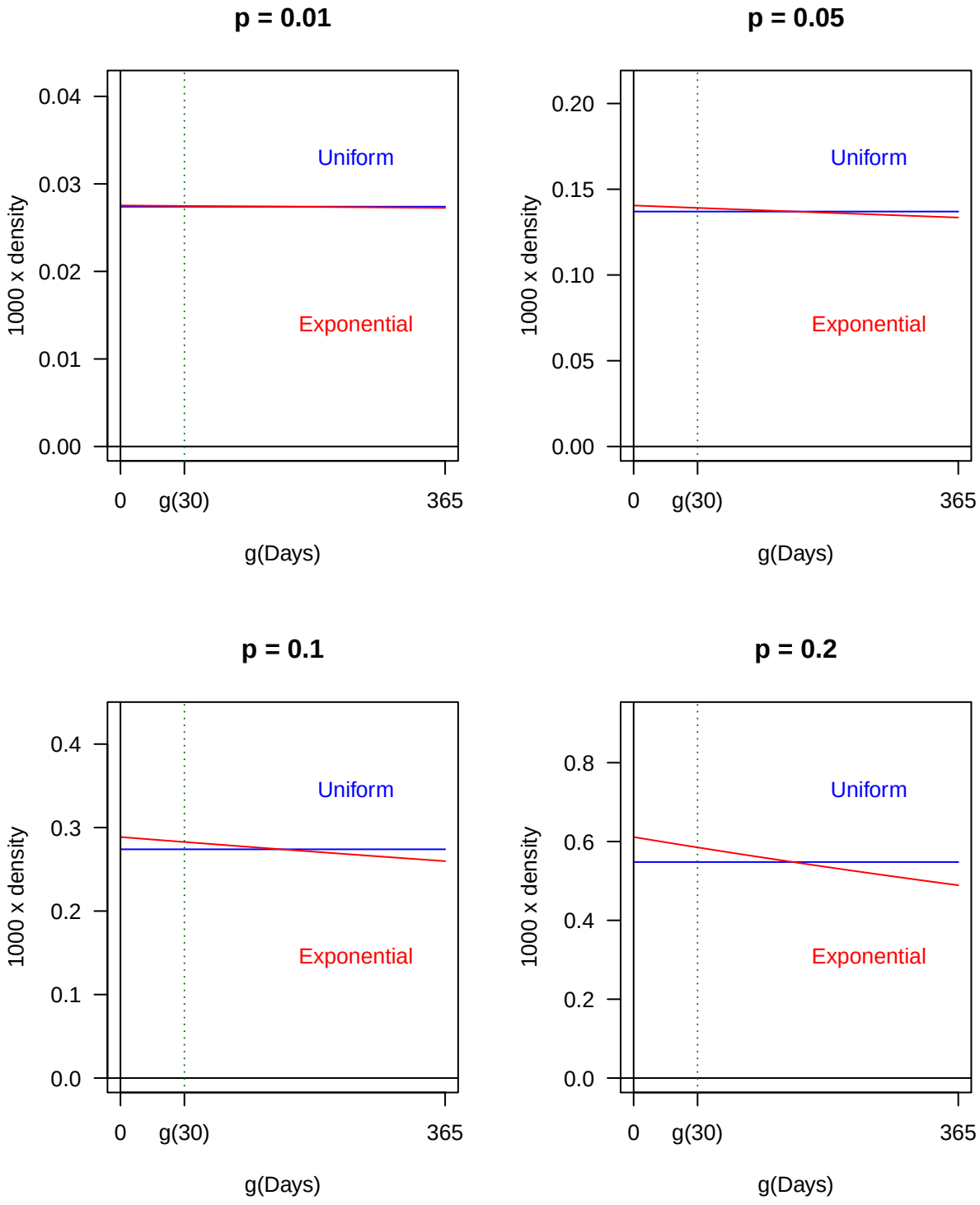

Figure 8: Densities for uniform and exponential distributions for varying death probability p.

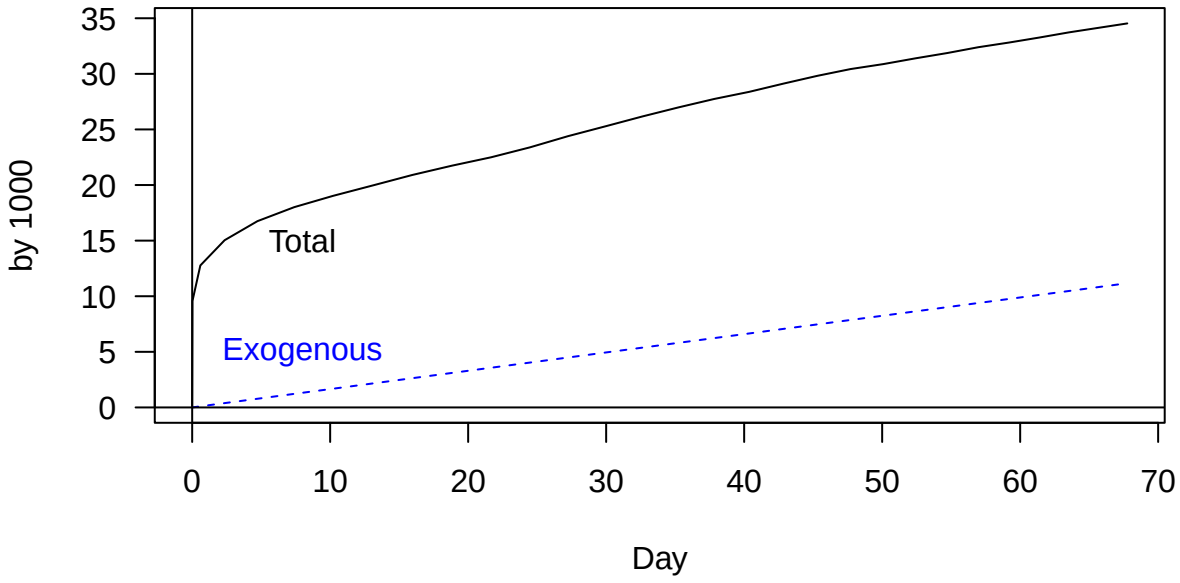

Figure 9: Cumulative hazard functions for neonatal mortality.

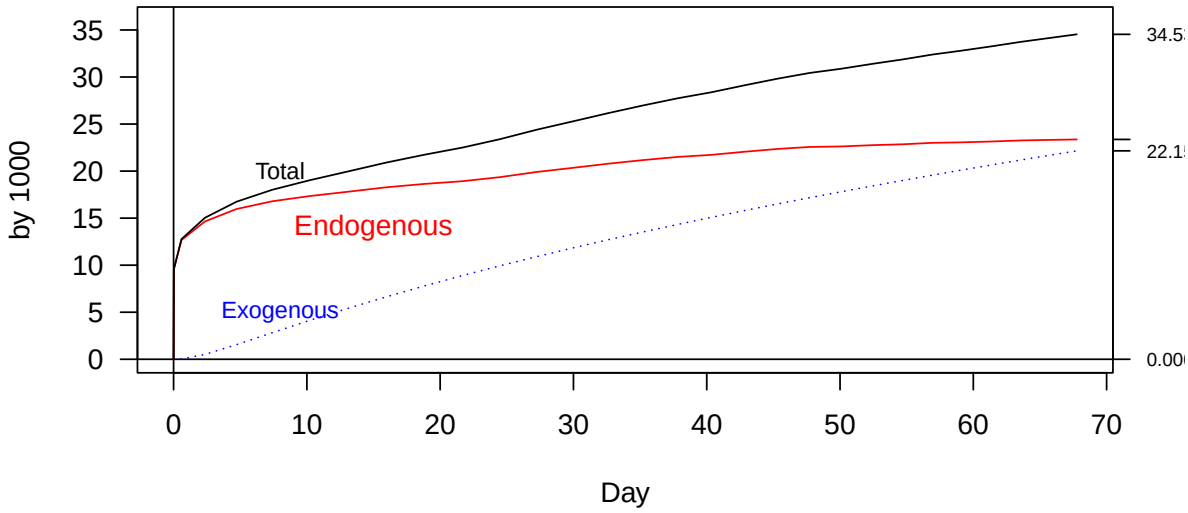

Figure 10: Cumulative hazard function for endogenous neonatal mortality.

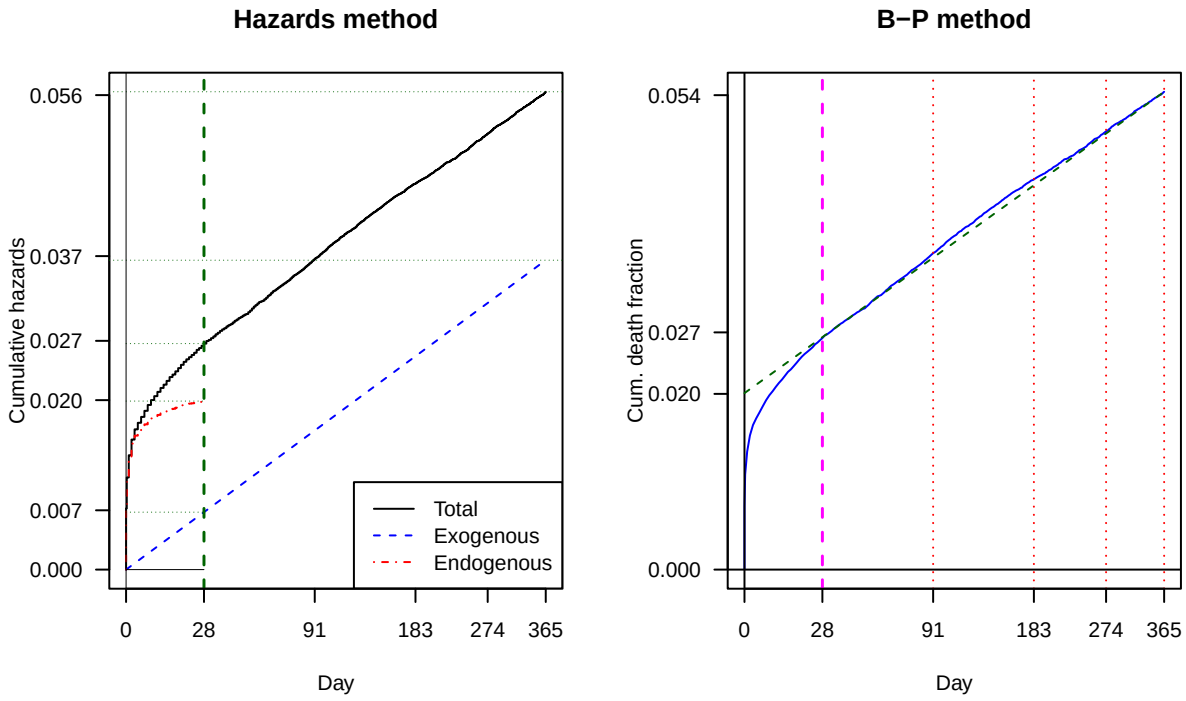

Figure 11: Västerbotten 1921–1950.

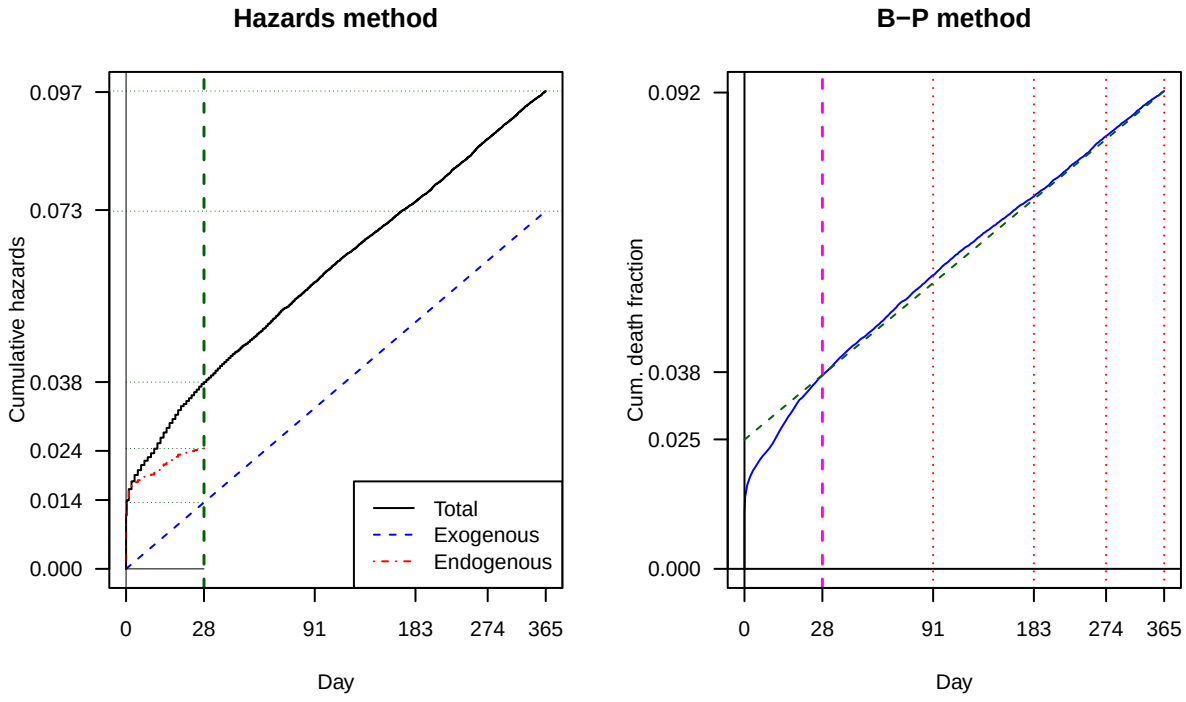

Figure 12: Västerbotten 1891–1920.

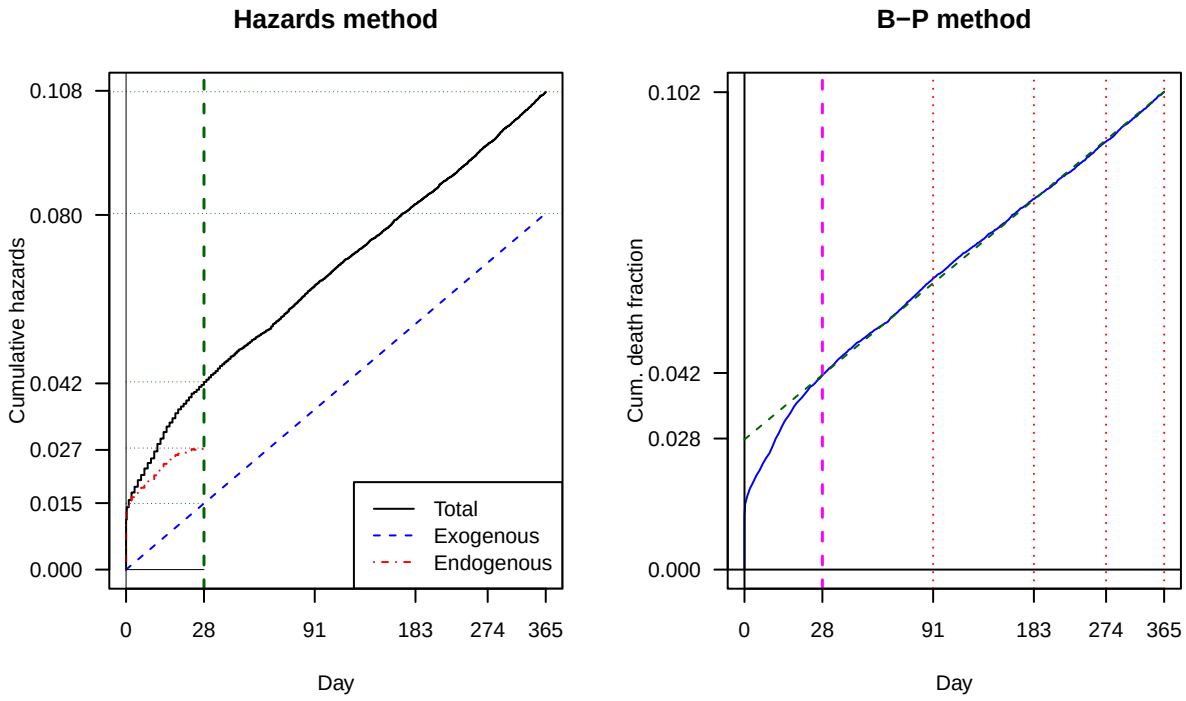

Figure 13: Västerbotten 1861–1890.

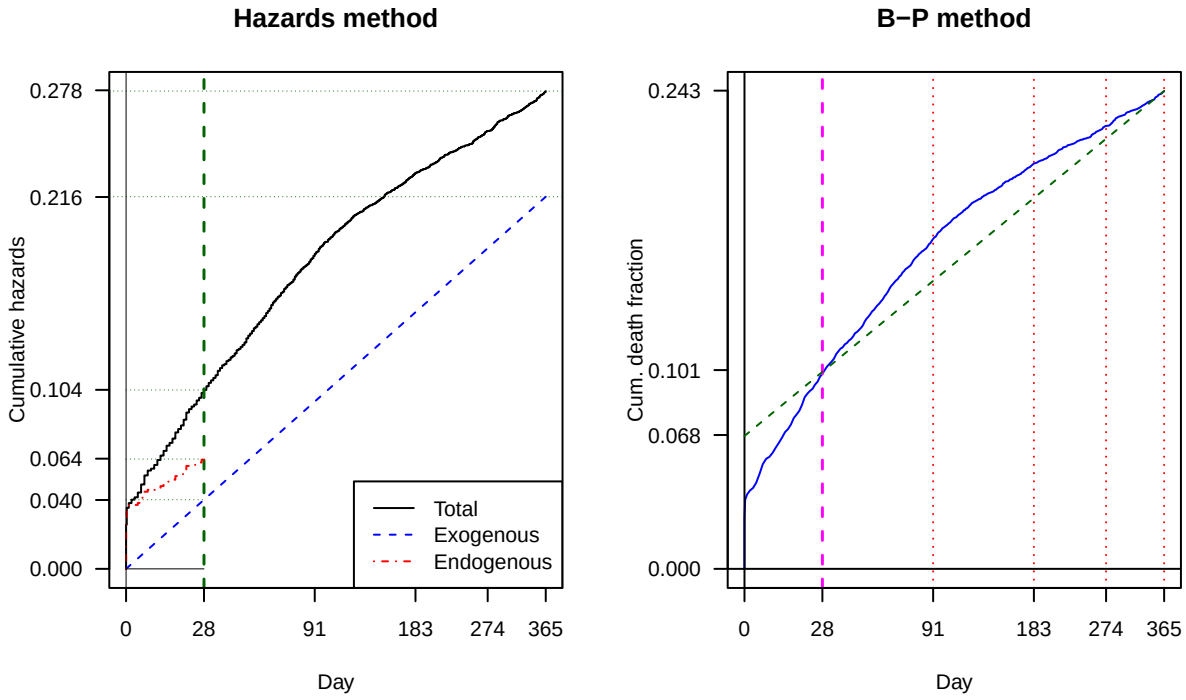

Figure 14: Scania 1710–1800.

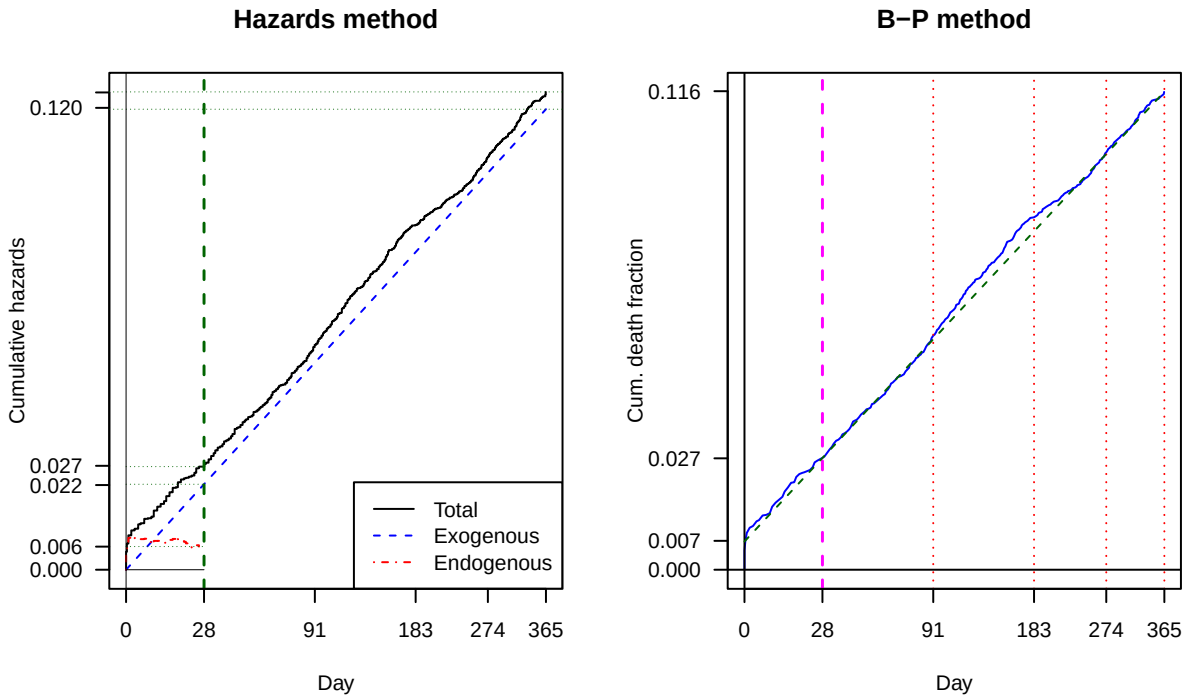

Figure 15: Skellefteå 1821–1838.