# Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model

PATRICK ROYSTON
*MRC Clinical Trials Unit, United Kingdom*

PAUL C. LAMBERT
*Department of Health Sciences, University of Leicester, United Kingdom and*
*Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden*

A Stata Press Publication
StataCorp LP
College Station, Texas



Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845
Typeset in LaTeX 2ε
Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

ISBN-10: 1-59718-079-3
ISBN-13: 978-1-59718-079-5

Library of Congress Control Number: 2011921921

# Contents

*(Pages omitted)*

# Preface

We would first like to quote from the preface of a well-known and respected Stata Press book on survival analysis in Stata (Cleves et al. 2010):

> This is a book about survival analysis for the professional data analyst, whether a health scientist, an economist, a political scientist, or any of a wide range of scientists who have found that survival analysis is applicable to their problems. This is a book for researchers who want to understand what they are doing and to understand the underpinnings and assumptions of the tools they use; in other words, this is a book for all researchers.

In a way, the aims of our book are similar to those of Cleves et al. (2010). We extend their book in particular directions: flexible, parametric, going beyond the standard models, particularly the Cox model. We include, for example, detailed treatments of time-dependent effects and relative survival. Our starting point is a basic understanding of survival analysis and how it is done in Stata. We would be surprised, for example, if a reader had not created and plotted Kaplan–Meier curves and fitted a Cox model in Stata. Our aim is that researchers can build on our examples to apply the methodology to their own investigations of survival data. To that end, we have provided the basic tools (ado-files) but also, in the examples, we present Stata code to do many of the analyses and produce many of the graphs. Indeed, presentation of the results of flexible parametric modeling is often best achieved by well-chosen graphs, and we regard that as an important message of our book.

Royston–Parmar models are a key tool in our approach; they are currently available only in Stata. (See section 1.10 for more information.) We would like to see their implementation in other software, such as R or SAS. However, we are very unlikely to implement this ourselves! If anyone has attempted such an implementation (or plans to do so) and would value our input, we would encourage them to contact us.

This book uses Stata version 12 throughout, but is fully compatible with Stata 11.1 or later, with only minor cosmetic differences across versions.

Finally, we would like to thank the folk who have contributed to our understanding of survival analysis and those who have undertaken the seemingly thankless task of commenting on our draft text. We are particularly grateful to

Therese Andersson, Karolinska Institute
Carol Coupland, University of Nottingham

Paul Dickman, Karolinska Institute
Sandra Eloranta, Karolinska Institute
Bobby Gutierrez, StataCorp
Hans van Houwelingen, University of Leiden
Bernard Rachet, London School of Hygiene and Tropical Medicine
Bill Rising, StataCorp
Mark Rutherford, University of Leicester
Willi Sauerbrei, University of Freiburg Medical Center
Michael Schemper, University of Vienna

*London and Leicester*
*April 2011*

Patrick Royston
Paul C. Lambert

# 1 Introduction

## 1.1 Goals

Most books on survival analysis devote a substantial section of their material to the Cox proportional hazards (PH) model (Cox 1972). The Cox model has played a vital role in applied survival analysis during the last three decades. The model and its software implementations have popularized survival analysis and made it accessible to researchers in varied disciplines who are not necessarily statisticians. It has been so successful that it is probably used in most practical analyses of the effects of covariates on survival.

Some years ago, Sir David Cox, in a revealing interview with Nancy Reid (Reid 1994), was asked what he thought of the cottage industry that had grown up around "his" model. He responded by saying that he would normally wish to attack a problem parametrically, because operations such as prediction were so much easier. Prediction (really estimation) of relevant features of survival data is a key theme in the present book.

Our main goals are to describe and to illustrate the use and applications of flexible parametric survival models, programmed in Stata, which in some important respects go beyond the Cox model and beyond the standard parametric survival models (such as the Weibull). These flexible models overcome the problems of potentially poor fit of standard parametric models and of the "noisy" estimates of the hazard and survival functions associated with the Cox model and with nonparametric estimators such as the Kaplan–Meier.

Flexible parametric survival models can help us in a number of ways. For example, they allow us to obtain an estimate of the baseline survival function and its uncertainty which vary smoothly over time. Prediction of survival probabilities and differences, hazard functions, hazard differences and ratios, time-dependent effects of covariates, and excess mortality rates in the context of relative survival are just some of the possible outputs from the models. Furthermore, the Stata commands are easy to use and to apply to real problems in a variety of settings.

Other than in chapter 1, we give extensive code showing how to implement the methods we describe in Stata. We present results graphically in many cases, but do not present code for all graphs because many are similar in style. More details of the structure and content of our book are outlined briefly in section 1.11.

## 1.2 A brief review of the Cox proportional hazards model

The Cox PH model is by far the most common model used in survival analysis. Many texts, some excellent, have been published on the model; we recommend, for example, Hosmer, Lemeshow, and May (2008) for a good, practical introduction and Grambsch and Therneau (2000) for extensions of the model. The quantities estimated from a Cox model are *hazard ratios* (HRs), which measure how much a covariate increases or decreases the rate of a particular event, assuming that it acts multiplicatively. For example, if the event were mortality and we applied a Cox model that estimated an HR of two for males compared with females, the mortality rate would be twice as high in males as in females.

A basic assumption of the Cox model is that the estimated parameters are not associated with time. In other words, we assume that any two hazard rates predicted by the model are proportional over time. In the above example, we assume that the doubling of the rate for males holds at 1 week, 1 month, 1 year, etc.

We can write the Cox model algebraically, as follows:

$$h_i(t|\mathbf{x}_i) = h_0(t) \exp(\mathbf{x}_i \beta)$$

The hazard function for the $i$th individual, $h_i(t|\mathbf{x}_i)$, is conditional on covariates $\mathbf{x}_i$, where $\beta = \beta_1, \ldots, \beta_k$ is the vector of regression coefficients. The baseline hazard function $h_0(t)$ is $h_i(t|\mathbf{x} = 0)$. One of the most recognized features of the Cox model is that we do not need to assume that the baseline hazard function has a specific shape. For this reason, the Cox model is often called *semiparametric*: we make parametric assumptions about the effects of covariates on the hazard function, but not about the shape of the hazard function itself. This is an important and appealing feature of the Cox model. If we were interested only in the HR, we could disregard distributional assumptions about the event times.

## 1.3 Beyond the Cox model

### 1.3.1 Estimating the baseline hazard

Consider arguably the simplest possible situation in survival modeling: a randomized, controlled clinical trial (say, in cancer) with right-censored time-to-event outcomes and a single covariate, `trt` (treatment), coded 0 for control or standard therapy and 1 for the experimental treatment arm. The data in the following example are taken from a Medical Research Council trial in 347 patients with advanced kidney cancer (Medical Research Council Renal Cancer Collaborators 1999). The primary outcome measure in the trial was all-cause mortality. The control and experimental treatments are the drugs medroxyprogesterone acetate (MPA) and interferon-$\alpha$ (IFN), respectively. To compare IFN with MPA, we fit a Cox model (`stcox trt`) with the following results:

```
. use kidney_ca
(kidney cancer data)
. stcox trt, nolog
         failure _d:  cens
   analysis time _t:  survtime/365.25
Cox regression -- Breslow method for ties
No. of subjects =          347                     Number of obs   =       347
No. of failures =          322
Time at risk    =  375.6769336
                                                   LR chi2(1)      =      6.81
Log likelihood  =   -1610.1366                     Prob > chi2     =    0.0091
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         trt |   .7464934   .0836699    -2.61   0.009     .5992665    .9298907
------------------------------------------------------------------------------
```

What do we get directly from the analysis? Principally, two things: an estimate (with confidence interval [CI]) of the HR comparing MPA with IFN, and a test of significance of the treatment effect. We can infer from the Stata output that IFN has reduced the mortality (hazard) rate by 25% with a 95% CI of (7%, 40%). The treatment effect is significant at the 1% level.

If (in Stata 11 or Stata 12) the `predict` command is used, or (in Stata 10 and earlier) we include certain options of the `stcox` command, we can obtain several additional outputs, including an estimate of the baseline survival function (in this case, $S(t)$ for the control arm), Schoenfeld residuals (which can be used to test the PH assumption), martingale residuals (which are useful for assessing the functional form for continuous predictors), and several other quantities. However, we find no option to get a useful estimate of the baseline hazard function when using `predict`.

Why is the hazard function useful? Because

- in medicine, it is a clinically meaningful measure of disease course, and
- it is the "ground" against which relative hazard effects are estimated.

The thicker pair of lines in figure 1.1 show an estimate of the hazard function in the control and experimental arms of the kidney cancer dataset. We estimated them under the PH assumption by fitting a Royston–Parmar (RP) model, a major theme of this book. RP models are implemented in the `stpm2` command (Lambert and Royston 2009). We outline the `stpm2` command in section 1.6 and describe the models in detail in chapter 5.

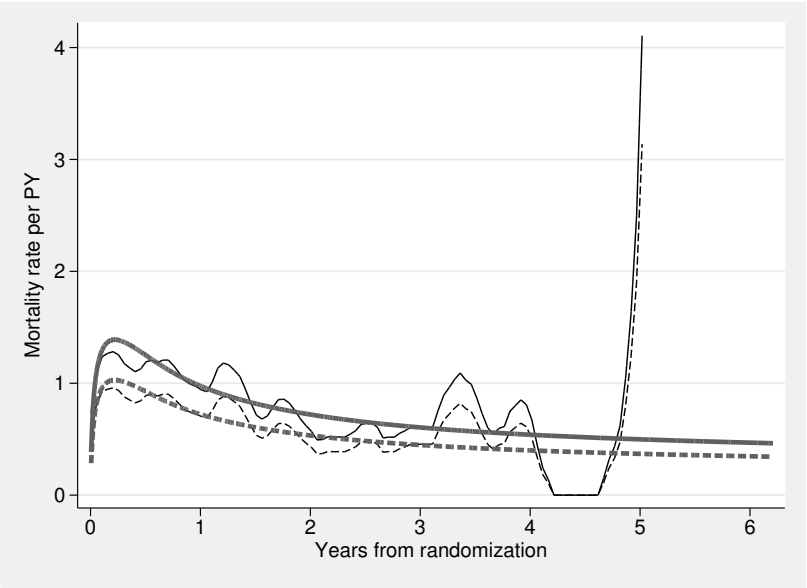


Figure 1.1. Kidney cancer data. Hazard functions in two treatment groups estimated under the PH assumption. Thick lines are from `stpm2` with two degrees of freedom for the baseline log cumulative-hazard function. Thin lines are from `stcurve` following `stcox`. Solid lines show the control group; dashed lines, the experimental group. PY stands for person-year.

The figure tells us the following:

- The death rate from advanced kidney cancer seems to be highest about 3 months after randomization, and it decreases after that time.
- The hazard is substantially reduced by the experimental treatment at all time points. (Under the PH assumption, the curves are forced to be proportional to each other.)
- Even after 4 years, the hazard is still substantial. The fact that it does not approach zero suggests that the disease is fatal, which is nearly always the case.

We have harvested quite a lot of useful information. Even if we relax the PH assumption, the plot of the ensuing hazard functions (not shown) is very similar to the thick lines in figure 1.1, so our conclusion about the treatment effect seems to be robust.

The thin lines in figure 1.1 are a nonparametric estimate of the hazard (mortality) rate. We estimated them with the `stcurve` command, which uses a technique known as kernel smoothing. The code that produced figure 1.1 is as follows:

```
. use kidney_ca
(kidney cancer data)
. stpm2 trt, df(2) scale(hazard)
Iteration 0:   log likelihood =  -564.1407
Iteration 1:   log likelihood = -564.09236
Iteration 2:   log likelihood = -564.09235
Log likelihood = -564.09235                     Number of obs   =        347
```

| | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| xb | | | | | | |
| trt | -.3005732 | .1118951 | -2.69 | 0.007 | -.5198836 | -.0812628 |
| _rcs1 | 1.224624 | .0629679 | 19.45 | 0.000 | 1.101209 | 1.348039 |
| _rcs2 | .1814725 | .0426715 | 4.25 | 0.000 | .0978379 | .265107 |
| _cons | -.4535717 | .0863545 | -5.25 | 0.000 | -.6228235 | -.2843199 |

```
. predict h0, at(trt 0) hazard
. predict h1, at(trt 1) hazard
. stcox trt, noshow nolog nohead
Cox regression -- Breslow method for ties
No. of subjects =          347                  Number of obs   =       347
No. of failures =          322
Time at risk    =  375.6769336
                                                LR chi2(1)      =      6.81
Log likelihood  =   -1610.1366                  Prob > chi2     =    0.0091
```

| _t | Haz. Ratio | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| trt | .7464934 | .0836699 | -2.61 | 0.009 | .5992665 | .9298907 |

```
. stcurve, hazard at1(trt=0) at2(trt=1) kernel(epan2)
> legend(off) lpattern(l -) title("") ylabel(0(1)4, angle(h))
> xscale(range(0 6.2)) xlabel(0(1)6) lwidth(medthin ..)
> addplot(line h0 h1 _t, sort lpattern(l -) lwidth(thick ..)
> lcolor(gs6 ..) xtitle("Years from randomization")
> ytitle("Mortality rate per person year"))
```

We obtained the curves after fitting the Cox model to the `trt` variable, assuming PH. Notice how wiggly and hard to interpret they are compared with those from `stpm2`. We think that the apparent sharp increase in mortality rate after four years is an artifact; the data there are sparse, the feature is not biologically plausible, and it is not seen in the curves from `stpm2`.

Finally, the HRs were 0.746 (standard error [SE] 0.084) and 0.740 (SE 0.083) according to the Cox and RP models, respectively—for practical purposes, they are identical.

### 1.3.2 The baseline hazard contains useful information

One of the consequences of a method that only estimates relative risk and not absolute risk is that users may ignore the importance of absolute risk. If we are told that the mortality rate is double for individuals with a particular exposure, then we want to

know what reference value this doubling refers to. In a survival model, the reference is usually the baseline hazard rate, which usually changes as a function of time. Thus even if the PH assumption is reasonable, the impact of a particular exposure in absolute terms depends on how long has passed since the time origin (diagnosis, randomization, start of treatment, etc.) and the magnitude of the underlying hazard rate.

An example to illustrate the importance of the baseline hazard is in survival from colon cancer. Figure 1.2(a) shows data from England and Wales where the time from diagnosis to death from colon cancer in those ages $< 50$ years has been modeled and smooth estimates of the hazard function derived for two time periods, 1981–1985 and 1986–1990. The event is death from any cause and thus the hazard rate can be considered as a mortality rate. The model assumes that the two hazard rates are proportional. The figure shows that the mortality rate is high in the first few months after diagnosis, but then decreases. By about 8 years, the mortality rate is very close to zero. We can infer that very few colon cancer patients who have survived to this time will actually die between 8 and 10 years. When the mortality rate associated with a diagnosis of a particular disease approaches zero, we have what is know as "statistical" or "population" cure (Lambert et al. 2007). The HR between the two time periods is 0.92, implying that the mortality rate is 8% lower in the more recent period. As the model assumes PH, the estimated relative effect is forced to be the same over the whole time period.

Figure 1.2(b) shows the difference in the hazard (mortality) rates. The absolute difference decreases with increasing follow-up time. Thus the 8% reduction in the mortality rate has little impact beyond about 6 years.

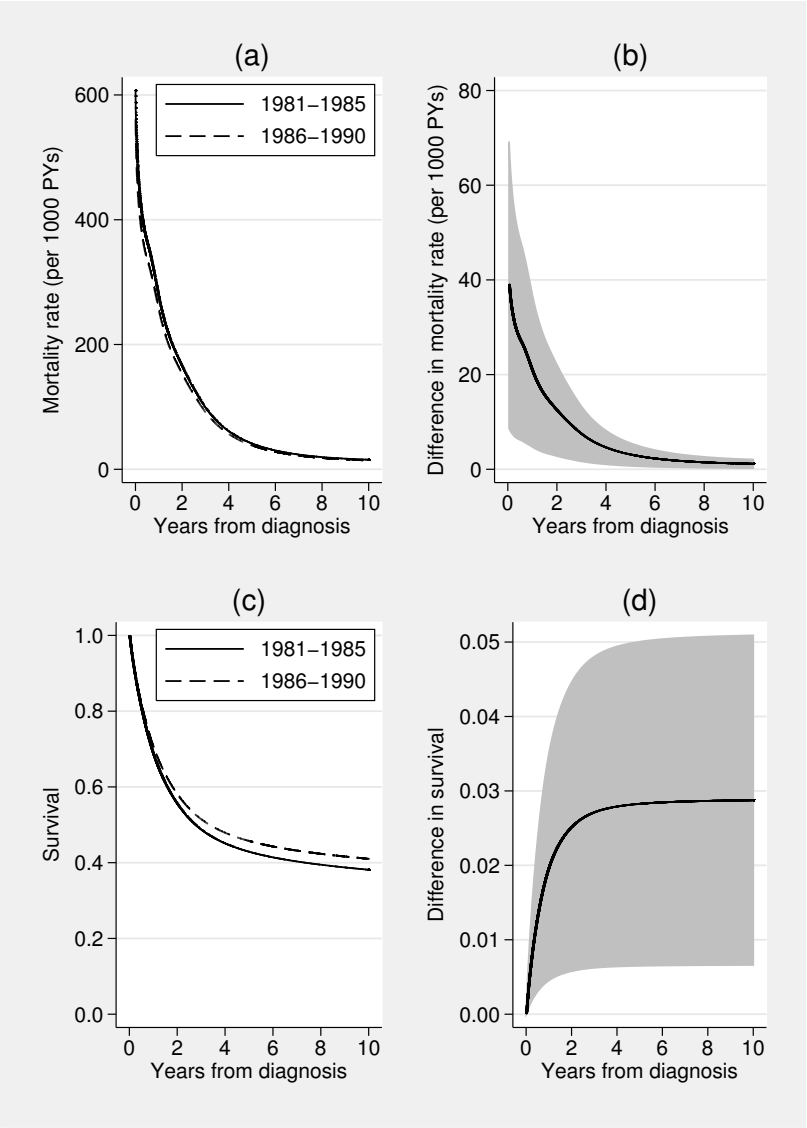


Figure 1.2. Cancer of the colon in England and Wales 1981–1985 and 1986–1990 for subjects aged $< 50$ years: (a) hazard rates, (b) difference in hazard rates, (c) survival functions, and (d) difference in survival functions. PYs stands for person-years.

Figure 1.2(c) shows the estimated survival functions and figure 1.2(d) shows the difference in the two survival curves. Figure 1.2(d) shows an improvement of just under 3% in absolute terms in survival in the more recent period. This should be expected given that the more recent period has a lower mortality rate. When we look at the

difference in the survival curves, we see that most of the improvement has been in the first 2–3 years.

We feel that the graphs shown in figure 1.2 give a better understanding of the disease and of the improvement in the more recent period than just quoting a hazard ratio of 0.92.

### 1.3.3 Advantages of smooth survival functions

A Kaplan–Meier plot of the survival function, $S(t)$, is an important feature of most survival analyses and is widely presented in publications of applied work. For the Cox model, Stata's `predict` command after `stcox` with the `basesurv()` option provides an estimate of the baseline survival function, $S_0(t) = S(t|\mathbf{x} = 0)$. From the baseline survival and the HR, we can predict the survival function for any combination of covariate values. However, all such survival functions are step functions and typically are not particularly smooth. However, it is reasonable to suppose that the underlying function is smooth. Also, the least precise parts of the curve get the most visual weight, a general criticism of Kaplan–Meier survival curves.

Kaplan–Meier-type estimates of $S(t)$ are composed of a sequence of point estimates of the survival function that are highly serially correlated. Accordingly, Kaplan–Meier plots tend to display "runs" of values that move away from and back toward the general trend, giving an undulating appearance. This may make the curve difficult to interpret and may lead to overemphasis of local features.

An example of these aspects, which is particularly a problem in smaller samples, appears in the kidney cancer data (see figure 1.3).

*(Pages omitted)*

possible to define such an average HR, we doubt its usefulness, because the issue of noninterpretability remains. The HR is by definition a ratio of hazard functions. For example, a HR function that starts $> 1$ for small $t$ and becomes $< 1$ for large $t$ is not meaningfully summarized by a single value near 1. We therefore regard the single HR as a meaningless summary under nonproportional hazards unless the departures from proportionality are so small as to be unimportant. We prefer to allow the HR to be a function of time, as described for some of the models in chapters 5 and 7.

## 1.4 Why parametric models?

### 1.4.1 Smooth baseline hazard and survival functions

Parametric survival models generally provide smooth estimates of the hazard and survival functions for any combination of covariate values. Exceptions are piecewise models—for example, the piecewise exponential (see section 4.3.1), for which the hazard function is a step function and the survival function has discontinuities in the first derivative.

### 1.4.2 Time-dependent HRs

With parametric models, we can obtain essentially any type of output—for example, a time-dependent HR (see section 7.6)—as a function of the estimated model parameters (the covariates and time). Furthermore, we can use Stata's powerful `predictnl` command, which implements the delta method using numeric derivatives, to get SEs and CIs quite easily (see section 1.9).

### 1.4.3 Modeling on different scales

Sometimes, a covariate whose effect is nonproportional on the hazards scale may be (much closer to) proportional on another scale, such as the odds or probit (inverse normal probability) scales (see chapter 5). We may be able to take advantage of the different possible scales to build a parsimonious and efficient alternative to a PH model.

### 1.4.4 Relative survival

In cancer survival, we often want to know the impact of covariates on the mortality rate for a particular cancer. However, because cancer is usually a disease of old age, many people may die of diseases other than the cancer they were originally diagnosed with. In relative survival models, we deal with this issue by incorporating expected mortality, which we can usually obtain from routine data sources. Traditionally, simple piecewise models have been used for relative survival, but all the advantages of standard parametric survival models also apply to relative survival models. See chapter 8 for details.

### 1.4.5 Prediction out of sample

The baseline survival function in a Cox model (estimated by `predict` *varname*, `basesurv()` following use of `stcox`) is available only in the estimation sample. To predict survival outside the estimation sample, we need special measures, such as interpolation or even extrapolation. Using special measures limits the applications of the Cox model in some situations. An important case arises when we wish to validate a survival model in an independent sample, a task that necessitates out-of-sample prediction (see section 6.8).

### 1.4.6 Multiple time scales

In a Cox model, we can consider only one time scale—for example, time from diagnosis of disease or time from randomization in a clinical trial. Sometimes, for example, in age–period–cohort models (Clayton and Schifflers 1987), we might want to consider more than one time scale. See section 7.9 for an example of using multiple time scales.

## 1.5 Why not standard parametric models?

We have outlined some advantages of working with parametric models. In chapter 13 of Cleves et al. (2010)—an excellent introduction to survival analysis in Stata—the authors describe six standard parametric survival models—exponential, Weibull, Gompertz, lognormal, loglogistic, and generalized gamma. The models, together with a rich set of extensions, are implemented in the portmanteau command `streg`. Cleves et al. (2010) give formulas for the hazard and survival functions for these models, together with detailed examples and their implementation in Stata. We do not repeat the material here.

With such riches available, why do we need to go beyond `streg`? There are two main reasons. First, the simpler parametric models in `streg` may not be flexible enough to adequately represent, say, the hazard function—in other words, they may not fit the data well enough. (Concern about possible lack of fit of parametric models is one of the main reasons for the popularity of the Cox model; the shape of the baseline distribution does not influence estimates of HRs.) For example, the main parametric PH model, the Weibull, has a hazard function that always goes in the same direction with time—up, down, or constant. Many real-life datasets have hazards that peak after some period of time and then decline, so the Weibull model can never fit such data well. Second, in our book, we present new classes of parametric models that include flexible PH models, but also flexible proportional odds (PO) and probit-scale models. These alternative models greatly extend the range of survival distributions that can be estimated.

As an example, consider figure 1.6.

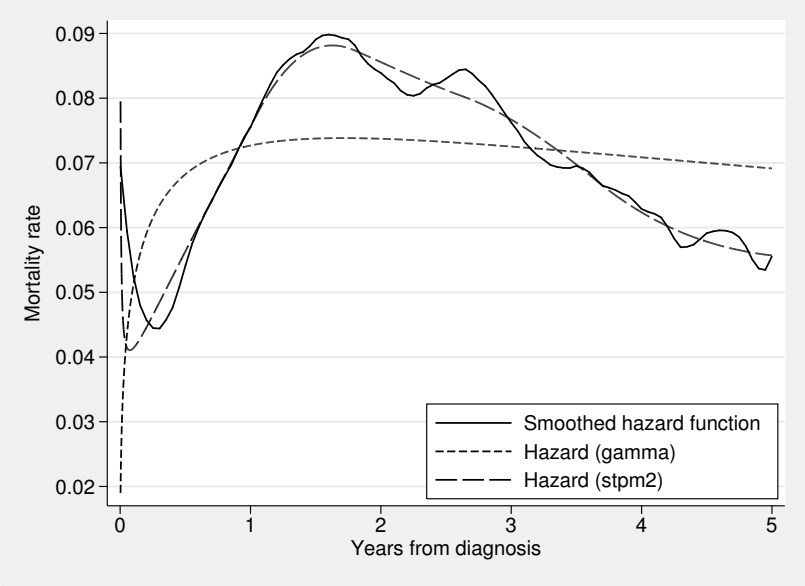


Figure 1.6. England and Wales breast cancer data. Three estimates of the hazard function (mortality rate).

The dataset relates to survival for 24,889 patients with breast cancer in England and Wales (see section 3.3). We have estimated the hazard function (here, the mortality rate) by a nonparametric smoothing technique (as in figure 1.1) and by two parametric survival models. The first model, the generalized gamma distribution, is the most complex parametric survival distribution supported by `streg`. Most of the other distributions are special cases of the gamma. The second is an RP model, which is implemented in `stpm2`. Notice how the shape of the gamma hazard function bears little resemblance to that derived from the other approaches, whereas the estimates from the nonparametric method and the RP model broadly agree (apart from the unconvincing "wiggles" in the nonparametric estimate).

One further issue with standard parametric models is that if a PH model is to be fit using `streg`, then the only choices are the exponential, Weibull, or Gompertz distributions. All of these have monotonic hazard functions in that they either increase or decrease over time (or stay constant, in the case of the exponential distribution). Thus if the underlying hazard function has a turning point it is not possible to find a well-fitting parametric PH survival model. The lognormal, loglogistic, and generalized gamma distributions do have hazard functions with turning points. However, it is not possible using these distributions in `streg` to have PH models because they can only be expressed in the accelerated failure-time metric. Although coefficients from accelerated