Journal of Statistical Software

May 2013, Volume 53, Issue 12. *http://www.jstatsoft.org/*

stgenreg: A Stata Package for General Parametric Survival Analysis

Michael J. Crowther
University of Leicester

Paul C. Lambert
University of Leicester

Abstract

In this paper we present the Stata package **stgenreg** for the parametric analysis of survival data. Any user-defined hazard function can be specified, with the model estimated using maximum likelihood utilising numerical quadrature. Models that can be fitted range from the Weibull proportional hazards model to the generalized gamma model, mixture models, cure rate models, accelerated failure time models and relative survival models. We illustrate the features of **stgenreg** through application to a cohort of women diagnosed with breast cancer with outcome all-cause death.

Keywords: survival analysis, parametric models, numerical quadrature, maximum likelihood, Stata.

1. Introduction

Parametric models remain a standard tool for the analysis of survival data. Through a fully parametric approach, we can not only obtain relative effects, such as hazard ratios in a proportional hazards model, but also clinically relevant absolute measures of risk, such as differences in survival proportions (Lambert, Dickman, Nelson, and Royston 2010). Parametric models are also useful where extrapolation is required, such as in the economic decision modelling framework (Weinstein *et al.* 2003).

The most popular tool for analysing survival data remains the Cox proportional hazards model (Cox 1972), which avoids making any assumptions for the shape of the baseline hazard function. One of the reasons the Cox model remains the prefered choice over parametric models is that standard parametric models available in standard software are often not flexible enough to capture the underlying shape of the hazard function seen in real data.

The traditional approach to estimation of parametric models is through maximum likelihood. This is relatively simply when using a known probability distribution function, such as the

Weibull or Gompertz. Many commonly used parametric survival models are implemented in a variety of software packages, such as the **streg** package in Stata (StataCorp. 2011), **survreg** (Therneau 2012) in R (R Core Team 2013) and **LIFEREG** in SAS (SAS Institute Inc. 2008). However, every parametric model has underlying assumptions, for example, the widely used Weibull proportional hazards model assumes a monotonically increasing or decreasing baseline hazard rate. Such assumptions can be considered restrictive, leading to the development of other more flexible parametric approaches (Royston and Parmar 2002; Royston and Lambert 2011).

In this paper we present the Stata command `stgenreg` which enables the user to fit general parametric models through specifying *any* baseline hazard function which can be written in a standard analytical form. This is implemented through numerical integration of the user-defined hazard function. This allows complex extensions to standard parametric models, for example, modelling the log baseline hazard function using splines or fractional polynomials, as well as complex time-dependent effects; methods that are unavailable in standard software. Time-varying covariates can also be incorporated through using multiple records per subject. We do not consider frailty (unobserved heterogeneity) in this article.

One of the key advantages of such a general framework for survival analysis is in the development of new models, for example in one line of code a parametric survival model can be fitted rather than having to directly program the likelihood evaluator.

2. Parametric survival analysis

Let T_i^* be the true event time of patient $i = 1, \dots, n$, and $T_i = \min(T_i^*, C_i)$ the observed survival time, with C_i the censoring time. Define an event indicator d_i, which takes the value of 1 if $T_i^* \leq C_i$ and 0 otherwise. We define the probability density function of T_i^* as

$$f(t) = \lim_{\delta \to 0} \frac{P(t \leq T^* \leq t + \delta)}{\delta}$$

where $f(t)$ is the unconditional probability of an event occuring in the interval $(t, t + \delta)$. We define the hazard and survival functions as

$$h(t) = \lim_{\delta \to 0} \frac{P(t \leq T^* \leq t + \delta | T^* \geq t)}{\delta} \qquad \text{and} \qquad S(t) = P(T^* \geq t)$$

such that $h(t)$ is the instantaneous failure rate at time t, and $S(t)$ is the probability of 'surviving' longer than time t. This leads to

$$f(t) = h(t)S(t) \tag{1}$$

We can further write

$$H(t) = \int_0^t h(u) \mathrm{d}u \qquad S(t) = \exp\{-H(t)\} \tag{2}$$

where $H(t)$ is the cumulative hazard function. When the integral in Equation 2 is analytically intractible, we can use numerical integration techniques to derive the cumulative hazard and thus still calculate the survival function.

2.1. Maximum likelihood estimation

The log-likelihood contribution of the i-th patient, allowing for right censoring and delayed entry (left truncation), using Equation 1 can be written as

$$\begin{aligned} l_i &= \log\left\{ f(t_i)^{d_i} \left(\frac{S(t_i)}{S(t_{0i})} \right)^{1-d_i} \right\} \\ &= d_i \log\{f(t_i)\} + (1-d_i)\log\{S(t_i)\} - (1-d_i)\log\{S(t_{0i})\} \end{aligned} \tag{3}$$

where t_{0i} and t_i are the observed entry and survival/censoring times for the i-th patient. If delayed entry is not present then the third term in Equation 3 can be dropped. Using Equation 3 we can directly maximize the log-likelihood if using known probability density and survival functions. Alternatively, using Equation 1 we can write

$$\begin{aligned} l_i &= \log\left\{ h(t_i)^{d_i} \frac{S(t_i)}{S(t_{0i})} \right\} \\ &= d_i \log\{h(t_i)\} + \log\{S(t_i)\} - \log\{S(t_{0i})\} \end{aligned}$$

and substituting Equation 2 this becomes

$$l_i = d_i \log\{h(t_i)\} - \int_{t_{0i}}^{t_i} h(u)\mathrm{d}u \tag{4}$$

We note from Equation 4 that the likelihood can also be maximized if only the hazard function is known. Of course, in standard parametric models, all 3 functions are known; however, given that often the hazard function is of most interest, specifying a complex hazard function can be advantageous. The maximization of such a specified hazard model relies on being able to evaluate the integral in Equation 4. If we propose to use such functions as fractional polynomials or splines to model a complex baseline hazard function, or incorporating complex time-dependent effects, then we have a situation where this integral cannot always be evaluated analytically, motivating alternative approaches.

2.2. Numerical integration

We propose to use numerical quadrature to evaluate the cumulative hazard, and hence maximize the likelihood in Equation 4, allowing the user to estimate a parametric survival model, specifying *any* function for the baseline hazard, satisfying $h(t) > 0$ for all $t > 0$.

Gaussian quadrature allows us to evaluate an analytically intractible integral through a weighted sum of a function evaluated at a set of pre-defined points, known as nodes (Stoer and Burlirsch 2002). We have

$$\int_{-1}^{1} g(x)\mathrm{d}x = \int_{-1}^{1} W(x)g(x)\mathrm{d}x \approx \sum_{i=1}^{m} w_i g(x_i)$$

where $W(x)$ is a known weighting function and $g(x)$ can be approximated by a polynomial function. The integral over $[t_{0i}, t_i]$ in Equation 4 must be changed to an integral over $[-1, 1]$

using the following rule

$$\begin{aligned}\int_{t_{0i}}^{t_i} h(x)\mathrm{d}x &= \frac{t_i - t_{0i}}{2}\int_{-1}^{1} h\left(\frac{t_i - t_{0i}}{2}x + \frac{t_{0i}+t_i}{2}\right)\mathrm{d}x \\ &\approx \frac{t_i - t_{0i}}{2}\sum_{i=1}^{m} w_i h\left(\frac{t_i - t_{0i}}{2}x_i + \frac{t_{0i}+t_i}{2}\right)\end{aligned}$$

This transformation allows the incorporation of delayed entry quite simply. The form of Gaussian quadrature depends on the choice of weighting function. The default within `stgenreg` is Gauss-Legendre quadrature, with weighting function, $W(x) = 1$.

The accuracy of the numerical integral depends on the number of quadrature nodes, m, with node locations dependent on the type of quadrature chosen. As with all methods which use numerical integration, the stability of maximum likelihood estimates should be established by using an increasing number of quadrature nodes.

2.3. Time-dependent effects and time-varying covariates

The presence of non-proportional hazards, i.e., time-dependent effects, is common in the analysis of time to event data (Jatoi, Anderson, Jeong, and Redmond 2011). This is frequently observed in registry data sources where follow-up time is often over many years (Lambert *et al.* 2011). Similarly in clinical trials, time-dependent treament effects are also observed (Mok *et al.* 2009). Time-dependent effects are incorporated seemlessly into our modelling framework, by allowing the user to interact any covariates with a specified function of time. We illustrate this in Section 4.2.1.

Time-varying covariates are a further often observed scenario in the analysis of survival data, where the value of a covariate for individual patients can change at various points in follow-up. For example in oncology clinical trials, patients will often switch treatment group when their condition progresses (Morden, Lambert, Latimer, Abrams, and Wailoo 2011), or biomarkers may be measured repeatedly over time, resulting in multiple records per subject (**?**). For this form of analysis the data is often set up into start and stop times, and since delayed entry (left truncation) is allowed, this again is incorporated into the described modelling framework. We illustrate through example in Section 4.4.

3. The Stata package stgenreg

The Stata package **stgenreg** is implemented as three Stata ado files. The primary shell program, `stgenreg.ado`, handles the syntax options for the package, which then calls the likelihood evaluator program `stgenreg_d0.ado`, described in Section 3.1. Finally, a variety of predictions can be obtained following estimation of a model using Stata's `predict` command, which calls the program `stgenreg_pred.ado`, described in Section 3.2.

3.1. Program implementation and syntax

The log-likelihood shown in Equation 4 is maximized using the Newton-Raphson algorithm, with first and second derivatives estimated numerically, as implemented in the `ml` command in Stata (Gould, Pitblado, and Poi 2010). As described in Section 2.1, the integral in Equation 4 is evaluated using m-point Gaussian quadrature.

The evaluator program has been optimized using Stata's matrix programming language, Mata. This provides computational benefits and use of the wide array of mathematical functions available for the user to specify in the hazard function. In addition, we have implemented specific functions which allow the incorporation of restricted cubic splines or fractional polynomials into the hazard or log hazard function (Durrleman and Simon 1989; Royston and Altman 1994).

When using `stgenreg` one of the options `loghazard()` or `hazard()` must be defined. These specify a user-defined log hazard or hazard function. The function must be defined in Mata code, with parameters specified in square brackets, for example `[ln_lambda]`. The use of Mata means that mathematical operations require a colon (`:`) prefix, for example `:+` instead of `+`. Time must be coded as `#t`. The user can specify covariates or functions of time within the linear predictor of any parameter, providing a highly flexible framework.

For example, we can specify a Weibull distribution using either the log hazard or hazard function. Each parameter is parameterized to contain the entire real number line, i.e., both λ and γ are restricted to be positive by modelling on the log scale.

```
. stgenreg, loghazard([ln_lambda] :+ [ln_gamma]                  ///
> :+ (exp([ln_gamma]) :- 1) :* log(#t))
. stgenreg, hazard(exp([ln_lambda]) :* exp([ln_gamma]) :*        ///
> #t :^ (exp([ln_gamma]) :- 1))
```

A linear predictor can be defined for any of the parameters, with the name of the option defined as the name of the parameter specified in the `loghazard()` or `hazard()` option. For example a proportional hazards Weibull model can be fitted with covariates treatment, age and sex by adding the option `ln_lambda(treatment age sex)`.

One of the key advantages of `stgenreg` is that we can incorporate a variety of functions (including functions of time) into the linear predictor of any parameter. For example, parameter `[ln_lambda]` has an available option `ln_lambda(comp1 | comp2 | ...| compn)`, which can contain a variety of component functions to increase complexity. Each `comp`*j* can contain a variety of functions described in Table 1.

Additionally, excess mortality (relative survival) models (Nelson, Lambert, Squire, and Jones 2007) can be fitted by use of the `bhazard(varname)` option. In these models a known expected mortality rate, $h^*(t)$, is included in the model as follows,

$$h(t) = h^*(t) + \lambda(t)$$

Here the `loghazard()` and `hazard()` options now refer to the modelling of $\lambda(t)$. Note that it is the expected mortality rate at the event time that needs to be supplied to the `bhazard()` option.

Finally, all standard options of the `ml` suite in Stata can be used when fitting a `stgenreg` model, such as `constraints()` which allow the user to constrain the value of any coefficient to be a particular constant.

3.2. Predictions

A variety of predictions can be obtained following the estimation of a model. These include the hazard, survival and cumulative hazard functions.

Component	Description
`varlist [, nocons]`	The user may specify a standard variable list within a component section, with an optional `nocons` option.
`g(#t)`	Where `g()` is any user defined function of `#t` written in Mata code, for example `#t:^2`.
`#rcs(`*options*`)`	Creates restricted cubic splines of either log time or time. Options include `df(int)`, the number of degrees of freedom, `noorthog` which turns off the default orthogonalisation, `time`, which creates splines using time rather than log time, the default, and `offset(varname)` to include an offset when calculating the splines. See **rcsgen** in Stata for more details.
`#fp(numlist [,`*options*`])`	Creates fractional polynomials of time with powers defined in `numlist`. If 0 is specified, log time is generated. The only current option is `offset()` which is consistent with that described in `#rcs()` above.
`varname:*f(#t)`	To include time-dependent effects, where `f(#t)` is one of `#rcs()`, `#fp()` or `g()`.

Table 1: Description of each component that can be included in the linear predictor of a parameter.

The standard Stata syntax to obatin predictions following a model fit is as follows

```
. predict newvarname, statistic
```

So for example, to obtain the fitted survival, hazard and cumulative hazard functions

```
. predict surv1, survival
. predict haz1, hazard
. predict cumhaz1, cumhazard
```

Extended prediction options unavilable in standard software include: `zeros` – obtains baseline predictions, `at()` – obtains predictions at specified covariate patterns, `timevar()` – obtains predictions at specified times. These options can be combined with standard choices of `hazard`, `cumhazard` and `survival`. Finally, the `ci` option can be used to obtain confidence intervals.

4. Analysis of example datasets using stgenreg

We illustrate `stgenreg` through use of a dataset comprising of 9721 women aged under 50 and diagnosed with breast cancer in England and Wales between 1986 and 1990. The event of interest is death from any cause, with follow-up restricted to 5 years. Deprivation was categorized into 5 levels; however, we have restricted the analyzes to comparing the most affluent and most deprived groups, for illustrative purposes. We therefore only consider a binary covariate, `dep5`, with 0 for the most affluent and 1 for the most deprived group.

We further illustrate how to incorporate a time-varying covariate through use of a dataset of 488 patients with liver cirrhosis (Anderson, Borgan, Gill, and Keiding 1993). A total of 251 patients were randomized to receive prednisone, with 237 randomized to receive a placebo. Prothrombin index was measured repeatedly, with between 1 and 17 measurements per subject, resulting in 2968 observations. Outcome was all-cause death.

4.1. Weibull proportional hazards model

We begin by fitting a Weibull proportional hazards model to the breast cancer dataset, investigating the effect of deprivation status. Given that Weibull models are available in all standard statistical software, we first illustrate the concept showing that the estimates agree with estimates derived using analytically tractible definitions of the hazard and survival functions. The baseline hazard and log hazard functions have the following form

$$h(t) = \lambda\gamma t^{\gamma-1}\exp(\beta X)$$

and

$$\log(h(t)) = \log(\lambda) + \log(\gamma) + (\gamma - 1)\log(t) + \beta X$$

where X is a vector of covariates, with corresponding regression coefficients β. In this case it is convenient to use the `loghazard()` option of `stgenreg`. We can investigate covariate effects by including deprivation status in the linear predictor of $\log(\lambda)$, using the option `ln_lambda`.

```
. stgenreg, loghazard([ln_lambda] :+ [ln_gamma] :+ ///
> (exp([ln_gamma]) :- 1) :* log(#t)) nodes(30) ln_lambda(dep5)

Log likelihood =  -8808.149                     Number of obs   =       9721

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ln_lambda    |
        dep5 |   .2698633   .0392017     6.88   0.000     .1930293    .3466972
       _cons |  -2.824814   .0370151   -76.32   0.000    -2.897362   -2.752265
-------------+----------------------------------------------------------------
ln_gamma     |
       _cons |   .0464514   .0179823     2.58   0.010     .0112068     .081696
------------------------------------------------------------------------------
 Quadrature method: Gauss-Legendre with 30 nodes
```

We observe a log hazard ratio of 0.270 (95% CI: 0.193, 0.347) and consequently a hazard ratio of 1.310 (95% CI: 1.213, 1.414), indicating a 31% increase in the mortality rate in the most deprived group compared to the most affluent. We could further adjust the γ parameter by deprivation status but adding the option `ln_gamma(dep5)`.

When fitting models which rely on numerical integration, it is important to establish the stability of maximum likelihood estimates by using an increasing number of quadrature nodes. In the case of a Weibull proportional hazards model, we can both compare with the optimized model using `streg` in Stata, and compare with an increasing number of quadrature nodes. Here we present results from fitting the `streg` model and `stgenreg` models with 15, 30, 50 and 100 nodes.

```
------------------------------------------------------------------------------
  Variable |    streg      stgenreg15   stgenreg30   stgenreg50   stgenreg100
-----------+------------------------------------------------------------------
#1         |
      dep5 |   .2698715    .26983514    .26986326    .26986899    .26987095
           |   .0392017    .03920178    .03920173    .03920172    .03920171
     _cons | -2.8252423   -2.8232443   -2.8248136   -2.8251059   -2.8252139
           |  .03694985    .03718485    .03701515    .03697471    .03695639
-----------+------------------------------------------------------------------
#2         |
     _cons |  .04673335    .04542627    .04645138    .04664313    .04671442
           |  .01792781    .01812554    .01798227    .01794843     .0179332
-----------+------------------------------------------------------------------
Statistics |
        ll | -8808.0854   -8808.3461    -8808.149   -8808.1075   -8808.0906
------------------------------------------------------------------------------
```

We obtain consistent parameter estimates to 3 decimal places with 30 nodes, and accuracy is improved when the number of nodes are increased. However, computation time will increase with an increasing number of nodes, for example using 15 nodes takes 7.4 seconds compared with 12.4 seconds using 100 nodes (on a HP laptop with Intel i5 2.5GHz processor with 8GB of RAM). In comparison, the fully optimized `streg` model took 0.4 seconds to converge. This difference is clearly expected as the `stgenreg` formulation of the Weibull model is not the most computationally efficient, as there is no need to use numerical integration when using the standard Weibull model.

4.2. Restricted cubic spline proportional hazards model

We now introduce a much more flexible proportional hazards survival model, modelling the baseline log hazard function using restricted cubic splines of log(time). We formulate the baseline log hazard function

$$\log(h(t)) = s(\log(t)) + X\beta \tag{5}$$

where $s(\log(t))$ is a restricted cubic spline function of $\log(t)$. This can be implemented by using the `#rcs` component option. We use the default knot locations, based on the centiles of the distribution of uncensored survival times.

This draws parallels with the flexible parametric model of Royston and Parmar (2002), implemented in Stata as the `stpm2` command (Royston and Lambert 2011), which uses restricted cubic splines to model the log *cumulative* hazard function

$$\log(H(t)) = s(\log(t)) + X\beta \tag{6}$$

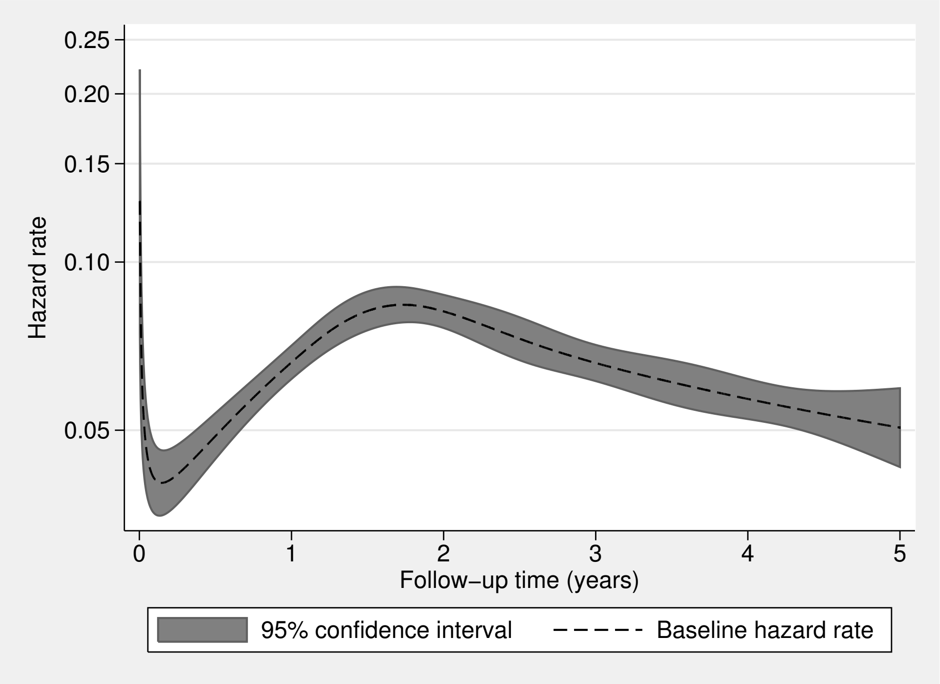

Figure 1: Predicted hazard function for the most affluent group with 95% confidence interval.

An advantage of modelling on the log hazard scale is that when there are multiple time dependent effects, the interpretation of the time-dependent hazard ratios is simplified as they do not depend on values of other covariates, which is the case when modelling on the cumulative hazard scale (Royston and Lambert 2011).

We apply the model in Equation 5 with 5 degress of freedom, i.e., 4 internal knots placed at the 20th, 40th, 60th and 80th percentiles of the distribution of log event times, and 2 boundary knots placed at the 0th and 100th percentiles.

```
. stgenreg, loghazard([xb]) xb(dep5 | #rcs(df(5))) nodes(30)

Log likelihood = -8756.2213                     Number of obs   =       9721

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dep5 |   .2693634   .0392018     6.87   0.000     .1925293    .3461976
_eq1_cp2_rcs1 | -.0621779   .0274602    -2.26   0.024    -.1159989    -.008357
_eq1_cp2_rcs2 |   .0784834   .0192975     4.07   0.000     .0406611    .1163057
_eq1_cp2_rcs3 |   .1158689   .0176746     6.56   0.000     .0812272    .1505106
_eq1_cp2_rcs4 | -.0251518   .0143719    -1.75   0.080    -.0533202    .0030165
_eq1_cp2_rcs5 |   .0012793   .0134076     0.10   0.924    -.0249991    .0275576
       _cons | -2.910463   .0607005   -47.95   0.000    -3.029434   -2.791492
------------------------------------------------------------------------------
 Quadrature method: Gauss-Legendre with 30 nodes
```

When using the component options stgenreg will create variables labelled by the equation number (indexed from left to right in the log hazard or hazard specification) and the component number (again counting from left to right in each parameter option). So variables _eq1_cp2_* contain the spline basis variables defined by the #rcs(df(5)) component. The estimate of the log hazard ratio for the effect of deprivation is very similar to the Weibull based estimate; however, we have now estimated 6 parameters to model the baseline hazard function, an intercept and 5 parameters associated with the spline terms. We can obtain the predicted baseline hazard function and 95% confidence interval as follows

```
. predict haz1, hazard ci zeros
```

We illustrate the fitted baseline hazard function in Figure 1.

Time-dependent effects

We now investigate the presence of a time-dependent effect due to deprivation status. Within the framework of restricted cubic splines, this can be investigated using the component form varname:*#rcs(df(num)), i.e., an interaction between the effect of time (using splines) and the deprivation group. We use 3 degrees of freedom for illustration.

```
. stgenreg, loghazard([xb]) nodes(30) ///
> xb(dep5 | #rcs(df(5)) | dep5 :* #rcs(df(3)))

Log likelihood = -8747.3275                      Number of obs   =       9721

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dep5 |   .0723415   .0924005     0.78   0.434    -.1087602    .2534433
_eq1_cp2_rcs1 | -.0108058   .0309504    -0.35   0.727    -.0714673    .0498558
_eq1_cp2_rcs2 |  .0672877   .0224852     2.99   0.003     .0232177    .1113578
_eq1_cp2_rcs3 |  .1128672   .0207167     5.45   0.000     .0722634    .1534711
_eq1_cp2_rcs4 | -.0261438   .0145455    -1.80   0.072    -.0546525     .002365
_eq1_cp2_rcs5 |  .0014202   .0134079     0.11   0.916    -.0248589    .0276992
_eq1_cp3_rcs1 | -.1464002   .0443983    -3.30   0.001    -.2334194   -.0593811
_eq1_cp3_rcs2 |  .0425164   .0333753     1.27   0.203     -.022898    .1079307
_eq1_cp3_rcs3 |  .0135896   .0322604     0.42   0.674    -.0496396    .0768187
       _cons | -2.849318   .0649361   -43.88   0.000    -2.976591   -2.722046
------------------------------------------------------------------------------
 Quadrature method: Gauss-Legendre with 30 nodes
```

In Figure 2 we compare the fit of the models with either time-independent or time-dependent hazard ratios for deprivation status, by overlaying the fitted survival functions onto the Kaplan-Meier curve, for each deprivation group. We observe a much improved fit to the Kaplan-Meier curve when modelling the time-dependent effect of deprivation group. We can predict the time-dependent hazard ratio using the partpred (Lambert 2010) command as follows.

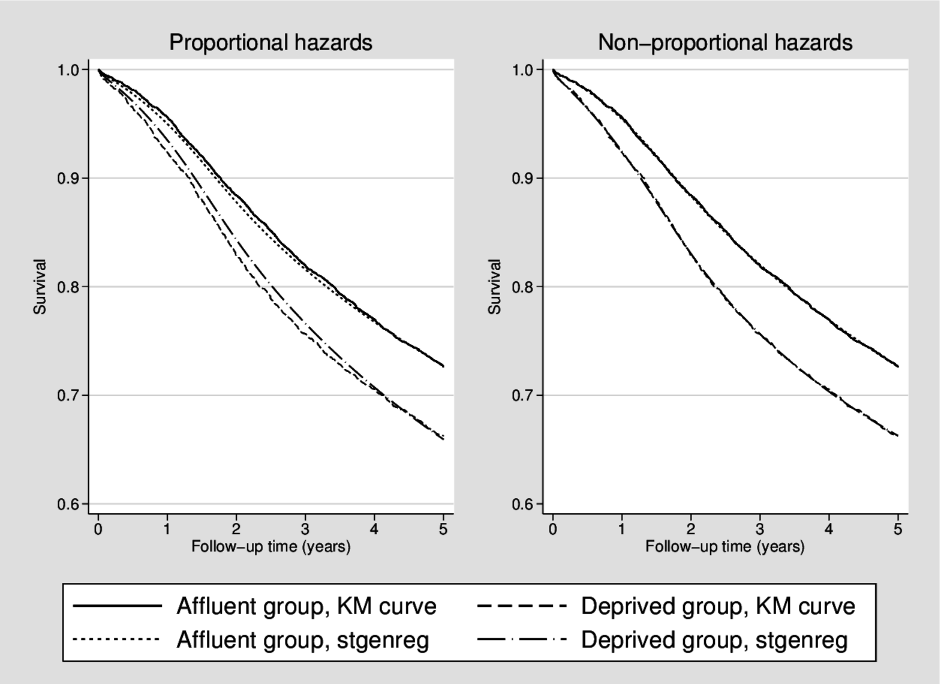

Figure 2: Kaplan-Meier estimates for the most affluent and most deprived groups, with predicted survival overlaid. The figure on the left shows predicted survival with a proportional effect of deprivation status, with the figure on the right allowing for non-proportional hazard sin the effect of deprivatin status.

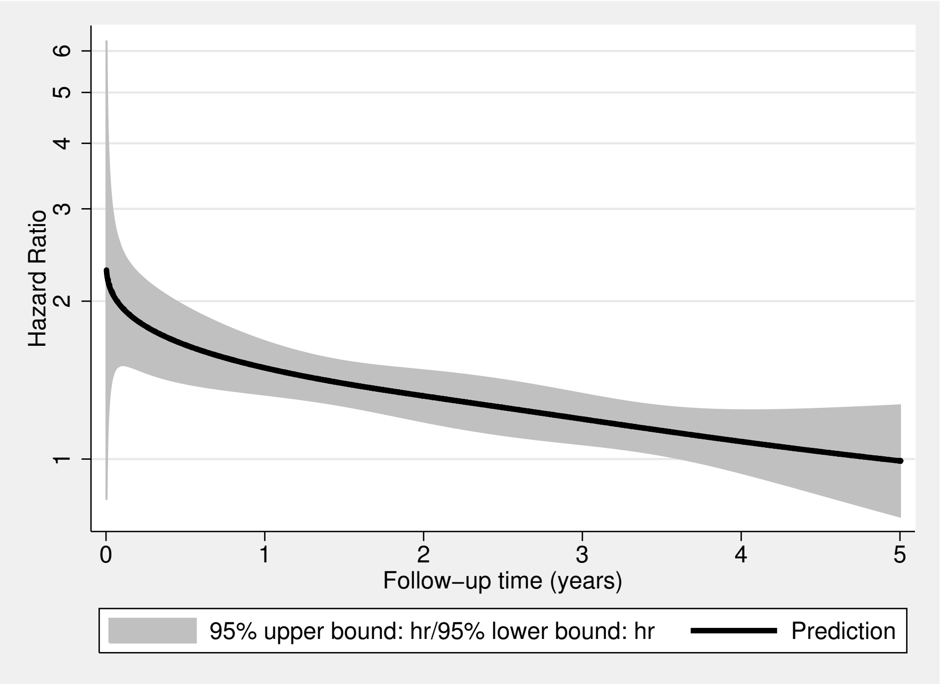

Figure 3: The estimated time-dependent hazard ratio for deprivation group and associated 95% confidence interval.

```
. partpred hr, for(dep5 _eq1_cp3*) ci(hr_uci hr_lci) eform
```

This is then plotted in Figure 3 which shows that the relative increase in the mortality rate is much larger at the start of follow-up and decreases to around one by 5 years.

4.3. Generalized gamma proportional hazards model

The generalized gamma (GG) is a 3-parameter parametric model implemented in a variety of statistical packages (Cox, Chu, Schneider, and Munoz 2007). However, it is parameterized as an accelerated failure time model in Stata. We can write the survival and density functions as

$$S_{GG}(t) = \begin{cases} 1 - I\left(\gamma, u\right) & \text{if } \kappa > 0 \\ 1 - \Phi\left(z\right) & \text{if } \kappa = 0 \\ I\left(\gamma, u\right) & \text{if } \kappa < 0 \end{cases} \tag{7}$$

and

$$f_{GG}(x) = \begin{cases} \frac{\gamma^{\gamma}}{\sigma t\sqrt{2\pi}} \exp(z\sqrt{}(\gamma) - u) & \text{if } \kappa \neq 0 \\ \frac{1}{\sigma t\sqrt{2\pi}} \exp(-z^2/2) & \text{if } \kappa = 0 \end{cases} \tag{8}$$

where $\gamma = |\kappa|^{-2}$, $z = \text{sign}\{\log(t) - \mu\}$, $\mu = \gamma \exp(|\kappa| z)$, $\Phi(z)$ is the standard normal cumulative distribution, and $I(a, x)$ is the incomplete gamma function.

Therefore using Equation 1, we can write down our baseline hazard function as the ratio of the probability distribution function to the survival function.

$$h_{GG}(t) = \frac{f_{GG}(t)}{S_{GG}(t)}$$

To invoke proportional hazards we can then simply multiply by the exponential of a parameter, the linear parameter of which is our vector of covariates

$$h_{GG}(t) = \frac{f_{GG}(t)}{S_{GG}(t)} \exp(X\beta) \quad \text{or} \quad \log(h_{GG}(t)) = \log\left(\frac{f_{GG}(t)}{S_{GG}(t)}\right) + X\beta$$

Where β is a vector of log hazard ratios. In terms of implementation, in the linear predictor for our $X\beta$ parameter we must specify the `nocons` option to ensure no intercept term, obtaining a proportional hazards formulation for the GG model. As this is a complex function, we can use Stata's local macros to build up the function.

```
. local mu [mu]
. local sigma exp([ln_sigma])
. local kappa [kappa]
. local gamma (abs(`kappa') :^ (-2))
. local z (sign(`kappa') :* (log(#t) :- `mu') :/ (`sigma'))
. local u ((`gamma') :* exp(abs(`kappa') :* (`z')))
. local surv1 (1 :- gammap(`gamma',`u')) :* (`kappa' :> 0)
. local surv2 (1 :- normal(`z')) :* (`kappa' :== 0)
. local surv3 gammap(`gamma',`u') :* (`kappa' :< 0)
. local pdf1 ((`gamma' :^ `gamma') :* exp(`z' :* sqrt(`gamma') :- `u') :/ ///
> (`sigma' :* #t :* sqrt(`gamma') :* gamma(`gamma'))) :* (`kappa' :! =0)
. local pdf2 (exp(-(`z' :^ 2) :/ 2) :/ (`sigma' :* #t :* sqrt(2 :* pi())))///
> :* (`kappa' :== 0)
. local haz (`pdf1' :+ `pdf2') :/ (`surv1' :+ `surv2' :+ `surv3')
. stgenreg, hazard(exp([xb]) :* (`haz')) nodes(30) xb(dep5,nocons)

Log likelihood = -8801.2754                     Number of obs   =       9721

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
        dep5 |   .2694578   .0391992     6.87   0.000     .1926289    .3462868
-------------+----------------------------------------------------------------
kappa        |
       _cons |   .6752793   .0749985     9.00   0.000      .528285    .8222735
-------------+----------------------------------------------------------------
mu           |
       _cons |   2.710497    .032793    82.65   0.000     2.646224    2.774771
-------------+----------------------------------------------------------------
ln_sigma     |
       _cons |   .1727204   .0521935     3.31   0.001     .0704231    .2750178
------------------------------------------------------------------------------
 Quadrature method: Gauss-Legendre with 30 nodes
```

Once again we obtain very similar estimates to the Weibull model, but now modelling the baseline with 3 parameters. This model formulation illustrates a powerful tool where by simply introducing an extra parameter we can implement a model not available in any software package.

4.4. Time-varying covariates

We now illustrate the data setup required for survival analysis incorporating a time-varying covariate. We use the liver cirrhosis dataset described above. Here we use the `enter()` and `id()` options of `stset` in Stata, to declare the data as multiple record per subject.

```
. stset stop, enter(start) id(id) failure(event=1)

                id:  id
     failure event:  event == 1
obs. time interval:  (stop[_n-1], stop]
 enter on or after:  time start
 exit on or before:  failure

------------------------------------------------------------------------------
     2968  total obs.
        0  exclusions
```

```
----------------------------------------------------------------------------
     2968  obs. remaining, representing
      488  subjects
      292  failures in single failure-per-subject data
 1777.749  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =  13.39393
```

We illustrate the data structure of 2 patients, where `_t0` represents the enter times at which prothrombin was measured

```
. list id pro trt _t0 _t _d if id==1 | id==111, noobs sepby(id)

  +-----------------------------------------------------+
  |  id   pro          trt         _t0          _t   _d |
  |-----------------------------------------------------|
  |   1    38      placebo           0    .2436754    0 |
  |   1    31      placebo    .2436754   .38057169    0 |
  |   1    27      placebo   .38057169   .41342679    1 |
  |-----------------------------------------------------|
  | 111    59   prednisone           0   .24641332    0 |
  | 111    60   prednisone   .24641332   .49830249    0 |
  | 111    87   prednisone   .49830249   .74471581    0 |
  | 111    59   prednisone   .74471581   1.1280254    0 |
  | 111    35   prednisone   1.1280254   1.1581426    1 |
  +-----------------------------------------------------+
```

We can now fit a `stgenreg` model using restricted cubic splines to model the baseline, adjusting for the proportional effects of treatment and prothrombin index.

```
. stgenreg, loghazard([xb]) xb(pro trt | #rcs(df(3))) nolog

Variables _eq1_cp2_rcs1 to _eq1_cp2_rcs3 were created

Log likelihood = -588.17466                      Number of obs   =     2968

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pro |  -.0349754   .0024771   -14.12   0.000    -.0398304   -.0301205
         trt |   .1325576   .1182068     1.12   0.262    -.0991235    .3642388
_eq1_cp2_rcs1 |   -.091006   .0579785    -1.57   0.116    -.2046419    .0226298
_eq1_cp2_rcs2 |  -.1354551   .0431334    -3.14   0.002     -.219995   -.0509151
_eq1_cp2_rcs3 |  -.2292129   .0499583    -4.59   0.000    -.3271295   -.1312964
       _cons |   .7376377   .1690535     4.36   0.000     .4062988    1.068977
------------------------------------------------------------------------------
 Quadrature method: Gauss-Legendre with 15 nodes
```

We observe a log hazard ratio of −0.35 (95% CI: −0.040, −0.030) indicating lower values of the biomarker are associated with an increased risk of death.

Alternatively `stgenreg` can be used in conjunction with Stata's `stsplit` command, to create at risk time intervals.

5. Discussion

We have presented the `stgenreg` command in Stata, for the general parametric analysis of survival data. Through specification of a user-defined hazard function, we have illustrated how to implement standard proportional hazards models, novel restricted cubic spline survival models and a generalized gamma model with proportional hazards. In essence, `stgenreg` may be used to implement a parametric survival model defined by anything from a very simple one parameter proportional hazards model, to models which contain highly flexible functions of time, for both the baseline and time-dependent effects. Any parameter defined in the hazard function can be dependent on complex functions of time, including fractional polynomials or restricted cubic splines.

The choice of the number of quadrature nodes is left to the user. An increasing number of quadrature nodes should be used to establish consistent parameter estimates.

As it is a general framework, it may not be the most computationally efficient; however, it is a useful tool for the development of novel models. For example, it may be useful to develop ideas and test new models, but then spend time developing more computationally efficient methods for specific cases.

In future developments we aim to allow for interval censoring, the extension to incorporate frailty and a post-estimation command to calculate the cumulative incidence function for competing risks. The package is available from the Statistical Software Components archive (Crowther and Lambert 2013) and can be installed from Stata by typing `ssc install stgenreg`.

Acknowledgments

Michael Crowther was funded by a National Institute for Health Research (NIHR) Doctoral Fellowship (DRF-2012-05-409).

The authors would like to thank two anonymous reviewers and an editor whose comments greatly improved the paper.

References

Anderson PK, Borgan Ø, Gill RD, Keiding N (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag.

Cox C, Chu H, Schneider MF, Munoz A (2007). "Parametric Survival Analysis and Taxonomy of Hazard Functions for the Generalized Gamma Distribution." *Statistics in Medicine*, **26**(23), 4352–4374.

Cox DR (1972). "Regression Models and Life-Tables." *Journal of the Royal Statistical Society B*, **34**(2), 187–220.

Crowther MJ, Lambert P (2013). "**stgenreg**: Stata Module to Fit General Parametric Survival Models." Statistical Software Components, Boston College Department of Economics. URL http://ideas.repec.org/c/boc/bocode/s457579.html.

Durrleman S, Simon R (1989). "Flexible Regression Models with Cubic Splines." *Statistics in Medicine*, **8**(5), 551–561.

Gould W, Pitblado J, Poi B (2010). *Maximum Likelihood Estimation with Stata*. 4th edition. Stata Press.

Jatoi I, Anderson WF, Jeong JH, Redmond CK (2011). "Breast Cancer Adjuvant Therapy: Time to Consider Its Time-Dependent Effects." *Journal of Clinical Oncology*, **29**(17), 2301–2304.

Lambert P (2010). "**partpred**: Stata Module to Generate Partial Predictions." Statistical Software Components, Boston College Department of Economics. URL http://ideas.repec.org/c/boc/bocode/s457176.html.

Lambert PC, Dickman PW, Nelson CP, Royston P (2010). "Estimating the Crude Probability of Death due to Cancer and other Causes using Relative Survival Models." *Statistics in Medicine*, **29**(7-8), 885–895.

Lambert PC, Holmberg L, Sandin F, Bray F, Linklater KM, Purushotham A, Robinson D, Møller H (2011). "Quantifying Differences in Breast Cancer Survival between England and Norway." *Cancer Epidemiology*, **35**(6), 526–533.

Mok TS, Wu YL, Thongprasert S, Yang CH, Chu DT, Saijo N, Sunpaweravong P, Han B, Margono B, Ichinose Y, Nishiwaki Y, Ohe Y, Yang JJ, Chewaskulyong B, Jiang H, Duffield EL, Watkins CL, Armour AA, Fukuoka M (2009). "Gefitinib or Carboplatin-Paclitaxel in Pulmonary Adenocarcinoma." *New England Journal of Medicine*, **361**(10), 947–957.

Morden JP, Lambert PC, Latimer N, Abrams KR, Wailoo AJ (2011). "Assessing Methods for Dealing with Treatment Switching in Randomised Controlled Trials: A Simulation Study." *BMC Medical Research Methodology*, **11**, 4.

Nelson CP, Lambert PC, Squire IB, Jones DR (2007). "Flexible Parametric Models for Relative Survival, with Application in Coronary Heart Disease." *Statistics in Medicine*, **26**(30), 5486–5498.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Royston P, Altman DG (1994). "Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling." *Journal of the Royal Statistical Society C*, **43**(3), 429–467.

Royston P, Lambert PC (2011). *Flexible Parametric Survival Analysis using Stata: Beyond the Cox model*. Stata Press.

Royston P, Parmar MKB (2002). "Flexible Parametric Proportional Hazards and Proportional Odds Models for Censored Survival Data, with Application to Prognostic Modelling and Estimation of Treatment Effects." *Statistics in Medicine*, **21**(15), 2175–2197.

SAS Institute Inc (2008). *SAS/STAT Software, Version 9.2.* Cary, NC. URL http://www.sas.com/.

StataCorp (2011). "Stata Data Analysis Statistical Software: Release 12." URL http://www.stata.com/.

Stoer J, Burlirsch R (2002). *Introduction to Numerical Analysis.* 3rd edition. Springer-Verlag.

Therneau T (2012). ***survival**: A Package for Survival Analysis in S.* R package version 2.36-14, URL http://CRAN.R-project.org/package=survival.

Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C, Luce BR (2003). "Principles of Good Practice for Decision Analytic Modeling in Health-Care Evaluation: Report of the ISPOR Task Force on Good Research Practices–Modeling Studies." *Value in Health*, **6**(1), 9–17.

Affiliation:

Michael J. Crowther
Department of Health Sciences
University of Leicester
Leicester, United Kingdom
E-mail: michael.crowther@le.ac.uk
URL: http://www2.le.ac.uk/departments/health-sciences/research/biostats/staff-pages/mjc76/

Paul C. Lambert
Department of Health Sciences
University of Leicester
Leicester, United Kingdom
and
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden
E-mail: paul.lambert@le.ac.uk
URL: http://www2.le.ac.uk/Members/pl4/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 53, Issue 12 *Submitted:* 2012-07-09
May 2013 *Accepted:* 2013-01-08