The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

The *Stata Journal* is indexed and abstracted in the following:

- CompuMath Citation Index®
- Current Contents/Social and Behavioral Sciences®
- RePEc: Research Papers in Economics
- Science Citation Index Expanded (also known as SciSearch®)
- Scopus™
- Social Sciences Citation Index®

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata, STATA, Stata Press, Mata, MATA, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2012)
12, Number 2, pp. 242–256

Faster estimation of a discrete-time proportional hazards model with gamma frailty

Michael G. Farnworth
University of New Brunswick
Fredericton, New Brunswick, Canada
mikefarn@unb.ca

Abstract. Fitting a complementary log-log model that accounts for gamma-distributed unobserved heterogeneity often takes a significant amount of time. This is in part because numerical derivatives are used to approximate the gradient vector and Hessian matrix. The main contribution of this article is the use of Mata and a `gf2` evaluator to express the gradient vector and Hessian matrix. Gradient vector expression allows one to use a few different options and postestimation commands. Furthermore, expression of the gradient vector and Hessian matrix increases the speed at which a likelihood function is maximized.

In this article, I present a complementary log-log model, show how the gamma distribution has been incorporated, and point out why the gradient vector and Hessian matrix can be expressed. I then discuss the speed at which a maximum is achieved, and I apply sampling weights that require an expression of the gradient vector. I introduce a new command for fitting this model. To demonstrate how this model can be applied, I will examine information on when young males first try marijuana.

Keywords: st0256, pgmhazgf2, hazard model, discrete duration analysis, complementary log-log model, gamma distribution, unobserved heterogeneity

1 Introduction

In the duration literature, survival times are often grouped into months, years, or some other width. Information on widely banded survival times is common in the social sciences, and in the discrete duration literature, a few different functional forms have been used (Jenkins 1995, 134). One option is a logistic that corresponds to a proportional odds model. Alternatively, a probit model can be fit. A third option is the complementary log-log proposed by Prentice and Gloeckler (1978). Jenkins (2008) provides an introduction to survival analysis that pays particular attention to these discrete-time models, and Sueyoshi (1995) shows that each of them is nested in a class of continuous-time hazard models.

When applying duration analysis, unobserved heterogeneity may be an important concern. In response, one can assume that heterogeneity across individuals is normally distributed with a mean of 0, and the `xtlogit` or `xtcloglog` command may be applied. Alternatively, one can fit a complementary log-log model under which unobserved heterogeneity is accounted for with a multinomial distribution that has a particular number of mass points (Heckman and Singer 1984). To fit this model, Jenkins (2004a) provides a program called `hshaz`.[1]

On the topic of unobserved heterogeneity, Abbring and van den Berg (2007) find that in a large class of hazard models with proportional unobserved heterogeneity, the distribution of the heterogeneity among survivors converges to a gamma distribution. Meyer (1990) incorporates this distribution into a complementary log-log functional form, and hence this is called a Prentice–Gloekler–Meyer (pgm) hazard model. This model can be fit when there are no left-truncated survival times and when each individual experiences a single spell. Ondrich and Rhody (1999) extend this method to examine multiple spells that an individual may experience. To examine single spells that are not left-truncated, Jenkins (2004b) provides a command called `pgmhaz8`.[2] This is a `d0` maximum likelihood program that is publicly available and runs under Stata 8.0 or higher. This program approximates the gradient vector and Hessian matrix with numerical derivatives, so maximization can take a significant amount of time.

To demonstrate the speed of the new `pgmhazgf2` command that uses an expression of the gradient vector and Hessian matrix, I examine information on when young males first try marijuana. This sample is from a 2006–2007 Canadian provincial Youth Smoking Survey of New Brunswick students in grades 6 through 12 (grade 5 students in some schools) conducted by the University of New Brunswick and the Université de Moncton in collaboration with the Centre for Behavioural Research and Program Evaluation at the University of Waterloo. The failure-indicator variable is based on the survey question "Have you ever tried marijuana or cannabis? (a joint, pot, weed, hash, ...)". If the answer is yes, individuals are also asked "How old were you when you first used marijuana or cannabis?" This briefly demonstrates how the estimation can be applied, and hence I just examine the sample of males.

Incorporating gradient and Hessian expressions within a maximum likelihood program increases the speed at which a maximum is arrived at and allows one to apply sampling `pweights` as well as postestimation, such as robust, clustering, and stratification. This maximum likelihood command also supports estimation with survey data (`svy:`). After estimation, the `predict` command with a `score` option can be used.

2 Model

Under a complementary log-log model, the hazard of individual i first trying marijuana at any point within age-level j is

$$h(\text{age}_{ij}) = 1 - \exp\left\{ - \int_{\text{beginning of age}_j}^{\text{end of age}_j} \lambda_0(t)dt \times \exp(x_{ij} \times \beta) \right\}$$

[1] This can be installed in Stata by typing `ssc install hshaz`. This program along with the associated help file is available at http://ideas.repec.org/c/boc/bocode/s444601.html.

[2] This can be installed in Stata by typing `ssc install pgmhaz8`. This program along with the associated help file is available at http://ideas.repec.org/c/boc/bocode/s438501.html.

The term in curly brackets is minus one times a cumulative hazard that has an unspecified baseline over continuous time within an age level. The cumulated baseline is multiplied by an exponentiated scalar equal to a row vector of covariates multiplied by a column vector of parameters. Each covariate is one general measurement of what takes place within age-level j.

Under this model, the exponential base raised to the power of one parameter is the cumulated hazard ratio associated with an age level. When the same proportional hazard model holds at each point in time, an exponentiated parameter is also a hazard ratio. This is a common interpretation of complementary log-log parameters.

As mentioned above, Meyer (1990) incorporates gamma-distributed unobserved heterogeneity across individuals into this complementary log-log model. After choosing a particular functional form for the baseline hazard across discrete-time periods at risk and incorporating gamma-distributed unobserved heterogeneity, the survival rate to the end of age-level j among individuals who report the same covariates is

$$\left\{1 + \sigma^2 \times \exp(x_{tc,i} \times \beta_{tc}) \times \exp(x_{tv,i} \times \beta_{tv})' \times J_i\right\}^{-1/\sigma^2}$$

Gamma distributions are typically expressed in terms of location along with shape and scale parameters. With location equal to zero, the shape parameter equal to $1/\sigma^2$, and the scale parameter equal to σ^2, unobserved heterogeneity has a mean of 1 and a variance of σ^2. The variable $x_{tc,i}$ is a row vector of time-constant covariates, and β_{tc} is the associated column vector of parameters. The term $x_{tv,i}$ is a matrix that has a row for each of individual i's time intervals examined, and each column is a covariate that may vary over time. This matrix is multiplied by a column vector of parameters, and the exponential base is applied to each row. Finally, the transposition of this column vector is multiplied by a column vector of ones (J_i). This survival rate is used to construct a likelihood function.

To express the gradient vector and Hessian matrix, the first derivative of survival with respect to the gamma variance parameter is calculated by applying the natural logarithm to the survival function. Furthermore, each parameter in β_{tv} is treated as a scalar within the likelihood program. Overall, the gradient and Hessian expressions can make the maximization process faster and easier to achieve, the gradient and Hessian may be more accurate, and postestimation commands that require an analytic gradient vector can be used.

3 Estimation, computer time, and parameter estimate differences

To examine the speed at which the likelihood function is maximized, I examine a sample of 6,247 individuals who are associated with 34,756 age-level time intervals. The covariates are listed in table 5, and complementary log-log parameter estimates along with a log-gamma variance of minus one are used as maximum likelihood starting values.

The likelihood function is maximized with Stata 11.1. As mentioned above, a Mata gf2 maximum likelihood program that requires gradient vector and Hessian matrix expressions is used. With a Dell 32-bit computer with Windows Vista,[3] fitting a standard complementary log-log model and then maximizing the likelihood function with the gf2 program takes approximately two minutes. With this same computer and sample, the pgmhaz8 command takes approximately 21.5 minutes to maximize.

When the likelihood function is maximized with gf1 rather than gf2, the Hessian matrix expression is not used and estimation takes approximately 39 minutes. Maximization with gf0 involves just the likelihood function, does not use the gradient or Hessian expressions, and takes approximately 6.5 hours.

With these different methods and the rounding that takes place, each set of estimates will be slightly different. In particular, the difference between pgmhaz8 and gf2 log likelihood levels is 4e–09, and the mreldif() function is used to compare the parameter vectors and variance–covariance matrices. For the parameter vectors, the largest relative difference is 9e–06, and for the variance–covariance matrix it is 4e–07.

All of these comparisons are made without sampling weights. When pweights are applied, gf2 maximization takes a few more seconds, and according to the mreldif() function, some of the parameter estimate and variance–covariance changes are rather large. In particular, the school district 2 parameter estimate goes from 0.05 without weights to −0.34 with weights, and the log-gamma variance parameter estimate goes from 0.70 without weights to 0.40 with weights. For the variance–covariance matrix, the largest relative difference is 0.06, and this is for the school district 12 parameter estimate variance.

4 Failure variable and covariates

The sample examined is from a 2006–2007 Canadian provincial Youth Smoking Survey of New Brunswick students. Self-reported student information is examined. The dataset is one cross-section that identifies individuals who have tried marijuana and, if so, the age at which this first took place. The first discrete-time period is at 9 years old, and the last possible full discrete-time period in the sample is at 17 years old.

When people are surveyed, they will have experienced their age level for less than 12 months. Birth dates as well as survey dates are not reported. For these reasons, only fully experienced age levels are examined. Individuals who have never tried marijuana are right-censored at the start of the age level during which they are surveyed. The sample potential time periods for each person are [start of 9 years old, end of 9 years old], (end of 9 years old, end of 10 years old], (end of 10 years old, end of 11 years old], . . . , (end of 16 years old, end of 17 years old]. The failure-indicator variable is equal to 0 for all age levels that an individual survives. For those who report failure at 9 years old or younger, the failure-indicator variable is equal to 1 at 9 years old. Among those who fail at older age levels, this variable is equal to 1 at the age level during which the individual tries marijuana for the first time.

[3] This computer contains an Intel Core 2 E8400 processor at 3.00 GHz with 4.00 GB of RAM.

The short form for the natural logarithm of the gamma variance is ln-varg. Reported covariates that do not change as a person grows older are the following:

aboriginal: Equal to 1 if First Nations, Métis, or Inuit and equal to 0 if not an Aboriginal or if there is no response.

parents smoke: Equal to 1 if at least one parent, step-parent, or guardian smokes and equal to 0 otherwise. This is based on current information in the sample and is assumed to be associated with factors that are constant over time for a child.

age relative to other students in a grade at school:

most common age in grade = reference category: Equal to 1 if a person's age is the most common one that people in his grade level report and equal to 0 otherwise. The sample does not report birth or survey dates. When people were surveyed, most of them were in a grade level equal to their age minus five; in this situation, this variable is equal to 1. When a person's age level minus five is not equal to his grade, this variable is equal to 0. This time-constant covariate is constructed under the assumption that students do not fail or skip any grades.

young in grade: Equal to 1 if a person is younger than most other students in his grade and equal to 0 otherwise.

old in grade: Equal to 1 if a person's age is older than most other students in his grade and equal to 0 otherwise.

school district 1 = reference category, 2, 3, . . . , 14: There are 14 school districts in the province of New Brunswick, and an indicator variable is constructed for each one. For confidentiality reasons, each district is arbitrarily assigned a number.

Reported covariates that can or must change as a person grows older are the following:

9 years old = reference category, 10, 11, . . . , 17: Each of these indicator variables identifies an age level.

1997–1998, 1998–1999, 1999–2000, . . . , 2005–2006 = reference category: Each of these indicator variables identifies a calendar year.

first time consuming five or more drinks of alcohol on one occasion:

never 5+ drinks at previous age or younger = reference category: Equal to 1 when a person has never consumed five or more drinks of alcohol on one occasion or first experienced this during the age level under consideration and equal to 0 otherwise.

first 5+ drinks at previous age: Equal to 1 when a person consumed five or more drinks of alcohol on one occasion for the first time during the previous age level and equal to 0 otherwise.

first 5+ drinks before previous age: Equal to 1 when a person consumed five or more drinks of alcohol on one occasion for the first time two or more age levels ago and equal to 0 otherwise.

first time trying a cigarette:

never tried cigarette at previous age or younger = reference category: Equal to 1 if a person has never tried a cigarette (even a few puffs) or first experienced this during the age level under consideration and equal to 0 otherwise.

first cigarette at previous age: Equal to 1 when a person tried a cigarette (even a few puffs) for the first time during the previous age level and equal to 0 otherwise.

first cigarette before previous age: Equal to 1 when a person has tried a cigarette (even a few puffs) for the first time two or more age levels ago and equal to 0 otherwise.

Each individual sampled responded to one of five potential modules of questions. One module surveyed physical activity, one surveyed healthy eating, and the other three focused on the topic of smoking. Certain schools in the province of New Brunswick were surveyed, and for each school the survey was conducted in the primary language of the majority of students, which was either French or English. Two of the modules on smoking report the information required for this article, and the associated sampling weight for each individual is used.

5 Results

The sample provides information on 7,702 males; each person is surveyed just once. After dropping individuals who were born outside of Canada or who are right-censored and do not report their age, there were 6,372 individuals left. For each age level, table 1 reports the number of individuals who had not failed or became right-censored during younger age levels, and hence they all had some chance of reporting failure. This table also reports the calendar year range associated with each age level examined. Finally, the number of uncensored individuals who survived into an age level is summed over the age levels for a total of 35,429.

Table 1. Sample size for each age level

Age level	Associated calendar year range	Number who survived to and were not right-censored before the beginning of the age level
9	1997–2005	6,372
10	1998–2006	6,230
11	1999–2006	6,148
12	2000–2006	5,587
13	2001–2006	4,779
14	2002–2006	3,383
15	2003–2006	1,952
16	2004–2006	836
17	2005–2006	142
sum over age levels:		35,429

Table 2 reports complementary log-log parameter estimates arrived at with just age covariates and sampling weights. Throughout this section, a single star indicates a p-value of less than 5%, two stars indicate less than 1%, and three stars indicate less than 0.1%. Apart from the summary statistics in tables 1 and 3, all the findings were arrived at with sampling weights.

The age parameter estimates in table 2 indicate that some individuals claim to have first tried marijuana at 9 years old or younger; the conditional likelihood of this first happening at age 10 or 11 is slightly lower. This could be due to some people claiming to have tried marijuana at age 9 or younger when in fact this may not be true. After 10 years old, the conditional likelihood of first trying marijuana consistently rises until age 17, which is the final age level examined. With these estimates, the survival rate to the end of age 17 can be predicted with

```
nlcom 1000*exp(-exp(_b[_cons])*                                     ///
       (1+exp(_b[age10])+exp(_b[age11])+exp(_b[age12])+exp(_b[age13])  ///
       +exp(_b[age14])+exp(_b[age15])+exp(_b[age16])+exp(_b[age17])))
```

which is equal to 364 individuals per thousand with a standard error of 22.

Table 2. Descriptive `cloglog`, males

	Parameter estimate	z
constant	−3.842***	−41.41
9 years old	reference category	
10 years old	−0.726***	−4.54
11 years old	−0.311*	−2.08
12 years old	0.645***	5.32
13 years old	1.368***	12.52
14 years old	1.920***	17.95
15 years old	2.215***	20.23
16 years old	2.289***	18.07
17 years old	2.584***	12.20

Note: These estimates are arrived at with just age covariates.

The next estimates are arrived at with both age and calendar year covariates. The sample was collected during a school year that goes from September to June. Given the absence of birth dates, the sample identifies one age level that takes place within an unknown part of the September to June time period. For each school year in the sample, table 3 reports the associated age range along with the number of individuals who have some chance of reporting failure.

An important thing to note is that during the year of 1997–1998, there is only information on 9-year-olds. This means that a 1997–1998 covariate could be associated with a factor that is unique to being 9 years old during this calendar year. It is also the case that at 17 years old, all individuals are experiencing the year of 2005–2006. Similar issues arise with wider time spans, and hence the age covariates could be associated with year covariates and vice versa. Table 4 reports these complementary log-log parameter estimates.

Table 3. Age levels associated with each calendar time range

School year calendar range	Age levels that individuals reported info on within the calendar range	Number who survived to and were not right-censored before the beginning of the age levels that take place within each calendar range
1 September 1997–1 July 1998	9	389
1 September 1998–1 July 1999	9–10	1,491
1 September 1999–1 July 2000	9–11	2,735
1 September 2000–1 July 2001	9–12	3,994
1 September 2001–1 July 2002	9–13	5,061
1 September 2002–1 July 2003	9–14	5,530
1 September 2003–1 July 2004	9–15	5,767
1 September 2004–1 July 2005	9–16	5,458
1 September 2005–1 July 2006	10–17	5,004
	sum over each calendar range:	35,429

Table 4. Descriptive `cloglog`, males

	Parameter estimate	z		Parameter estimate	z
constant	−4.922***	−31.25	1997–1998	2.059***	7.53
9 years old	reference category		1998–1999	1.417***	6.18
10 years old	−0.431*	−2.54	1999–2000	1.428***	7.83
11 years old	0.266	1.64	2000–2001	0.976***	6.17
12 years old	1.447***	9.54	2001–2002	0.551***	4.47
13 years old	2.310***	15.16	2002–2003	0.297**	2.97
14 years old	2.946***	18.89	2003–2004	0.102	1.15
15 years old	3.290***	20.31	2004–2005	−0.017	−0.23
16 years old	3.373***	18.99	2005–2006	reference category	
17 years old	3.663***	14.85			

Note: These estimates are arrived at with just age and calendar time covariates.

With these estimates, the survival rate to the end of age 17 among those who were 17 in 2005–2006 can be predicted with

```
nlcom 1000*exp(-exp(_b[_cons])*(exp(_b[y1997_98])                ///
        +exp(_b[age10]+_b[y1998_99])+exp(_b[age11]+_b[y1999_00]) ///
        +exp(_b[age12]+_b[y2000_01])+exp(_b[age13]+_b[y2001_02]) ///
        +exp(_b[age14]+_b[y2002_03])+exp(_b[age15]+_b[y2003_04]) ///
        +exp(_b[age16]+_b[y2004_05])+exp(_b[age17])))
```

which is equal to 295 per thousand with a standard error of 20.

Finally, the log-gamma variance parameter and all the covariates are included. All the findings thus far are based on 6,372 individuals. Because of missing covariate information, a few observations have been deleted; hence, what follows is based on 6,247 males. Among these males, the number of uncensored individuals who survived into an age level, summed over age levels, is 35,429.

The log-gamma variance estimate is 0.40, which implies a gamma variance of 1.49. The Stata short-form for this parameter estimate is `_b[ln_gvar:_cons]`. The command

```
testnl exp([ln_gvar]:_cons)=0
```

along with the resulting *p*-value divided by two for a one-tailed test conveys statistical significance at 0.1%. To interpret this variance estimate, consider a hazard ratio between two individuals who are the same except that the unobserved term takes on a particular value for the individual in the numerator and is at the normalized level of 1 for the other person. The distribution of this hazard ratio can be plotted with the following command:

```
twoway function y = gammaden(1/exp(_b[ln_gvar:_cons]),exp(_b[ln_gvar:_cons]),0,x)
```

Furthermore, the percentage of individuals predicted to have a level of the unobserved heterogeneity term at 0.5 or less can be identified with

```
display gammap(1/exp(_b[ln_gvar:_cons]),0.5/exp(_b[ln_gvar:_cons]))
```

These commands are used to plot the density function in figure 1. The log-gamma estimate implies that 47% of individuals have a degree of unobserved heterogeneity that is half the mean of 1 or less, and 66% have a degree of unobserved heterogeneity of 1 or less. In contrast, the variance estimate implies that 15% have a degree of unobserved heterogeneity that is more than twice the mean. This suggests that there are a small number of people who for reasons not reported are highly likely to try marijuana for the first time.

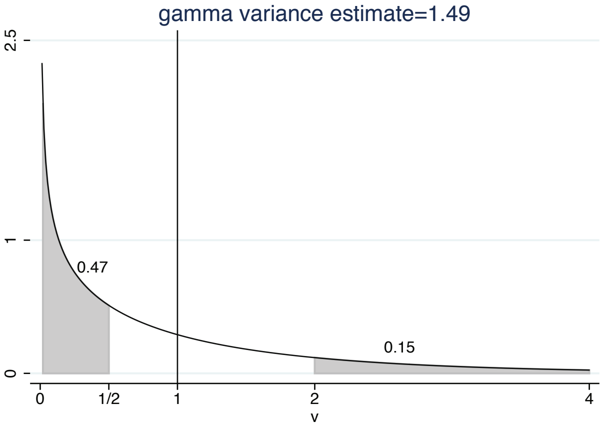

Figure 1. Gamma variance estimate

For each covariate, the hazard ratio associated with a 1-unit increase in one variable is reported in table 5. A hazard ratio of 1 implies that the associated parameter is equal to 0, and hence the z statistic associated with the null hypothesis that the parameter is equal to 0 is also reported. The intercept estimate is -5.03, and the associated z statistic is -23.36.

Table 5. `pgmhaz` males

Time constant	Hazard ratio	$z^\dagger$	Time varying	Hazard ratio	$z^\dagger$
aboriginal	1.92***	3.67	9 years old	reference category	
not aboriginal	reference category		10 years old	0.46***	−4.47
parents smoke	1.94***	7.19	11 years old	0.88	−0.75
parents			12 years old	2.79***	6.12
don't smoke	reference category		13 years old	6.33***	9.99
most common			14 years old	12.83***	11.79
age in grade	reference category		15 years old	20.92***	11.50
young in grade	1.17	0.67	16 years old	27.01***	9.81
old in grade	1.31**	2.72	17 years old	33.55***	7.86
school district 1	reference category		1997–1998	4.15***	4.24
school district 2	0.71	−1.36	1999–1999	2.92***	4.03
school district 3	0.83	−1.07	1999–2000	3.04***	4.89
school district 4	0.37***	−4.03	2000–2001	2.01***	3.48
school district 5	1.20	1.45	2001–2002	1.42*	2.14
school district 6	0.81	−1.67	2002–2003	1.12	0.84
school district 7	1.34	1.33	2003–2004	1.01	0.13
school district 8	0.88	−0.83	2004–2005	0.94	−0.72
school district 9	0.70	−1.88	2005–2006	reference category	
school district 10	0.90	−0.63			
school district 11	0.66**	−2.67			
school district 12	0.21***	−3.91			
school district 13	1.24	1.14			

Time varying	Hazard ratio	$z^\dagger$
never 5+ drinks at previous age or younger	reference category	
first 5+ drinks at previous age	4.40***	10.69
first 5+ drinks before previous age	5.48***	7.10
never tried cigarette at previous age or younger	reference category	
first cigarette at previous age	5.77***	12.51
first cigarette before previous age	5.60***	10.37

† Each z statistic is based on the null hypothesis that the associated parameter is equal to 0, which implies a hazard ratio of 1.

For reasons discussed above, each time-varying covariate parameter is a scalar within the likelihood program. Hence, the age 10 parameter estimate is called `_b[age10:_cons]` and the 1997–1998 parameter estimate is called `_b[y1997_98:_cons]`. The age and calendar year parameter estimates along with all other variables set equal to 0 imply the following survival rate to the end of 17 years old among those who were 9 years old in 1997–1998:

```
nlcom 1000*(1+exp(_b[ln_gvar:_cons]+_b[fmari_d:_cons])          ///
        *(exp(_b[y1997_98:_cons])                               ///
        +exp(_b[age10:_cons]+_b[y1998_99:_cons])                ///
        +exp(_b[age11:_cons]+_b[y1999_00:_cons])                ///
        +exp(_b[age12:_cons]+_b[y2000_01:_cons])                ///
        +exp(_b[age13:_cons]+_b[y2001_02:_cons])                ///
        +exp(_b[age14:_cons]+_b[y2002_03:_cons])                ///
        +exp(_b[age15:_cons]+_b[y2003_04:_cons])                ///
        +exp(_b[age16:_cons]+_b[y2004_05:_cons])                ///
        +exp(_b[age17:_cons])))^(-exp(-_b[ln_gvar:_cons]))
```

This survival rate is 600 per thousand with a standard error of 31.

The reference categories other than age and calendar time are the following: not aboriginal, no parent smokes, age level is the most common in grade at school, attends a school in district 1, have not consumed five or more drinks on one occasion for the first time during the previous age or earlier, and have not tried a cigarette for the first time during the previous age or earlier. Each of the time-constant covariates is separately changed from 0 to 1, and table 6 reports the associated survival rates to the end of being 17 years old. In this table, the school district survival rates are listed from highest to lowest. Among all the time constant variables, the survival rate falls to the lowest level when the covariate indicating that a parent smokes goes from 0 to 1; changing the aboriginal covariate from 0 to 1 has a similar effect.

Table 6. Survival rate to the end of 17 years old per thousand

	Survival rate	Std. error		Survival rate	Std. error
reference categories	600	31			
aboriginal	458	44	school district 2	671	58
parents smoke	457	31	school district 6	644	29
young in grade	566	59	school district 3	640	40
old in grade	541	30	school district 8	626	35
school district 12	867	44	school district 10	622	39
school district 4	790	41	school district 5	560	31
school district 11	687	33	school district 13	553	44
school district 9	675	43	school district 7	536	53

The reference categories are not aboriginal, parents don't smoke, most common age level in grade, school district 1, never 5+ drinks at previous age or younger, and never tried cigarette at previous age or younger.

The remaining results involve previous alcohol and cigarette consumption. When individuals consume five or more drinks of alcohol on one occasion for the first time at age 16, the survival rate to the end of 17 years old falls from 600 per thousand to 453 per thousand (standard error = 42). When this volume of drinks is first consumed at age 15, the survival rate falls to 367 per thousand (standard error = 38), and when such drinking takes place at age 14, it falls to 322 per thousand (standard error = 33).

Similarly, when individuals first try a cigarette at age 16, the survival rate to the end of 17 years old falls from 600 per thousand to 415 per thousand (standard error = 43); when first trying a cigarette at age 15, the survival rate falls to 348 per thousand (standard error = 33); and with a first cigarette at 14, it falls to 308 per thousand (standard error = 29). These previous choices appear to have the strongest relationship with the hazard of trying marijuana for the first time. In the future, it will be valuable to examine what motivates people to make each of these choices and to compare choices among males and females.

6 Summary

Expressing the gradient vector and Hessian matrix within a computer program increases the speed at which a complementary log-log model with gamma-distributed unobserved homogeneity can be fit. Furthermore, one can use estimation options and postestimation commands that require the expression of a gradient vector.

The findings suggest that there are a small number of individuals who are very likely to first try marijuana because of unobserved factors. The findings also suggest that if a young male has not yet tried marijuana, the hazard of such initiation is rather high after he first consumes a large amount of alcohol on one occasion or tries a cigarette for the first time. In the future, it may be informative to examine the most common order of marijuana, cigarette, and alcohol initiation, and the time between each of these choices among both males and females.

7 Acknowledgments

This research was in part made possible and supported by the Health & Education Research Group at the University of New Brunswick. The coexecutive director, Dr. Bill Morrison, was particularly patient and helpful. Financial support was also received from the Canadian Institute for Health Research.

While developing this computer program and when writing generally about the duration literature, advice from professor Stephen P. Jenkins was essential, and his comments were very helpful. Isabel Canette, who has a PhD in statistics/data analysis and is a senior statistician with Stata, was also very helpful in answering questions I had while working on this. I would also like to thank professors Ted McDonald, Mehmet Dalkir, and David Haardt for their advice and encouragement.

8 References

Abbring, J. H., and G. J. van den Berg. 2007. The unobserved heterogeneity distribution in duration analysis. *Biometrika* 94: 87–99.

Heckman, J., and B. Singer. 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 52: 271–320.

Jenkins, S. P. 1995. Easy estimation methods for discrete-time duration models. *Oxford Bulletin of Economics and Statistics* 57: 129–136.

———. 2004a. hshaz: Stata module to estimate discrete time (grouped data) proportional hazards models. Statistical Software Components S444601, Department of Economics, Boston College. http://ideas.repec.org/c/boc/bocode/s444601.html.

———. 2004b. pgmhaz8: Stata module to estimate discrete time (grouped data) proportional hazards models. Statistical Software Components S438501, Department of Economics, Boston College. http://ideas.repec.org/c/boc/bocode/s438501.html.

———. 2008. Survival analysis with Stata. University of Essex module EC968. http://www.iser.essex.ac.uk/resources/survival-analysis-with-stata-module-ec968.

Meyer, B. D. 1990. Unemployment insurance and unemployment spells. *Econometrica* 58: 757–782.

Ondrich, J., and S. E. Rhody. 1999. Multiple spells in the Prentice–Gloeckler–Meyer likelihood with unobserved heterogeneity. *Economics Letters* 63: 139–144.

Prentice, R. L., and L. A. Gloeckler. 1978. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* 34: 57–67.

Sueyoshi, G. T. 1995. A class of binary response models for grouped duration data. *Journal of Applied Econometrics* 10: 411–431.

About the author

Michael G. Farnworth is an associate professor in the Department of Economics at the University of New Brunswick.