UNIVERSITY OF ESSEX
INSTITUTE FOR SOCIAL AND ECONOMIC RESEARCH
Professor Stephen P. Jenkins <stephenj@essex.ac.uk>

Essex Summer School course 'Survival Analysis'
and
EC968. Part II: Introduction to the analysis of spell duration data

Lesson 5. Estimation: (i) continuous time models (parametric and Cox)

Contents

1 Aim

The aim of this lesson is to illustrate how to use Stata to estimate multivariate continuous time survival time models. These include the parametric models (with hazard functions of the type discussed in Lesson 2) and the semi-parametric Cox model.

2 Introduction

Stata provides an extensive suite of estimators. Parametric regression survival-time models (including the piece-wise constant exponential model) are estimated by maximum likelihood using **streg**. Models corresponding to six types of parametric distribution can be estimated: Exponential, Weibull, Log-logistic, Gompertz, Lognormal, and Generalised Gamma. We will focus on the first three (discussed in Lesson 2). Cox's partial likelihood model (the 'Cox model') is estimated using **stcox**.

To use these programs, you must **stset** the data first, as discussed in Lesson 3.

I discuss and illustrate **streg** and **stcox** in turn, using the Cancer data set assumed to be **stset** already. At the end I ask you, as an exercise, to repeat parts of the analysis with alternative models or with different data sets.

Note that typing **streg** by itself after estimating a model with **streg**, or typing **stcox** by itself after estimation with **stcox**, will result in the previous estimates being replayed on the screen.

3 Estimation using streg (and plotting fitted curves with stcurv)

The different parametric models estimated by **streg** share a common command syntax – the different distributions are chosen via option specifications. See **help streg** for the full command syntax and all the options available. We will ignore the **frailty(.)** option the moment. Frailty (unobserved heterogeneity) models are considered separately in Lesson 7.

The basic syntax is

```
streg [varlist], dist(distname) nohr time tr nolog
```

dist(distname) specifies the survival model to be estimated. **distname** is one of the following: **exponential**, **weibull**, **gompertz**, **lognormal**, **loglogistic** or **gamma**. Abbreviations are allowed (to the minimum, as underlined).

As Stata's on-line help says (this is text modified from Stata version 7, which still applies):

```
'nohr' specifies that coefficients rather than exponentiated coefficients are to be displayed
       or, said differently, coefficients rather than hazard ratios.  This option is valid
       only for models with a proportional hazard ratio parameterization: exponential,
       Weibull, and Gompertz.

'hr', which can be specified when the model is estimated or when redisplaying results, states
       that the underlying log relative hazard coefficients are to be displayed.  This option
       affects only how results are displayed, not how they are estimated.

'time' specifies that the model is to be estimated in the accelerated failure-time metric
       rather than the log relative-hazard metric.  This option is only valid for the
       exponential and Weibull models since they have both a hazard ratio and an accelerated
       failure-time parameterization.  For these two models, in the log relative-hazard
       metric, estimates of (B,s) are produced and in the accelerated failure-time metric,
       estimates of (-B*s,s) are produced.
       Regardless of metric, the likelihood function is the same and models are equally
       appropriate viewed in either metric; it is just a matter of changing interpretation.
       'time' must be specified when the model is estimated.

'tr' is appropriate only for the log-logistic, lognormal, and gamma models, or for the
       exponential and Weibull models when estimated in the log expected time metric.  'tr'
       specifies that exponentiated coefficients are to be displayed, which have the
       interpretation of time ratios.  'tr' may be specified when the model is estimated or
       when results are redisplayed

'nolog' prevents streg from showing the iteration log.
```

stcurve and **predict** are commands used after having run **streg**. See below.

Recall that for models which can be written in the *proportional hazards* (PH) metric, the hazard function for person *i* can be written

$$h_i(t, X_i) = h_0(t).\lambda_i, \quad \text{where } \lambda_i \equiv \exp(\beta' X_i), \text{ or}$$

$$\log[h_i(t\, X_i)] = \log[h_0(t)] + X_i\beta$$

where $h_0(t)$ is the baseline hazard, X_i is a vector of individual characteristics, and β is a vector regression coefficients and includes an intercept term. In a PH model, λ_i scales the baseline hazard multiplicatively by the same amount at each value of *t*.

For PH models Stata reports estimates for covariate *k* of either β_k (use the **nohr** option) or of the 'hazard ratio', $\exp(\beta_k)$, for which you use the **hr** option. The PH form is referred to as the 'log relative hazard' in Stata output.

Models which can be written in the *accelerated failure time* (AFT) metric are of the form:

$$\ln(t_i) = X_i\beta^* + z_i \text{ , or}$$

$$\ln(t_i\psi_i) = z_i \text{ , or}$$

$$t_i = \psi_i\exp(z_i)$$

where $\psi_i \equiv \exp(-\beta^{*\prime} X_i)$ and $z_i = \sigma u_i$ is a generalised error term (u_i is an error term, and σ is a scale factor).

The ψ_i is a survival time scaling factor: values of $\psi_i > 1$ accelerate failure (reduce survival time) whereas values of $\psi_i < 1$ decelerate failure (increase survival time).

For AFT models Stata reports estimates for covariate *k* of either β_k^* (the default) or of $\exp(\beta_k^*)$, for which you use the **tr** (time ratio) option.

The Weibull and Exponential models are the only ones which are both PH and AFT. In this case, the relationship between the PH and AFT representations is

$$\beta^* = -\sigma\beta$$

where $\sigma = 1/\alpha$ and α is the Weibull parameter (see Lesson 2). Stata refers to α as '*p*'.

The Gompertz model is PH only. The lognormal, log-logistic, and generalised Gamma models are AFT only.

The illustrations in this Lesson focus on the Weibull and Log-logistic models, though the Exercises encourage you to explore other specifications.

The empirical illustration uses the Cancer data set, which has already been **stset**. Recall that there are two variables in the data set which are available to be used as covariates: age and drug. I shall recode the drug variable from three categories into a simpler binary variable summarising whether the subjects receive the drug or not.

```
. recode drug 1=0 2/3=1
(drug: 48 changes made)
. lab var drug "receives drug?"
. lab def drug 0 "placebo" 1 "drug"
. lab val drug drug
```

4 Estimation of the Weibull model

The Weibull model estimates are:

```
. streg drug age, dist(weibull) nolog nohr

         failure _d:  died
   analysis time _t:  studytim


Weibull regression -- log relative-hazard form

No. of subjects =           48                     Number of obs   =        48
No. of failures =           31
Time at risk    =          744
                                                   LR chi2(2)      =     35.39
Log likelihood  =   -42.931335                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      _t |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    drug |  -2.196936   .4087791     -5.374   0.000      -2.998129   -1.395744
     age |   .1202027   .0371599      3.235   0.001       .0473707    .1930348
   _cons |  -10.58396   2.326271     -4.550   0.000      -15.14337   -6.024553
---------+--------------------------------------------------------------------
   /ln_p |   .5204297   .1389037      3.747   0.000       .2481834     .792676
------------------------------------------------------------------------------
       p |   1.682751   .2337403                          1.281695    2.209301
     1/p |   .5942651   .0825456                           .452632    .7802168
------------------------------------------------------------------------------
```

The **nohr** option meant that coefficient estimates were shown. We can show the corresponding hazard ratio estimates by simple replaying the command and adding the **hr** option (if we had wished, instead we could have used the **nohr** option and replayed using the **hr** option). Interpretation of the estimates follows the display of the estimates in hazard ratio form.

```
. streg, hr

Weibull regression -- log relative-hazard form

No. of subjects =           48                     Number of obs   =        48
No. of failures =           31
Time at risk    =          744
                                                   LR chi2(2)      =     35.39
Log likelihood  =   -42.931335                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      _t | Haz. Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    drug |   .1111431    .045433     -5.374   0.000       .0498803    .2476487
     age |   1.127725   .0419062      3.235   0.001       1.048511    1.212925
---------+--------------------------------------------------------------------
   /ln_p |   .5204297   .1389037      3.747   0.000       .2481834     .792676
------------------------------------------------------------------------------
       p |   1.682751   .2337403                          1.281695    2.209301
     1/p |   .5942651   .0825456                           .452632    .7802168
------------------------------------------------------------------------------
```

The estimates suggest that the hazard rate is increasing over time at a decreasing rate: note that $1 < p < 2$ (see Lesson 2). In the Weibull model, the ratio of the hazard rate at survival time t to the hazard rate at survival time u, given the same X, is given by $(t/u)^{\alpha-1}$. Thus according to our model estimates, the ratio of the hazard rate at time 10 to that at time 5 is 1.6, and the ratio of the hazard rate at time 30 to that at time 5 is 3.4:

```
. di "h(10,X)/h(5,X) =  "  (10/5)^(e(aux_p)-1)
h(10,X)/h(5,X) =  1.6051972

. di "h(30,X)/h(5,X) =  "  (30/5)^(e(aux_p)-1)
h(30,X)/h(5,X) =  3.3984681
```

The coefficient estimates indicate that those receiving the drug have lower hazard rates *ceteris paribus* (i.e. lower conditional death rates and hence longer survival times). Note the negative (and statistically significant) coefficient for drug in the **nohr** representation and the hazard ratio for drug less that one in the **hr** representation: 0.11 = exp(–2.2). The estimates imply that, at each survival time, the hazard rate for those who received the drug is only 11% of the hazard rate for those who received the placebo. The output also shows that there is a positive association between age and the hazard rate: older people die earlier. In fact a one year rise in age is associated with a 13% higher hazard rate.

The elasticity of the hazard rate with respect to a one unit change in the kth explanatory variable is given by $\beta_k X_{ik}$; for age, it is therefore (0.1202027)*age$_i$. (If the explanatory variable had instead been ln(age) rather than age, the estimated coefficient on ln(age) would be the elasticity of the hazard with respect to age.) Here are the elasticities:

```
. * Elasticity of hazard w.r.t. age (age covariate in levels) = b_age * age
. ge elas_age = _b[age]*age
. su elas_age, detail

                          elas_age
-------------------------------------------------------------
      Percentiles      Smallest
 1%     5.649528      5.649528
 5%     5.769731      5.769731
10%     5.889934      5.769731       Obs                  48
25%     6.070238      5.889934       Sum of Wgt.          48

50%     6.731353                     Mean           6.716327
                       Largest       Std. Dev.      .6802519
75%     7.212164      7.813178
90%     7.813178      8.053583       Variance       .4627426
95%     8.053583      8.053583       Skewness       .3161068
99%     8.053583      8.053583       Kurtosis       2.125197
```

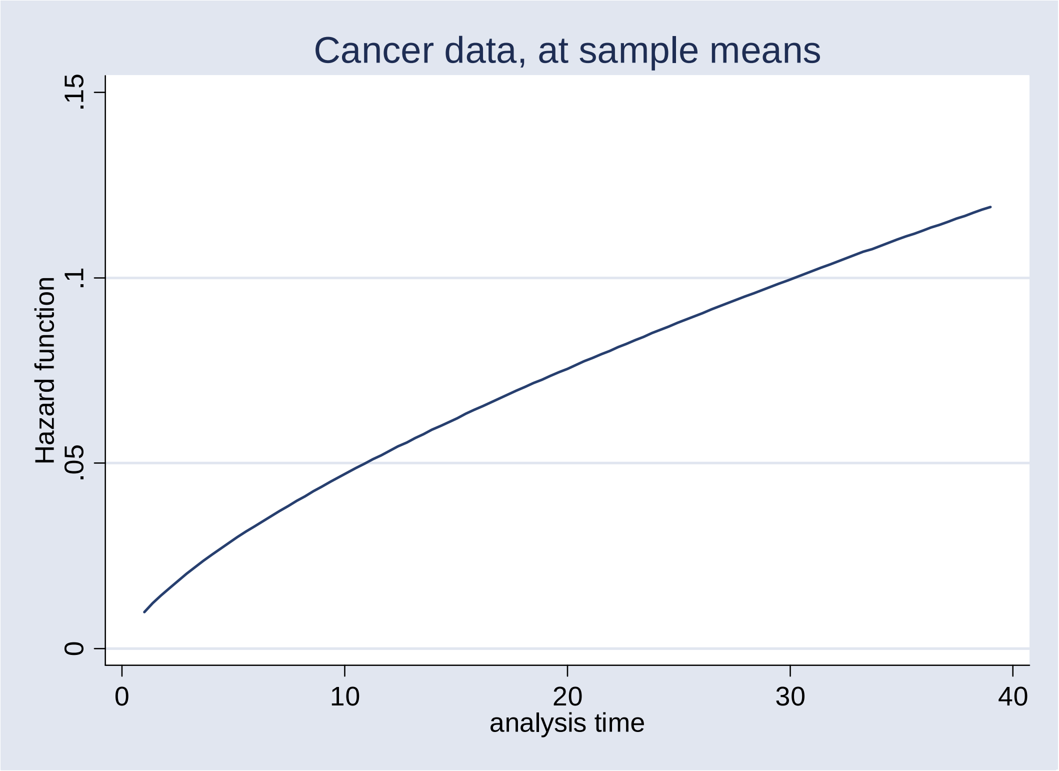
Observe the way in which we can retrieve and refer to the estimated model coefficients: **_b[*something*]** refers to the estimated coefficient on the explanatory variable with name *something* in the last regression that was run. (One can also refer to many other estimates after running regressions, including estimated standard errors, log-likelihood values, and so on: see the User's Guide.)

More generally, hazard rate ratios at each survival time are related to absolute differences in characteristics: $h(t,X_1)/h(t,X_2) = \exp[\beta'(X_1–X_2)]$. Thus a ten year difference in age, other things equal, is associated with a hazard rate ratio of some 3.3. Some one aged y+10 and who is receiving the drug has a hazard ratio that is 37% of some one aged y who gets the placebo:

```
. di "h(t;age=y+10,drug=x)/h(t;age=y,drug=x) = "  exp(_b[age]*10)
h(t;age=y+10,drug=x)/h(t;age=y,drug=x) = 3.3268546

. di "h(t;age=y+10,drug=1)/h(t;age=y,drug=0) = "  exp(_b[age]*10 + _b[drug])
h(t;age=y+10,drug=1)/h(t;age=y,drug=0) = .36975709
```

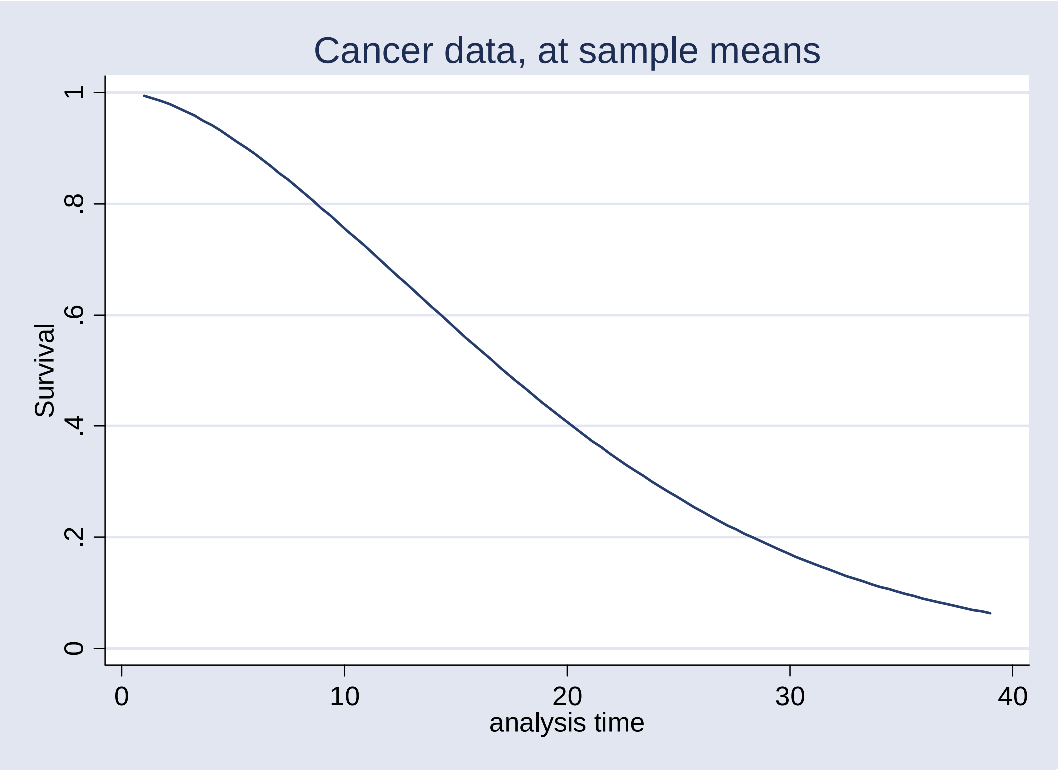
Let us now look at the estimated hazard and survivor functions graphically. We can do this using **stcurv**, run after **streg**. For example:

```
. stcurv, hazard title("Cancer data, at sample means") ///
>         saving(streg2,replace)
```



Note the monotonically rising hazard. The corresponding survival curve is as follows

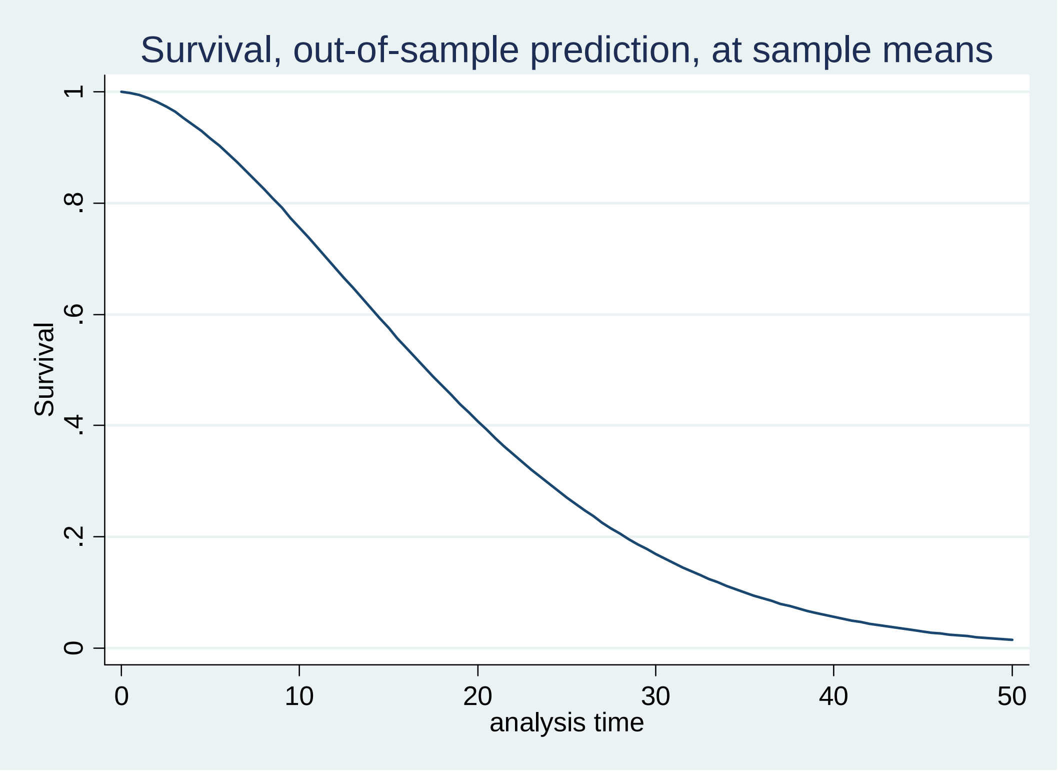
```
. stcurv, survival title("Cancer data, at sample means") /*
>         */ saving(streg1,replace)
```

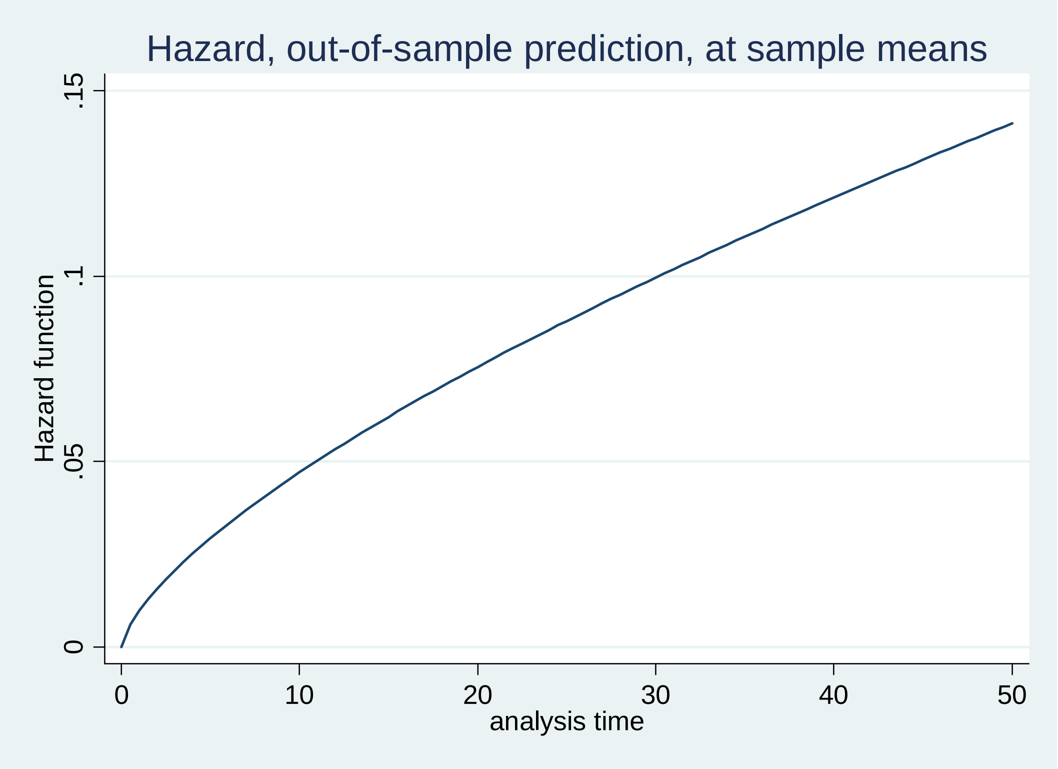


The median survival time is about 17.

Observe that **stcurv** can also be used for out-of-sample projections, i.e. showing what the estimated functions look like at survival times beyond the range that exists in the estimation sample. To do this use the **range(# #)** option to **stcurve**. Here's what the previous hazard and survival curves look like if the analysis time axis is extended to 50.

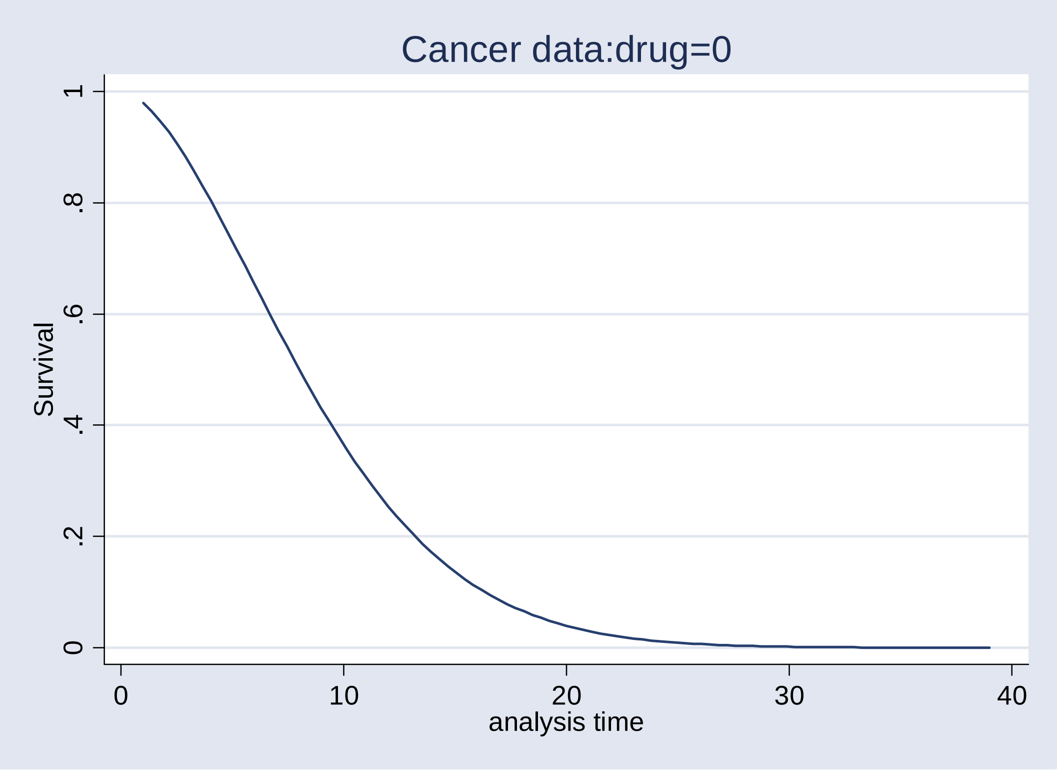
```
stcurv, survival title("Survival, out-of-sample prediction, at sample means") ///
        saving(streg1a,replace) range(0 50)
```



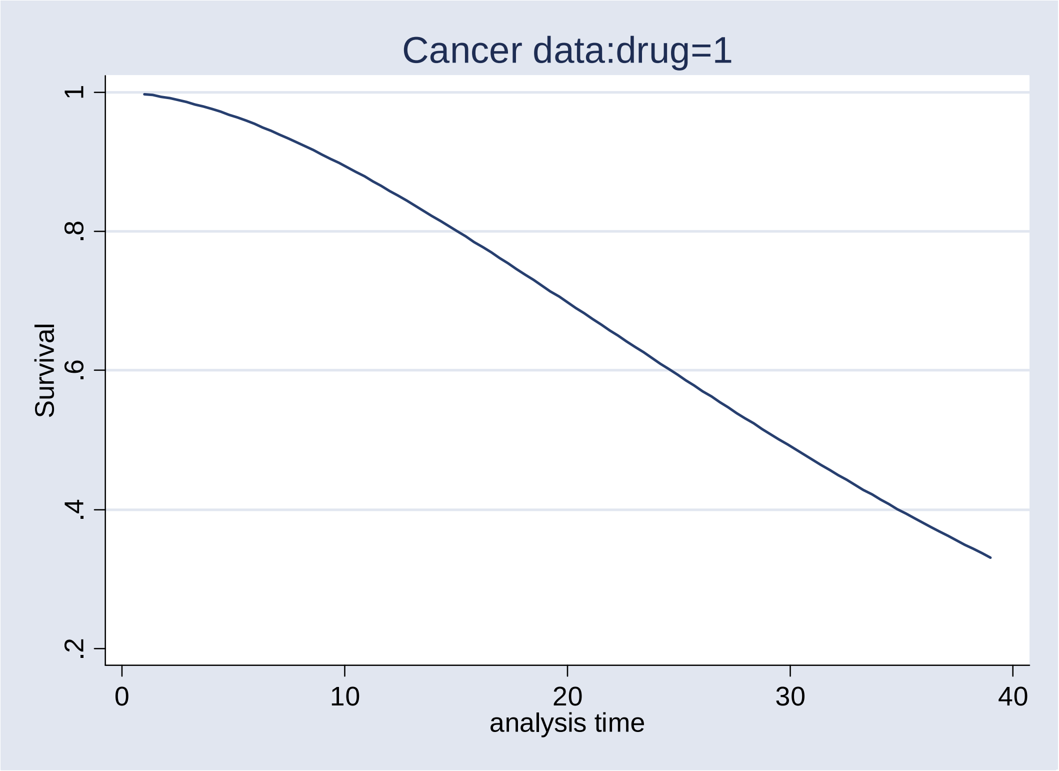


The last four graphs were, by default, drawn with the covariates set at their mean values. This does not make a lot of intuitive sense for categorical covariates. Compare instead the survivor curves for persons with drug = 0 and drug = 1 (and mean age), making use of the **at(.)** option.

```
. . stcurv, survival title("Cancer data:drug=0") at(drug=0) ///
>         saving(streg5,replace)
```

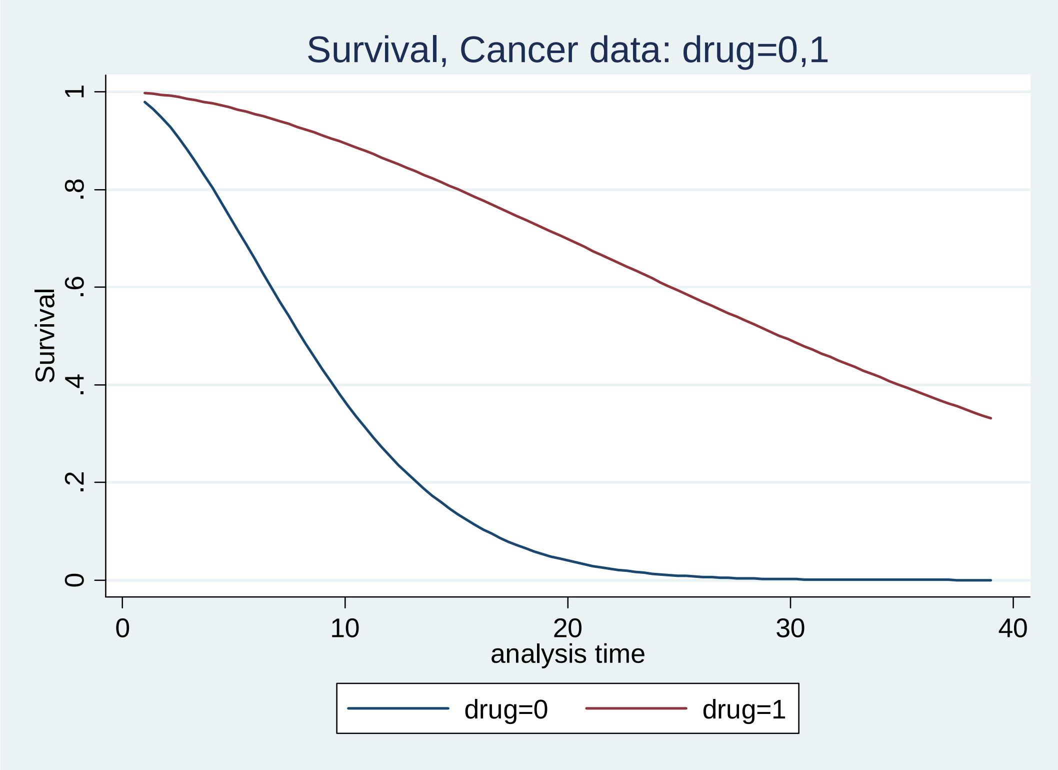


```
. stcurv, survival title("Cancer data:drug=1") at(drug=1) ///
>         saving(streg6,replace)
```



Clearly survival times are much lower for placebo recipients. This can be seen even more clearly if we take advantage of the fact that **stcurv** allows us to use multiple **at#(.)** options in order to draw several lines on one graph:

```
stcurv, survival title("Survival, Cancer data: drug=0,1") at(drug=0) at1(drug=1) ///
        saving(streg5a,replace)
```



Let us now *calculate* the median and mean survival times, using the formulae discussed in Lesson 2. To do this we have to specify the values of the covariate vector (*X*), and thence can derive λ_i. (In Chapter 2, we simulated values for a particular value of λ.)

The code below shows first how to calculate the Weibull median and mean for the case when the covariates are set at the sample average values. Second it shows how to calculate the median and mean for each person in the sample – we can then examine the values for particular covariate combinations using **list**.

In both cases, the derivations use the **predict** command after **streg**. The **xb** option with **predict** generates a new variable equal to the estimate $\beta'X_i$ for each person *i*. The other calculations also use other automatically saved results, such as the mean after a **summarize**, **r(mean)**, and the estimate of the Weibull shape parameter *p* as **e(aux_p)**. I have also used a couple of local macros to hold scalar results to use in other calculations (see **help macro**). In fact, the calculation of mean and mean medians can be done directly using **predict**.

[Note: after all '**e**stimation class' commands, examples of which are mostly regression commands (including **streg**, **logit**, **cloglog**), Stata saves a variety of results in objects with names **e(*something*)**. You can find the full list of saved results by typing **ereturn list** after an estimation command. Examples include **e(b)** which is a vector containing the parameter estimates, and **e(V)** which is a matrix containing the variance-covariance matrix of the parameter estimates. Different commands save extra results relevant to their model; e.g. after a Weibull regression, **e(aux_p)** contains the shape parameter α. Results are also saved after commands like **summarize** and **tabulate** in objects with names like **r(*something*)**. You can find the full list of saved results by typing **return list** after one of these 'rclass' commands. E.g. after a **summarize**, **r(mean)** contains the estimate of the mean. For a more complete discussion of saved results, see the Manuals. Finally, observe that virtually all estimation commands may be followed with a **predict** command that generates predictions for the observations in memory, based on the parameter estimates of the most recent model. The sorts of things that one can predict depends, of course, on the command. See the Manual entries for the relevant command about **predict** for that command.]

First is the code for predictions for the case when the covariates are set at the sample average values. The trick here is to note that our calculations require the value $\beta' X_m$ where X_m is a vector containing the sample mean values of the characteristics. Instead of first calculating X_m and then $\beta' X_m$, we take advantage of the fact that $\beta' X_m$ is equal to the mean of the individual $\beta' X_i$ for each subject *i* in the sample. But I know that **predict** will produce the $\beta' X_i$ so all we have to do is generate that and take its mean. Then we can feed the result into our calculation of the mean and median spell lengths.

```
. predict xb, xb

. su xb

Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
      xb |      48   -5.149179   1.303897  -7.131369  -2.530378

. di "Pred. Median [at sample mean X] = "  (ln(2)*exp(-r(mean)))^(1/e(aux_p))
Pred. Median [at sample mean X] = 17.15298

. di "Pred. Mean [at sample mean X] = " exp(-r(mean)/e(aux_p))*exp(lngamma(1+1/
> e(aux_p)))
Pred. Mean [at sample mean X] = 19.042575
```

Now, second, we examine how to calculate the estimated mean and median survival time for every person in the sample.

```
. * median duration for each person in sample
. ge mediand = (ln(2)*exp(-xb))^(1/e(aux_p))

. * expected (mean) duration for each person in sample
. ge meand = exp(-xb/e(aux_p))*exp(lngamma(1+1/e(aux_p)))
```

In fact, Stata allows you to calculate these variables directly, using **predict** after **streg**, rather than calculating them by hand. Here's how:

```
. predict mediandS, median time
. predict meandS, mean time
```

Let's confirm that we get the same results for both methods of derivation:

```
. su mediand mediandS meand meandS

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     mediand |        48    22.36656     15.3532   3.617957   55.70768
    mediandS |        48    22.36656     15.3532   3.617957   55.70768
       meand |        48    24.83049    17.04453   4.016515   61.84451
      meandS |        48    24.83049    17.04453   4.016516   61.84451
```

Now let's look and see how the estimated means and median differ for individuals with different characteristics. Among those aged 50–60 years, we compare those who received the drug with those who received the placebo.

```
. sort age drug

. list id age drug mediand meand if age >50 & age <60 & drug ==0 , noobs

  +---------------------------------------+
  | id   age   drug    mediand      meand |
  |---------------------------------------|
  | 17    51      0   11.34556    12.5954 |
  |  4    52      0   10.56339   11.72706 |
  | 11    52      0   10.56339   11.72706 |
  | 20    52      0   10.56339   11.72706 |
  | 14    55      0   8.525801   9.465014 |
  |---------------------------------------|
  |  9    56      0   7.938026    8.81249 |
  |  5    56      0   7.938026    8.81249 |
  | 19    57      0   7.390773    8.20495 |
  |  8    58      0   6.881248   7.639295 |
  | 10    58      0   6.881248   7.639295 |
  |---------------------------------------|
  |  3    59      0    6.40685   7.112638 |
  +---------------------------------------+

. list id age drug mediand meand if age >50 & age <60 & drug ==1 , noobs

  +---------------------------------------+
  | id   age   drug    mediand      meand |
  |---------------------------------------|
  | 34    52      1   38.97641    43.2701 |
  | 48    52      1   38.97641    43.2701 |
  | 36    54      1   33.78753   37.50961 |
  | 41    55      1    31.4582   34.92368 |
  | 35    55      1    31.4582   34.92368 |
  |---------------------------------------|
  | 31    55      1    31.4582   34.92368 |
  | 24    56      1   29.28944   32.51601 |
  | 44    56      1   29.28944   32.51601 |
  | 42    57      1   27.27021   30.27434 |
  | 23    58      1   25.39019   28.18721 |
  |---------------------------------------|
  | 32    58      1   25.39019   28.18721 |
  | 39    58      1   25.39019   28.18721 |
  +---------------------------------------+
```

Finally, let's compare the estimated means and medians for two (hypothetical) persons, call them *i* and *j*, each of which has an age equal to the sample mean age, but one received the drug and the other didn't. (These comparisons parallel those that we undertook in Lesson 2.) First we drop the previous variables, then we find the mean age using **summarize** and place its value into a **local** macro that we can refer to later. The next steps compute $\beta'X_i$ and $\beta'X_j$ for the two individuals *i* and *j*, and then finally we substitute these values into the formula for the mean and median for the Weibull model.

```
. drop xb mediand meand

.
. su age

Variable |     Obs        Mean    Std. Dev.       Min        Max
---------+-----------------------------------------------------
     age |      48      55.875    5.659205         47         67

. local meana = r(mean)
```

We now have the mean age. Now follows the calculations for the placebo recipient.

```
. local xb0 = _b[_cons] + _b[age]*`meana' + _b[drug]*0

. di "Mean age = " `meana' " ,_b[_cons] + _b[age]*(mean age) + _b[drug]*0 = " `
> xb0'
Mean age = 55.875 ,_b[_cons] + _b[age]*(mean age) + _b[drug]*0 = -3.8676331

. di "Pred. Median [mean(age), drug=0] = "  (ln(2)*exp(-`xb0'))^(1/e(aux_p))
Pred. Median [mean(age), drug=0] = 8.0092224

. di "Pred. Mean [mean(age),drug=0] = " exp(-`xb0'/e(aux_p))*exp(lngamma(1+1/e(
> aux_p)))
Pred. Mean [mean(age),drug=0] = 8.8915291
```

Here are the calculations for the drug recipient.

```
. local xb0 = _b[_cons] + _b[age]*`meana' + _b[drug]*1

. di "Mean age = " `meana' " ,_b[_cons] + _b[age]*(mean age) + _b[drug]*1 = " `
> xb0'
Mean age = 55.875 ,_b[_cons] + _b[age]*(mean age) + _b[drug]*1 = -6.0645694

. di "Pred. Median [mean(age), drug=1] = "  (ln(2)*exp(-`xb0'))^(1/e(aux_p))
Pred. Median [mean(age), drug=1] = 29.552145

. di "Pred. Mean [mean(age),drug=1] = " exp(-`xb0'/e(aux_p))*exp(lngamma(1+1/e(
> aux_p)))
Pred. Mean [mean(age),drug=1] = 32.807649
```

The results highlight again the very large difference in the survival time distribution between drug and placebo recipients. Observe too the difference between the mean and median durations.

Exactly the same principles as described here could be used if you had a model with a large number of explanatory variables rather than simply two.

The **predict** command after **streg** can be used to create other types of variables. E.g. residuals (Cox-Snell and martingale-like) for analysis of specification. See the Reference Manuals.

AFT representation

To complete the discussion of the Weibull model, consider now the AFT representation of the results, showing either coefficients or time-ratios.

```
. streg drug age, dist(weibull) nolog time

         failure _d:  died
   analysis time _t:  studytim


Weibull regression -- accelerated failure-time form

No. of subjects =           48                     Number of obs   =        48
No. of failures =           31
Time at risk    =          744
                                                   LR chi2(2)      =     35.39
Log likelihood  =   -42.931335                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      _t |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    drug |   1.305563   .2369046      5.511   0.000       .8412383    1.769887
     age |  -.0714323   .0217129     -3.290   0.001      -.1139888   -.0288758
   _cons |   6.289679   1.220494      5.153   0.000       3.897554    8.681804
---------+--------------------------------------------------------------------
   /ln_p |   .5204297   .1389037      3.747   0.000       .2481834     .792676
------------------------------------------------------------------------------
       p |   1.682751   .2337403                          1.281695    2.209301
     1/p |   .5942651   .0825456                           .452632    .7802168
------------------------------------------------------------------------------

.
. * Weibull model, AFT, exp(coefficients), via replay of earlier
. streg, tr

Weibull regression -- accelerated failure-time form

No. of subjects =           48                     Number of obs   =        48
No. of failures =           31
Time at risk    =          744
                                                   LR chi2(2)      =     35.39
Log likelihood  =   -42.931335                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      _t |  Tm. Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    drug |   3.689765    .874122      5.511   0.000       2.319237     5.87019
     age |   .9310593    .020216     -3.290   0.001       .8922679    .9715372
---------+--------------------------------------------------------------------
   /ln_p |   .5204297   .1389037      3.747   0.000       .2481834     .792676
------------------------------------------------------------------------------
       p |   1.682751   .2337403                          1.281695    2.209301
     1/p |   .5942651   .0825456                           .452632    .7802168
------------------------------------------------------------------------------
```

Recall that the estimated coefficient on drug in PH version of the model was 2.20, and that –(–2.20)(0.59) = 1.31 which is indeed the estimated coefficient on drug in the AFT version of the model, as expected. Similarly, for age, –(0.12)(0.59) = –0.07, and for the constant term –(–10.58)(0.59) = 6.29. The values of the AFT coefficients can be interpreted as saying that drug recipients have longer (log) survival times, and older people have shorter ones.

5 Estimation of the Log-logistic model

Estimation of this alternative model is quite straightforward, in the sense that little modification of the commands used earlier is required. The model is an AFT one (but not as a PH one).

```
. streg drug age, dist(logl) nolog

         failure _d:  died
   analysis time _t:  studytim


Log-logistic regression -- accelerated failure-time form

No. of subjects =           48                     Number of obs   =        48
No. of failures =           31
Time at risk    =          744
                                                   LR chi2(2)      =     35.14
Log likelihood  =    -43.21698                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      _t |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    drug |   1.420237   .2502148      5.676   0.000       .9298251    1.910649
     age |  -.0803289   .0221598     -3.625   0.000      -.1237614   -.0368964
   _cons |   6.446711   1.231914      5.233   0.000       4.032204    8.861218
------------------------------------------------------------------------------
 /ln_gam |  -.8456552   .1479337     -5.716   0.000        -1.1356   -.5557105
------------------------------------------------------------------------------
   gamma |    .429276   .0635044                          .3212293    .5736646
------------------------------------------------------------------------------
```

Qualitatively, the estimates are similar to those derived from the Weibull model. Those receiving the drug, and younger, have longer survival times. It turns out that the estimated median duration (at covariate means), 16, is a little smaller than that predicted by the Weibull model (17). Also the hazard function (evaluated at covariate means) is also rather different: note the negative slope at longer durations.

What about the estimates of the median and mean duration. (The latter can be calculated because the estimate of gamma is less than one – see Lesson 2.) We proceed as we did with the Weibull model: First are the estimates for the (hypothetical) person with characteristics corresponding to sample mean values.

```
. predict xb, xb

. su xb

Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
      xb |      48    2.786804   .8510832   1.064673   4.091489

. di "Pred. Median [at sample mean X] = "  exp(r(mean))
Pred. Median [at sample mean X] = 16.229071

. di "Pred. Mean [at sample mean X] = "  exp(r(mean))*(_pi*e(gamma))/sin(_pi*e(
> gamma))
Pred. Mean [at sample mean X] = 22.438269
```

Here, second are calculations of the mean and median for each person in the sample. We are going to this 'by hand'. Derivation of the values using **predict** is left as an exercise.

```
. * median duration for each person in sample
. ge mediand = exp(xb)
. * mean duration for each person in sample
. ge meand = exp(xb)*(_pi*e(gamma))/sin(_pi*e(gamma))
```

Now we list the values to compare them across individuals with different characteristics.

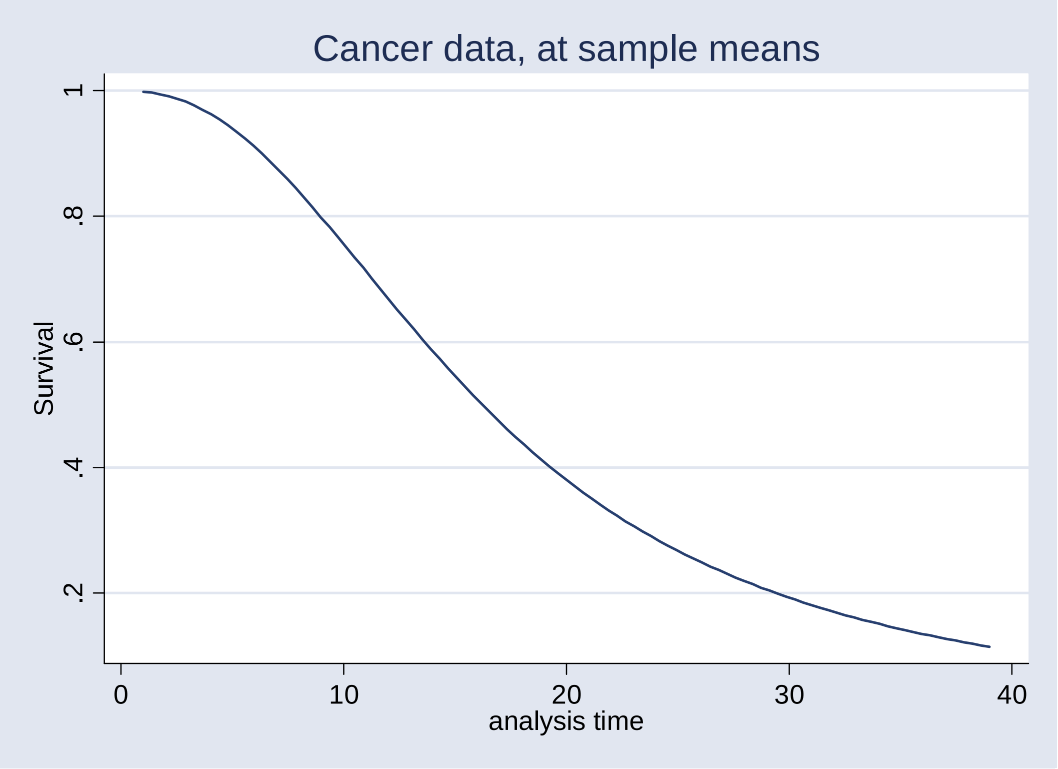
```
. sort age drug

. list id age drug mediand meand if age >50 & age <60 & drug ==0 , noobs

       id       age      drug    mediand      meand
       17        51         0    10.4849   14.49639
       11        52         0   9.675598   13.37746
        4        52         0   9.675598   13.37746
       20        52         0   9.675598   13.37746
       14        55         0   7.603588    10.5127
        5        56         0   7.016689   9.701255
        9        56         0   7.016689   9.701255
       19        57         0    6.47509   8.952441
        8        58         0   5.975296   8.261427
       10        58         0   5.975296   8.261427
        3        59         0    5.51408   7.623751

. list id age drug mediand mean if age >50 & age <60 & drug ==1, noobs

       id       age      drug    mediand      meand
       34        52         1    40.0386   55.35727
       48        52         1    40.0386   55.35727
       36        54         1   34.09621   47.14133
       41        55         1   31.46442   43.50262
       31        55         1   31.46442   43.50262
       35        55         1   31.46442   43.50262
       44        56         1   29.03576   40.14477
       24        56         1   29.03576   40.14477
       42        57         1   26.79458   37.04611
       32        58         1   24.72638   34.18663
       23        58         1   24.72638   34.18663
       39        58         1   24.72638   34.18663
```

Finally, let's compare the estimated means and medians for two (hypothetical) persons, call them *i* and *j*, each of which has an age equal to the sample mean age, but one received the drug and the other didn't.

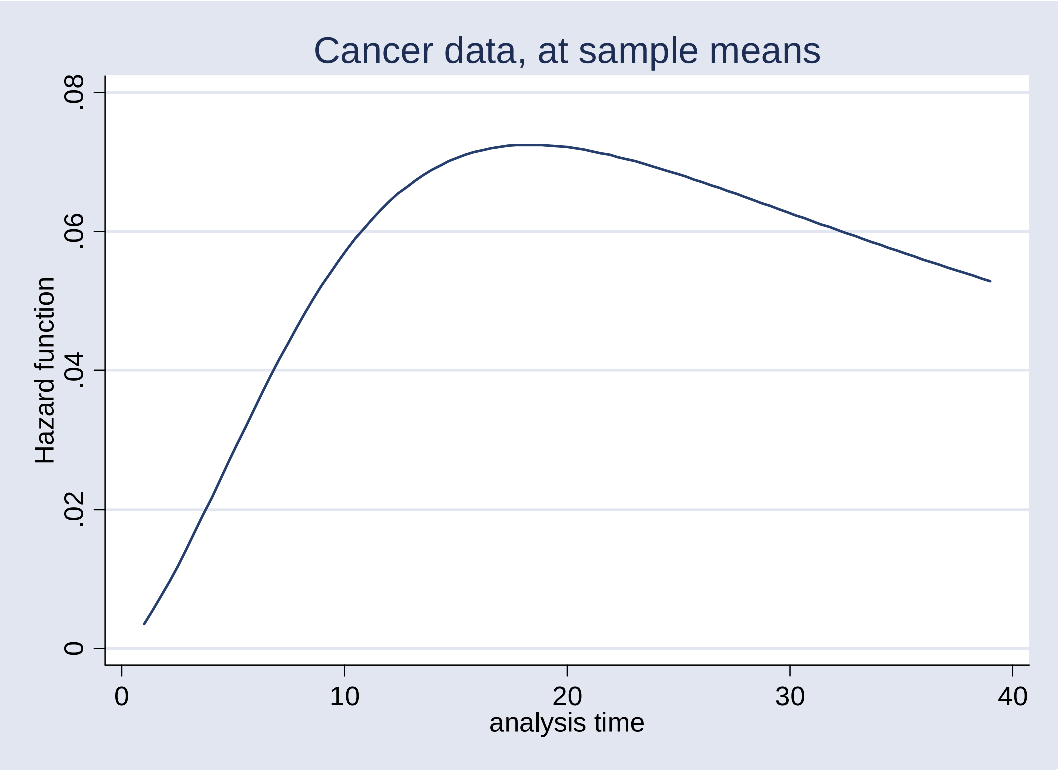
```
. drop xb mediand meand

.
. * mean age is already in local macro meana
.
. local xb0 = _b[_cons] + _b[age]*`meana' + _b[drug]*0

. di "Mean age = " `meana' " ,_b[_cons] + _b[age]*(mean age) + _b[drug]*0 = " `
> xb0'
Mean age = 55.875 ,_b[_cons] + _b[age]*(mean age) + _b[drug]*0 = 1.9583325

. di "Pred. Median [mean(age), drug=0] = "  exp(`xb0')
Pred. Median [mean(age), drug=0] = 7.0874991

. di "Pred. Mean [mean(age), drug=0] = "  exp(`xb0')*(_pi*e(gamma))/sin(_pi*e(g
> amma))
Pred. Mean [mean(age), drug=0] = 9.7991566

. local xb0 = _b[_cons] + _b[age]*`meana' + _b[drug]*1

. di "Mean age = " `meana' " ,_b[_cons] + _b[age]*(mean age) + _b[drug]*1 = " `
> xb0'
Mean age = 55.875 ,_b[_cons] + _b[age]*(mean age) + _b[drug]*1 = 3.3785696

. di "Pred. Median [mean(age), drug=1] = "  exp(`xb0')
Pred. Median [mean(age), drug=1] = 29.328789

. di "Pred. Mean [mean(age), drug=1] = "  exp(`xb0')*(_pi*e(gamma))/sin(_pi*e(g
> amma))
Pred. Mean [mean(age), drug=1] = 40.549902
```

We can also look at the estimated survivor and hazard functions using **stcurv**:

```
. * use stcurve to look at estimated survivor and hazard functions
. *  for person with sample mean values of covariates

. stcurv, survival title("Cancer data, at sample means") ///
>         saving(streg7,replace)
```



```
. stcurv, hazard  title("Cancer data, at sample means") ///
>         saving(streg8,replace)
```



The **at(.)** and **range(.)** options could of course also been used here (see the earlier discussion).

6 Estimation of the Cox PH model using stcox

All the models estimated so far used a parametric specification for the relationship between hazard rates and characteristics (PH models) or survival times and characteristics (AFT models). Forcing the hazard function to take a particular shape may be a disadvantage.

Cox's partial likelihood model allows derivation of estimates of the slope coefficients within the vector β from a PH model, but places no restrictions at all on the shape of the baseline hazard. Let us apply the Cox model using the same covariates as used when estimating the parametric models: The command in Stata is **stcox** and, as for estimation using **streg** models, the data must have first been **stset**. Assuming that has been done, here are the model estimates. First, we make **stcox** display the coefficient estimates (the slope coefficients within the vector β) using the **nohr** option; second, we display the hazard ratios exp(β_k) for each regressor k. (Observe how typing the estimation command name again 'replays' the results.) Alternatively, we could have estimated the model with default display format (hazard ratios) and looked at the coefficients by redisplaying the results and using the **nohr** option.)

```
. stcox drug age, nohr

         failure _d:  died
   analysis time _t:  studytim

Iteration 0:   log likelihood = -99.911448
Iteration 1:   log likelihood = -83.551879
Iteration 2:   log likelihood = -83.324009
Iteration 3:   log likelihood = -83.323546
Refining estimates:
Iteration 0:   log likelihood = -83.323546

Cox regression -- Breslow method for ties

No. of subjects =           48                     Number of obs   =        48
No. of failures =           31
Time at risk    =          744
                                                   LR chi2(2)      =     33.18
Log likelihood  =   -83.323546                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      _t |
      _d |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    drug |  -2.254965   .4548338     -4.958   0.000      -3.146423   -1.363507
     age |   .1136186   .0372848      3.047   0.002       .0405416    .1866955
------------------------------------------------------------------------------
```

Now replay the results:

```
. stcox, hr

Cox regression -- Breslow method for ties

No. of subjects =           48                     Number of obs   =        48
No. of failures =           31
Time at risk    =          744
                                                   LR chi2(2)      =     33.18
Log likelihood  =   -83.323546                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      _t |
      _d | Haz. Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    drug |   .1048772   .0477017     -4.958   0.000       .0430057    .2557622
     age |   1.120325   .0417711      3.047   0.002       1.041375     1.20526
------------------------------------------------------------------------------
```

The coefficient (and hazard ratio) estimates are similar to the Weibull ones derived earlier.

If one wants to create the non-parametric estimates of the baseline survivor and cumulative hazard functions, one simply uses the **basesurvival** and **basechazard** options as follows (note how they can be abbreviated), and then one can **graph** (or **list**) the estimates:

The Cox model does not fit a baseline hazard function – that is not identified. However, all proportional hazards models, including the Cox model, satisfy the properties that $S(t,X) = [S_0(t,X)]^{\lambda}$ and $H(t,X) = \lambda.H_0(t,X)$. The baseline survivor and integrated hazard functions are derived using the methods described in Lesson 4, and then these are scaled using functions of the slope coefficients estimated for the Cox model (part of λ).

```
. stcox drug age, nohr bases(s0) basech(ch0)

<output omitted>
```

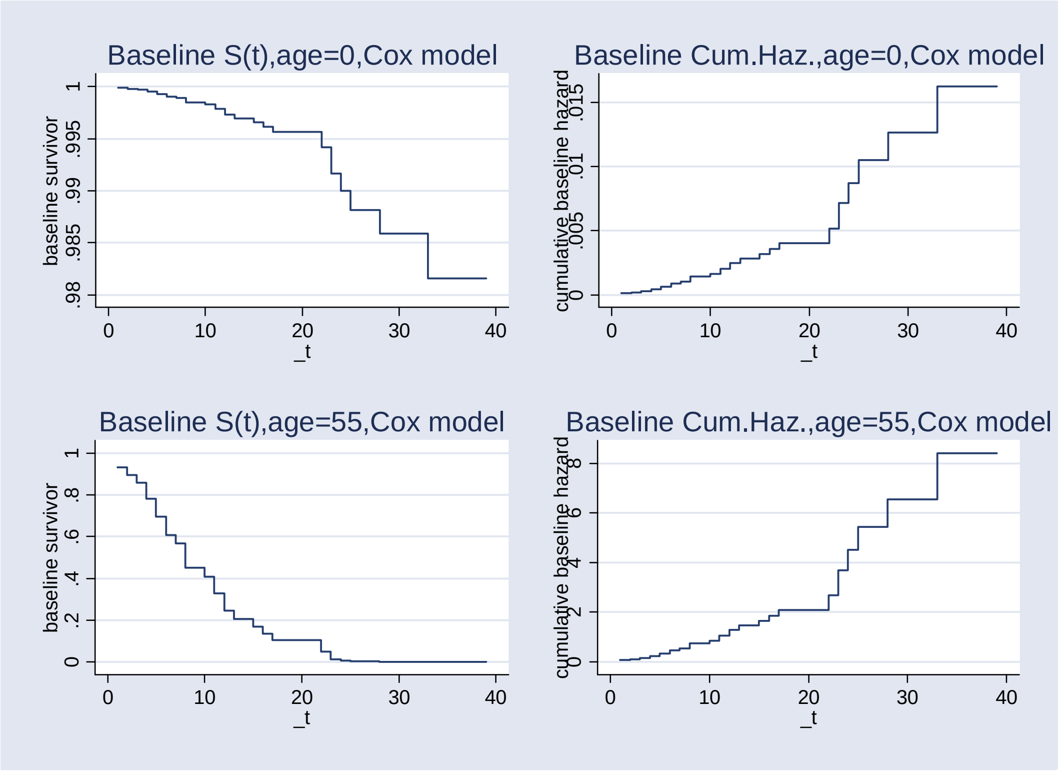
```
. twoway line s0 _t, sort connect(J) title("Baseline S(t),age=0,Cox model") ///
>         saving(stcox1, replace)
(file stcox1.gph saved)

. twoway line ch0 _t, sort connect(J) title("Baseline Cum.Haz.,age=0,Cox model") ///
>         saving(stcox2, replace)
(file stcox2.gph saved)
```

These graphs may not be very appealing in this form. The problem is that the baseline curves are generated setting all the covariates equal to zero. But values of the variable age range from 47 to 67 years in our data set. It is more sensible to calculate baseline curves using values lying within the covariate range. (Remember that the general shape of the hazard function will be the same – this is a PH model.). So let us redo the estimates for age = 55 (and drug = sample mean): To do this, we first calculate a new age variable equal to age minus 55; this ‘recentres’ the estimates. Then we re-estimate the model using the new regressor.

```
. ge age55 = age-55

. stcox drug age55, nohr bases(s1) basech(ch1)


         failure _d:  died
   analysis time _t:  studytim

Iteration 0:   log likelihood = -99.911448
Iteration 1:   log likelihood = -83.551879
Iteration 2:   log likelihood = -83.324009
Iteration 3:   log likelihood = -83.323546
Refining estimates:
Iteration 0:   log likelihood = -83.323546

Cox regression -- Breslow method for ties

No. of subjects =           48                     Number of obs   =        48
No. of failures =           31
Time at risk    =          744
                                                   LR chi2(2)      =     33.18
Log likelihood  =   -83.323546                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      _t |
      _d |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    drug |  -2.254965   .4548338     -4.958   0.000      -3.146423   -1.363507
   age55 |   .1136186   .0372848      3.047   0.002       .0405416    .1866955
------------------------------------------------------------------------------


. twoway line s1 _t, sort connect(J) title("Baseline S(t),age=55,Cox model") ///
>         saving(stcox3, replace)
(file stcox3.gph saved)

. twoway line ch1 _t, sort connect(J) title("Baseline Cum.Haz.,age=55,Cox model") ///
>         saving(stcox4, replace)
(file stcox4.gph saved)
```

The model estimates are just the same. But compare the baseline curves for the uncentred and centred results. I use the **graph combine** command to put all the graphs into one picture.

```
. graph combine stcox1.gph stcox2.gph stcox3.gph stcox4.gph, saving(stcox5, replace)
(file stcox5.gph saved)
```



The median duration from the revised model now corresponds with those derived earlier.

stcox has several other facilities. E.g. there are several alternative methods for dealing with tied survival time. (Recall that the theoretical model was derived assuming one event per survival time. The default option uses the ubiquitous Breslow method for handling ties – this was reported in the estimation output. Other methods are also available: **help stcox**.) It is possible to partition ('stratify') the sample into subgroups and estimate a model which allows a separate non-parametric baseline hazard for each subgroup. There are also possibilities for residual analysis. See **help stcox**.

7 Estimation of the piece-wise constant exponential model using streg

The parametric models that we have considered make strong assumptions about the shape of the hazard function, and the Cox model makes none. Sometimes an in between approach is more appealing, in which we fit a semi-parametric hazard. The piece-wise constant exponential model is the model most commonly used for doing this (in a continuous time modelling framework). The hazard is assumed constant within pre-specified survival time intervals but the constants may differ for different intervals.

The model is simple to estimate using **streg, dist(exponential)** but first requires some reorganisation of the data and creation of some time-varying covariates.

Recall that the exponential model is

$$h_i(t) = h_0.\lambda_i, \quad \text{where } \lambda_i \equiv \exp(\beta'X_i), \text{ or}$$
$$\log[h_i(t)] = \log(h_0) + \beta'X_i$$

since, in this case $h_0(t) = h_0$, a constant.

We can generalise this specification to have a constant hazard within each of K intervals along the survival time axis:

$$\log[h_i(t)] = \log(h_{01}) + \beta'X_i, \quad t \in (0, \tau_1]$$
$$\log[h_i(t)] = \log(h_{02}) + \beta'X_i, \quad t \in (\tau_1, \tau_2]$$
$$\ldots$$
$$\log[h_i(t)] = \log(h_{0\mathrm{K}}) + \beta'X_i, \quad t \in (\tau_{\mathrm{K}-1}, \tau_{\mathrm{K}}]$$

All we need to estimate the model is to generate variables which allow the constant term in the hazard regression to differ from interval to interval. This we do by changing the organisation of the data set (**expand**ing it or **stsplit**ting it) and specifying the variables using appropriate dummy variables. Reread the relevant section in Lesson 3.

Suppose the estimates above lead us to wish to allow the baseline hazard to differ over three intervals (0,8], (8, 17] and (17, 39]. In Lesson 3 we showed how to split episodes and create dummy variables that linked the (new) episodes with these time intervals – we called these variables, e1, e2, and e3, respectively. We then have two possible strategies in estimation: either we include all three variables (e1, e2, and e3) as regressors in the model and exclude the constant term, or we can include a constant term but only two of the variables. (We can't include all three variables plus a constant because that would introduce a collinearity between the regressors, and the model could not be estimated.) I prefer the second display because it allows us to look directly at how the baseline hazard for the two intervals in question differs from that of the interval corresponding to the excluded variable. (This is a presentational or interpretational issue – the models are the same models and would generate the same predictions.)

Here, first, are the estimates for the case where all three variables are included, there is no constant term, and we want coefficients displayed rather than hazard ratios.

```
. streg drug age e1 e2 e3, dist(exp) nolog nocons nohr

         failure _d:  died
   analysis time _t:  studytim
                id:  id

Exponential regression -- log relative-hazard form

No. of subjects =           48                     Number of obs   =        98
No. of failures =           31
Time at risk    =          744
                                                   Wald chi2(5)    =    252.59
Log likelihood  =   -46.134703                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        drug |  -2.010659   .4069367    -4.94   0.000     -2.80824   -1.213077
         age |   .1031196   .0357331     2.89   0.004      .033084    .1731553
          e1 |  -8.178818   2.071186    -3.95   0.000    -12.23827   -4.119368
          e2 |  -7.720299   2.006316    -3.85   0.000    -11.65261   -3.787991
          e3 |  -7.098881   2.000018    -3.55   0.000    -11.01884   -3.178916
------------------------------------------------------------------------------
```

Now, second, see what happens if, instead, you estimate the model including the last two variables and a constant term:

```
. streg drug age e2 e3, dist(exp) nolog nohr

         failure _d:  died
   analysis time _t:  studytim
                id:  id

Exponential regression -- log relative-hazard form

No. of subjects =           48                     Number of obs   =        98
No. of failures =           31
Time at risk    =          744
                                                   LR chi2(4)      =     30.42
Log likelihood  =   -46.134703                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        drug |  -2.010659   .4069367    -4.94   0.000     -2.80824   -1.213077
         age |   .1031196   .0357331     2.89   0.004      .033084    .1731553
          e2 |   .4585195   .4406342     1.04   0.298    -.4051077    1.322147
          e3 |   1.079937   .4924212     2.19   0.028     .1148094    2.045065
       _cons |  -8.178818   2.071186    -3.95   0.000    -12.23827   -4.119368
------------------------------------------------------------------------------
```

Finally here is that same model again, but now with hazard ratios displayed.

```
. streg drug age e2 e3, dist(exp) nolog

         failure _d:  died
   analysis time _t:  studytim
                 id:  id

Exponential regression -- log relative-hazard form

No. of subjects =           48                     Number of obs   =         98
No. of failures =           31
Time at risk    =          744
                                                   LR chi2(4)      =      30.42
Log likelihood  =   -46.134703                     Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        drug |   .1339005    .054489    -4.94   0.000     .0603111     .297281
         age |   1.108624   .0396146     2.89   0.004     1.033637    1.189051
          e2 |    1.58173   .6969645     1.04   0.298      .666905    3.751466
          e3 |   2.944495   1.449932     2.19   0.028      1.12166    7.729663
------------------------------------------------------------------------------
```

The hazard for the third interval (17, 39] is 2.94 times higher than the hazard for the first interval (the reference category). That it is higher is what we would expect from the non-parametric estimates. The hazard ratios for drug and age are similar to those estimated by the other PH models.

8 Exercise 5.1

(i) Repeat all the derivations above, but using the marriage data set (duration.dta) with the sex and married variables as covariates (you need to create a dummy variable from married first). Compare the estimates from the Weibull, Log-logistic and Cox models. Consider the impact of the covariates, the shape of the hazard function, and the median partnership survival time for persons of different legal status (single, married) and sex (man, woman).

(ii) For the Weibull model and the Cancer data, we computed the estimated mean and median for each individual both by hand and directly using Stata's **predict** command. For the Log-logistic model, we only used the by-hand method. Repeat the derivation using **predict**, and show that it yields the same results.

(iii) Return to the Cancer data set and estimate the lognormal model using the same covariates as before (drug, age). Then run the following three commands:

```
        ge cens = 1-died
        ge lntime = ln(studytim)
        cnreg lntime drug age, censored(cens)
```

Compare the results from the **cnreg** and **streg, d(lognormal)** regressions. Can you explain the relationship between them? Now examine how badly OLS does. Rerun the regression first simply ignoring censoring (`reg lntime drug age`) and then excluding the censored cases (`reg lntime drug age if died==1`). How do the results compare with the earlier ones?

(iv) Now estimate the generalised Gamma mode using the Cancer data and the same regressors. What do the hazard and survivor functions look like? Compare the median durations for persons with drug = 0 and drug = 1 (use **stcurv**).

(v) [Harder] The generalized Gamma model has ancillary parameters kappa and sigma. For particular values of these parameters, the model reduces to models considered above (they are nested). If kappa = 1: Weibull model. If kappa = 1, sigma = 1: Exponential model. If kappa = 0: Lognormal model. Do a Wald test of the null hypothesis that kappa = 0, and a Wald test of the null hypothesis that kappa = 1. Compare the test statistics with $\chi^2(1)$. Alternatively use a likelihood ratio test based on the log-likelihood values derived from estimation of the generalized Gamma model and a Weibull model.