VI. HAZARD REGRESSION ANALYSIS OF SURVIVAL DATA

- Extend simple proportional hazards regression to models with multiple covariates
- Model parameters, hazard ratios and relative risks
- Similarities between hazard regression and linear regression
  - Categorical variables, multiplicative models, models with interaction
  - Estimating the effects of two risk factors on a relative risk
  - Calculating 95% CIs for relative risks derived from multiple parameter estimates.
  - Adjusting for confounding variables
- Restricted cubic splines and survival analysis
- Stratified proportional hazards regression models
- Using age as the time variable in survival analysis
- Checking the proportional hazards assumption
  - Comparing Kaplan-Meier plots to analogous plots drawn under the proportional hazards assumption
  - Log-log plots
- Hazards regression models with time-dependent covariates
  - Testing the proportional hazards assumption

1. The Model

The simple proportional hazards model generalizes to a multiple regression model in much the same way as for linear and logistic regression.

Suppose we have a cohort of n people. Let

t_i = the time from entry to exit for the i^{th} patient,

$$f_i = \begin{cases} 1: & i^{th} \text{ patient dies at exit} \\ 0: & i^{th} \text{ patient alive at exit} \end{cases}$$

$x_{i1}, x_{i2}, ..., x_{iq}$ be the value of q covariates for the i^{th} patient.

Let $\lambda_0[t]$ be the hazard function for patients with covariates

$$x_{i1} = x_{i2} = ... = x_{iq} = 0$$

Then the proportional hazards model assumes that the hazard function for the i^{th} patient is

$$\lambda_i[t] = \lambda_0[t] \exp[\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_q x_{iq}].$$

a) Relative risks and hazard ratios

Suppose that patients in risk groups 1 and 2 have covariates $x_{11}, x_{12}, ..., x_{1q}$ and $x_{21}, x_{22}, ..., x_{2q}$, respectively.

Then the relative risk of patients in Group 2 with respect to those in Group 1 in the time interval (t, t+Δt) is

$$\frac{\lambda_2[t]\Delta t}{\lambda_1[t]\Delta t}$$

$$= \frac{\lambda_0[t]\exp\left[x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2q}\beta_q\right]}{\lambda_0[t]\exp\left[x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1q}\beta_q\right]}$$

$$= \exp\left[(x_{21} - x_{11})\beta_1 + (x_{22} - x_{12})\beta_2 + \cdots (x_{2q} - x_{1q})\beta_q\right]$$

Note that $\lambda_0[t]$ drops out of this equation, and that this instantaneous relative risk remains constant over time.

Thus, if the proportional hazards model is reasonable, we can interpret

$$(x_{21} - x_{11})\beta_1 + (x_{22} - x_{12})\beta_2 + \cdots (x_{2q} - x_{1q})\beta_q$$

as being the log relative risk associated with being in Group 2 as compared to being in Group 1.

2. Analyzing Multiple Hazard Regression Models

The analysis of hazard regression models is very similar to that for logistic regression. A great strength of Stata is that the commands for analyzing these two models are almost identical. The key difference is in how we interpret the coefficients: in logistic regression

$$\exp\left[(x_{21} - x_{11})\beta_1 + (x_{22} - x_{12})\beta_2 + \cdots (x_{2q} - x_{1q})\beta_q\right]$$

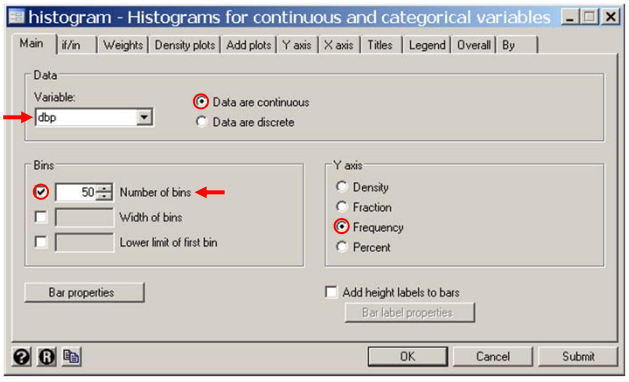
estimates an odds ratio, while in proportional hazards regression this expression estimates a relative risk.

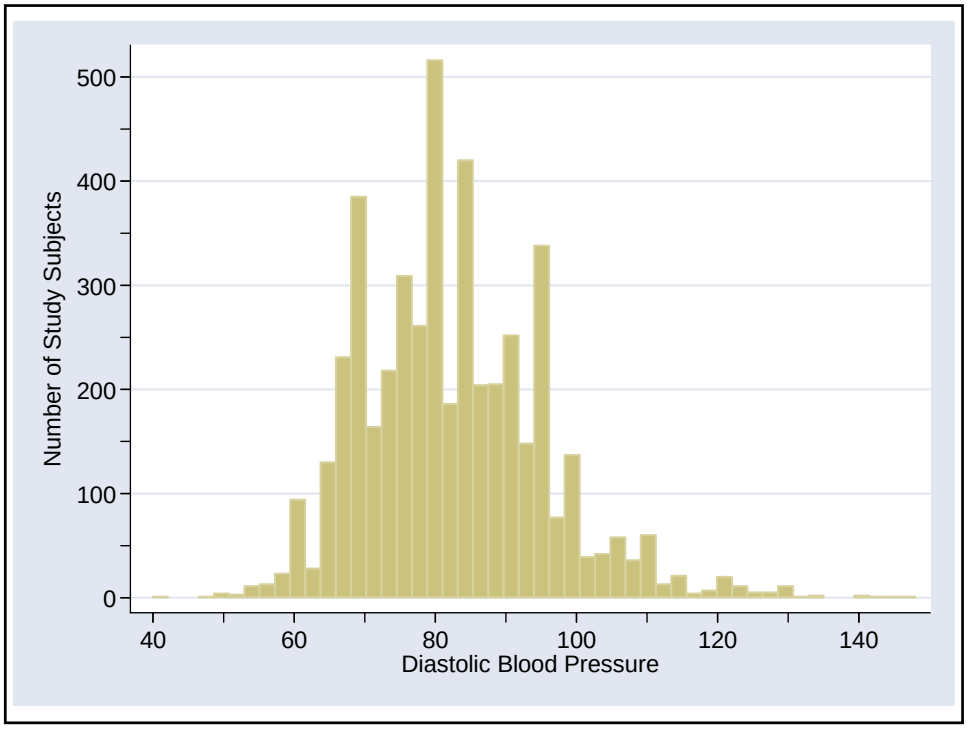
b) Example: Diastolic blood pressure and gender on risk of coronary heart disease

The Framingham data set (Levy 1999) also contains follow-up data on coronary heart disease. Consider the following survival analysis.

```
. *  7.6.Framingham.ClassVersion.log
. *
. *  Proportional hazards regression analysis of the effect of gender and
. *  baseline diastolic blood pressure (DBP) on coronary heart disease (CHD)
. *  adjusted for age, body mass index (BMI) and serum cholesterol (SCL)
. *  (Levy 1999).
. *
. use C:\WDDtext\2.20.Framingham.dta, clear
.
. *  Univariate analysis of the effect of DBP on CHD
. *
. * Graphics > Histogram
. histogram dbp, bin(50) frequency xlabel(40(20)140) xtick(40(10)140)    /// {1}
>     ylabel(0(100)500, angle(0)) ytick(0(50)500)                       ///
>     ytitle("Number of Study Subjects")
(bin=50, start=40, width=2.16)
```

{1} This command draws the histogram on the next slide. **bin** specifies the number of bars. **frequency** specifies that the *y*-axis is to be number of patients rather than proportion of patients.






```
. generate dbpgr = recode(dbp,60,70,80,90,100,110,111)              {2}

. * Statistics > Summaries... > Tables > Two-way tables with measures...
. tabulate dbpgr chdfate                                            {3}

           |    Coronary Heart
           |        Disease
     dbpgr |  Censored        CHD |     Total
-----------+----------------------+----------
        60 |       132         18 |       150
        70 |       592        182 |       774
        80 |     1,048        419 |     1,467
        90 |       863        404 |     1,267
       100 |       417        284 |       701
       110 |       125        110 |       235
       111 |        49         56 |       105
-----------+----------------------+----------
     Total |     3,226      1,473 |     4,699
```

{2} Define ***dbpgr*** to be a **categorical** variable based on *dbp*. This ***recode*** function sets *drpgr* equal to
60 for all patients with $dbp \leq 60$,
70 for all patients with $60 < dbp \leq 70$,
80 for all patients with $70 < dbp \leq 80$,
.
.
110 for all patients with $100 < dbp \leq 110$,
111 for all patients with $110 < dbp$.

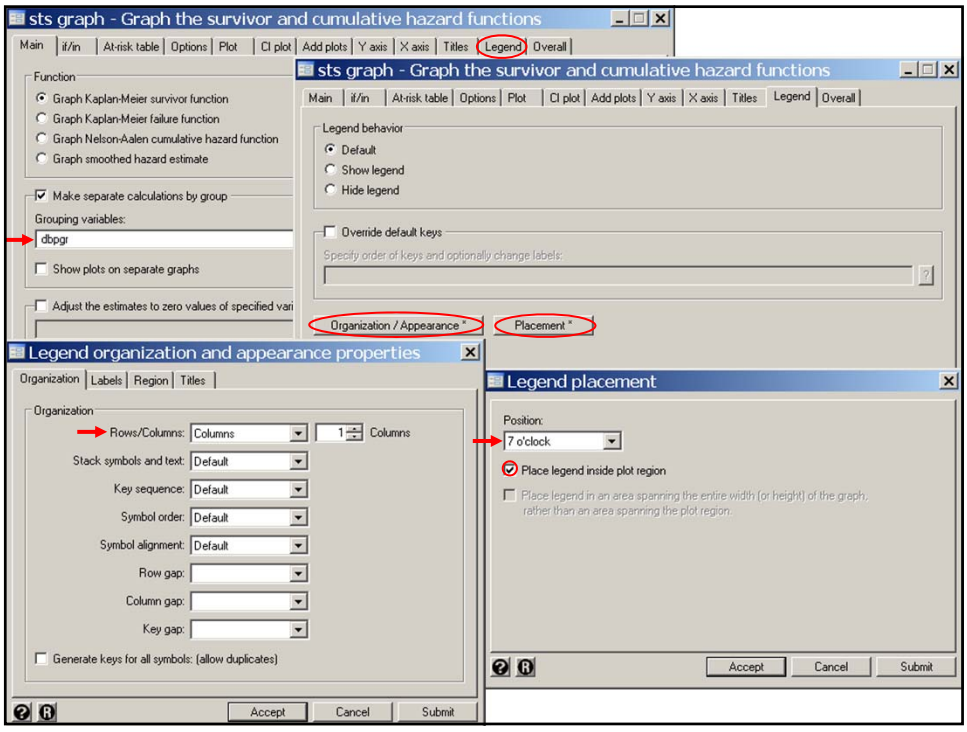
{3} This ***tabulate*** statement shows that the preceding ***recode*** statement **worked**. Subjects with DBPs less than 61 or greater than 110 are rare. However, the database is large enough to provide **255** such subjects.

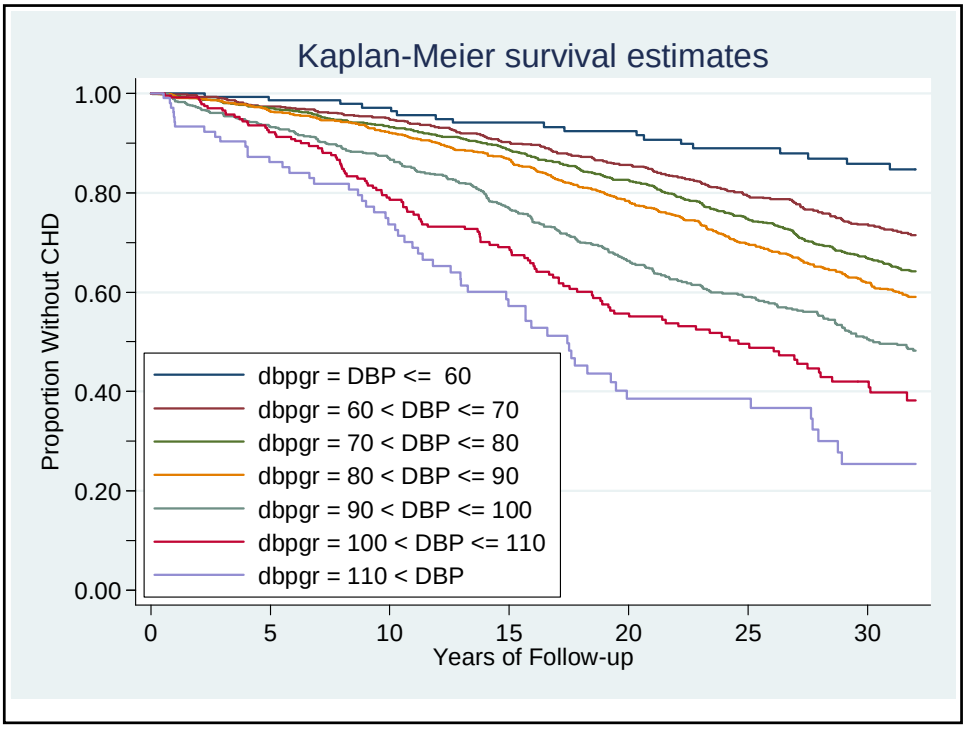
```
. * Variables Manager
. label define dbp 60  "DBP <=  60"      70  "60 < DBP <= 70"         ///
>                  90  "80 < DBP <= 90"  80  "70 < DBP <= 80"         ///
>                 100 "90 < DBP <= 100" 110 "100 < DBP <= 110" 111 "110 < DBP"

. label variable dbpgr "DBP level"

. label values dbpgr dbp

. generate time= followup/365.25                                          {4}

. label variable time "Follow-up in Years"

. * Statistics > Survival... > Setup... > Declare data to be survival...
. stset time, failure(chdfate)

     failure event:  chdfate != 0 & chdfate < .
obs. time interval:  (0, time]
 exit on or before:  failure
------------------------------------------------------------------------------
     4699  total obs.
        0  exclusions
------------------------------------------------------------------------------
     4699  obs. remaining, representing
     1473  failures in single record/single failure data
 103710.1  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =        32
```

{4} We define time to be follow-up in years to make graphs more intelligible.

```
. * Graphics > Survival analysis graphs > Kaplan-Meier survivor function
. sts graph, by(dbpgr) ytitle(Proportion Without CHD)                ///
>     ylabel(0(.2)1, angle(0)) ytick(.0(.1)1) xlabel(0(5)30)         ///
>     xtitle("Years of Follow-up") legend(ring(0) position(7) col(1))     {5}

        failure _d:  chdfate
  analysis time _t:  time
```

{5} These **legend** sub-options have the following effects. **ring(0)** specifies that the legend is to be inside the graph axes. **position** specifies the clock position of the legend: 12 is top center, 3 is left center, 6 is bottom center, 7 is bottom left, etc. **col(1)** specifies that the legend is to be given in a single column.



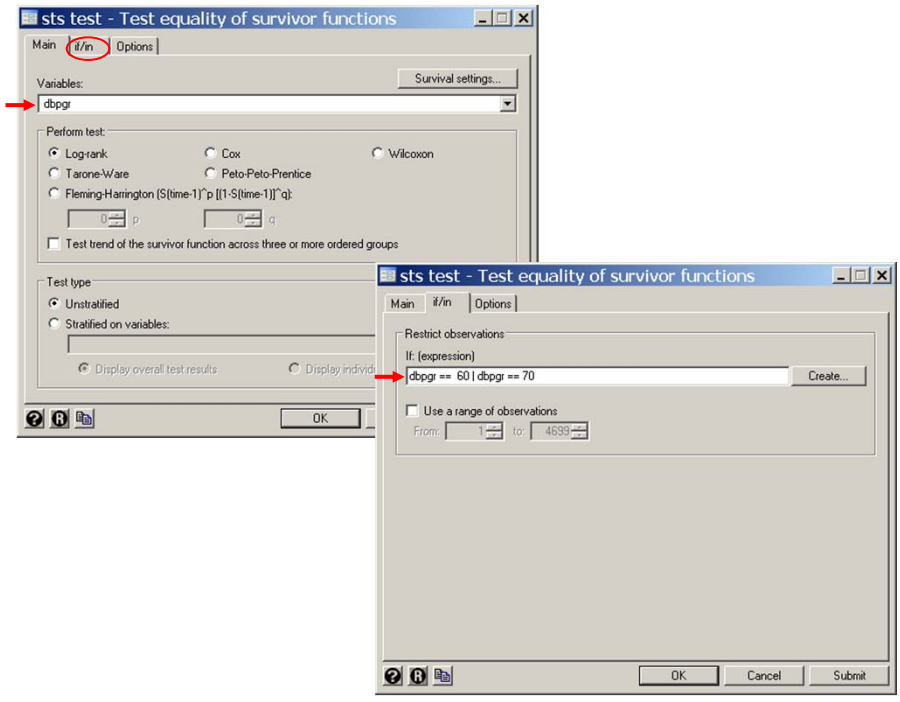



```
. * Statistics > Survival... > Summary... > Test equality of survivor...
. sts test dbpgr                                                        {6}

         failure _d:  chdfate
   analysis time _t:  time


Log-rank test for equality of survivor functions

                  |   Events         Events
dbpgr             |  observed       expected
------------------+------------------------
DBP <=  60        |        18          53.63
60 < DBP <= 70    |       182         275.72
70 < DBP <= 80    |       419         489.41
80 < DBP <= 90    |       404         395.62
90 < DBP <= 100   |       284         187.97
100 < DBP <= 110  |       110          52.73
110 < DBP         |        56          17.94
------------------+------------------------
Total             |      1473        1473.00

                      chi2(6) =     259.71
                      Pr>chi2 =     0.0000
```

{6} This command tests the null hypotheses that the CHD free survival curves for all 7 baseline DBP groups are equal

```
. * Statistics > Survival... > Summary... > Test equality of survivor...
. sts test dbpgr if dbpgr ==  60 | dbpgr == 70                          {7}

         failure _d:  chdfate
   analysis time _t:  time

Log-rank test for equality of survivor functions

               |   Events         Events
dbpgr          |  observed       expected
---------------+-------------------------
DBP <=  60     |        18          32.58
60 < DBP <= 70 |       182         167.42
---------------+-------------------------
Total          |       200         200.00

                     chi2(1) =       7.80
                     Pr>chi2 =     0.0052
```

{7} This command tests the null hypotheses that the CHD free survival curves for the two lowest baseline DBP groups are equal.



```
. sts test dbpgr if dbpgr ==  70 | dbpgr == 80                    {8}
. sts test dbpgr if dbpgr ==  80 | dbpgr == 90
                                                  Pr>chi2 =     0.0090
. sts test dbpgr if dbpgr ==  90 | dbpgr == 100
                                                  Pr>chi2 =     0.0000
. sts test dbpgr if dbpgr == 100 | dbpgr == 110
                                                  Pr>chi2 =     0.0053
. sts test dbpgr if dbpgr == 110 | dbpgr == 111
                                                  Pr>chi2 =     0.0215
```

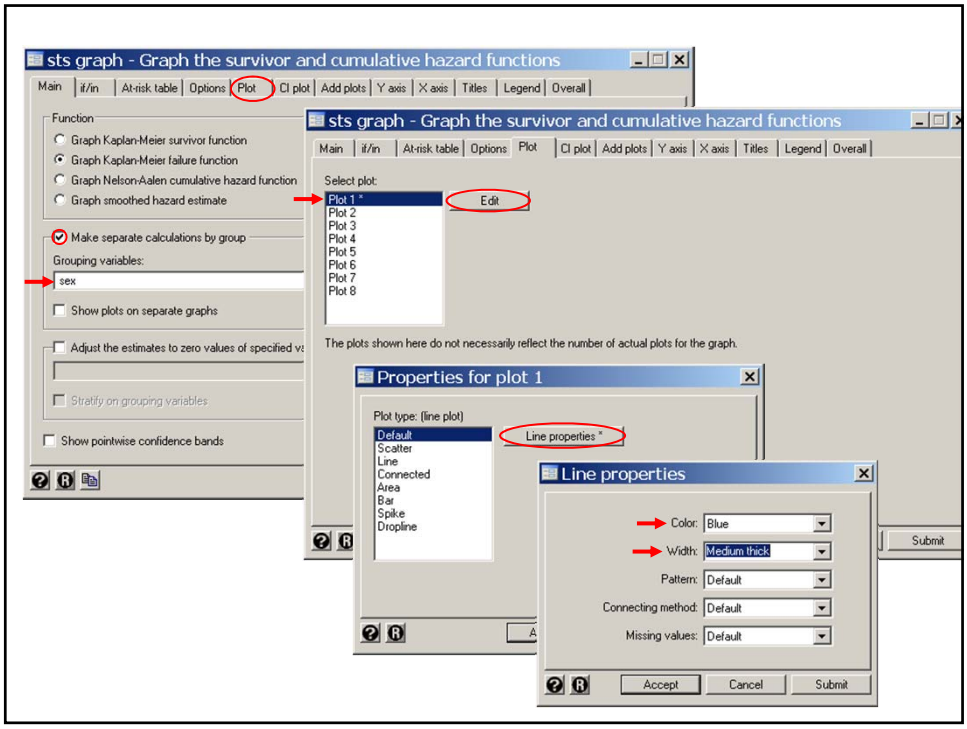
{8} All pair-wise logrank tests of adjacent DBP group levels are not statistically significant (output deleted).

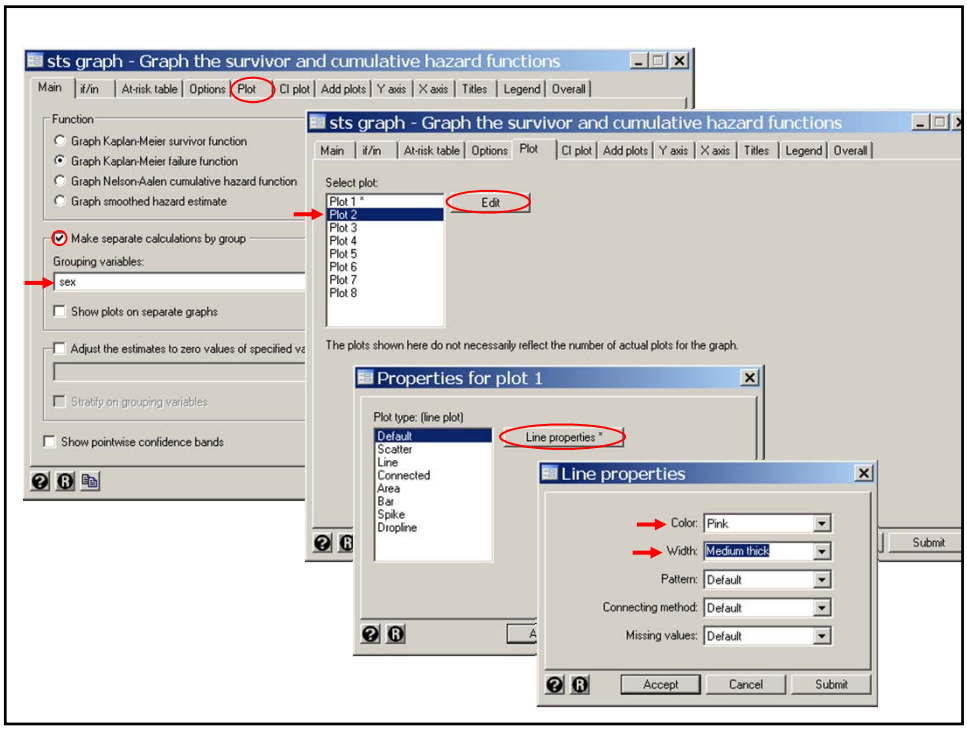
```
. *
. *  Univariate analysis of the effect of gender on CHD
. *
. * Graphics > Survival analysis graphs > Kaplan-Meier survivor function
. sts graph, by(sex) plot1opts(color(blue) lwidth(medthick) )     ///  {9}
>    plot2opts(color(pink)  lwidth(medthick))                     ///
>    ytitle(Cumulative CHD Morbidity)                             ///
>    xtitle(Years of Follow-up) xlabel(0(5)30) failure            ///  {10}
>    ylabel(0(.1).5, angle(0)) legend(ring(0) position(11) col(1))

         failure _d:  chdfate
   analysis time _t:  time
```

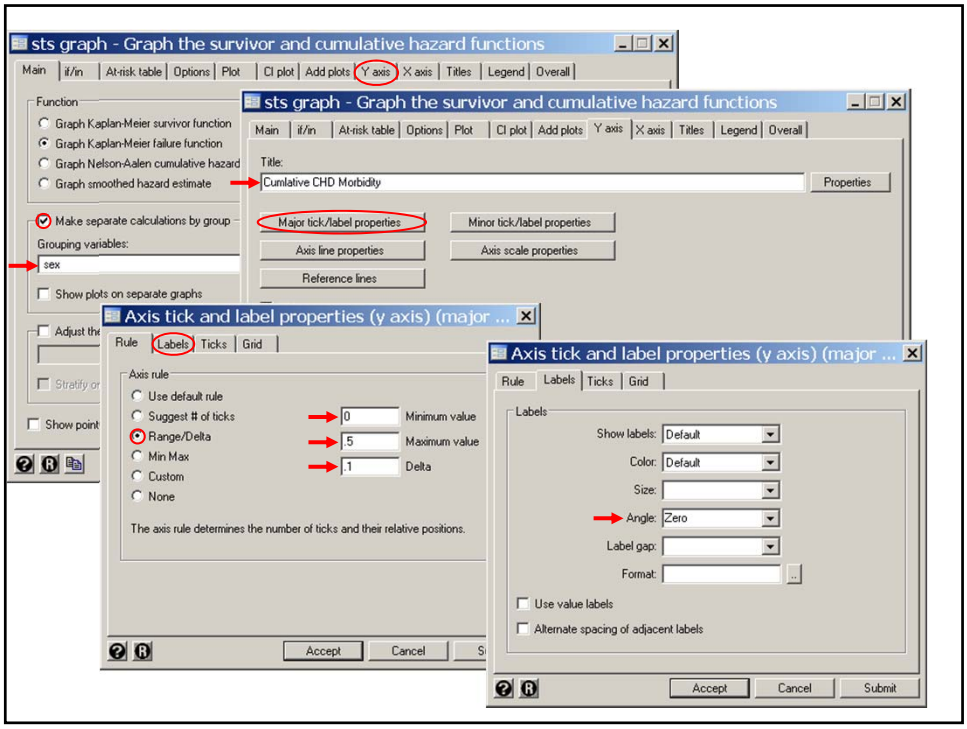
{9} The ***plot1opts*** and ***plot2opts*** options control the appearance of the first and second plot, respectively. The ***color*** and ***lwidth*** suboptions control the color and width of the plotted lines. In this example blue and pink curves of medium thickness are plotted for men and women, respectively.

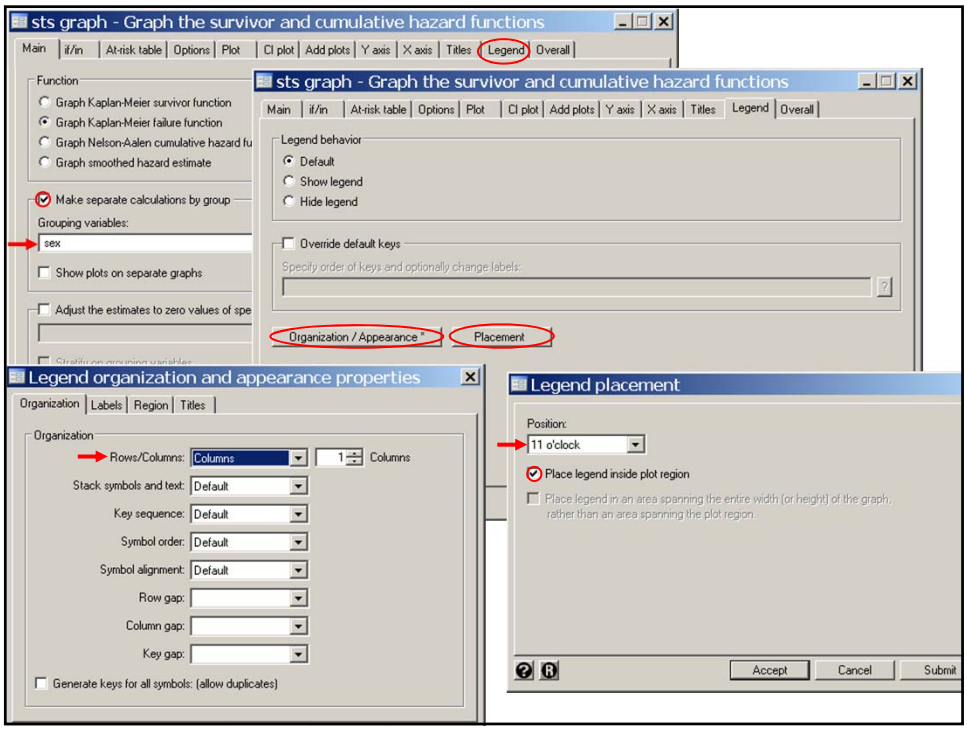
{10} The failure option converts a standard survival curve into a cumulative morbidity curve.

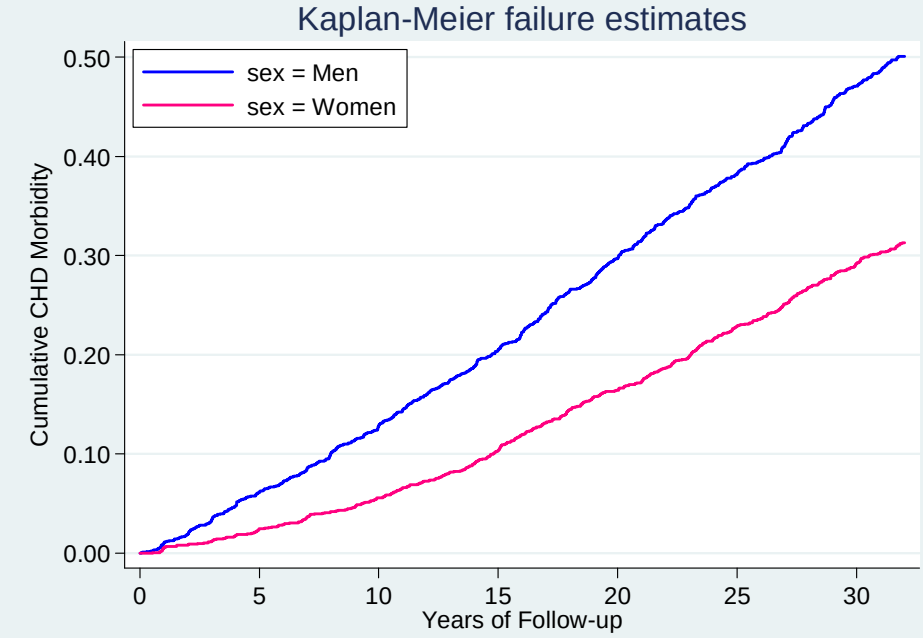
Cumulative morbidity plots are particularly **effective** when a large proportion of subjects **never** suffer the **event** of interest. Note that in this plot of CHD morbidity by sex that the *y*-axis only extends to 0.5


sts graph - Graph the survivor and cumulative hazard functions
Main | if/in | At-risk table | Options | Plot | CI plot | Add plots | Y axis | X axis | Titles | Legend | Overall
Function
Graph Kaplan-Meier survivor function
Graph Kaplan-Meier failure function
Graph Nelson-Aalen cumulative hazard function
Graph smoothed hazard estimate
Make separate calculations by group
Grouping variables:
sex
Show plots on separate graphs
Adjust the estimates to zero values of specified va
Stratify on grouping variables
Show pointwise confidence bands
sts graph - Graph the survivor and cumulative hazard functions
Select plot:
Plot 1 *
Plot 2
Plot 3
Plot 4
Plot 5
Plot 6
Plot 7
Plot 8
Edit
The plots shown here do not necessarily reflect the number of actual plots for the graph.
Properties for plot 1
Plot type: (line plot)
Default
Scatter
Line
Connected
Area
Bar
Spike
Dropline
Line properties *
Line properties
Color: Blue
Width: Medium thick
Pattern: Default
Connecting method: Default
Missing values: Default
Accept
Cancel
Submit


sts graph - Graph the survivor and cumulative hazard functions
Main | if/in | At-risk table | Options | Plot | CI plot | Add plots | Y axis | X axis | Titles | Legend | Overall
Function
Graph Kaplan-Meier survivor function
Graph Kaplan-Meier failure function
Graph Nelson-Aalen cumulative hazard function
Graph smoothed hazard estimate
Make separate calculations by group
Grouping variables:
sex
Show plots on separate graphs
Adjust the estimates to zero values of specified va
Stratify on grouping variables
Show pointwise confidence bands
sts graph - Graph the survivor and cumulative hazard functions
Select plot:
Plot 1 *
Plot 2
Plot 3
Plot 4
Plot 5
Plot 6
Plot 7
Plot 8
Edit
The plots shown here do not necessarily reflect the number of actual plots for the graph.
Properties for plot 1
Plot type: (line plot)
Default
Scatter
Line
Connected
Area
Bar
Spike
Dropline
Line properties *
Line properties
Color: Pink
Width: Medium thick
Pattern: Default
Connecting method: Default
Missing values: Default
Accept
Cancel
Submit


sts graph - Graph the survivor and cumulative hazard functions
Main | if/in | At-risk table | Options | Plot | CI plot | Add plots | Y axis | X axis | Titles | Legend | Overall
Function
Graph Kaplan-Meier survivor function
Graph Kaplan-Meier failure function
Graph Nelson-Aalen cumulative hazard
Graph smoothed hazard estimate
Make separate calculations by group
Grouping variables:
sex
Show plots on separate graphs
Title:
Cumulative CHD Morbidity
Properties
Major tick/label properties
Minor tick/label properties
Axis line properties
Axis scale properties
Reference lines
Axis tick and label properties (y axis) (major ...
Rule | Labels | Ticks | Grid
Axis rule
Use default rule
Suggest # of ticks
Range/Delta
Min Max
Custom
None
0
Minimum value
.5
Maximum value
.1
Delta
The axis rule determines the number of ticks and their relative positions.
Accept
Cancel
Labels
Show labels: Default
Color: Default
Size:
Angle: Zero
Label gap:
Format:
Use value labels
Alternate spacing of adjacent labels
Submit


sts graph - Graph the survivor and cumulative hazard functions
Main | if/in | At-risk table | Options | Plot | CI plot | Add plots | Y axis | X axis | Titles | Legend | Overall
Function
Graph Kaplan-Meier survivor function
Graph Kaplan-Meier failure function
Graph Nelson-Aalen cumulative hazard fu
Graph smoothed hazard estimate
Make separate calculations by group
Grouping variables:
sex
Show plots on separate graphs
Adjust the estimates to zero values of spe
Legend behavior
Default
Show legend
Hide legend
Override default keys
Specify order of keys and optionally change labels:
Organization / Appearance...
Placement
Legend organization and appearance properties
Organization | Labels | Region | Titles
Organization
Rows/Columns: Columns
1 Columns
Stack symbols and text: Default
Key sequence: Default
Symbol order: Default
Symbol alignment: Default
Row gap:
Column gap:
Key gap:
Generate keys for all symbols: (allow duplicates)
Legend placement
Position:
11 o'clock
Place legend inside plot region
Place legend in an area spanning the entire width (or height) of the graph, rather than an area spanning the plot region.
Accept
Cancel
Submit



A survival plot with a *y*-axis that runs from **0 to 1.0** would leave a lot of **blank space** on the graph and would less clearly indicate the difference in morbidity between men and women.

A survival plot with a *y*-axis that runs from **0.5 to 1.0** might leave some readers with **false impression** of the magnitude of the difference in CHD morbidity between men and women.

```
. * Statistics > Survival... > Summary... > Test equality of survivor...
. sts test sex

         failure _d:  chdfate
   analysis time _t:  time


Log-rank test for equality of survivor functions

      |   Events         Events
sex   |  observed       expected
------+-------------------------
Men   |       823         589.47
Women |       650         883.53
------+-------------------------
Total |      1473        1473.00

            chi2(1) =     154.57
            Pr>chi2 =     0.0000
```

CHD cumulative morbidity curves for men and women differ with a high level of statistical significance

```
. codebook sex

sex ---------------------------------------------------- Sex
                  type:  numeric (float)
                 label:  sex

                 range:  [1,2]          units:  1
         unique values:  2       coded missing:  0 / 4699

            tabulation:  Freq.   Numeric  Label
                         2049         1  Men
                         2650         2  Women
. generate male = sex==1                                                    {11}

. * Statistics > Summaries... > Tables > Two-way tables with measures...
. tabulate male sex

           | Sex
      male |       Men      Women |     Total
-----------+----------------------+----------
         0 |         0       2650 |      2650
         1 |      2049          0 |      2049
-----------+----------------------+----------
     Total |      2049       2650 |      4699
```

{11} In the database men and women are coded 1 and 2, respectively. I have decided to treat **male sex** as a **positive** risk factor in our analyses. To do this we need to give men a higher code than women. (Otherwise, female sex would be a protective risk factor.) The logical value ***sex*==1** is true (equals **1**) when the subject is a **man** (*sex*=1), and is false (equals **0**) when she is a **woman** (*sex*=2). Hence the effect of this statement is to define the variable *male* as equaling 0 or 1 for women and men, respectively. The following tabulate command shows that *male* has been defined correctly.

```
. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox male                                                                {12}

 failure _d:  chdfate
   analysis time _t:  time

Iteration 0:   log likelihood = -11834.856
Iteration 1:   log likelihood = -11759.624
Iteration 2:   log likelihood = -11759.553
Refining estimates:
Iteration 0:   log likelihood = -11759.553

Cox regression -- Breslow method for ties

No. of subjects =         4699                     Number of obs   =      4699
No. of failures =         1473
Time at risk    =  103710.0917
                                                   LR chi2(1)      =    150.61
Log likelihood  =   -11759.553                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   1.900412   .0998308    12.22   0.000     1.714482    2.106504
------------------------------------------------------------------------------
```

{12} This statement fits the simple hazard regression model

$$\lambda(t, male) = \lambda_0(t)\exp(\beta \times male)$$

The estimate of the risk of CHD for **men** relative to **women** is

$$e^{\hat{\beta}} = \mathbf{1.90}$$

If we had fitted the model $\lambda(t, sex) = \lambda_0(t)\exp(\beta \times sex)$ we would have got that the estimated risk of CHD for **women** relative to **men** is

$$e^{\hat{\beta}} = 1/1.9004 = \mathbf{0.526.}$$

```
. *
. *  To simplify the analyses let us use fewer DBP groups
. *
. generate dbpg2 = recode(dbp,60,90,110,111)

. * Statistics > Summaries, tables and tests > Tables > One-way tables
. tabulate dbpg2

      dbpg2 |      Freq.     Percent        Cum.
------------+-----------------------------------
         60 |        150        3.19        3.19
         90 |      3,508       74.65       77.85
        110 |        936       19.92       97.77
        111 |        105        2.23      100.00
------------+-----------------------------------
      Total |      4,699      100.00
```

```
. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox  i.dbpg2                                                          {13}
i.dbpg2           _Idbpg2_60-111      (naturally coded; _Idbpg2_60 omitted)

         failure _d:  chdfate
   analysis time _t:  time

Cox regression -- Breslow method for ties

No. of subjects =         4699                     Number of obs   =      4699
No. of failures =         1473
Time at risk    =  103710.0917
                                                   LR chi2(3)      =    188.25
Log likelihood  =   -11740.729                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dbpg2 |
          90 |   2.585841   .6149551     3.99   0.000     1.622454    4.121273
         110 |   4.912658   1.184529     6.60   0.000     3.062505    7.880545
         111 |   9.435655   2.559389     8.27   0.000     5.544808    16.05675
------------------------------------------------------------------------------
```

{13} The ***i.*** prefix is used in the same way as in **logisitc** regression. Recall that *dbpg2* takes the values 60, 90, 110, and 111. *i.dbpg2* defines the following three indicator variables:

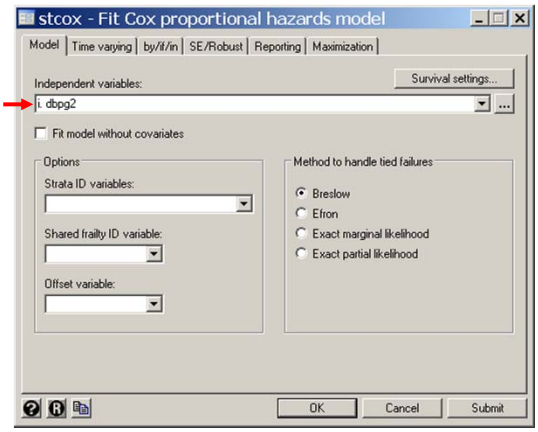
90.dbpg2 = 1 if *dbpg2* = 90, and = 0 otherwise;
110.dbpg2 = 1 if dbpg2 = 110, and = 0 otherwise;
111.dbpg2 = 1 if *dbpg2* = 111, and = 0 otherwise.

Our model is

$$\lambda(t, 90.dbpg2, 110.dpbg2, 111.dpbg2) =$$
$$\lambda_0(t)\exp(\beta_1 \times 90.dbpg2 + \beta_2 \times 110.dbpg2 + \beta_3 \times 111.dbpg2)$$

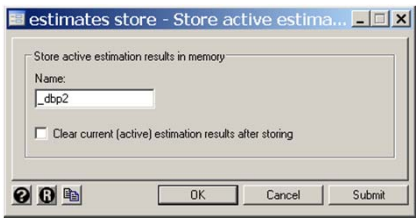
This allows us to obtain the following **relative risk** estimates for CHD compared to people with DBP≤60.

$e^{\hat{\beta}_1} = 2.58$ = risk of people with 60<DBP≤90
$e^{\hat{\beta}_2} = 4.91$ = risk of people with 90<DBP≤100
$e^{\hat{\beta}_3} = 9.44$ = risk of people with 100<DBP




```
. *
. *  Store estimates from this model for future likelihood ratio
. *  tests (tests of change in model deviance).
. *
. *  Statistics > Postestimation > Manage estimation results > Store in memory
. estimates store _dbpg2                                          {14}
```

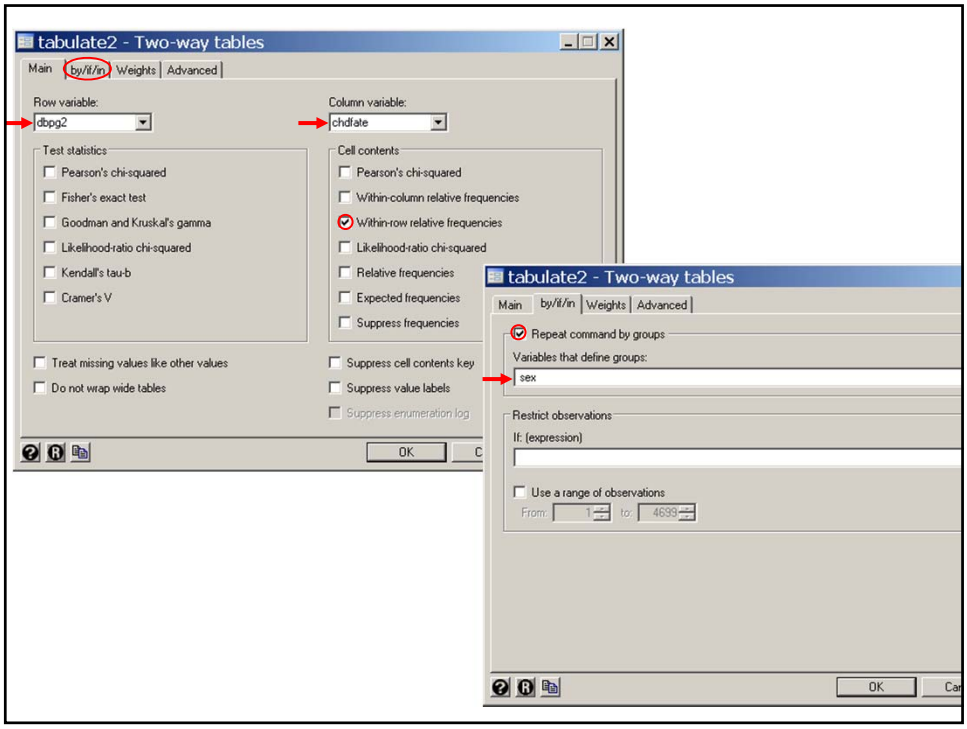
{14} The maximum value of the log likelihood function (as well as other statistics) from this model is stored under the name ***_dbpg2***



```
. sort sex
. * Statistics > Summaries... > Tables > Two-way tables with measures...
. by sex: tabulate dbpg2 chdfate ,row                                    {15}

-> sex=      Men                                Women
           | Coronary Heart Disease          | Coronary Heart Disease
   dbpg2   |  Censored        CHD |  Total   |  Censored        CHD |     Total
-----------+----------------------+----------+----------------------+----------
  DBP<= 60 |        40          9 |     49   |        92          9 |       101
           |     81.63      18.37 | 100.00   |     91.09       8.91 |    100.00   {16}
-----------+----------------------+----------+----------------------+----------
  60<DBP90 |       933        568 |   1501   |      1570        437 |      2007
           |     62.16      37.84 | 100.00   |     78.23      21.77 |    100.00
-----------+----------------------+----------+----------------------+----------
  90DBP110 |       232        227 |    459   |       310        167 |       477
           |     50.54      49.46 | 100.00   |     64.99      35.01 |    100.00
-----------+----------------------+----------+----------------------+----------
  110< DBP |        21         19 |     40   |        28         37 |        65
           |     52.50      47.50 | 100.00   |     43.08      56.92 |    100.00
-----------+----------------------+----------+----------------------+----------
     Total |      1226        823 |   2049   |      2000        650 |      2650
           |     59.83      40.17 | 100.00   |     75.47      24.53 |    100.00
```

{15} The ***row*** option on the tabulate statements shows row percentages. For example 9 of 49 (18.4%) of men with DBP≤60 develop CHD. I have edited the table produced by this command to show the results for men and women on the same rows.

{16} Note the evidence of **interaction** between the effects of **sex** and **DBP** on CHD. Among people with DBP≤60 men have twice the risk of CHD than women (18.4 vs. 8.9). Among people with DBP>110, women have more CHD than men. We need to be able to account for this in our models.




```
. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox  i.dbpg2 male                                                    {17}
i.dbpg2           _Idbpg2_60-111      (naturally coded; _Idbpg2_60 omitted)

        failure _d:  chdfate
  analysis time _t:  time

No. of subjects =         4699                     Number of obs    =      4699
No. of failures =         1473
Time at risk    =  103710.0917
                                                   LR chi2(4)       =    325.65
Log likelihood  =   -11672.032                     Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dbpg2 |
         90  |    2.42989   .5780261     3.73   0.000     1.524409    3.873217
        110  |    4.44512   1.072489     6.18   0.000       2.7702     7.13273
        111  |   9.156554   2.483587     8.16   0.000     5.380908    15.58147
             |
        male |   1.848482   .0972937    11.67   0.000     1.667297    2.049358
------------------------------------------------------------------------------

. display 2*(11740.729   -11672.032)                                     {18}
137.394

. display chi2tail(1, 137.394)                                           {19}
9.888e-32
```

```
Log likelihood  =   -11740.729                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Idbpg2_90 |   2.585841   .6149551     3.99   0.000     1.622454    4.121273
 _Idbpg2_110 |   4.912658   1.184529     6.60   0.000     3.062505    7.880545
 _Idbpg2_111 |   9.435655   2.559389     8.27   0.000     5.544808    16.05675
------------------------------------------------------------------------------
```

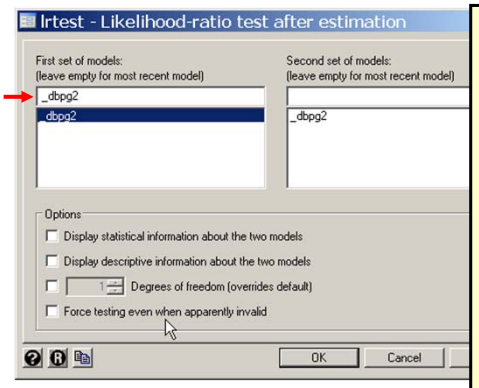
{17} We next fit a **multiplicative** model of **gender** and our four **DBP** groups. That is we fit a model without gender-DBP interaction terms.

{18} The ***display*** command can be used as a pocket calculator for quick calculations. The previous model is **nested** within the model with only the diastolic blood pressure terms. The **difference** in model **deviance** between these models is **137**.

{19} *chi2tail(df,*χ^2) gives the ***P* value** for a chi-squared statistic χ^2 with ***df*** degrees of freedom.

For example, the the distribution of a chi-squared statistic with one degree of freedom is the same as the square of a standard normal distribution, and hence *chi2tail*(1, 1.96^2) = 0.05.

```
. * Statistics > Postestimation > Tests > Likelihood-ratio test
. lrtest _dbpg2 .                                                {20}

Likelihood-ratio test                                 LR chi2(1)  =    137.40
(Assumption: _dbpg2 nested in .)                      Prob > chi2 =    0.0000
```



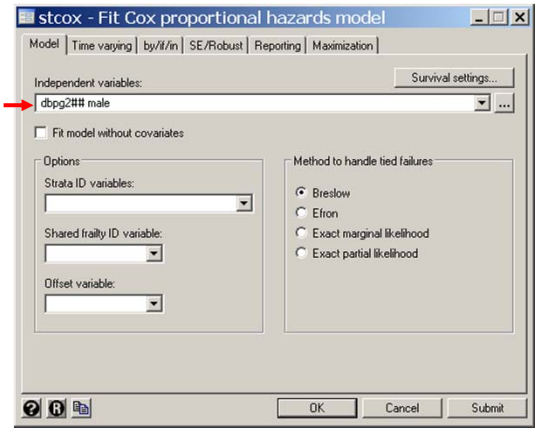
{20} The ***lrtest*** command performs the same change in model deviance calculation that we just did by hand. ***_dbpg2*** is the name that we assigned to the parameter estimates in the model with just the ***i.dbpg2*** covariates. The period refers to the most recent regression command. This command performs the likelihood ratio test associated with the change in model deviance between these two models. It is the responsibility of the user to ensure that these models are nested.

```
. *  Statistics > Postestimation > Manage estimation results > Store in memory
. estimates store dbp_male
```

```
. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox  dbpg2##male                                                          {21}

No. of subjects =         4699                     Number of obs    =      4699
No. of failures =         1473
Time at risk    =  103710.0917
                                                   LR chi2(7)       =    335.16
Log likelihood  =   -11667.275                     Prob > chi2      =    0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dbpg2 |
          90 |   2.608528   .8784348     2.85   0.004     1.348184    5.047099
         110 |   5.410225   1.851724     4.93   0.000     2.766177    10.58159
         111 |   13.58269   5.051908     7.01   0.000     6.552275    28.15654
             |
      1.male |   2.371498   1.117948     1.83   0.067     .9413644    5.974309
             |
  dbpg2#male |
      90 1   |   .8469065    .402857    -0.35   0.727     .3333768    2.151471
     110 1   |   .6818294   .3288495    -0.79   0.427     .2649338    1.754746
     111 1   |   .4017463   .2207453    -1.66   0.097     .1368507    1.179388
------------------------------------------------------------------------------
```

{21} We next add **three** interaction terms,

90.dbp2#1.male = *90.dbp2* × *1.male*,
110.dbp2#1.male = *110.dbp*× *1.male*, and
111.dbp2#1.male = *111.dbp* ×*1.male*.



```
. * Statistics > Postestimation > Tests > Likelihood-ratio test
. lrtest dbp_male .                                                    {22}

Likelihood-ratio test                                 LR chi2(3)  =      9.51
(Assumption: dbp_male nested in .)                    Prob > chi2 =    0.0232
. *  Statistics > Postestimation > Manage estimation results > Store in memory
. estimates store dbp_maleInteract
```

{22} Adding these terms significantly **improves** the model **deviance** with P < **0.023**. Note that the change in deviance has **3** degrees of freedom because we are adding 3 parameters to the model.

```
. lincom 90.dbpg2  + 1.male +  90.dbpg2#1.male, hr                     {23}

 ( 1)  110.dbpg2 + 1.male + 110.dbpg2#1.male = 0

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   5.239064   1.760301     4.93   0.000     2.711777     10.1217
------------------------------------------------------------------------------

. lincom 110.dbpg2 + 1.male + 110.dbpg2#1.male, hr

  ( 1)  110.dbpg2 + 1.male + 110.dbpg2#1.male = 0

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   8.748101   2.974112     6.38   0.000     4.492922     17.0333
------------------------------------------------------------------------------

. lincom 111.dbpg2 + 1.male + 111.dbpg2#1.male, hr

 ( 1) 111.dbpg2 + 1.male + 111.dbpg2#1.male= 0

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   12.94078   5.238924     6.32   0.000     5.852767    28.61274
------------------------------------------------------------------------------
```

{23} This ***lincom*** post estimation command calculates the relative **risk** of a **man** in DBP stratum **2** relative to a **woman** from DBP stratum **1.**

The **hr** option states that the linear combination is to be exponentiated and listed under the heading **Haz. Ratio**

The preceding results allow us to construct the following table:

Table 6.1. Effect of Gender and Baseline DBP on Coronary Heart Disease
Model with all 2-Way Interaction Terms

Diastolic Blood Pressure	Gender			
	Women		Men	
	Relative Risk	95% CI	Relative Risk	95% CI
≤ 60 mm hg	1.0*		2.37	(0.94 - 6.0)
61 - 90 mm hg	2.61	(1.3 - 5.0)	5.24	(2.7 - 10)
91 - 110 mm hg	5.41	(2.8 - 11)	8.75	(4.5 - 17)
> 110 mm hg	13.6	(6.6 - 28)	12.9	(5.9 - 29)

* Denominator of relative risk

Note the pronounced interaction between DBP and sex. These relative risks are consistent with the incidence rates given above.

We next investigate whether age, body mass index, and serum cholesterol confound these results.

7.6.Framingham.ClassVersion.log continues as follows:

```
. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox  dbpg2##male age                                                {1}

No. of subjects =         4699                     Number of obs   =      4699
No. of failures =         1473
Time at risk    =  103710.0917
                                                                    612.05
Log likelihood  =   -11528.829                                      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dbpg2 |
          90 |   2.129403   .7175801     2.24   0.025     1.100055    4.121937
         110 |   3.289324    1.12979     3.47   0.001     1.677811    6.448672
         111 |    8.04123   2.999755     5.59   0.000     3.870656    16.70554
             |
      1.male |   2.119083   .9990903     1.59   0.111      .841065    5.339081
             |
  dbpg2#male |
       90 1  |   .9753056    .464017    -0.05   0.958     .3838559    2.478068
      110 1  |    .984806   .4754774    -0.03   0.975      .382278     2.53701
      111 1  |   .5050973   .2776141    -1.24   0.214      .172002    1.483258
             |
         age |   1.056687   .0034809    16.74   0.000     1.049886    1.063531
------------------------------------------------------------------------------
```

{1} We first add ***age*** to the model.

```
. * Statistics > Postestimation > Tests > Likelihood-ratio test
. lrtest dbp_maleInteract .                                             {2}

Likelihood-ratio test                                 LR chi2(1)  =    276.89
(Assumption: dbp_maleInte~t nested in .)              Prob > chi2 =    0.0000

. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox  dbpg2##male age if !missing(bmi) & !missing(scl)  {3}
. *  Statistics > Postestimation > Manage estimation results > Store in memory
. estimates store dbp_maleInteract_age
```

{2} The improvement to the model **deviance** has overwhelming statistical significance.

{3} Some patients have missing values of ***bmi*** and ***scl***. These patients will be excluded from our next model that included these variables. In order to keep the next model nested within the last we refit the last model excluding patients with missing values of ***bmi*** and ***scl***. This will ensure that the same patients are in both models, that the models are properly nested, and that our next likelihood ratio test is valid.

```
. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox  i.dbpg2##male age bmi scl

                                                LR chi2(10)     =    736.95
Log likelihood  =   -11390.412                  Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dbpg2 |
          90 |   1.708285   .5771462     1.58   0.113     .8810103    3.312377
         110 |   2.198904   .7613688     2.28   0.023     1.115522    4.334451
         111 |   5.166759    1.94896     4.35   0.000     2.466808    10.82184
             |
      1.male |    1.97694    .932211     1.45   0.148     .7845418    4.981626
             |
  dbpg2#male |
       90 1  |   1.052562   .5009358     0.11   0.914     .4141362    2.675173
      110 1  |    1.16722   .5641426     0.32   0.749     .4526355    3.009933
      111 1  |   .6184658   .3403661    -0.87   0.383     .2103129    1.818718
             |
         age |   1.049249   .0035341    14.27   0.000     1.042345    1.056198
         bmi |   1.040017   .0069042     5.91   0.000     1.026572    1.053637
         scl |    1.00584   .0005845    10.02   0.000     1.004695    1.006986
------------------------------------------------------------------------------
```

```
. * Statistics > Postestimation > Tests > Likelihood-ratio test
. lrtest dbp_maleInteract_age .

Likelihood-ratio test                                 LR chi2(2)  =    132.73
(Assumption: dbp_maleInte~e nested in .)              Prob > chi2 =    0.0000 {4}
```

{4} Adding BMI and serum cholesterol greatly improves the model fit.

The parameters from the preceding model can be converted into a relative risk table in the same way as Table 6.1. This table follows:

Table 6.2. Effect of Gender and Baseline DBP on Coronary Heart Disease
Model with all 2-Way Interaction Terms

Diastolic Blood Pressure	Gender			
	Women		Men	
	Relative Risk†	95% CI	Relative Risk	95% CI
60 mm hg	1.0*		1.98	(0.78 - 5.0)
61 - 90 mm hg	1.71	(0.88 - 3.3)	3.55	(1.8 - 6.9)
91 - 110 mm hg	2.19	(1.1 - 4.3)	5.07	(2.6 - 10)
> 110 mm hg	5.17	(2.5 - 11)	6.32	(2.8 - 14)

* Denominator of relative risk
†Adjusted for Age BMI and Serum Cholesterol

```
. lincom  90.dbpg2  + 1.male +  90.dbpg2#1.male, hr

  ( 1)  90.dbpg2 + 1.male + 90.dbpg2#1.male = 0

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   3.554688   1.197825     3.76   0.000     1.836419     6.88068
------------------------------------------------------------------------------

. lincom  110.dbpg2 + 1.male + 110.dbpg2#1.male, hr

 ( 1)  110.dbpg2 + 1.male + 110.dbpg2#1.male = 0

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   5.074023   1.735763     4.75   0.000     2.595174    9.920611
------------------------------------------------------------------------------

. lincom  111.dbpg2 + 1.male + 111.dbpg2#1.male, hr

 ( 1)  111.dbpg2 + 1.male + 111.dbpg2#1.male = 0

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    6.31724   2.572047     4.53   0.000     2.844219     14.0311
------------------------------------------------------------------------------
```

Comparing these tables shows that the adjusted risks of DBP and sex on CHD are far less than the crude risks. Our analyses show that age, BMI and serum cholesterol are CHD risk factors in their own right which are positively correlated with DBP and sex and hence inflate the apparent effects of these risk factors on CHD.

Diastolic Blood Pressure	Gender			
	Women		Men	
	Relative Risk	95% CI	Relative Risk	95% CI
Unadjusted				
60 mm hg	1.0		2.37	(0.94 - 6.0)
61 - 90 mm hg	2.61	(1.3 - 5.0)	5.24	(2.7 - 10)
91 - 110 mm hg	5.41	(2.8 - 11)	8.75	(4.5 - 17)
> 110 mm hg	13.6	(6.6 - 28)	12.9	(5.9 - 29)
Adjusted for Age BMI and Serum Cholesterol				
60 mm hg	1.0		1.98	(0.78 - 5.0)
61 - 90 mm hg	1.71	(0.88 - 3.3)	3.55	(1.8 - 6.9)
91 - 110 mm hg	2.19	(1.1 - 4.3)	5.07	(2.6 - 10)
> 110 mm hg	5.17	(2.5 - 11)	6.32	(2.8 - 14)

The preceding example covers the following topics…

c) Interaction terms in hazard regression models

See also Chapter IV, Section 14 on logistic regression analysis.

d) Estimating the joint effects of two risk factors on a relative risk

See also Chapter IV, Sections 13 and 14 on logistic regression.

e) Calculating 95% CIs for relative risks derived from multiple parameter estimates.

See also Chapter IV, Section 10 on logistic regression, respectively.

f) Adjusting for confounding variables

See also Chapter II, Sections 2 and 6 on linear regression.

3. Restricted Cubic Splines and Survival Analysis

Restricted cubic splines can be used in much the same way as for linear or logistic regression. Suppose that x_i is a continuous covariate of interest. Then a k knot model gives covariates

$$x_{i1}, x_{i2}, \ldots, x_{i,k-1}$$

The relative risk of a patient with covariate x_i compared to covariate x_j is

$$\frac{\lambda_0[t]\exp\left[x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{i,k-1}\beta_{k-1}\right]}{\lambda_0[t]\exp\left[x_{j1}\beta_1 + x_{j2}\beta_2 + \cdots + x_{j,k-1}\beta_{k-1}\right]}$$

$$= \exp\left[\left(x_{i1} - x_{j1}\right)\beta_1 + \left(x_{i2} - x_{j2}\right)\beta_2 + \cdots \left(x_{2,k-1} - x_{1k-1}\right)\beta_{k-1}\right]$$

We can directly estimate the log relative risk

$$\left(x_{i1} - x_{j1}\right)\beta_1 + \left(x_{i2} - x_{j2}\right)\beta_2 + \cdots + \left(x_{i,k-1} - x_{jk-1}\right)\beta_{k-1} \qquad \{6.1\}$$

However, we also wish to calculate confidence intervals for relative risks. Stata does not provide a *predict* post-estimation command to do this directly.

Suppose that the reference value of x_j is less than the first knot. Let this value be c.

Let $y_i = x_i - c$ and $y_{ij} = x_{ij} - c$ be the analogous spline covariates for y_i

Then when $x_j = c$ we have $y_{i1} = y_i = 0$, and $y_{j2} = y_{j3} = \cdots = y_{j,k-1} = 0$ because 0 is smaller than the smallest y-knot. Hence,

{6.1} can be rewritten

$$y_{i1}\beta_1 + y_{i2}\beta_2 + \cdots + y_{i,k-1}\beta_{k-1}$$

which is the linear predictor of the model as well as the log relative risk of interest. Regressing survival against y_i allows us to use Stata's post estimation commands to calculate 95% confidence bands for relative risks.

N.B. If it is difficult or inconvenient to make the model's linear predictor equal the desired log relative risk then we could always use the *predictnl* postestimation command to calculate the log relative risk and its associated standard error.

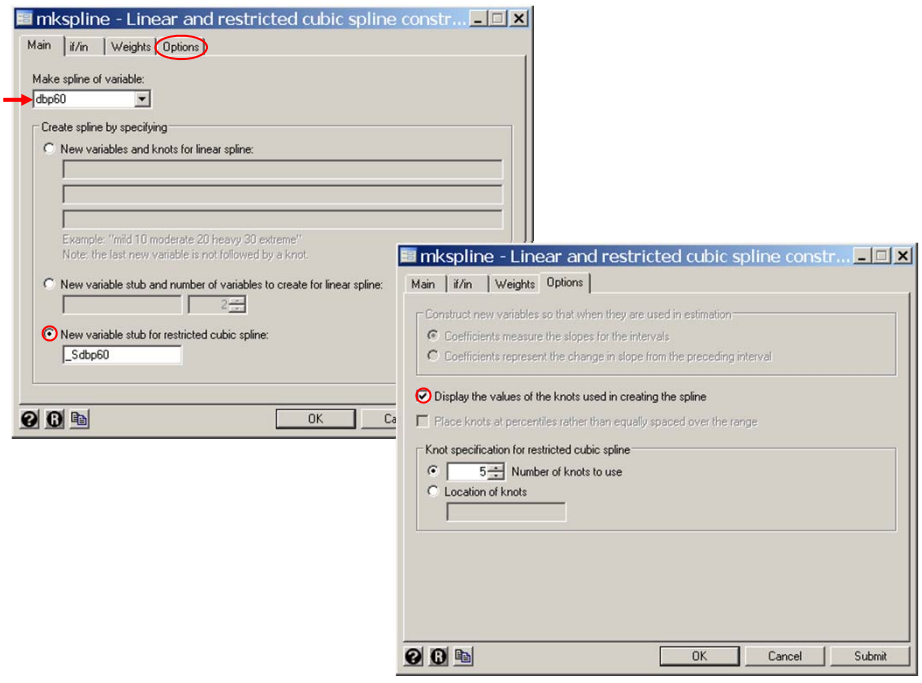
4. Fitting a Cubic Spline Model for the effect of DBP on CHD

```
. *  Framingham.Spline.log
. *
. *  Proportional hazards regression analysis of the effect of gender and
. *  baseline diastolic blood pressure (DBP) on coronary heart disease (CHD)
. *  Use restricted cubic splines to model the effect of DBP on CHD risk.
. *  We will use a DBP of 60 as the denominator of our relative risk estimates.
. *
. use C:\WDDtext\2.20.Framingham.dta, clear
. generate time= followup/365.25
. label variable time "Follow-up in Years"
. * Statistics > Survival... > Setup... > Declare data to be survival...
. stset time, failure(chdfate)
                                                                {Output omitted}
. sort dbp
. generate dbp60 = dbp - 60                                                {1}
. * Data > Create... > Other variable-creation... > linear and cubic...
. mkspline _Sdbp60 = dbp60, cubic displayknots                             {2}

             |     knot1      knot2      knot3      knot4      knot5
-------------+--------------------------------------------------------
       dbp60 |         4         14         20       29.5         45
```

{1} Note that ***dbp60*** = 0 when DBP = 60

{2} Calculate cubic spline covariates for the default 5 knot model. Note that the biggest knot is at DBP = 60+45 = 105 which is well below the extreme observed blood pressures. Note also that the smallest knot is at DBP = 64 > 60. This means that when DBP = 60, all of the spline covariates will equal 0.

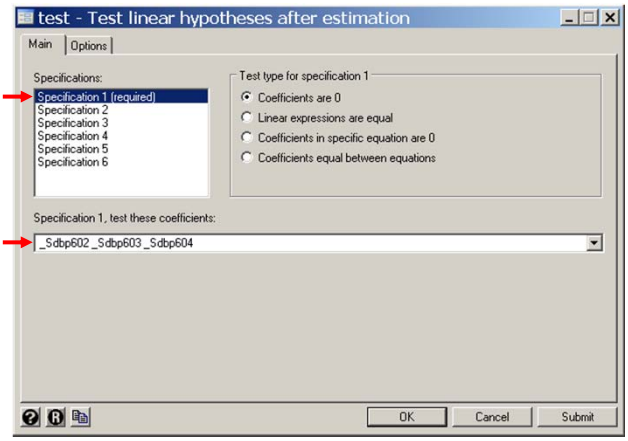
This command generates spline covariates named *_Sdbp601*, *_Sdbp602*, *_Sdbp603*, and *_Sdbp604*.



```
. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox  _S*, nohr                                                        {3}
                                                          {Output omitted}
No. of subjects =         4699                     Number of obs    =      4699
No. of failures =         1473
Time at risk    =  103710.0917
                                                   LR chi2(4)       =    246.93
Log likelihood  =   -11711.393                     Prob > chi2      =    0.0000
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Sdbp601 |   .0618603    .016815     3.68   0.000     .0289035     .094817
    _Sdbp602 |  -.2268319   .1120642    -2.02   0.043    -.4464737   -.0071902
    _Sdbp603 |     .93755   .4547913     2.06   0.039     .0461754    1.828925
    _Sdbp604 |   -.982937   .4821521    -2.04   0.041    -1.927938   -.0379362
------------------------------------------------------------------------------

. * Statistics > Postestimation > Tests > Test linear hypotheses
. test _Sdbp602 _Sdbp603 _Sdbp604                                         {4}

 ( 1)  _Sdbp602 = 0
 ( 2)  _Sdbp603 = 0
 ( 3)  _Sdbp604 = 0

           chi2(  3) =    4.66
         Prob > chi2 =    0.1984
```

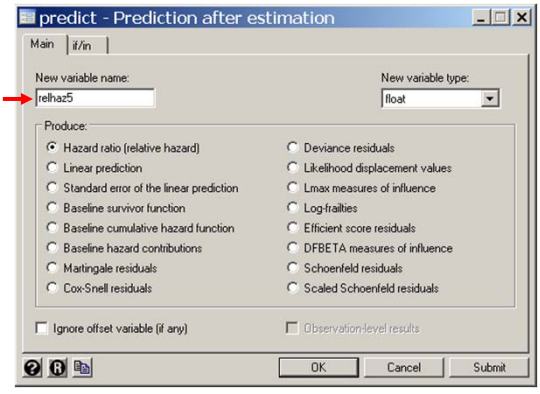
{3} Do a proportional hazards regression of CHD morbidity against the spline covariates. The ***nohr*** option causes the parameter estimates to be displayed.

{4} Test if the second, third and fourth spline covariates are all zero. That is, test the hypothesis that the relationship between DBP and log relative risk is linear. This hypothesis can not be rejected (P = 0.20)




```
. predict relhaz5, hr                                                    {5}
```

{5} Define relhaz5 to equal the exponentiated linear predictor for this model. That is, relhaz5 is the log relative hazard compared with a patient whose DBP = 60.



```
. *
. *  Experiment with fewer knots
. *
. *  Variables Manager
. drop _S*
. * Data > Create... > Other variable-creation... > linear and cubic...
. mkspline _Sdbp60 = dbp60, nknots(3) cubic displayknots                    {6}

             |     knot1      knot2      knot3
-------------+---------------------------------
       dbp60 |         8         20         40

. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox  _S*, nohr
                                                                  {Output omitted}
Log likelihood  =   -11713.643                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Sdbp601 |   .0347213   .0057337     6.06   0.000     .0234835    .0459592
    _Sdbp602 |  -.0041479   .0070762    -0.59   0.558    -.0180169    .0097212 {7}
------------------------------------------------------------------------------

. predict relhaz3, hr                                                       {8}
```

{6} Calculate spline covariates for three knots at their default locations

{7} The second spline covariate is not significantly different from zero. This means we cannot reject the model with dbp60 as the only raw covariate.

{8} *relhaz3* is the relative hazard for CHD associated with DBP from this model.

```
. *
. *  How about no knots?
. *
. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox dbp60
                                                                {Output omitted}
Log likelihood  =   -11713.816                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dbp60 |   1.032064   .0019926    16.35   0.000     1.028166    1.035977
------------------------------------------------------------------------------

. predict relhaz0, hr                                                    {9}
. *  Variables Manager
. drop _S*

. summarize dbp60, detail

                          dbp60
-------------------------------------------------------------
      Percentiles      Smallest
 1%           -2            -20
 5%            4            -12
10%            8            -10       Obs                4699
25%           14            -10       Sum of Wgt.        4699

50%           20                      Mean             22.5416
                        Largest       Std. Dev.        12.73732
75%           30             80
90%           40             82       Variance         162.2394
95%           45             84       Skewness         .6941674          {10}
99%           60             88       Kurtosis         4.147346
```

{9} *relhaz0* is the relative hazard for CHD associated with DBP from this model.

{10} 5% of the observations are greater than *dbp60* = 45 or DBP = 105. The largest observation is DBP = 88 + 60 = 148. Hence, our model may be going wrong for very high blood pressures even though we cannot reject the single covariate model. Lets experiment with a 3 knot model with a higher value of the last knot.

```
. *
. *  Add a knot at DBP60 = 60 and remove the knot at DBP60 = 8
. *
. * Data > Create... > Other variable-creation... > linear and cubic...
. mkspline _Sdbp60 = dbp60, knots(20 40 60) cubic displayknots

             |     knot1      knot2      knot3
-------------+------------------------------------
       dbp60 |        20         40         60

. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox  _S*, nohr
                                                              {Output omitted}
Log likelihood  =   -11713.127                     Prob > chi2     =    0.0000 {11}

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Sdbp601 |   .0342387   .0030075    11.38   0.000     .0283442    .0401333
    _Sdbp602 |  -.0063964   .0055413    -1.15   0.248    -.0172571    .0044642 {12}
------------------------------------------------------------------------------

. predict relhaz3a, hr
```

{11} The log likelihood increases by a modest 0.69.

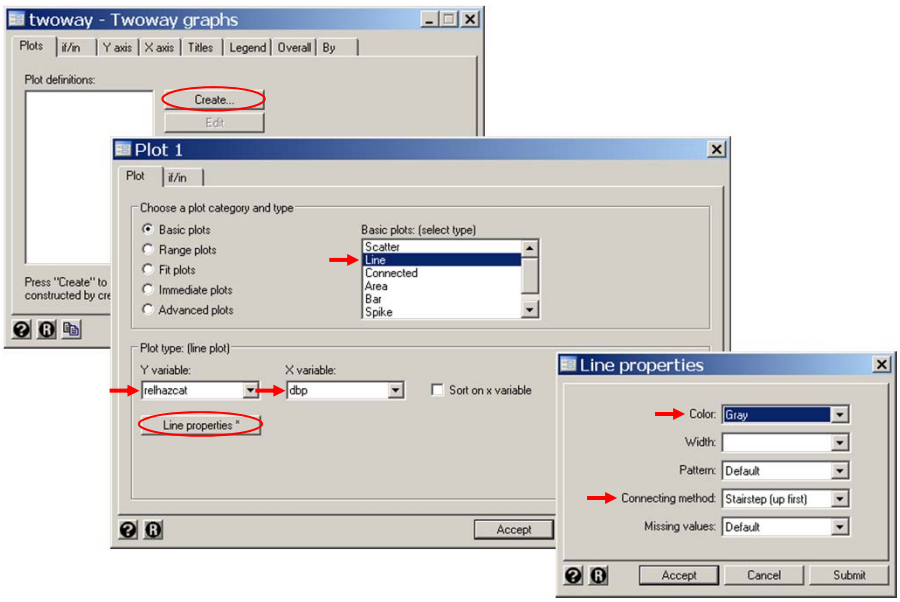
{12} The second spline covariate is not significantly different from zero.

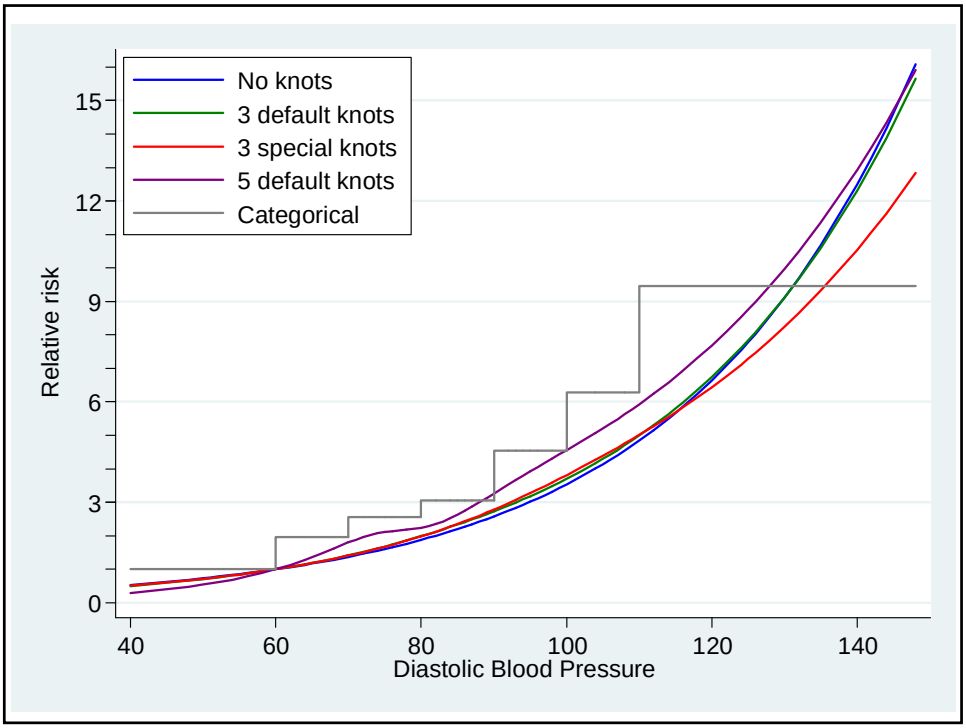
```
. *
. *  Calculate the relative hazard from model 7.12 in the text
. *
. generate relhazcat = 1

. replace relhazcat = 1.97 if dbp > 60
(4549 real changes made)

. replace relhazcat = 2.56 if dbp > 70
(3775 real changes made)

. replace relhazcat = 3.06 if dbp > 80
(2308 real changes made)

. replace relhazcat = 4.54 if dbp > 90
(1041 real changes made)

. replace relhazcat = 6.29 if dbp > 100
(340 real changes made)

. replace relhazcat = 9.46 if dbp > 110
(105 real changes made)
```

```
. *
. *  Plot relative hazards estimated so far
. *
. line relhaz0 relhaz3 relhaz3a relhaz5 dbp                ///
>     , color(blue green red purple)                       ///
>     || line relhazcat dbp, connect(stepstair) color(gray) ///  {13}
>     , legend(ring(0) position(11) col(1)                 ///
>         order(1 "No knots"  2 "3 default knots"          ///
>           3 "3 special knots" 4 "5 default knots"        ///
>           5 "Categorical")) ytitle(Relative risk)        ///
>       ytick(1(1)16) ylabel(0(3)15, angle(0))
```

{13} The *connect(stepstair)* option joins two consecutive points by rising or falling vertically from the first to the second y value and then moving horizontally to the second x value.

Dialogue boxes for drawing a step-stair graph



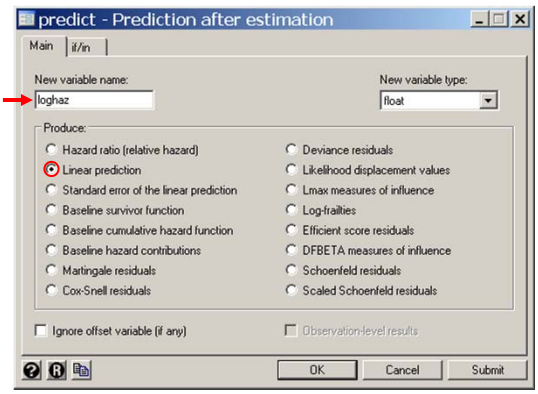


Note that the categorical model has all patients with a DBP $\leq$ 60 in the denominator of the relative risk while for the other models this denominator is patients with DBP = 60. This explains why the categorical relative risks are higher than the risks for the other models.

The no knot and default 3 knot models are in remarkably close agreement. The 3 special knot model agrees with the other two up to DBP = 120 and then gives lower risks. The no knot model may overestimate relative risks associated with extreme DBPs.

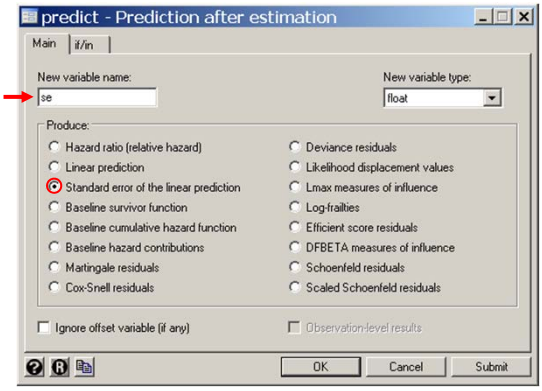
```
. predict loghaz, xb                                                {14}
```

{14} *loghaz* is the linear predictor for the 3 special knot model. It is also the log relative risk.




```
. predict se, stdp                                                  {15}
```

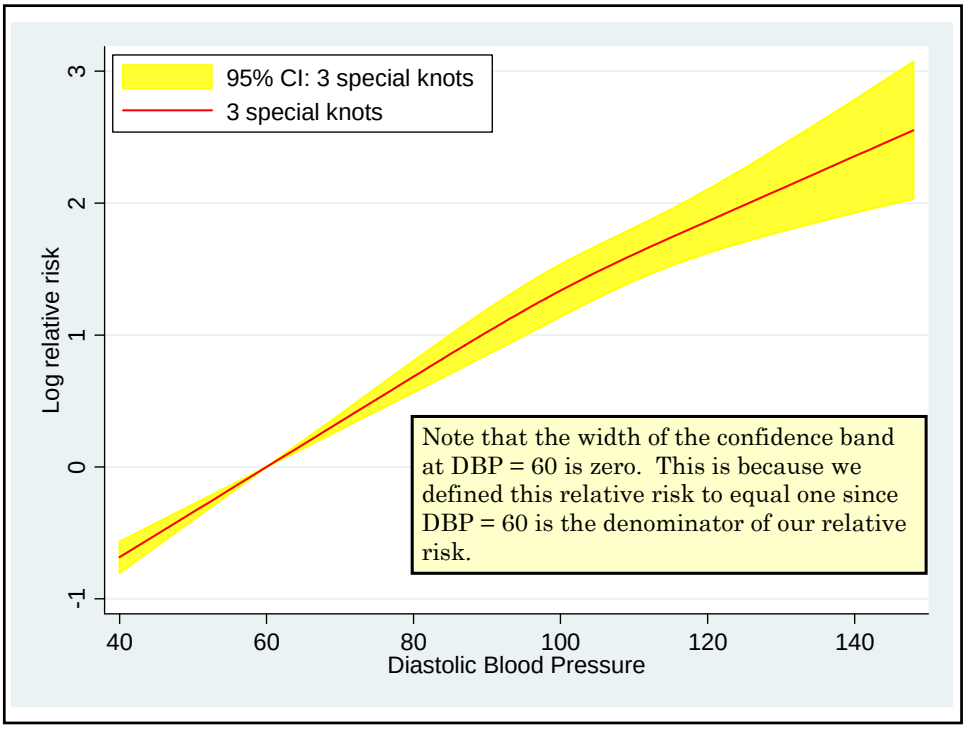
{15} *se* is the standard error of *loghaz*.



```
. generate logcil = loghaz - 1.96*se                                    {16}

. generate logciu = loghaz + 1.96*se                                    {16}

. twoway rarea logcil logciu dbp, color(yellow)                  /// {17}
>     || line loghaz dbp, color(red)                             ///
>     , legend(ring(0) position(11) col(1)                       ///
>         order(1 "95% CI: 3 special knots"                      ///
>         2 "3 special knots")) ytitle(Log relative risk)
```

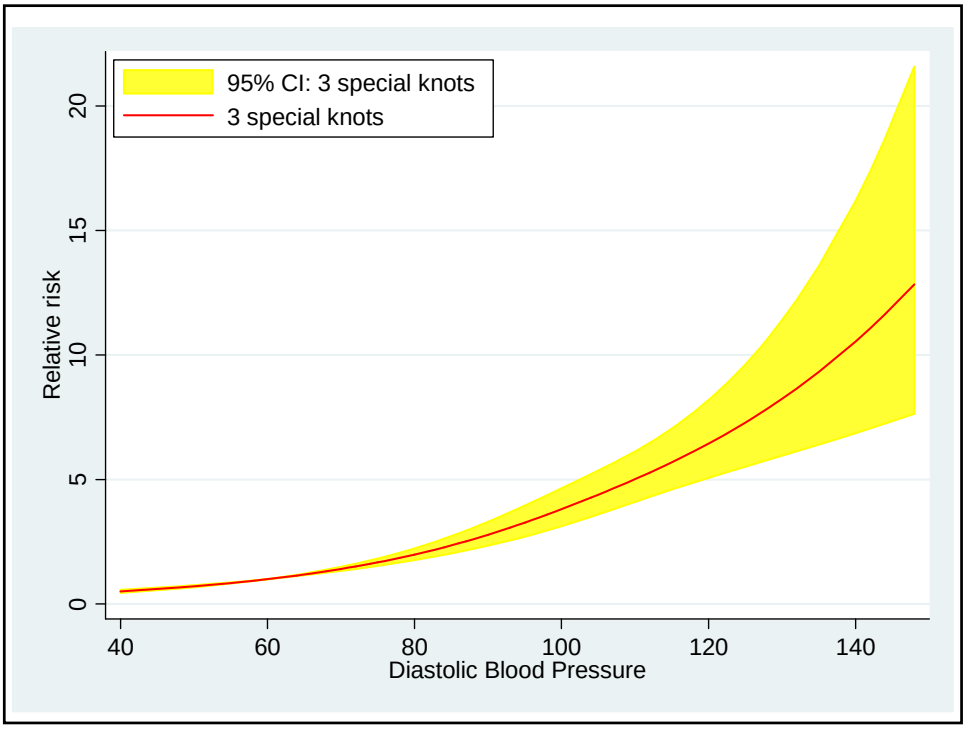
{16} *logcil* and *logciu* are the 95% confidence bands for *loghaz*.

{17} Plot the log relative risk of CHD together with its 95% confidence band.



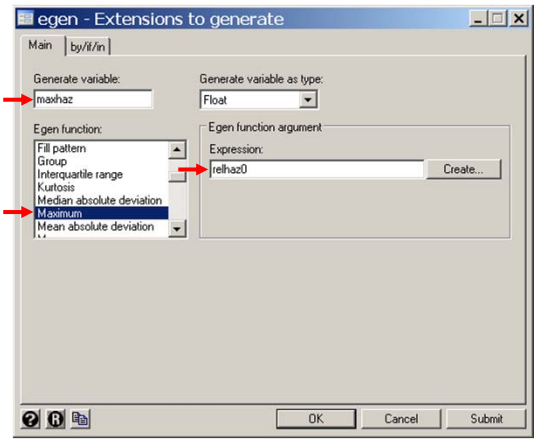
```
. generate cil3a = exp(logcil)

. generate ciu3a = exp(logciu)

. twoway rarea cil3a ciu3a dbp, color(yellow)                ///  {18}
>     || line relhaz3a dbp, color(red)                       ///
>     , legend(ring(0) position(11) col(1)                   ///
>         order(1 "95% CI: 3 special knots"                  ///
>           2 "3 special knots" )) ytitle(Relative risk)
```

{18} Lets repeat the previous graph on the linear scale.



```
. *
. *  Plot results from the no knot model and the preceding
. *  model together.  Truncate the upper error bounds.
. *
. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox dbp60
. *  Variables Manager
. drop loghaz se logcil logciu

. predict loghaz, xb

. predict se, stdp

. generate logcil = loghaz - 1.96*se

. generate logciu = loghaz + 1.96*se

. generate cil0 = exp(logcil)

. generate ciu0 = exp(logciu)

. * Data > Create or change data > Create new variable (extended)
. egen maxhaz = max(relhaz0)                                    {19}
```

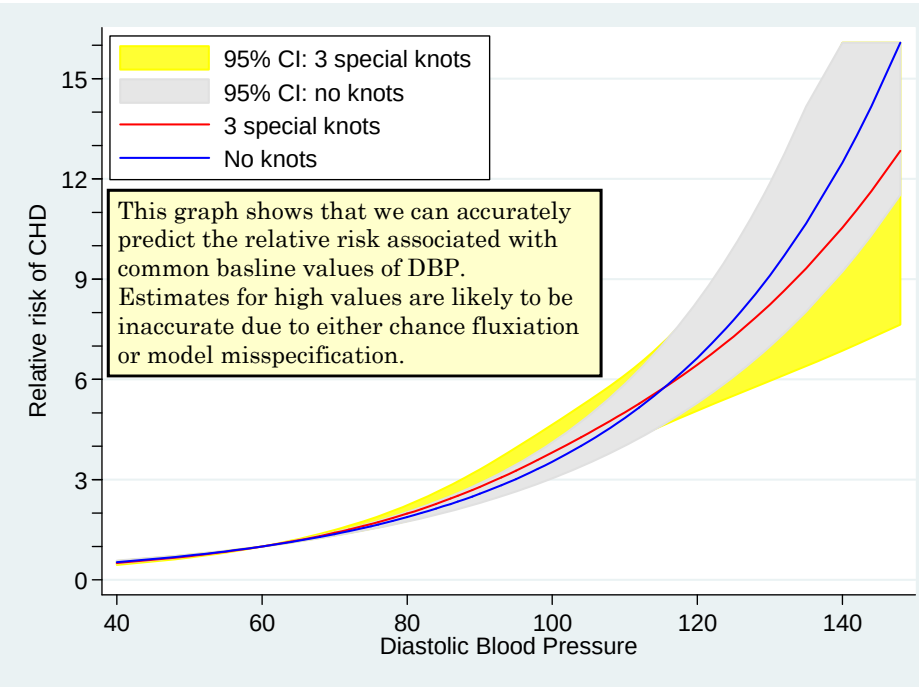
{19} This command defined *maxhaz* to equal the maximum value of *relhaz0* in the entire data set.



```
. generate ciu3a_chop =  min(ciu3a,maxhaz)                    {20}
. generate ciu0_chop =   min(ciu0,maxhaz)
. twoway rarea cil3a ciu3a_chop dbp, color(yellow)            ///
>     || rarea cil0 ciu0_chop dbp, color(gs14)                ///
>     || line relhaz3a dbp, color(red)                        ///
>     || line relhaz0  dbp, color(blue)                       ///
>     , legend(ring(0) position(11) col(1)                    ///
>         order(1 "95% CI: 3 special knots"                   ///
>           2 "95% CI: no knots" 3 "3 special knots"          ///
>           4 "No knots")) ytitle(Relative risk of CHD)       ///
>       ytick(1(1)16) ylabel(0(3)15, angle(0))
```

{20} *ciu3a_chop* is the upper bound of the confidence interval for the 3 special knot model truncated at *maxhaz*.

Plot the relative risks and confidence bands from both models together.



```
. *
. *  In our final graphs we will want to truncate the upper
. *  error bands at the top of the graph.  This can cause
. *  linear extrapolation errors due to sparse blood pressures
. *  at the extreme upper range.  To correct this we add
. *  dummy records to fill in some of these blood pressures.
. *
. set obs 4739                                                  {21}
obs was 4699, now 4739

. replace dbp = 135 +(_n - 4699)*0.1 if _n > 4699               {22}
(40 real changes made)

. replace dbp60 = dbp - 60
(40 real changes made)

. sort dbp
. *  Variables Manager
. drop loghaz se logciu maxhaz ciu0

. predict loghaz, xb

. predict se, stdp

. generate logciu = loghaz +1.96*se

. generate ciu0 = exp(logciu)
. * Data > Create or change data > Create new variable (extended)
. egen maxhaz = max(relhaz0)

. replace ciu0_chop =   min(ciu0,maxhaz)                        {23}
(40 real changes made)
```

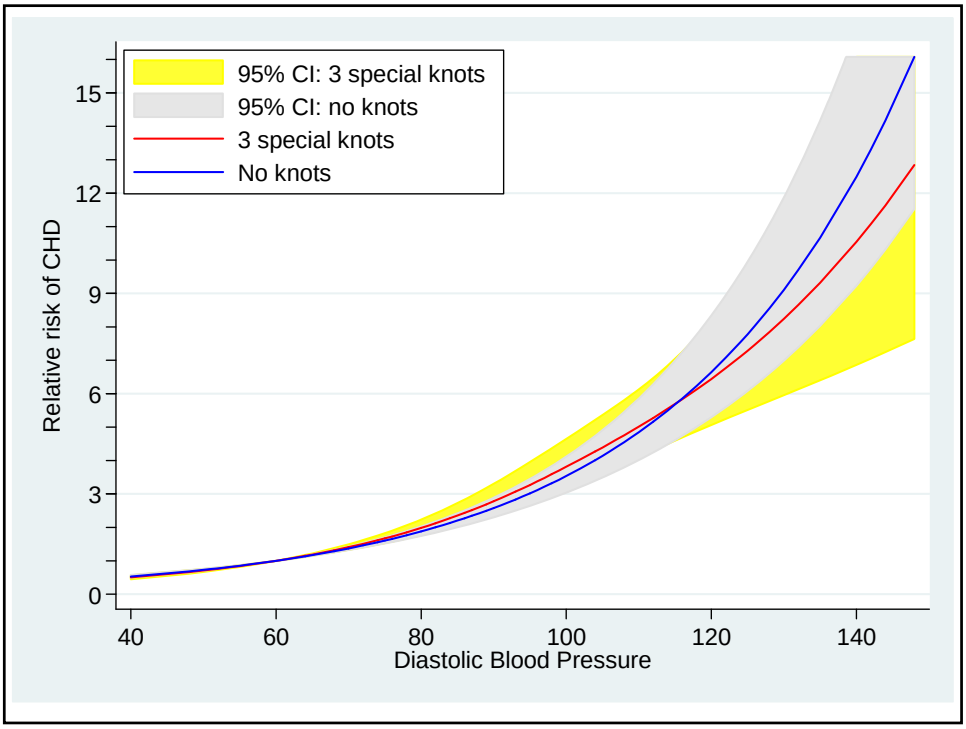
{21} Increase the number of records in the data set to 4739 by adding 40 dummy records. All of the variables in these records will be missing.

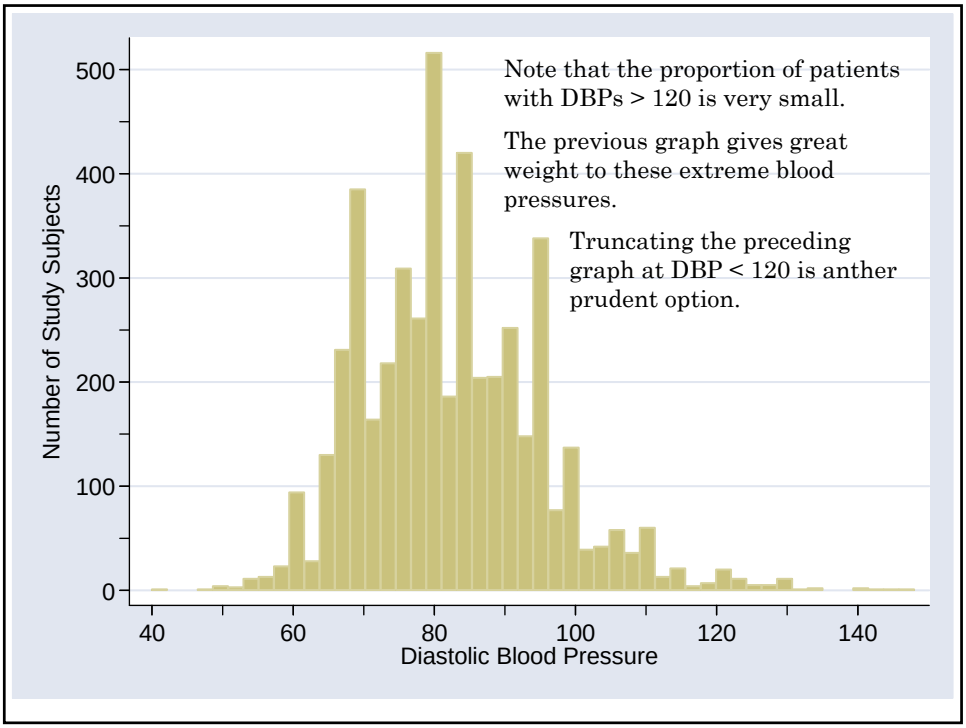
{22} There are no real blood pressures observed between 135 and 140. In these new records define *dbp* to range from 135.1 to 139 in increments of 0.1

{23} Define the upper confidence bound of the no knot model for these dummy records.

```
. twoway rarea ci13a ciu3a_chop dbp, color(yellow)          ///
>     || rarea ci10 ciu0_chop dbp, color(gs14)              ///
>     || line relhaz3a dbp, color(red)                      ///
>     || line relhaz0  dbp, color(blue)                     ///
>     , legend(ring(0) position(11) col(1)                  ///
>         order(1 "95% CI: 3 special knots"                 ///
>           2 "95% CI: no knots" 3 "3 special knots"        ///
>           4 "No knots")) ytitle(Relative risk of CHD)     ///
>       ytick(1(1)16) ylabel(0(3)15, angle(0))
```

Repeat the previous plot.





5. Stratified Proportional Hazard Regression Models

One way to weaken the proportional hazards assumption is to subdivide the patients into $j = 1, \ldots, J$ strata defined by the patient's covariates. We then define the hazard for the i^{th} patient from the j^{th} stratum at time t to be

$$\lambda_{ij}[t] = \lambda_{0j}[t]\exp\left[\beta_1 x_{ij1} + \beta_2 x_{ij2} + \ldots + \beta_q x_{ijq}\right] \qquad \{6.3\}$$

where $x_{ij1}, x_{ij2}, \ldots, x_{ijq}$ are the covariate values for this patient, and

$\lambda_{0j}[t]$ is the baseline hazard for patients from the j^{th} stratum.

Model 6.3 makes no assumptions about the shapes of the J baseline hazard functions. Within each strata the proportional hazards assumption applies. However, patients from different strata need not have proportional hazards.

For example, suppose that we were interested in the risk of CHD due to smoking in women and men. We might stratify the patients by gender, letting $j = 1$ or 2 designate men or women, respectively. Let

$$x_{ij} = \begin{cases} 1: \text{ if } i^{th} \text{ patient from } j^{th} \text{ stratum smokes} \\ 0: \text{ otherwise} \end{cases}, \text{ and}$$

$\lambda_{ij}[t]$ be the CHD hazard for the i^{th} patient from the j^{th} stratum.

Then Model 6.3 reduces to

$$\lambda_{ij}[t] = \lambda_{0j}[t]\exp[\beta x_{ij}] \qquad \{6.4\}$$

Model 6.4 makes no assumptions about how CHD risk varies with time among non-smoking men or women. It does, however, imply that the relative CHD risk of smoking is the same among men is it is among women.

The within strata relative risk of CHD in smokers relative to non-smokers is e^{β}. That is, smoking women have e^{β} times the CHD risk of non-smoking women while smoking men have e^{β} times the CHD risk of non-smoking men.

In this model $\lambda_{01}[t]$ and $\lambda_{02}[t]$ represent the CHD hazard for men and women who do not smoke, while $\lambda_{01}[t]e^{\beta}$ and $\lambda_{02}[t]e^{\beta}$ represents this hazard for men and women who do.

In Stata, a stratified proportional hazards model is indicated by the *strata(varnames)* option of the *stcox* command. Model {6.4} might be implemented by a command such as

stcox *smoke*, strata(*sex*)

where *smoke* = 1 or 0 for patients who did or did not smoke, respectively.

6. Survival Analysis with Ragged Study Entry

Usually the time variable in a survival analysis measures follow-up time from some event. This event may be recruitment into a cohort, diagnosis of cancer, et cetera. In such studies everyone is at risk at time zero, when they enter the cohort.

Sometimes, however, we may wish to use the patient's age as the time variable rather than follow-up time. Both Kaplan-Meier survival curves and hazard regression analyses can be easily adapted to this situation. The key difference is that when age is the time variable, patients are not at risk of failure until they reach the age at which they enter the cohort. Hence, no one may be at risk at age zero, and subjects will enter the analysis at different "times" when they reach their age at recruitment.

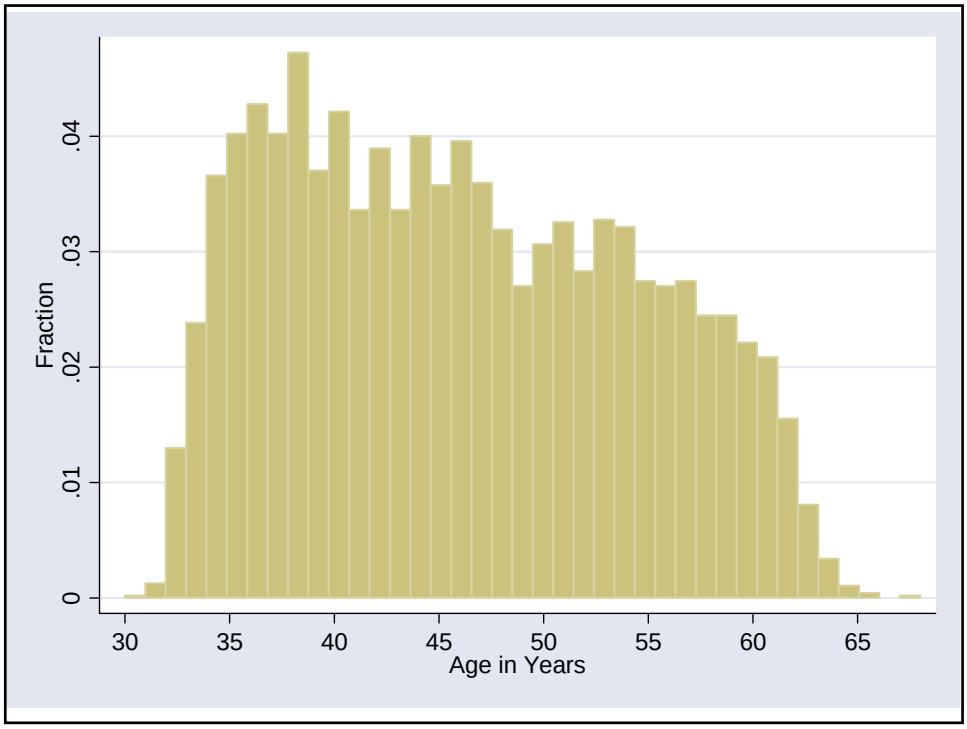
These analyses must be interpreted as the effect of age and other covariates on the risk of failure conditioned on the fact that each patient had not failed prior to her age of recruitment.

a) Example: Kaplan-Meier Survival Curves as a Function of Age

```
. *  Framingham.age.log
. *
. *  Plot Kaplan-Meier cumulative CHD morbidity curves as a function of age.
. *  Patients from the Framingham Heart Study enter the analysis when they
. *  reach the age of their baseline exam.
. *
. use C:\WDDtext\2.20.Framingham.dta, clear

. * Graphics > Histogram
. histogram age, bin(39) fraction ylabel(0(.01).04) xlabel(30(5)65)          {1}
(bin=39, start=30, width=.97435897)

. generate time= followup/365.25

. label variable time "Follow-up in Years"
```

{1} The **age** of study subjects at recruitment in the Framingham Heart Study ranged from **30** to **68** years.

In this **histogram** command, fraction indicates that the *y*-axis is to be the proportion of patients at each age.




```
. generate exitage = time + age                                          {2}

. label variable exitage Age
. * Statistics > Survival... > Setup... > Declare data to be survival...
. stset exitage, enter(time age) failure(chdfate)                        {3}

    failure event:  chdfate != 0 & chdfate < .
obs. time interval:  (0, exitage]
 enter on or after:  time age
 exit on or before:  failure

------------------------------------------------------------------------------
     4699  total obs.
        0  exclusions
------------------------------------------------------------------------------
     4699  obs. remaining, representing
     1473  failures in single record/single failure data
 103710.1  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =        30
                                  last observed exit t =        94
```

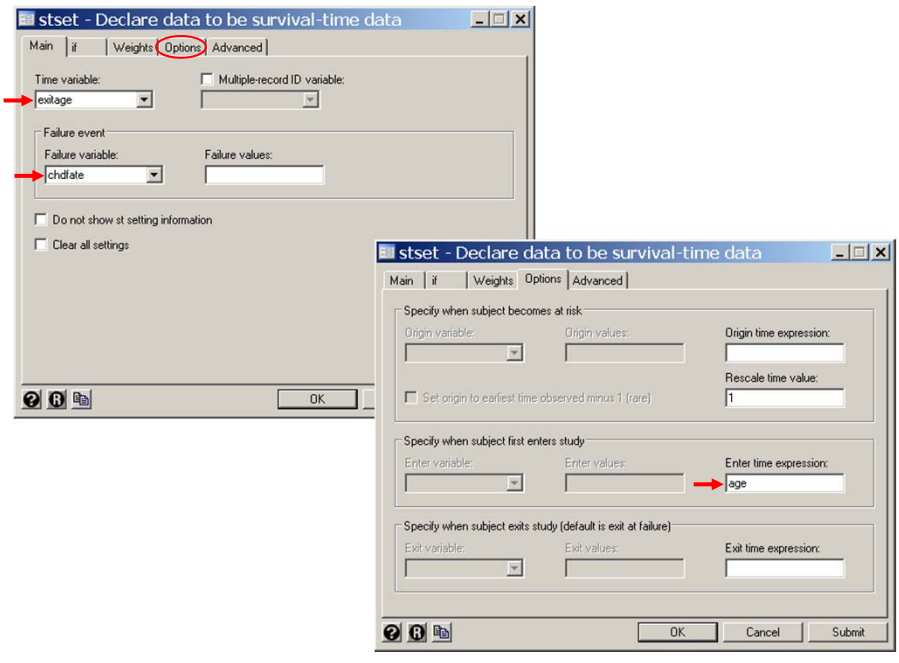
{2} We define ***exitage*** to be the patient's age at exit.

{3} This command changes the survival-time variable from time since recruitment to age.

exitage is the patient's time of exit. That is, it is the time (age) when the subject either suffers CHD or is censored.

chdfate is the subject's fate at exit.

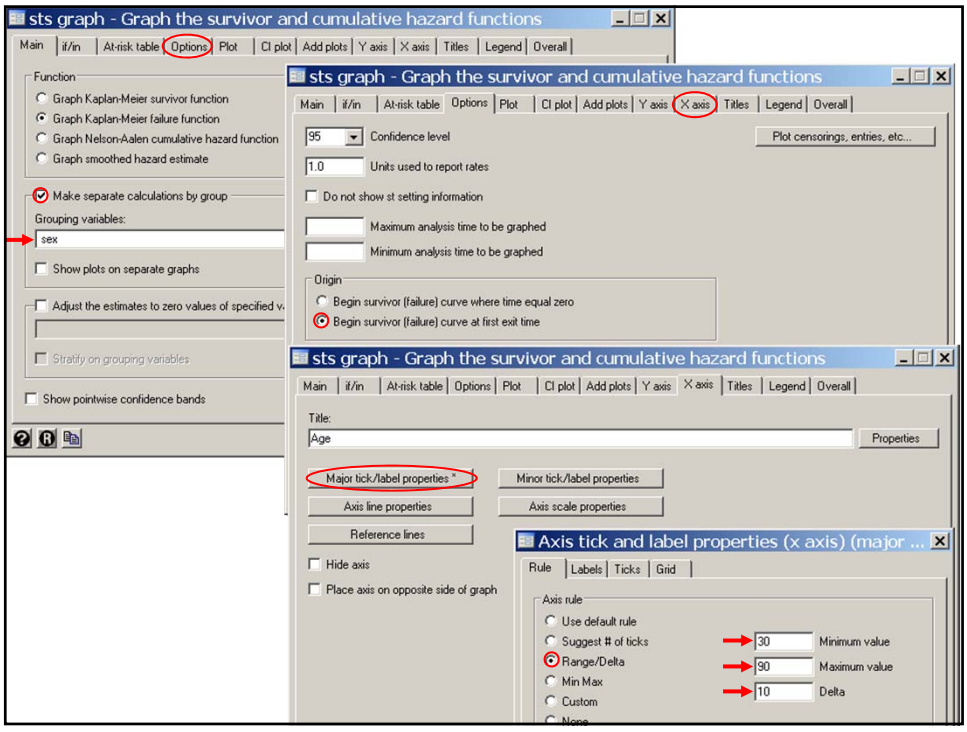
enter(time age) defines age to be the patient's entry time. That is, patients enter the analysis when they reach the age of their baseline exam. We know that all patients were free from CHD at that time.

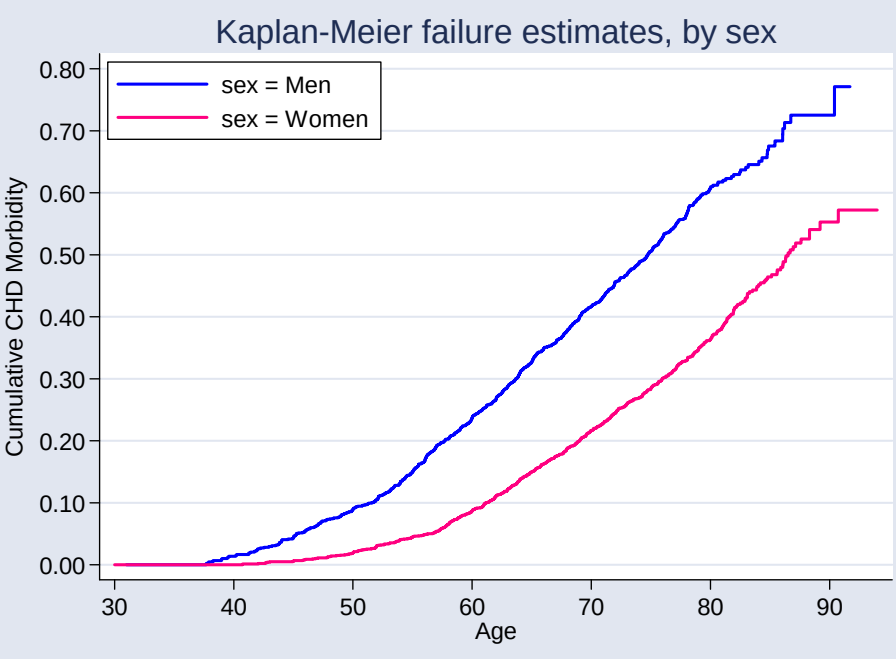


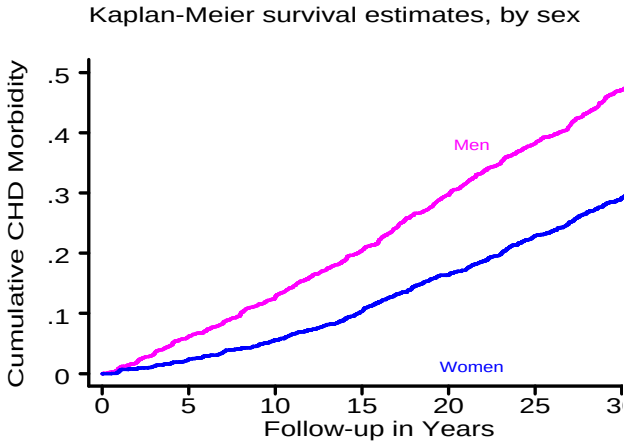
```
. * Graphics > Survival analysis graphs > Kaplan-Meier failure function
. sts graph, by(sex) failure ytitle(Cumulative CHD Morbidity) xtitle(Age) /// {4}
>     ylabel(0(.1).8, angle(0)) legend(ring(0) position(11) col(1))       ///
>     plot1opts(color(blue) lwidth(medthick))                             ///
>     plot2opts(color(pink) lwidth(medthick)) xlabel(30(10)90) noorigin

        failure _d:  chdfate
  analysis time _t:  exitage
enter on or after:  time age
```

{4} This command plots cumulative **CHD** morbidity as a function of **age** for **men** and **women**. **noorigin** specifies that the morbidity curves starts at the first exit age

Strictly speaking these plots are for people who are free of CHD at age 30, since this is the earliest age at recruitment. However, since CHD is rare before age 30 these plots closely approximate the cumulative morbidity curves from birth.

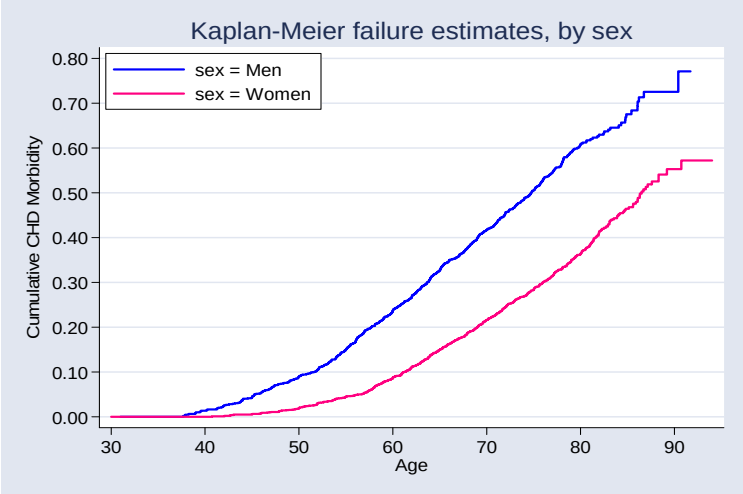






Compare the preceding graph with the analogous graph that we plotted as a function of time since recruitment. In the former graph, the morbidity curves continually widen, which indicates that men remain at greater risk than women regardless of the number of years since recruitment.

This interaction between age and sex on CHD is not apparent in the Kaplan-Meier curves that were plotted as a function of time since recruitment.



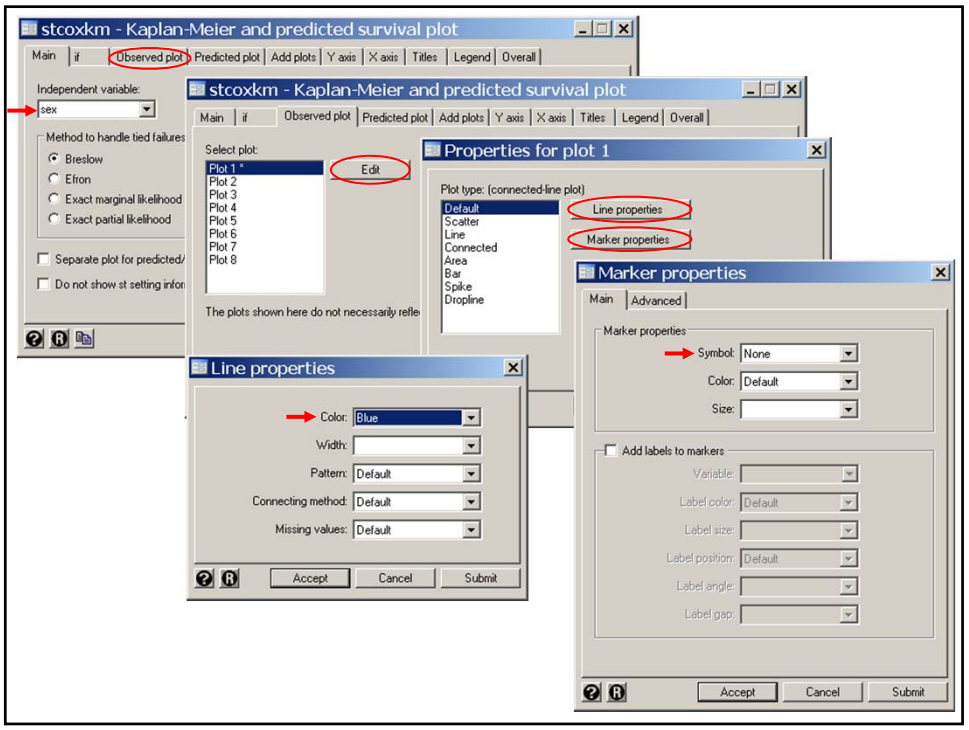
In the latter graph the curves for men and women separate rapidly as women approach the age of menopause. After age 70, however, the curves become parallel, which indicates a similar age-specific incidence for men and women. Hence this analysis is consistent with the hypothesis that the protective effect of female gender is related to premenopausal endocrine function.

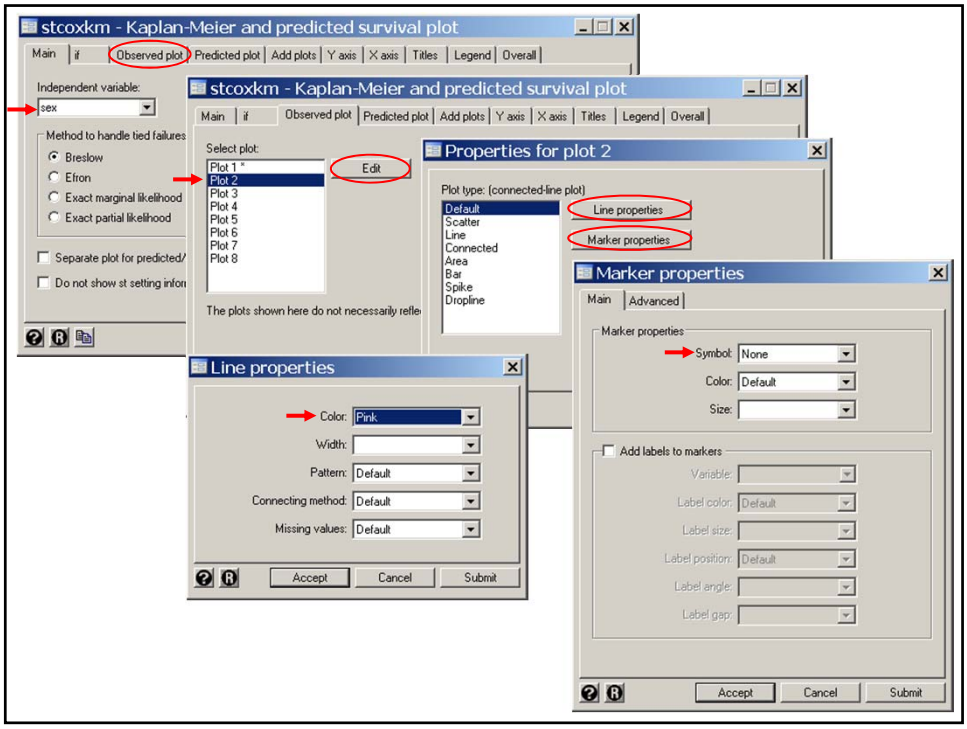
```
. *
. *  Compare Kaplan-Meier curve with best fitting survival curves under the
. *  proportional hazards model.
. *
. * Graphics > Survival analysis graphs > Compare Kaplan-Meier and Cox survival...
. stcoxkm, by(sex) obs1opts(symbol(none) color(blue))     ///        {5}
>     pred1opts(symbol(none) color(blue) lpattern(dash))  ///        {6}
>     obs2opts( symbol(none) color(pink))                 ///        {7}
>     pred2opts(symbol(none) color(pink) lpattern(dash))  ///
>     legend(ring(0) position(7) col(1))

        failure _d:  chdfate
  analysis time _t:  exitage
enter on or after:  time age
```

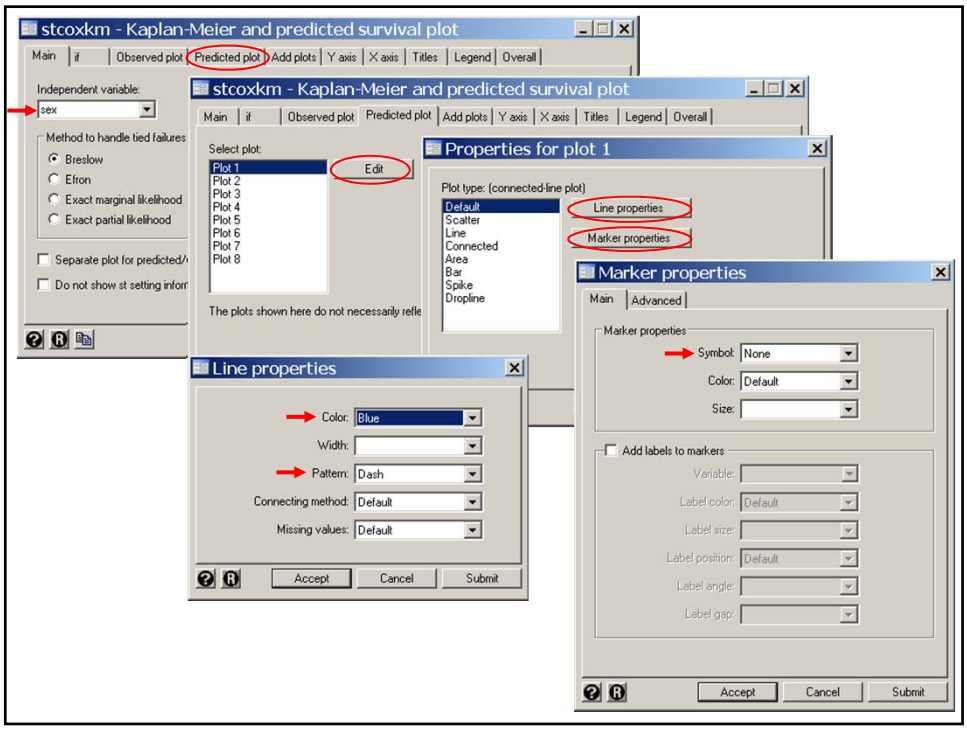
{5} This command plots the Kaplan-Meier survival curves for each sex together with the best fitting survival curves for each gender under the proportional hazards model.

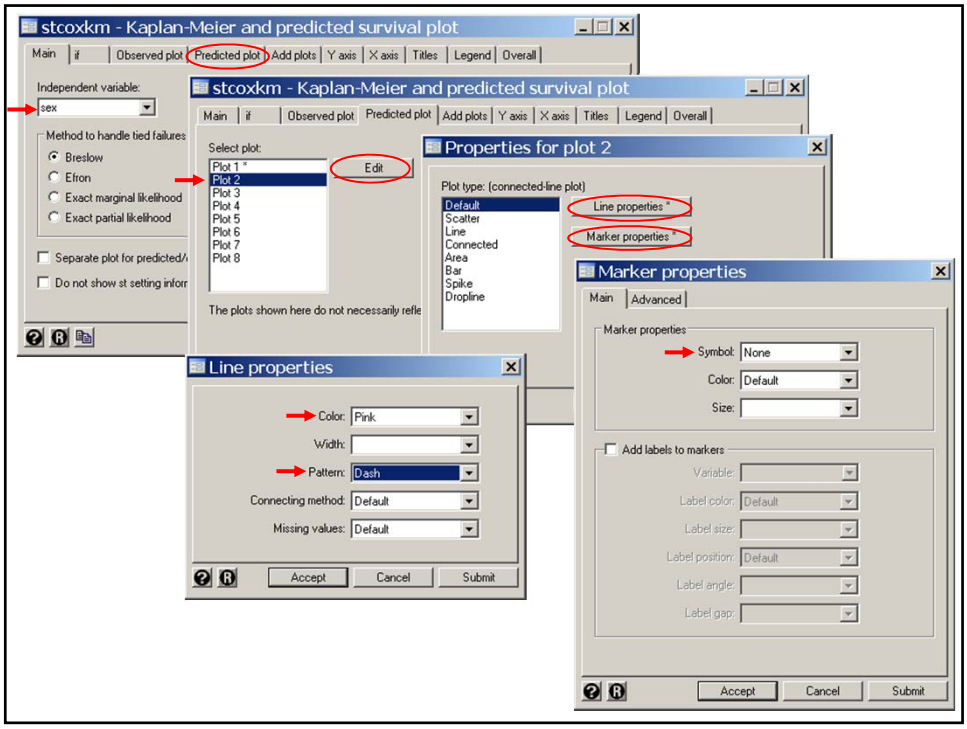
{6} The ***obs1opts*** and ***pred1opts*** options specify the characteristics of the observed and predicted male survival curves, respectively. The suboptions of these options are similar to those of the ***plot1opts*** option ***sts graph*** command. By default, ***stcoxkm*** plots a symbol at each exit time. The ***symbol(none***) suppresses these symbols.

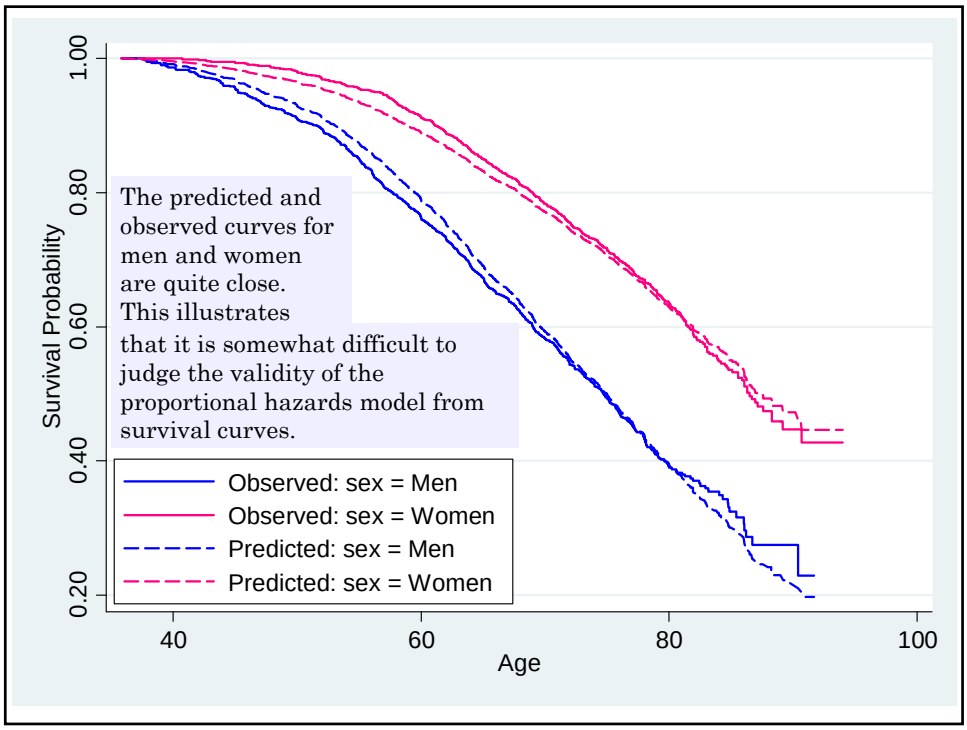
{7} The characteristics of the observed and predicted survival curves for women are similarly defined by the ***obs2opts*** and ***pred2opts*** respectively; ***obs1opts*** and ***obs2opts*** refer to men and women, respectively because the coded value of ***sex*** = 1 for men is less than that for women (***sex*** = 2).




stcoxkm - Kaplan-Meier and predicted survival plot
Main
if
Observed plot
Predicted plot
Add plots
Y axis
X axis
Titles
Legend
Overall
Independent variable:
sex
Method to handle tied failures
Breslow
Efron
Exact marginal likelihood
Exact partial likelihood
Separate plot for predicted/
Do not show st setting infor
Select plot:
Plot 1 *
Plot 2
Plot 3
Plot 4
Plot 5
Plot 6
Plot 7
Plot 8
Edit
The plots shown here do not necessarily refle
Properties for plot 2
Plot type: (connected-line plot)
Default
Scatter
Line
Connected
Area
Bar
Spike
Dropline
Line properties
Marker properties
Marker properties
Main
Advanced
Marker properties
Symbol: None
Color: Default
Size:
Add labels to markers
Variable:
Label color: Default
Label size:
Label position: Default
Label angle:
Label gap:
Accept
Cancel
Submit
Line properties
Color: Pink
Width:
Pattern: Default
Connecting method: Default
Missing values: Default
Accept
Cancel
Submit


stcoxkm - Kaplan-Meier and predicted survival plot
Main
if
Observed plot
Predicted plot
Add plots
Y axis
X axis
Titles
Legend
Overall
Independent variable:
sex
Method to handle tied failures
Breslow
Efron
Exact marginal likelihood
Exact partial likelihood
Separate plot for predicted/
Do not show st setting infor
Select plot:
Plot 1
Plot 2
Plot 3
Plot 4
Plot 5
Plot 6
Plot 7
Plot 8
Edit
The plots shown here do not necessarily refle
Properties for plot 1
Plot type: (connected-line plot)
Default
Scatter
Line
Connected
Area
Bar
Spike
Dropline
Line properties
Marker properties
Marker properties
Main
Advanced
Marker properties
Symbol: None
Color: Default
Size:
Add labels to markers
Variable:
Label color: Default
Label size:
Label position: Default
Label angle:
Label gap:
Accept
Cancel
Submit
Line properties
Color: Blue
Width:
Pattern: Dash
Connecting method: Default
Missing values: Default
Accept
Cancel
Submit


stcoxkm - Kaplan-Meier and predicted survival plot
Main
Observed plot
Predicted plot
Add plots
Y axis
X axis
Titles
Legend
Overall
Independent variable:
sex
Method to handle tied failures
Breslow
Efron
Exact marginal likelihood
Exact partial likelihood
Select plot:
Plot 1 *
Plot 2
Plot 3
Plot 4
Plot 5
Plot 6
Plot 7
Plot 8
Edit
Properties for plot 2
Plot type: (connected-line plot)
Default
Scatter
Line
Connected
Area
Bar
Spike
Dropline
Line properties *
Marker properties *
Line properties
Color: Pink
Width:
Pattern: Dash
Connecting method: Default
Missing values: Default
Accept
Cancel
Submit
Marker properties
Main
Advanced
Marker properties
Symbol: None
Color: Default
Size:
Add labels to markers
Variable:
Label color: Default
Label size:
Label position: Default
Label angle:
Label gap:
Accept
Cancel
Submit


Survival Probability
1.00
0.80
0.60
0.40
0.20
40
60
80
100
Age
The predicted and observed curves for men and women are quite close. This illustrates that it is somewhat difficult to judge the validity of the proportional hazards model from survival curves.
Observed: sex = Men
Observed: sex = Women
Predicted: sex = Men
Predicted: sex = Women

Under the proportional hazards assumption the survival function for the i^{th} patient is

$$S_i[t] = \exp\left[-\exp\left[\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_1 x_{iq}\right]\int_0^t \lambda_0[x]\,dx\right]$$

Hence,

$$\log\left[S_i[t]\right] = -\exp\left[\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_1 x_{iq}\right]\int_0^t \lambda_0[x]\,dx$$

$$\log\left[-\log\left[S_i[t]\right]\right] = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_1 x_{iq} + \log\left[\int_0^t \lambda_0[x]\,dx\right]$$

$$= \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_1 x_{iq} + f[t]$$

for some function $f[t]$.

This means that if the proportional hazards assumption is true then plots of $\log\left[-\log\left[S_i[t]\right]\right]$ for different covariate values should be parallel. That is, they should differ by $\beta_1\left(x_{i1} - x_{j1}\right) + \beta_2\left(x_{i2} - x_{j2}\right) + \cdots + \beta_1\left(x_{iq} - x_{jq}\right)$.

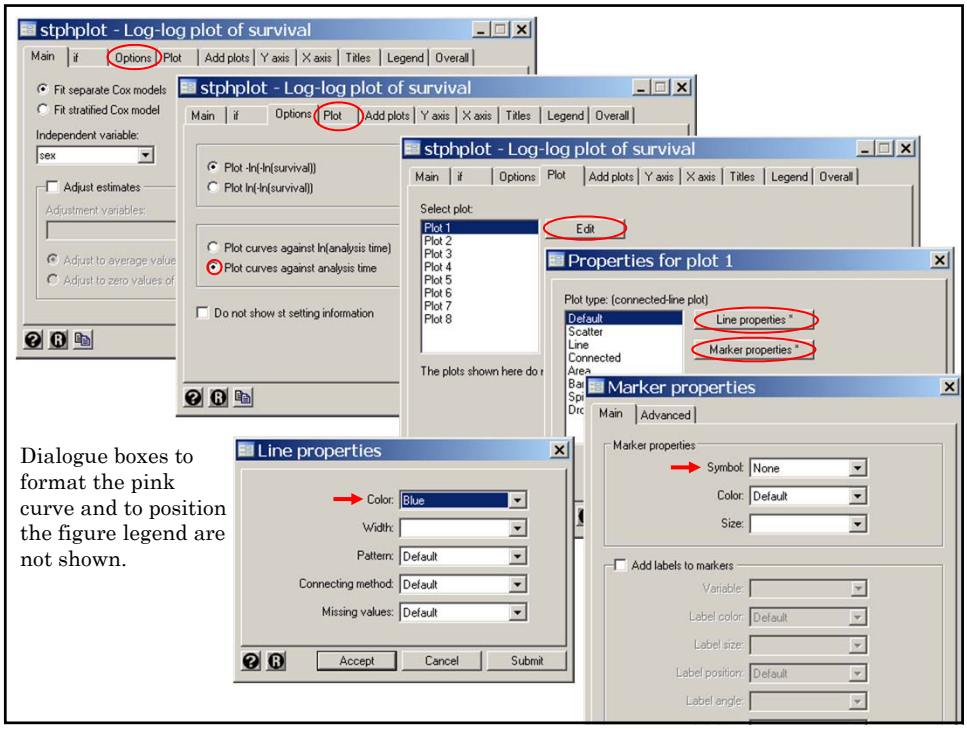
We draw such plots to visually evaluate the proportinal hazards assumption. *Framingham.age.log* continues as follows:

```
. *
. *  Draw log-log plots to assess the proportional hazards assumption.
. *
. * Graphics > Survival analysis graphs > Assess proportional-hazards ...
. stphplot, by(sex) nolntime                                    /// {8}
      plot1opts(symbol(none) color(blue))                       ///
>     plot2opts(symbol(none) color(pink))                       ///
>     legend(ring(0) position(2) col(1))

        failure _d:  chdfate
  analysis time _t:  exitage
```

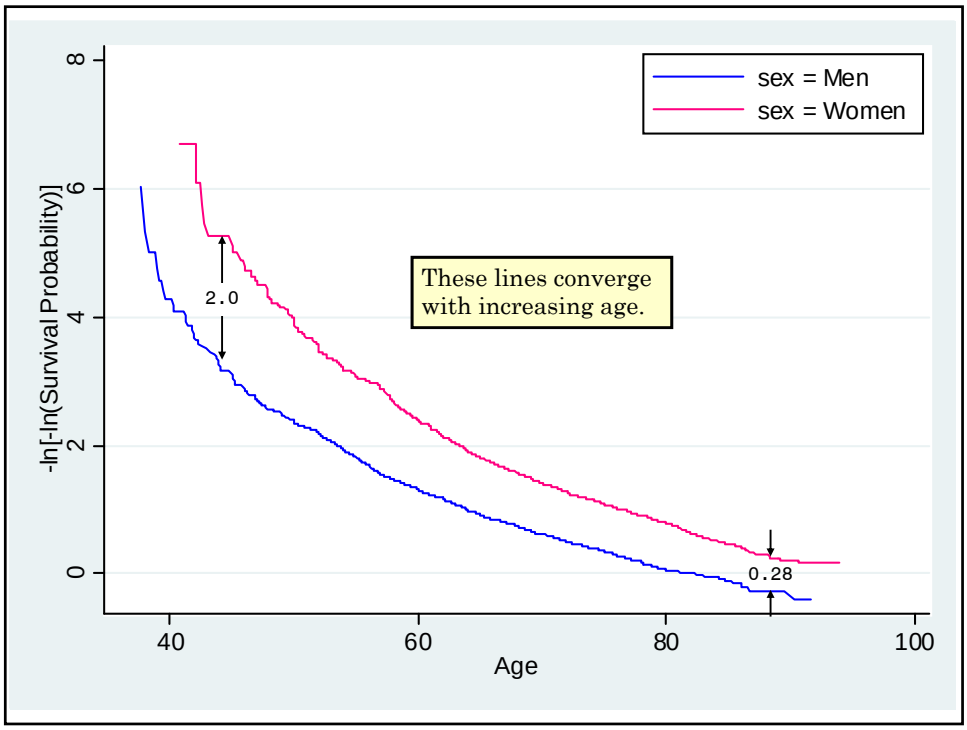
{8} The *stphplot* command draws log-log plots for each unique value of the covariate specified with the *by* option (in this example *sex*). It fits a proportional hazards model regressing *chdfate* against *sex* as defined by the previous *stset* command.

nolintime causes the *x*-axis to be analysis time (*exitage*) rather than the default which is log analysis time.

We can also use the *adjust(varlist)* option to graph log-log plots for patients with average values of the variables in *varlist*.



Dialogue boxes to format the pink curve and to position the figure legend are not shown.



7. Hazard Regression Models with Time Dependent Covariates

The proportional hazards assumption can be weakened by using **time-dependent covariates**. That is, we assume that the i^{th} patient has q covariates

$$x_{i1}[t], x_{i2}[t], \ldots, x_{iq}[t]$$

that are themselves functions of time t, and that the hazard function for this patient is

$$\lambda_i[t] = \lambda_0[t]\exp[x_{i1}[t]\beta_1 + x_{i2}[t]\beta_2 + \ldots + x_{iq}[t]\beta_q]$$

The simplest time dependent covariates are **step-functions**.

For example, in the preceding graph of cumulative CHD morbidity by sex we saw strong evidence that the protective effect of being a woman varies with age. To estimate how the relative risk of being male varies with age we could define the following covariate functions.

$$x_{i1}(age) = \begin{cases} 1: & i^{th} \text{ patient is a man } \leq \text{ age } 50 \\ 0: & \text{Otherwise} \end{cases}$$

$$x_{i2}(age) = \begin{cases} 1: & i^{th} \text{ patient is a man aged } 50 - 60 \\ 0: & \text{Otherwise} \end{cases}$$

$$x_{i3}(age) = \begin{cases} 1: & i^{th} \text{ patient is a man aged } 60 - 70 \\ 0: & \text{Otherwise} \end{cases}$$

$$x_{i4}(age) = \begin{cases} 1: & i^{th} \text{ patient is a man aged } 70 - 80 \\ 0: & \text{Otherwise} \end{cases}$$

$$x_{i5}(age) = \begin{cases} 1: & i^{th} \text{ patient is a man age } > 80 \\ 0: & \text{Otherwise} \end{cases}$$

$x_{ij}(age)$ are called step-functions because they are constant and equal 1 on the specified age intervals and then step down to 0 for larger or smaller values of *age*.

The hazard regression model is then

$$\lambda_i[age] = \lambda_0[age]\exp[x_{i1}[age]\beta_1 + x_{i2}[age]\beta_2 + \cdots + x_{i5}[age]\beta_5]$$

The functions $x_{i1}(age), x_{i2}(age), \cdots, x_{i5}(age)$ are associated with five parameters $\beta_1, \beta_2, \cdots \beta_5$ that assess the effect of male gender on CHD risk before age 50, from age 50 to 60, 60 to 70, 70 to 80 and above 80, respectively.

Note that β_1 has no effect on CHD hazard after *age* 50 since $x_{i1}(t) = 0$ regardless of the patient's sex.

Similarly, the other β coefficients have no effect on CHD hazard on ages where their covariate functions are uniformly zero.

Hence $\beta_1, \beta_2, \cdots \beta_5$ are the log relative risks of CHD in men, before age 50, from age 50 to 60, 60 to 70, 70 to 80 and above 80, respectively.

a) Analyzing time-dependent covariates in Stata

Stata can handle hazard regression models with time dependent covariates that are step-functions. To do this we first must define multiple data records per patient in such a way that the covariate functions for the patient are constant for the period covered by each record. This is best explained by an example.

Suppose that a man with study ID 924 enters the Framingham study at age 32 and exits with CHD at age 63. Then

id = 924
age = 32
exitage = 63, and
chdfate = 1.

We replace the record for this patient with three records. One that describes his covariates for age 32 to age 50, another that describes his covariated from age 50 to 60, and a third that describes his covariates from age 60 to 63.

Let *male1*, *male2*, …, *male5* denote $x_{i1}(age)$, $x_{i2}(age)$, …, $x_{i5}(age)$, respectively, and let *enter*, *exit* and *fate* be new variables which we define in the following table.

id	*male1*	*male2*	*male3*	*enter*	*exit*	*fate*
924	1	0	0	32	50	0
924	0	1	0	50	60	0
924	0	0	1	60	63	1

These records describe the patient in **three** age epochs: before age 50, between age 50 and 60, and after age 60. The patient enters the first epoch at age 32 when he enters the study and exits this epoch at age 50. During this time *male1* = 1 and *male2* = *male3* = 0; *fate* = 0 since he has not suffered CHD. He enters the second epoch at age 50 and exits at age 60 without CHD. Hence, for this epoch *male1* = *male3* = 0, *male2* = 1 and *fate* = 0. He enters the third epoch at age 60 and exits at age 62 with CHD. Hence, *male1* = *male2* = 0, *male3* = 1 and *fate* = 1. *male4* = *male5* = 0 in all records since the patient never reaches age 70.

Time dependent analyses must have an ID variable that allows Stata to keep track of which records belong to which patients.

The following log file illustrates how to create and analyze these records.

```
. * Framingham.TimeDependent.log
. *
. *  Perform hazard regressions of gender on CHD risk
. *  using age as the time variable.  Explore models
. *  with time dependent covariates for sex
. *

. use C:\WDDtext\2.20.Framingham.dta, clear

. generate time= followup/365.25

. generate male = sex==1

. label define male 0 "Women" 1 "Men"

. label values male male
```

```
. *
. *  Calculate the relative risk of CHD for men relative to women using
. *  age as the time variable.
. *
. generate exitage = age+time

. * Statistics > Survival... > Setup... > Declare data to be survival...
. stset exitage, enter(time age) failure(chdfate)

     failure event:  chdfate != 0 & chdfate < .
obs. time interval:  (0, exitage]
 enter on or after:  time age
 exit on or before:  failure

------------------------------------------------------------------------------
     4699  total obs.
        0  exclusions
------------------------------------------------------------------------------
     4699  obs. remaining, representing
     1473  failures in single record/single failure data
 103710.1  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =        30
                                  last observed exit t =        94
```

```
. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox male                                                          {1}

        failure_d: chdfate
  analysis time_t: exitage
enter on or after: time age

Cox regression - Breslow method for ties

No. of subjects =        4699                     Number of obs    =   4699
No. of failures =        1473
Time at risk    = 103710.0914
                                                  LR chi2(1)       = 177.15
Log likelihood =  -11218.785                      Prob > chi2      = 0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   2.011662   .1060464     13.26   0.000       1.814192    2.230626
------------------------------------------------------------------------------
```

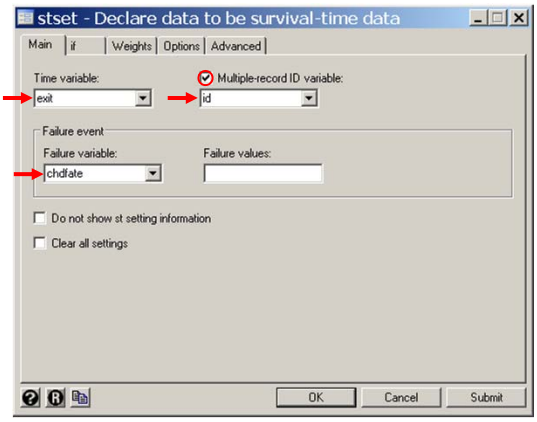
{1} First, we run the proportional hazards analysis of the effect of gender on CHD. This analysis estimates that men have **2.01** times the CHD risk of women, with overwhelming statistical significance.

```
. *
. * Perform hazard regression with time dependent covariates for sex
. *
. tabulate chdfate male                                                {2}

Coronary   |
Heart      |         male
Disease    |         0          1 |     Total
-----------+----------------------+----------
  Censored |      2000       1226 |      3226
       CHD |       650        823 |      1473
-----------+----------------------+----------
     Total |      2650       2049 |      4699
```

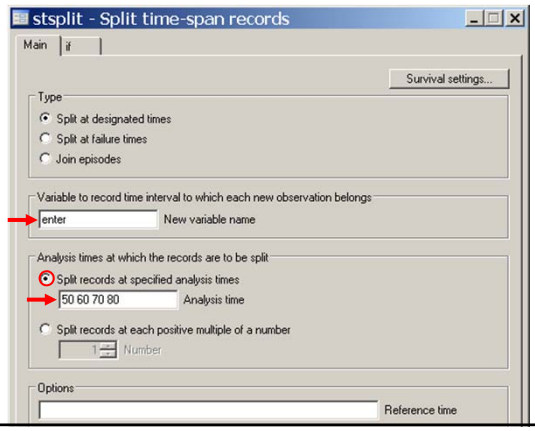
{2} The next few commands will create the multiple records that we need. It is **prudent** to be cautious doing this and to create **before** and **after tables** to confirm that we have done what we intended to do.

```
. *
. *  Split each patient's record into one or more records so that each
. *  record describes one epoch with constant covariates for the epoch.
. *
. generate exit = exitage

. * Statistics > Survival... > Setup... > Declare data to be survival...
. stset exit, id(id) enter(time age) failure(chdfate)                {3}

                id:  id
     failure event:  chdfate != 0 & chdfate < .
obs. time interval:  (exit[_n-1], exit]
 enter on or after:  time age
 exit on or before:  failure

------------------------------------------------------------------------------
     4699  total obs.
        0  exclusions
------------------------------------------------------------------------------
     4699  obs. remaining, representing
     4699  subjects
     1473  failures in single failure-per-subject data
 103710.1  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =        30
                                  last observed exit t =        94
```

{3} This is similar to the previous *stset* except that the exit variable is now *exit* rather than *exitage*. We will define *exit* to denote the patient's fate at the end of each epoch. Also the *id* option defines the variable *id* to be the patient identification variable. It is needed to link multiple records from the same patient in different epochs together.



```
. * Data > Describe data > List data
. list id male age exit chdfate if id == 924

     +--------------------------------------+
     |  id   male   age       exit   chdfate |
     |--------------------------------------|
3182. | 924    Men    32   63.23888       CHD |
     +--------------------------------------+

. * Statistics > Survival... > Setup... > Split time-span records
. stsplit enter, at(50 60 70 80)                                    {4}
(8717 observations (episodes) created)
```




```
. * Data > Describe data > List data
. list id male age exit chdfate if id == 924

     +--------------------------------------+
     |  id   male   age       exit   chdfate |
     |--------------------------------------|
3182. | 924    Men    32   63.23888       CHD |
     +--------------------------------------+

. * Statistics > Survival... > Setup... > Split time-span records
. stsplit enter, at(50 60 70 80)                                    {4}
(8717 observations (episodes) created)
```

```
. list id male enter exit chdfate if id == 924

      +---------------------------------------+
      |  id   male   enter       exit   chdfate |
      |---------------------------------------|
 7940. | 924    Men       0         50         . |
 7941. | 924    Men      50         60         . |
 7942. | 924    Men      60   63.23888       CHD |
      +---------------------------------------+
```

{4} This command creates up to 5 epochs for each patient: before age 50, between 50 and 60, 60 and 70, 70 and 80, and after age 80.

- For each patient, a separate record is created for each epoch that the patient experienced during follow-up.
- The *newvar* variable, (in this example *enter*) is set equal to the start of the patient's first epoch. That is, to the start of the latest epoch that is less than *age*. Stata considers the first epoch to start at age zero.
- The *timevar* of the last *stset* command, (in this example *exit*) is changed to equal the end of the epoch for all but the last record.
- The fate variable of the last *stset* command, (in this example *chdfate*) is set to missing for all but each patient's last record. *stcox* will treat patients with missing fate variables as being censored at the end of the epoch.

```
. replace enter=age if id~=id[_n-1]                                    {5}
(4451 real changes made)

. generate male1 = male*(              exit <= 50)                     {6}

. generate male2 = male*(enter >= 50 & exit <= 60)                     {7}

. generate male3 = male*(enter >= 60 & exit <= 70)

. generate male4 = male*(enter >= 70 & exit <= 80)

. generate male5 = male*(enter >= 80)
. * Data > Describe data > List data
. list id male? enter exit chdfate if id == 924                        {8}

      +-------------------------------------------------------------------+
      |  id  male1  male2  male3  male4  male5   enter       exit   chdfate |
      |-------------------------------------------------------------------|
7940. | 924      1      0      0      0      0      32         50         . |
7941. | 924      0      1      0      0      0      50         60         . |
7942. | 924      0      0      1      0      0      60   63.23888       CHD |
      +-------------------------------------------------------------------+
```

{5} Replace *enter* by the patient's age of entry for each patient's first record. This correction must be made whenever we have ragged entry since *stsplit* assumes that all patients enter at time zero.

{6} ***male1*** = 1 if and only if the subject is **male** and we are in the **first** epoch.

{7} ***male2*** = 1 if and only if the subject is **male** and we are in the **second** epoch. *male3*, *male4* and *male5* are similarly defined.

{8} ***male?*** Designates all variables that start with "male" and end with exactly one character. I.e. *male1*, *male2*, … , *male5*. Note that these covariates are now correctly defined and are constant within each epoch.

```
. generate testmale = male1 + male2 + male3 + male4 + male5
. * Statistics > Summaries... > Tables > Two-way tables with measures...
. tabulate chdfate testmale, missing                                    {9}
  Coronary |
     Heart |       testmale
   Disease |         0          1 |     Total
-----------+----------------------+----------
  Censored |     2,000      1,226 |     3,226
       CHD |       650        823 |     1,473
         . |     5,217      3,500 |     8,717
-----------+----------------------+----------
     Total |     7,867      5,549 |    13,416

last observed exit t =        94
```

{9} **No** subject has more than one value of ***male1, male2, male3, male4*** or ***male5*** equal to **1** in the same epoch.

- There are **2000 + 650 women** with all of these covariates equal 0, which agrees with the preceding table.
- The **8717 new records** have missing values of *chdfate* indicating censoring at the end of these epochs.
- This table shows that there are **650** records for women showing CHD and **823** such records for men. This is the same as the number of women and men who had CHD. Thus, we have not added or removed any CHD events by the previous manipulation.

```
. * Statistics > Summaries... > Tables > Two-way tables with measures...
. tabulate chdfate male
Coronary   |
Heart      |          male
Disease    |         0          1 |     Total
-----------+----------------------+----------
  Censored |      2000       1226 |      3226
       CHD |       650        823 |      1473
-----------+----------------------+----------
     Total |      2650       2049 |      4699
```

```
. * Statistics > Survival... > Setup... > Declare data to be survival...
. stset exit, id(id) enter(time enter) failure(chdfate)              {10}

                id:  id
     failure event:  chdfate != 0 & chdfate < .
obs. time interval:  (exit[_n-1], exit]
 enter on or after:  time enter
 exit on or before:  failure

--------------------------------------------------------------------------------
    13416  total obs.
        0  exclusions
--------------------------------------------------------------------------------
    13416  obs. remaining, representing
     4699  subjects
     1473  failures in single failure-per-subject data
 103710.1  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =        30
                                  last observed exit t =        94
```

{10} We define ***id*** to be the patient ID variable,
enter to be the patient's age at entry,
exit to be the exit time, and
chdfate to be the fate indicator.

The ***stset*** command also **checks** the data for **errors** or inconsistencies in the definition of these variables.

```
. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox male?                                                          {11}

        failure _d:  chdfate
  analysis time _t:  exit
  enter on or after:  time enter
                id:  id

Cox regression -- Breslow method for ties

No. of subjects =         4699                     Number of obs    =     13416
No. of failures =         1473
Time at risk    =  103710.0914
                                                   LR chi2(5)       =    203.92
Log likelihood  =   -11205.396                     Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       male1 |    4.22961   .9479718     6.43   0.000      2.72598    6.562631
       male2 |   2.480204    .264424     8.52   0.000     2.012508    3.056591
       male3 |   1.762634   .1465087     6.82   0.000     1.497652    2.074499
       male4 |   1.880939   .2127479     5.59   0.000     1.506946     2.34775
       male5 |   1.048225   .2579044     0.19   0.848     .6471809    1.697788
------------------------------------------------------------------------------
```

{11} Finally we perform a hazard regression analysis with the **time dependent** covariates ***male1, male2, … , male5***. Note how the relative risks for men drop with increasing age.

The data management commands in the preceding example were

```
generate exit = exitage
stset exit, id(id) enter(time age) failure(chdfate)
stsplit enter, at(50 60 70 80)
replace enter=age if id~=id[_n-1]
generate male1 = male*(              exit <= 50)
generate male2 = male*(enter >= 50 & exit <= 60)
generate male3 = male*(enter >= 60 & exit <= 70)
generate male4 = male*(enter >= 70 & exit <= 80)
generate male5 = male*(enter >= 80)
stset exit, id(id) enter(time age) failure(chdfate)
```

The highlighted lines are needed because of the ragged entry into the study. If all patients entered the study at time 0 (in this example birth) and were followed until time ***follow*** then the analogous commands would be

```
generate exit = follow
stset exit, id(id) failure(chdfate)
stsplit enter, at(50 60 70 80)
generate male1 = male*(              exit <= 50)
generate male2 = male*(enter >= 50 & exit <= 60)
generate male3 = male*(enter >= 60 & exit <= 70)
generate male4 = male*(enter >= 70 & exit <= 80)
generate male5 = male*(enter >= 80)
stset exit, id(id) failure(chdfate)
```

Note that by default ***stsplit*** sets the beginning of the first epoch to 0, which is what we want when time measures time since recruitment.

8. Testing the Proportional Hazards Assumption

In the preceding example, suppose that $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta$

Then our model is

$$\begin{aligned}\lambda_i[age] &= \lambda_0[age]\exp[x_{i1}[age]\beta_1 + x_{i2}[age]\beta_2 + \cdots + x_{i5}[age]\beta_5] \\ &= \lambda_0[age]\exp[(x_{i1}[age] + x_{i2}[age] + \cdots + x_{i5}[age])\beta] \\ &= \lambda_0[age]\exp[male \times \beta]\end{aligned}$$

which obeys the proportional hazards assumption.

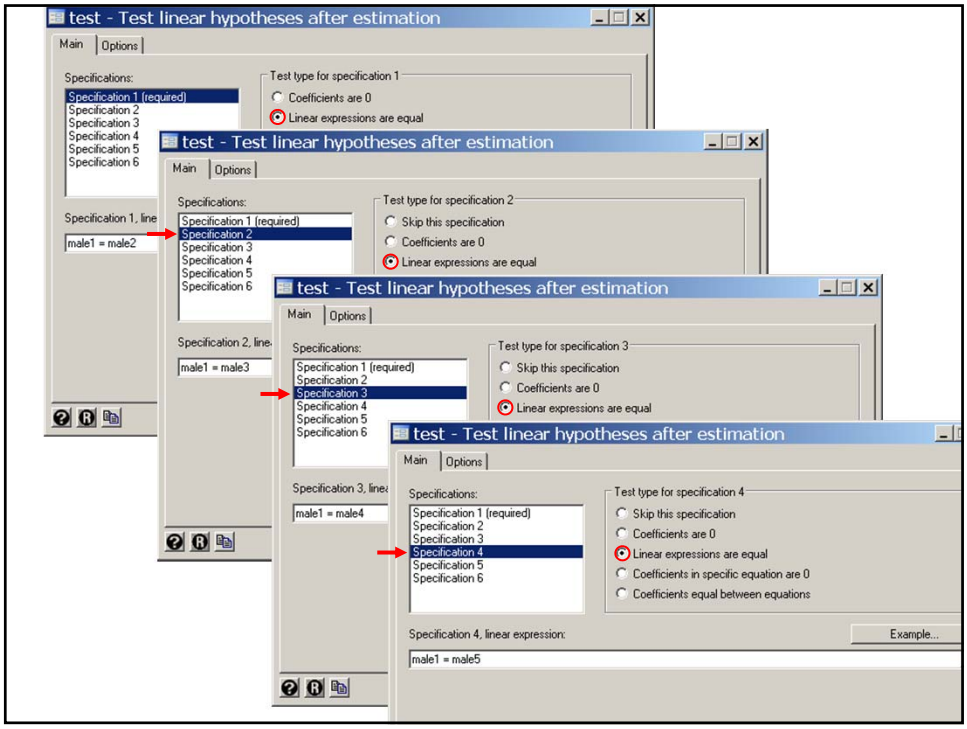
Hence, we can test the proportional hazards assumption by testing whether $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$

We can test this hypothesis in Stata using the *test* post estimation command.

We illustrate this test in *Framingham.TimeDependent.log,* which continues as follows:

```
. * Statistics > Postestimation > Tests > Test linear hypotheses
. test male1 = male2 = male3 = male4 = male5                          {12}

 ( 1)  male1 - male2 = 0
 ( 2)  male1 - male3 = 0
 ( 3)  male1 - male4 = 0
 ( 4)  male1 - male5 = 0

           chi2(  4) =   24.74
         Prob > chi2 =    0.0001
```

{12} This test that the five model parameters are equal had four degrees of freedom and can be rejected with overwhelming significance. Hence, the proportional hazards assumption is clearly false.



The *test* command can also test whether pairs of parameters are simultaneously equal. For example, if `x1, x2, x3` and `x4` are covariates associated with model parameters β_1, β_2, β_3, and β_4 then

```
. test (x1 = x2) (x3 = x4)
```

tests the joint hypothesis that $\beta_1 = \beta_2$ and $\beta_3 = \beta_4$.

```
. lincom male1 - male2                                          {13}

 ( 1)  male1 - male2 = 0

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .5337688   .2481927     2.15   0.032     .0473199    1.020218
------------------------------------------------------------------------------

. lincom male2 - male3                                          {14}

 ( 1)  male2 - male3 = 0

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .3415319   .1351862     2.53   0.012     .0765719    .6064919
------------------------------------------------------------------------------
```

{14} The relative risk for men aged **50 -- 60** is significantly different than for men aged **60 -- 70** (P = 0.01).

{13} The relative risk for men **before** age **50** is significantly different than for men aged **50 -- 60** (P = 0.03).

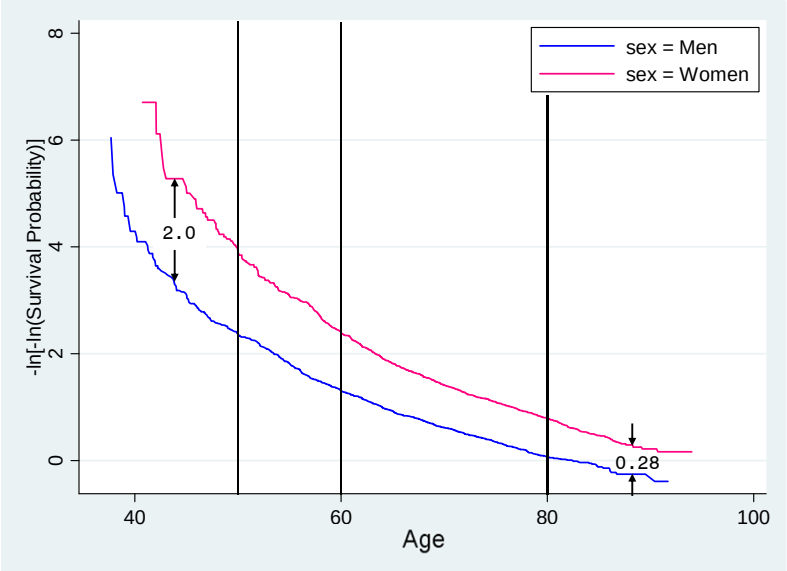
```
. lincom male3 - male4                                                  {15}

 ( 1)  male3 - male4 = 0

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0649622    .140364    -0.46   0.643    -.3400706    .2101463
------------------------------------------------------------------------------

. lincom male4 - male5                                                  {15}

 ( 1)  male4 - male5 = 0

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .5846729   .2707924     2.16   0.031     .0539295    1.115416
------------------------------------------------------------------------------

. generate male34 = male3 + male4                                       {16}
```

{15} The relative risks for men do not differ between epochs 3 and 4 but are significantly different between epocs 4 and 5.

{16} Lets combine the third and fourth epochs and reanalyze the data.

```
. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox male1 male2 male34 male5

        failure _d:  chdfate
  analysis time _t:  exit
  enter on or after:  time enter
                id:  id

No. of subjects =         4699                     Number of obs   =     13416
No. of failures =         1473
Time at risk    =  103710.0914
                                                   LR chi2(4)      =    203.71
Log likelihood  =   -11205.503                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       male1 |    4.22961   .9479718     6.43   0.000      2.72598    6.562631
       male2 |   2.480204    .264424     8.52   0.000     2.012508    3.056591
      male34 |   1.803271   .1208478     8.80   0.000     1.581309    2.056387
       male5 |   1.048225   .2579044     0.19   0.848     .6471809    1.697788
------------------------------------------------------------------------------
. * Statistics > Postestimation > Tests > Test linear hypotheses
. test male1 = male2 = male34 = male5

 ( 1)  male1 - male2 = 0
 ( 2)  male1 - male34 = 0
 ( 3)  male1 - male5 = 0

           chi2(  3) =   24.52
         Prob > chi2 =    0.0000
```

```
. lincom male1 - male2

 ( 1)  male1 - male2 = 0

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .5337688   .2481927     2.15   0.032     .0473199    1.020218
------------------------------------------------------------------------------

. lincom male2 - male34

 ( 1)  male2 - male34 = 0

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    .318739   .1259271     2.53   0.011     .0719264    .5655516
------------------------------------------------------------------------------

. lincom male34 - male5

 ( 1)  male34 - male5 = 0

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .5425036   .2550027     2.13   0.033     .0427074      1.0423
------------------------------------------------------------------------------
```



An alternate way to explore the effect of age on relative risk will be explained in Chapter 7 on Poisson regression.

9. What we have covered

- Extend simple proportional hazards regression to models with multiple covariates
- Model parameters, hazard ratios and relative risks
- Similarities between hazard regression and linear regression
  - Categorical variables, multiplicative models, models with interaction
  - Estimating the effects of two risk factors on a relative risk
  - Calculating 95% CIs for relative risks derived from multiple parameter estimates.
  - Adjusting for confounding variables
- Restricted cubic splines and survival analysis
- Stratified proportional hazards regression models
- Using age as the time variable in survival analysis
  - Ragged study entry: the *enter(time varname)* option of the *stset* command
- Checking the proportional hazards assumption
  - Comparing Kaplan-Meier plots to analogous plots drawn under the proportional hazards assumption: the *stcoxkm* command
  - Log-log plots: the *stphplot* command
- Hazards regression models with time-dependent covariates
  - Testing the proportional hazards assumption: the *test* command

Cited Reference

Levy D, National Heart Lung and Blood Institute., Center for Bio-Medical Communication. *50 Years of Discovery : Medical Milestones from the National Heart, Lung, and Blood Institute's Framingham Heart Study*. Hackensack, N.J.: Center for Bio-Medical Communication Inc.; 1999.

For additional references on these notes see.

Dupont WD. *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data. 2nd ed.* Cambridge, U.K.: Cambridge University Press; 2009.