BIOSTATISTICS II

Statistical Modeling for Biomedical Researchers

William D. Dupont, Ph.D.

Volume II, Sections 5-11

M.P.H. Program
Vanderbilt University School of Medicine
March 2011

V. INTRODUCTION TO SURVIVAL ANALYSIS

- Survival data: time to event
  - Right censored data
- Kaplan-Meier survival curves
- Kaplan-Meier cumulative mortality curves
  - Greenwood confidence bands for survival and mortality curves
  - Displaying censoring times and numbers of patients at risk
- Estimating survival probabilities
- Censoring and biased Kaplan-Meier survival curves
- Log rank test for comparing survival curves
- Hazard functions and cumulative mortality
- Simple proportional hazards regression model
  - Hazard rate ratios and relative risk
  - Estimating relative risks from proportional hazards models
- Tied failure times and biased relative risk estimates

1. Survival and Cumulative Mortality Functions

Suppose we have a cohort of n people.

Let

t_i be the age that the i^{th} person dies,

$m[t]$ be the number of patients for whom $t < t_i$, and

$d[t]$ be the number of patients for whom $t_i \le t$.

Then the **survival function** is

$S[t] = \Pr[t_i > t]$ = the probability of surviving until at least age t.

The **cumulative mortality function** is

$D[t] = \Pr[t_i \le t]$ = the probability of dying before age t.

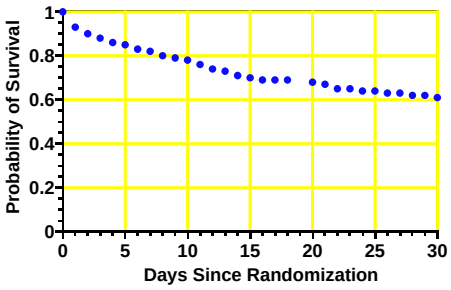
If t_i is known for all members of the cohort we can estimate $S(t)$ and $D(t)$ by

$\hat{S}[t] = m[t]/n$ the proportion of subjects who are alive at age t, and

$\hat{D}[t] = d[t]/n$ the proportion who have died by age t.

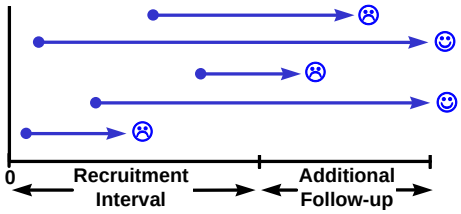
a) Example: Survival among sepsis patients

Days Since Entry	Number of Patients Alive	Number of Deaths	Proportion Alive
0	n = m(0) = 455	0	m(0) / n = 1.00
1	m(1) = 423	32	m(1) / n = 0.93
2	m(2) = 410	45	m(2) / n = 0.90
3	m(3) = 400	55	m(3) / n = 0.88
4	m(4) = 392	63	m(4) / n = 0.86
5	m(5) = 386	69	m(5) / n = 0.85
6	m(6) = 378	77	m(6) / n = 0.83
7	m(7) = 371	84	m(7) / n = 0.82
8	m(8) = 366	89	m(8) / n = 0.80
9	m(9) = 360	95	m(9) / n = 0.79
10	m(10) = 353	102	m(10) / n = 0.78
.	.	.	.
.	.	.	.
.	.	.	.
21	m(21) = 305	150	m(21) / n = 0.67
22	m(22) = 296	159	m(22) / n = 0.65
23	m(23) = 295	160	m(23) / n = 0.65
24	m(24) = 292	163	m(24) / n = 0.64
25	m(25) = 290	165	m(25) / n = 0.64
26	m(26) = 288	167	m(26) / n = 0.63
27	m(27) = 286	169	m(27) / n = 0.63
28	m(28) = 283	172	m(28) / n = 0.62
29	m(29) = 280	175	m(29) / n = 0.62
30	m(30) = 279	176	m(30) / n = 0.61



2. Right Censored Data

In clinical studies, patients are typically recruited over a recruitment interval and then followed for an additional period of time.



Let

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

and

f_i = 1: i^{th} patient dies at exit
0: i^{th} patient alive at exit

Patients who are alive at exit are said to be **right censored**. This means that we know that they survived until at least time t_i but do not know how much longer they lived thereafter.

With censored data, the proportion of patients who are known to have died by time t underestimates the true cumulative mortality since some patients will die after their censoring times.

3. Kaplan-Meier (Product Limit) Survival Curves

Suppose that we have censored survival data on a cohort of patients. We divide the follow-up time into intervals that are small enough that few patients die in any one interval.

Suppose this interval is days.

Let

n_i be the number of patients known to be at risk at the beginning of day i.

d_i be the number of patients who die on day i

Then for patients alive at the beginning of the i^{th} day, the estimated probability of surviving the day is

$$p_i = \frac{n_i - d_i}{n_i}$$

The probability that a patient survives the first t days is the joint probability of surviving days 1, 2, …,t which is estimated by

$$\hat{S}[t] = p_1 p_2 p_3 \ldots p_t$$

Note that $p_i = 1$ on all days that no deaths are observed. Hence, if t_k denotes the k^{th} day on which deaths are observed then

$$\hat{S}[t] = \prod_{\{k:t_k<t\}} p_k \qquad \{7.1\}$$

This estimate is the **Kaplan-Meier survival curve.**

The **Kaplan-Meier cumulative mortality curve** is

$$\hat{D}[t] = 1 - \hat{S}[t]$$

a) Example: Survival in lymphoma patients

Armitage et al. (2002: p. 579) discuss the following data on patient survival after recruitment into a clinical of patients with diffuse histiocytic lymphoma (KcKelvey et al. *Cancer* 1976; **38**: 1484 – 93).

Follow-up (days)							
Dead at end of follow-up				Alive at end of follow-up			
Stage 3							
6	19	32	42	43	126	169	211
42	94	207	253	227	255	270	310
				316	335	346	
Stage 4							
4	6	10	11	41	43	61	61
11	11	13	17	160	235	247	260
20	20	21	22	284	290	291	302
24	24	29	30	304	341	345	
30	31	33	34				
35	39	40	45				
46	50	56	63				
68	82	85	88				
89	90	93	104				
110	134	137	169				
171	173	175	184				
201	222						

4. Drawing Kaplan-Meier Survival Curves in Stata

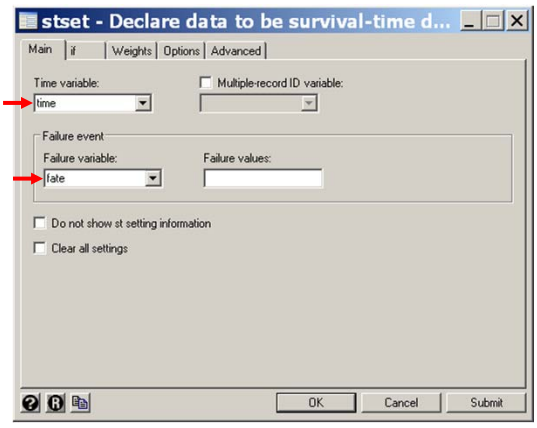
```
. *  Lymphoma.log
. *
. *  Plot Kaplan-Meier Survival curves of lymphoma
. *  patients by stage of tumor.  Perform log-rank test.
. *
. *  See Armitage et al.  2002, Table 17.3.
. *      McKelvey et al., 1976.
. *
. use  "f:/mph/data/armitage/lymphoma.dta", clear

. * Data > Describe data > List data
. list  in 1/7
     +-----------------------------+
     | id     stage   time    fate |                                    {1}
     |-----------------------------|
  1. |  1   Stage 3      6    Dead |
  2. |  2   Stage 3     19    Dead |
  3. |  3   Stage 3     32    Dead |
  4. |  4   Stage 3     42    Dead |
  5. |  5   Stage 3     42    Dead |
     |-----------------------------|
  6. |  6   Stage 3     43   Alive |
  7. |  7   Stage 3     94    Dead |
     +-----------------------------+
```

{1} Two variables must be defined to give each patient's length of **follow-up** and **fate** at exit. In this example, these variables are called *time* and *fate* respectively.

```
. * Data > Describe data > Describe data contents (codebook)
. codebook fate
fate ---------------------------------------- (unlabeled)
                  type:  numeric (float)
                 label:  fate

                 range:  [0,1]            units:  1
         unique values:  2         coded missing:  0 / 80

            tabulation:  Freq.   Numeric  Label
                            26         0  Alive                    {2}
                            54         1  Dead
. * Statistics > Survival... > Setup... > Declare data to be survival...
. stset time, failure (fate)                                        {3}

     failure event:  fate != 0 & fate < .
obs. time interval:  (0, time]
 exit on or before:  failure

--------------------------------------------------------------------------
       80  total obs.
        0  exclusions
--------------------------------------------------------------------------
       80  obs. remaining, representing
       54  failures in single record/single failure data
     9718  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =       346
```

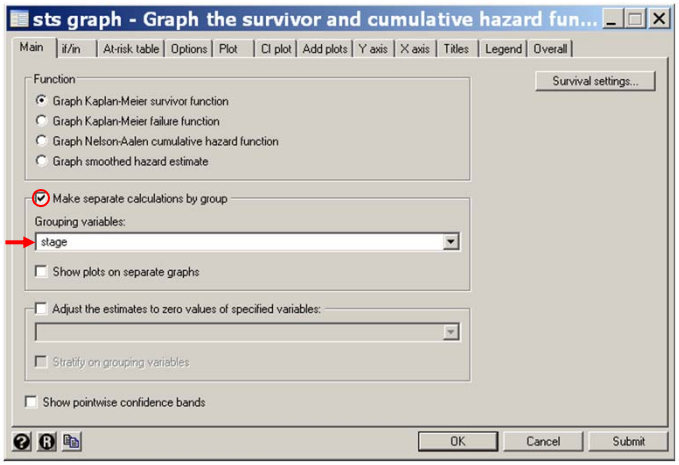
{2} The ***fate*** variable is coded as 0 = alive and 1 = dead at exit

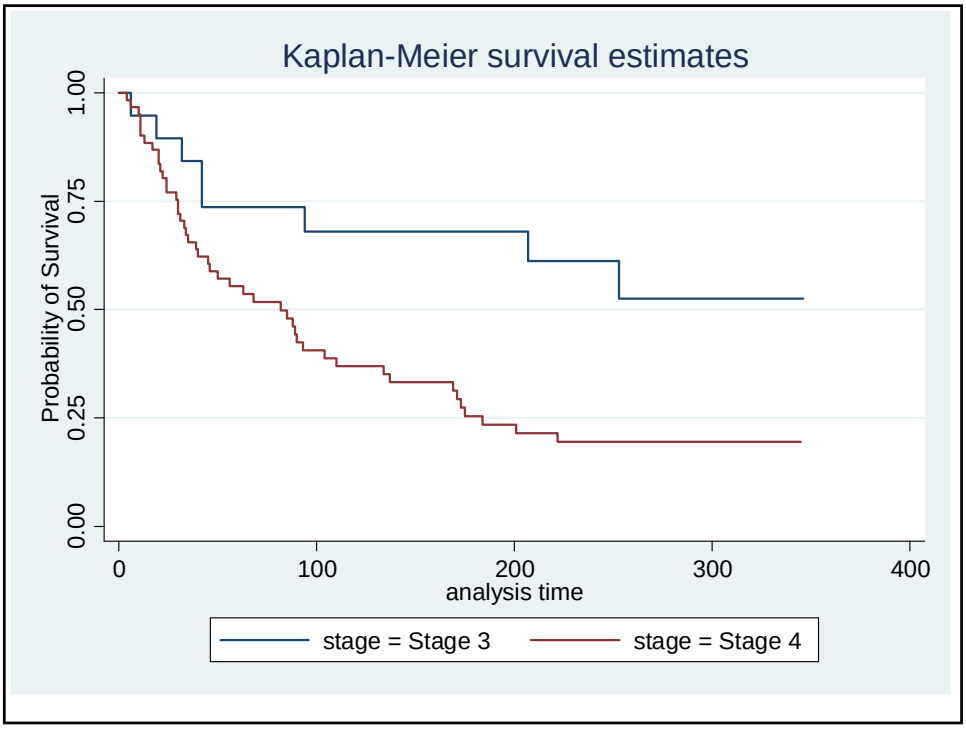
{3} ***stset*** specifies that the data set contains survival data, with each patient's **exit time** denoted by ***time*** and **status at exit** denoted by ***fate***. Stata interprets ***fate*** = 0 to mean that the patient is **censored** at exit and **fate** ≠ 0 to mean that she suffered the **event** of interest at exit.



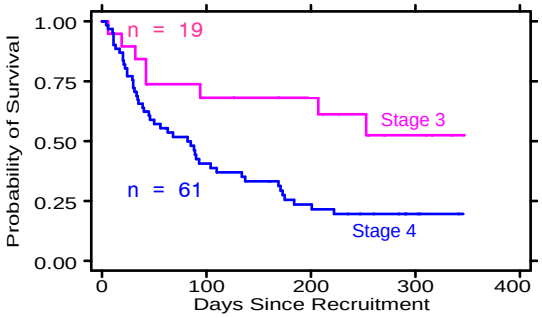
```
. * Graphics > Survival analysis graphs > Kaplan-Meier survivor function
. sts graph, by(stage) ytitle(Probability of Survival)              {4}

     failure time:  time
   failure/censor:  fate
```

{4} ***sts graph*** plots Kaplan-Meier survival curves. ***by(stage)*** specifies that **separate** plots will be generated for each value of *stage*. The *y*-axis title is *Probability of Survival*.



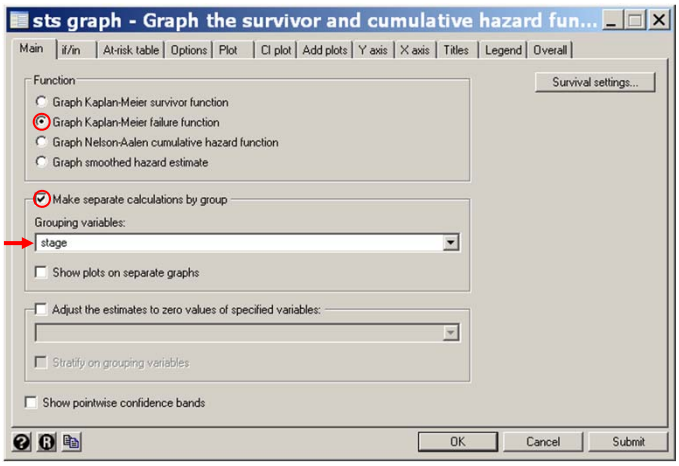


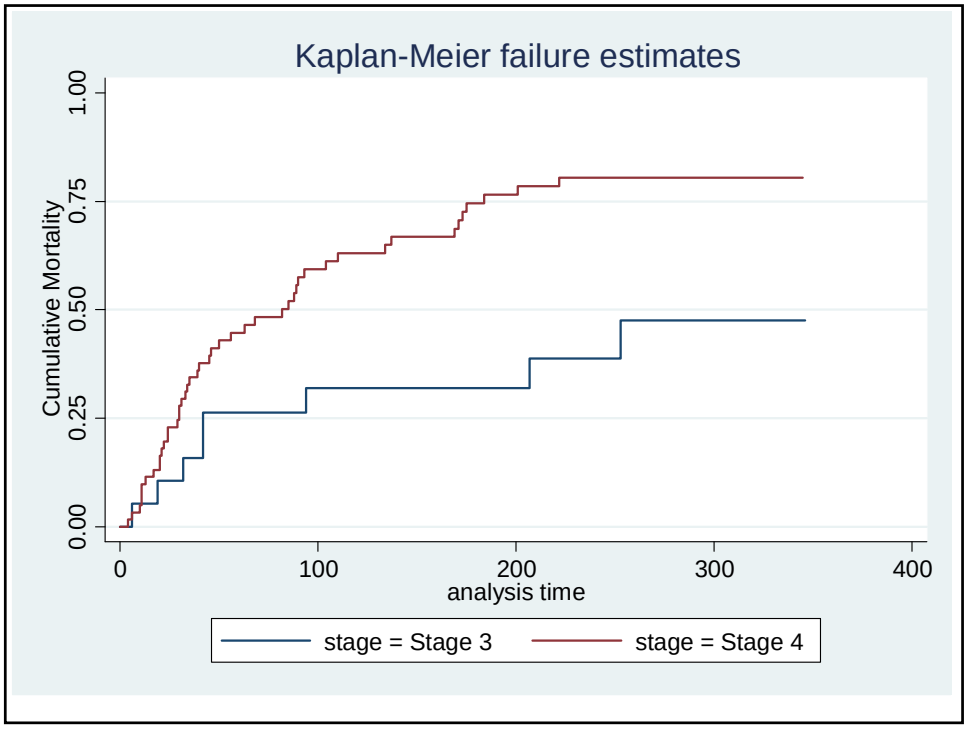
- In the preceding graph, $\hat{S}(t)$ is **constant** over days when **no deaths** are observed and **drops** abruptly on days when **deaths occur.**
- If the time interval is short enough that there is rarely more than one death per interval, then the **height** of the drop at each death day indicates the **size** of the cohort remaining on that day.
- The **accuracy** of the survival curve gets **less** as we move towards the right, as it is based on **fewer** and fewer **patients.**



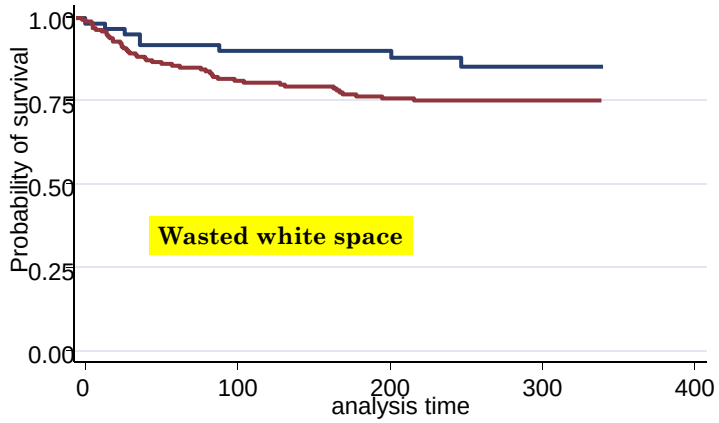
We can also plot the cumulative mortality curve using the **failure** option as follows

```
. * Graphics > Survival analysis graphs > Kaplan-Meier failure function
. sts graph, by(stage) ytitle(Cumulative Mortality) failure
```

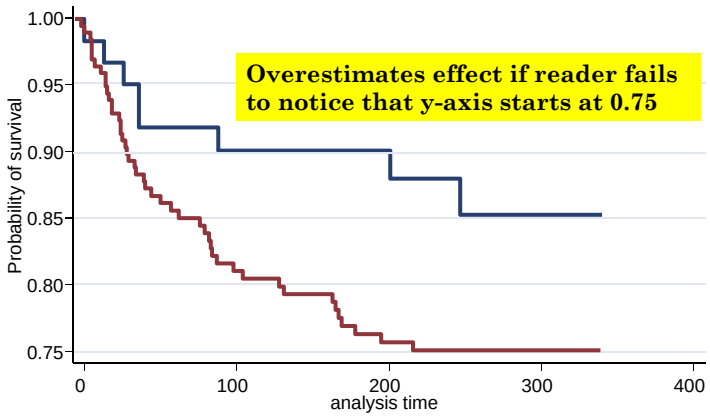




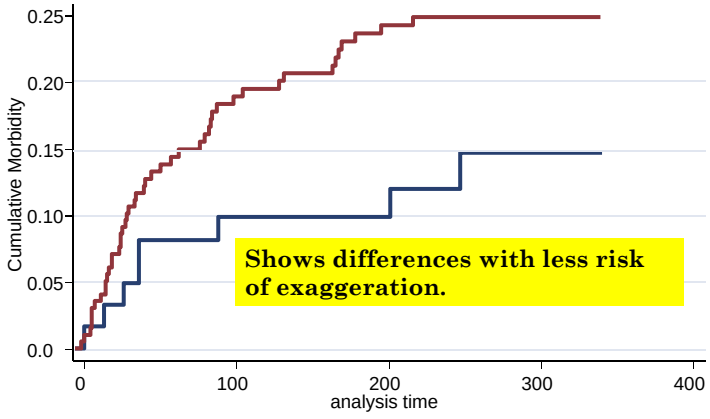
Cumulative morbidity plots are often better than survival plots when the overall survival is high.



Cumulative morbidity plots are often better than survival plots when the overall survival is high.



Cumulative morbidity plots are often better than survival plots when to overall survival is high.

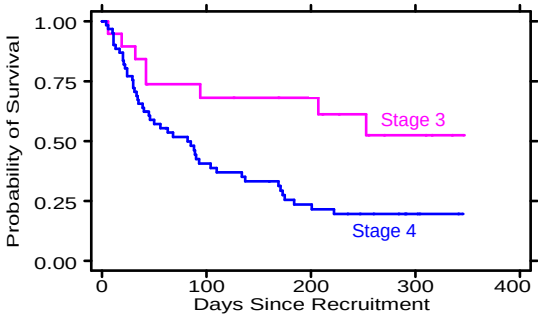


- If there is no censoring and there a q death days before time t then

$$\hat{S}(t) = \left(\frac{n_1 - d_1}{n_1}\right)\left(\frac{n_2 - d_2}{n_1 - d_1}\right)\ldots\left(\frac{n_q - d_q}{n_{q1} - d_{q1}}\right)$$

$$= \frac{n_q - d_q}{n_1} = \frac{m(t)}{n}$$

Hence the **Kaplan-Meier** survival curve reduces to the **proportion** of patients alive at time t if there is no **censoring**.



a) Life Tables

A life table is a table that gives estimates of $S(t)$ for different values of t. The term is slightly old fashioned but is still used.

5. 95% Confidence Intervals for Survival Functions

The variance of $\hat{S}(t)$ is estimated by Greenwood's formula

$$s^2_{\hat{S}(t)} = \hat{S}(t)^2 \sum_{\{k:t_k<t\}} \frac{d_k}{n_k(n_k - d_k)} \qquad \{7.2\}$$

A 95% confidence interval for $S(t)$ could be estimated by

$$\hat{S}(t) \pm 1.96 s_{\hat{S}(t)}$$

However, this interval does not optimal when $\hat{S}(t)$ is near 0 or 1 since this statistic will have a skewed distribution near these extreme values (the true survival curve is never less than 0 or greater than 1).

The variance of $\log\left[-\log\left[\hat{S}(t)\right]\right]$ has variance

$$\hat{\sigma}^2(t) = \frac{\sum_{\{k:t_k<t\}} \frac{d_k}{n_k(n_k-d_k)}}{\left[\sum_{\{k:t_k<t\}} \log\left[\frac{(n_k-d_k)}{d_k}\right]\right]^2} \qquad \{7.3\}$$

and a 95% confidence interval $\log\left[-\log\left[\hat{S}(t)\right]\right] \pm 1.96\hat{\sigma}(t)$.

Exponentiating twice gives a 95% confidence interval for $\hat{S}(t)$ of

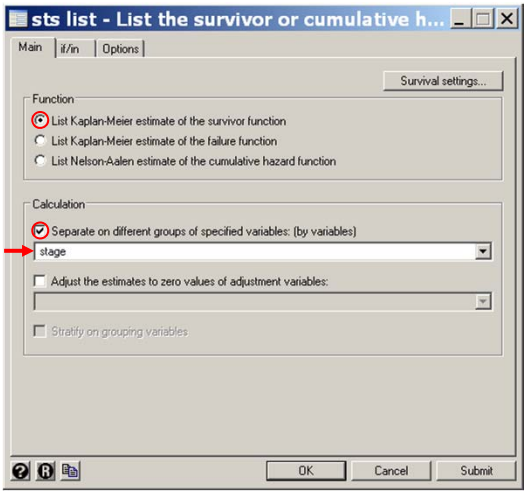
$$\hat{S}(t)^{\exp(\mp 1.96\hat{\sigma}(t))} \qquad \{7.4\}$$

which behaves better for extreme values of $\hat{S}(t)$. We can either list or plot these values with Stata. *Lymphoma.log* continues as follows:

```
. *
. *  List survival statistics
. *
. * Statistics > Survival... > Summary statistics... > List survivor...
. sts list, by(stage)                                                        {1}
    failure time:  time
  failure/censor:  fate
           Beg.          Net         Survivor      Std.
  Time     Total   Fail   Lost        Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
stage=3
     6        19      1      0          0.9474    0.0512     0.6812    0.9924
    19        18      1      0          0.8947    0.0704     0.6408    0.9726
    32        17      1      0          0.8421    0.0837     0.5865    0.9462
    42        16      2      0          0.7368    0.1010     0.4789    0.8810
    43        14      0      1          0.7368    0.1010     0.4789    0.8810
    94        13      1      0          0.6802    0.1080     0.4214    0.8421  {2}
         .
         .
         .
   335         2      0      1          0.5247    0.1287     0.2570    0.7363
   346         1      0      1          0.5247    0.1287     0.2570    0.7363
```

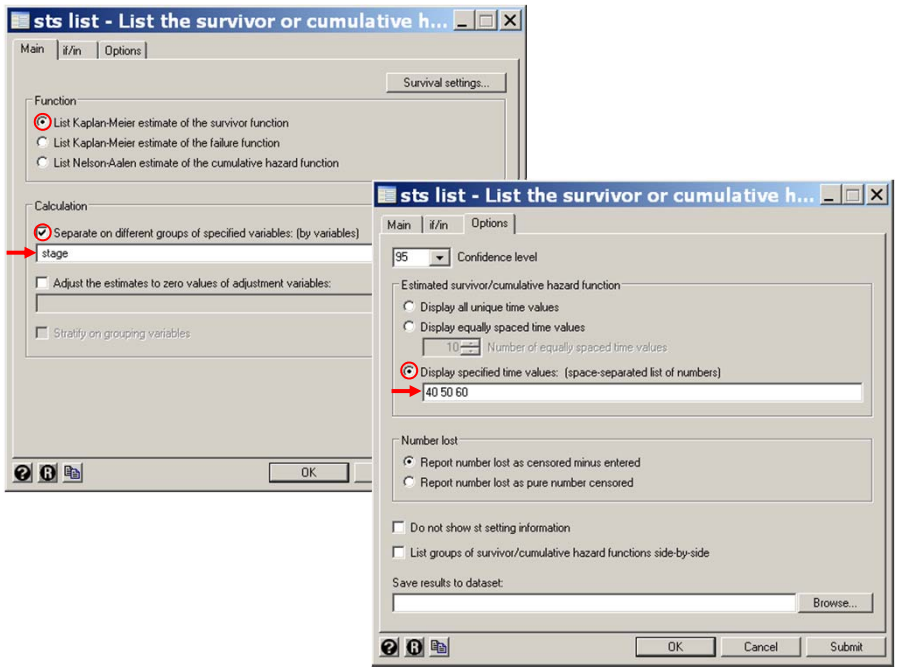
{1} ***sts list*** provides the same data that is plotted by *sts graph*.

{2} For example, of the original **19** stage three patients there are **13** still alive at the beginning of the 94 days of follow-up. There were **5 deaths** in this group before day **94** and one death on day **94**. The survivor Function $\hat{S}(94) = 0.68$, with standard error $s_{\hat{S}(t)} = 0.11$. The 95 % confidence interval for $\hat{S}(94)$ is (0.42, 0.84)



```
stage=4
     4       61       1       0          0.9836    0.0163     0.8893    0.9977
     6       60       1       0          0.9672    0.0228     0.8752    0.9917
                   .
                   .
                   .
   341        2       0       1          0.1954    0.0542     0.1026    0.3102
   345        1       0       1          0.1954    0.0542     0.1026    0.3102
-------------------------------------------------------------------------------

. * Statistics > Survival... > Summary statistics... > List survivor...
. sts list, by(stage) at(40 50 60) failure                                {3}

         failure _d:  fate
   analysis time _t:  time

             Beg.                    Failure       Std.
    Time    Total    Fail            Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
Stage 3
      40       17       3              0.1579    0.0837     0.0538    0.4135
      50       14       2              0.2632    0.1010     0.1190    0.5211
      60       14       0              0.2632    0.1010     0.1190    0.5211
Stage 4
      40       39      23              0.3770    0.0621     0.2690    0.5108
      50       34       3              0.4290    0.0637     0.3156    0.5630
      60       33       1              0.4463    0.0641     0.3315    0.5800
-------------------------------------------------------------------------------
Note:  Failure function is calculated over full data and evaluated at
       indicated times; it is not calculated from aggregates shown at left.
```

{3} The preceding **sts list** command can generate a very large listing for large data sets. If we want to know the survival function at specific values we can obtain them using the **at** option. If we wish cumulative morbidity rates rather than survival rates we can use the **failure** option. These options are illustrated with this command.




```
. *
. *  Kaplan-Meier survival curves by stage with 95% CIs
. *
. * Graphics > Survival analysis graphs > Kaplan-Meier survivor function
. sts graph, by(stage)   ci censored(single) separate     ///        {4}
>     xlabel(0 (50) 350) xmtick(0 (25) 350)               ///
>     byopts(title(, size(0)) legend(off))                ///        {5}
>     ytitle(Probability of Survival)                     ///
>     ylabel(0 (.1) 1, angle(0)) ciopts(color(yellow))    ///        {6}
>     xtitle(Days Since Recruitment)  ymtick(0 (.05) 1)
```

{4} Stata also permits users to graph confidence bounds for $\hat{S}(t)$ and to indicate when subjects lost to follow-up with tick marks. This is done with the ***ci*** and ***censored(single)*** options, respectively. The ***separate*** option causes the survival curves to be drawn in separate panels.

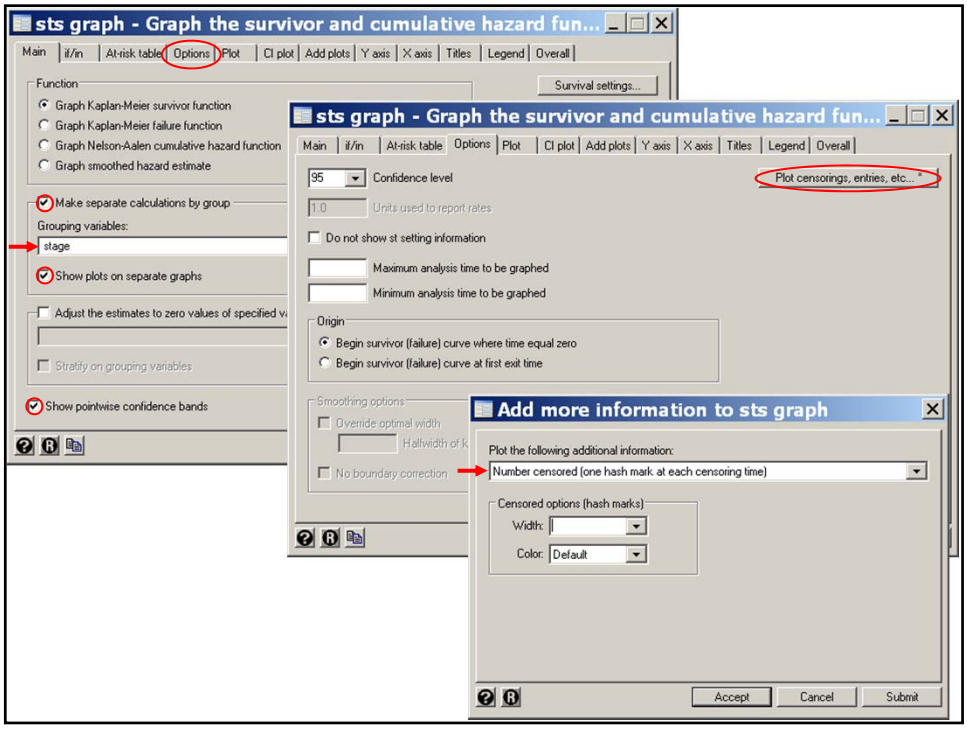
{5} The ***byopts*** option controls attributes related to having multiple curves on the same graph; ***title(" ", size(0))*** suppresses the graph's default title; ***legend(off)*** suppresses the legend. When the ***separate*** option is given ***title*** and ***legend*** must be suboptions of ***byopts*** rather than separate options.

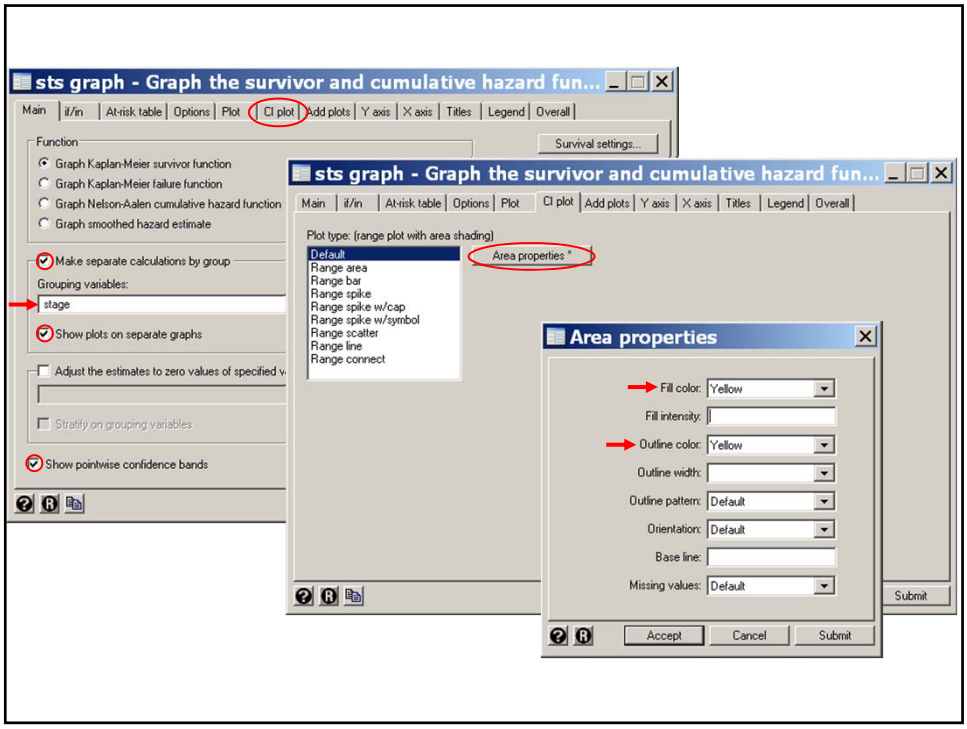
{6} The **ciopts** option allows control of the confidence bands. Here we choose yellow bands.

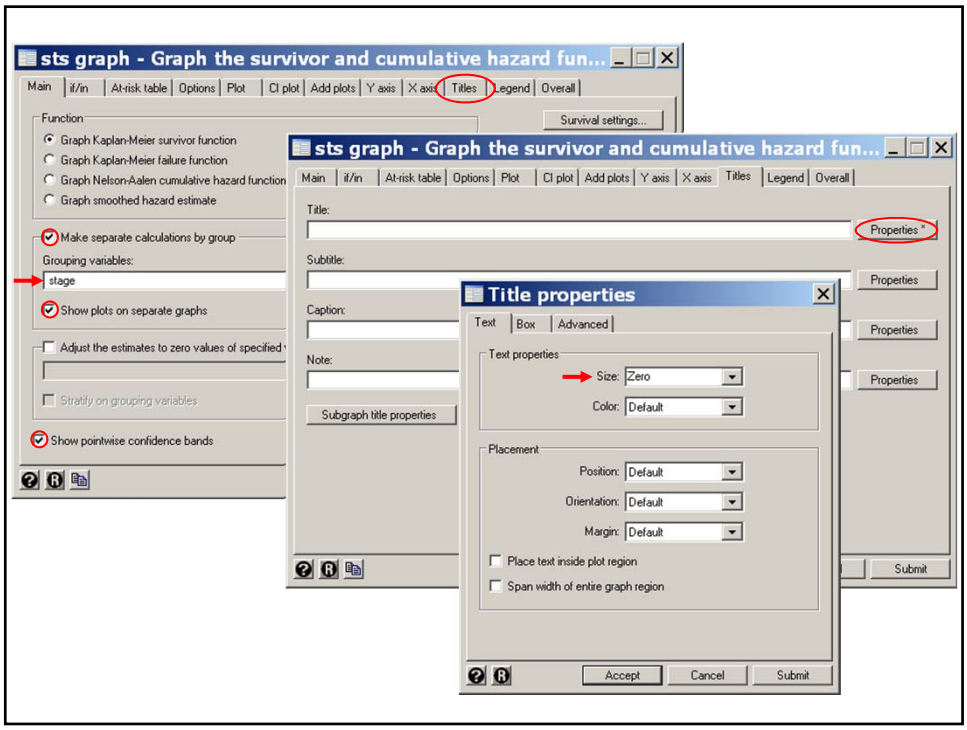
{4} Stata also permits users to graph confidence bounds for $\hat{S}(t)$ and to indicate when subjects lost to follow-up with tick marks. This is done with the ***ci*** and ***censored(single)*** options, respectively. The ***separate*** option causes the survival curves to be drawn in separate panels.

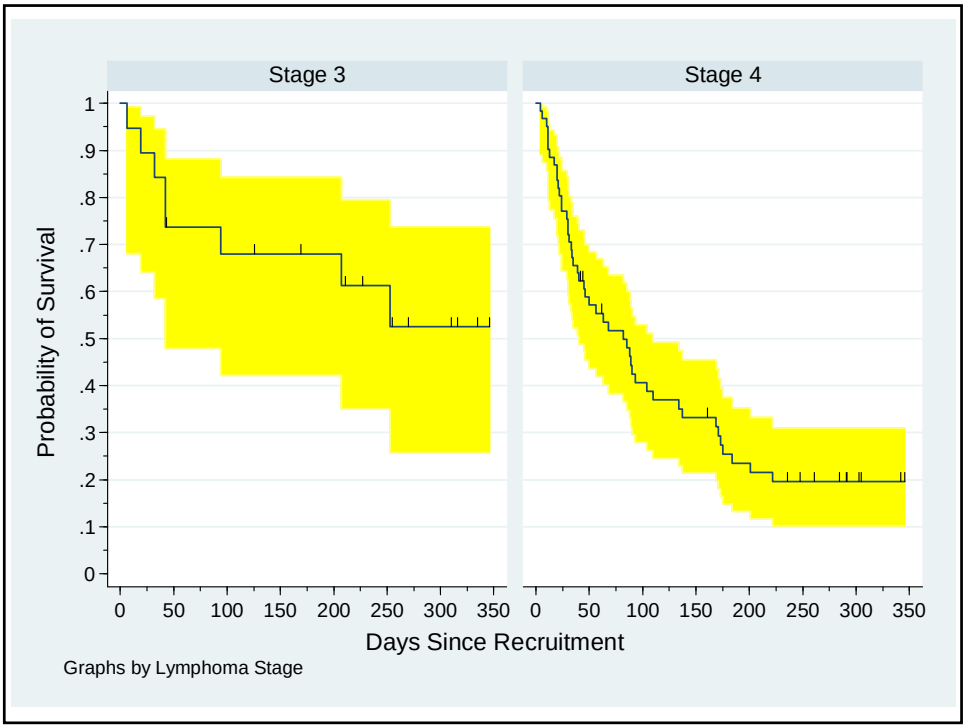
{5} The ***byopts*** option controls attributes related to having multiple curves on the same graph; ***title(" ", size(0))*** suppresses the graph's default title; ***legend(off)*** suppresses the legend. When the ***separate*** option is given ***title*** and ***legend*** must be suboptions of ***byopts*** rather than separate options.

{6} The **ciopts** option allows control of the confidence bands. Here we choose yellow bands.




sts graph - Graph the survivor and cumulative hazard fun...
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
Survival settings...
Make separate calculations by group
Grouping variables:
stage
Show plots on separate graphs
Adjust the estimates to zero values of specified
Stratify on grouping variables
Show pointwise confidence bands
Plot type: (range plot with area shading)
Default
Range area
Range bar
Range spike
Range spike w/cap
Range spike w/symbol
Range scatter
Range line
Range connect
Area properties *
Area properties
Fill color: Yellow
Fill intensity:
Outline color: Yellow
Outline width:
Outline pattern: Default
Orientation: Default
Base line:
Missing values: Default
Accept
Cancel
Submit

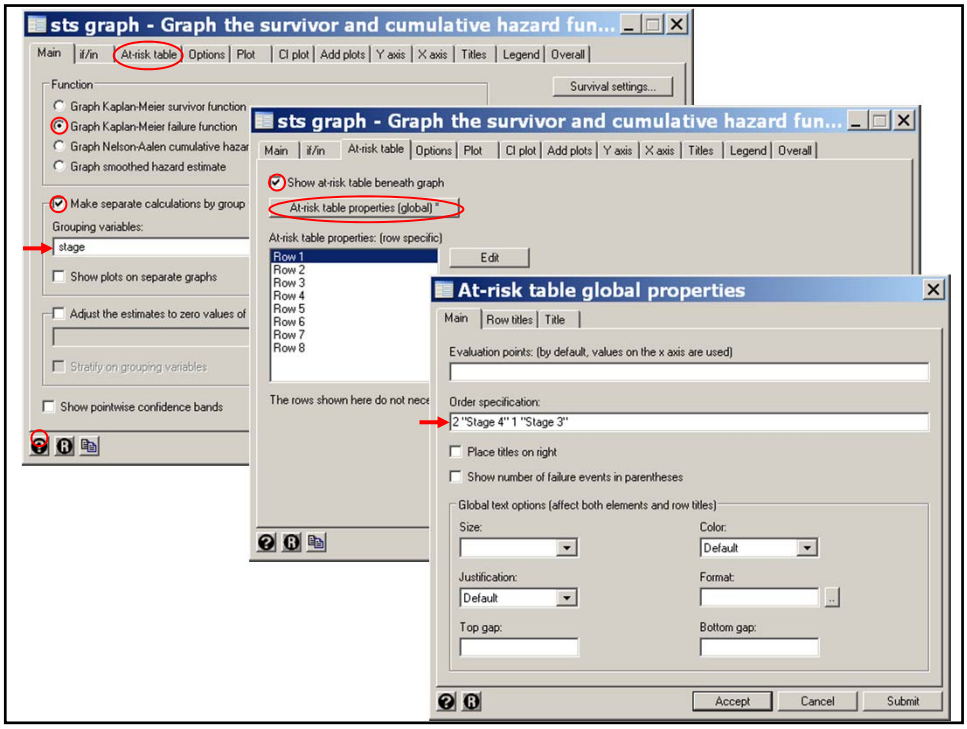

sts graph - Graph the survivor and cumulative hazard fun...
Main | if/in | At-risk table | Options | Plot | CI plot | Add plots | Y axis | X axis | Titles | Legend | Overall
Function
Graph Kaplan-Meier survivor function
Graph Kaplan-Meier failure function
Graph Nelson-Aalen cumulative hazard function
Graph smoothed hazard estimate
Survival settings...
Make separate calculations by group
Grouping variables:
stage
Show plots on separate graphs
Adjust the estimates to zero values of specified
Stratify on grouping variables
Show pointwise confidence bands
Title:
Properties *
Subtitle:
Caption:
Note:
Subgraph title properties
Title properties
Text | Box | Advanced
Text properties
Size: Zero
Color: Default
Placement
Position: Default
Orientation: Default
Margin: Default
Place text inside plot region
Span width of entire graph region
Accept
Cancel
Submit

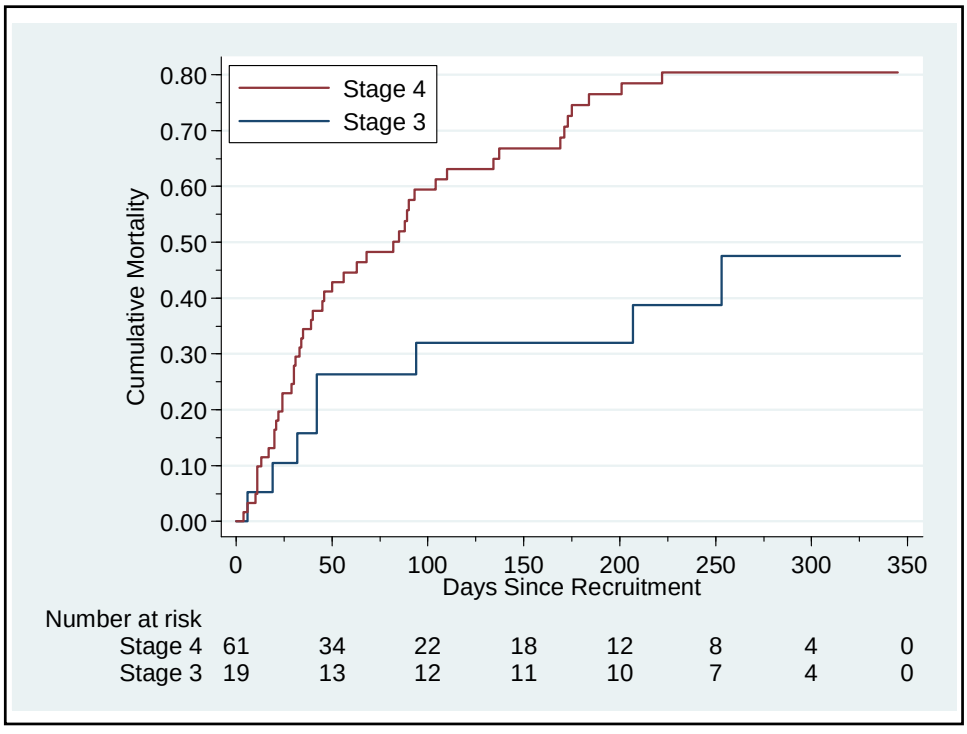


Some journals require a table showing the number of subjects at risk at different survival times given below the survival curve. In Stata this can be done as follows.

```
. *
. *  Kaplan-Meier morbidity curves by stage with risk table
. *
. * Graphics > Survival analysis graphs > Kaplan-Meier failure function
. sts graph, by(stage)  failure                                    ///
>     risktable(,order(2 "Stage 4" 1 "Stage 3"))                   /// {7}
>     ytitle(Cumulative Mortality)                                 ///
>     xlabel(0 (50) 350) xmtick(0 (25) 350)                        ///
>     ylabel(0 (.1) .8, angle(0))                                  ///
>     xtitle(Days Since Recruitment)  ymtick(0 (.05) .8)           ///
>     title(" ",size(0)) legend(ring(0) cols(1)                    ///
>        position(11) order(2 "Stage 4" 1 "Stage 3"))
```

{7} The ***risktable*** option creates a risk table below the graph with one row for each curve that is drawn. The ***order*** suboption orders and labels these rows. Its syntax is identical to that of the ***order*** suboption of the ***legend*** option.


sts graph - Graph the survivor and cumulative hazard fun...
Main
if/in
At-risk table
Options
Plot
CI plot
Add plots
Y axis
X axis
Titles
Legend
Overall
Function
Survival settings...
Graph Kaplan-Meier survivor function
Graph Kaplan-Meier failure function
Graph Nelson-Aalen cumulative hazar
Graph smoothed hazard estimate
Make separate calculations by group
Grouping variables:
stage
Show plots on separate graphs
Adjust the estimates to zero values of
Stratify on grouping variables
Show pointwise confidence bands
sts graph - Graph the survivor and cumulative hazard fun...
Show at-risk table beneath graph
At-risk table properties (global) *
At-risk table properties: (row specific)
Edit
Row 1
Row 2
Row 3
Row 4
Row 5
Row 6
Row 7
Row 8
The rows shown here do not nece
At-risk table global properties
Main
Row titles
Title
Evaluation points: (by default, values on the x axis are used)
Order specification:
2 "Stage 4" 1 "Stage 3"
Place titles on right
Show number of failure events in parentheses
Global text options (affect both elements and row titles)
Size:
Color:
Default
Justification:
Default
Format:
Top gap:
Bottom gap:
Accept
Cancel
Submit


Stage 4
Stage 3
Cumulative Mortality
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0
50
100
150
200
250
300
350
Days Since Recruitment
Number at risk
Stage 4 61 34 22 18 12 8 4 0
Stage 3 19 13 12 11 10 7 4 0

6. Censoring and Bias

Kaplan-Meier survival curves will be unbiased estimates of the true survival curve as long as

1. The patients are representative of the underlying population and
2. Patients who are censored have the same risk of suffering the event of interest as are patients who are not.

If censored patients are more likely to die than uncensored patients with equal follow-up then our survival estimates will be biased.

Such bias can occur for many reasons, not the least of which is that dead patients do not return for follow-up visits.

Survival curves are often derived for some endpoint other than death. In this case, some deaths may be treated as censoring events.

For example, if the event of interest is developing of breast cancer, then we may treat death due to heart disease as a censoring event. This is reasonable as long as there is no relationship between heart disease and breast cancer. That is, when we censor a woman who died of heart disease, we are assuming that she would have had the same subsequent risk of breast cancer as other women if she had lived.

If we were studying lung cancer, then treating death from heart disease as a censoring event would bias our results since smoking increases the risk of both lung cancer morbidity and cardiovascular mortality and patients who die of heart disease are more likely to have smoked and hence would have been more likely to develop lung cancer if they had not died of heart disease first.

7. Log-Rank Test

a) Mantel-Haenszel test for survivorship data

Suppose that two treatments have survival curves $S_1[t]$ and $S_2[t]$

We wish to test the **null hypothesis** that

$$H_0: S_1[t] = S_2[t] \quad \text{for all } t$$

Suppose that on the k^{th} death day that there are n_{1k} and n_{2k} patients at risk on treatments 1 and 2 and that d_{1k} and d_{2k} deaths occur in these groups on this day.

Let $D_k = d_{1k} + d_{2k}$

$N_k = n_{1k} + n_{2k}$

Then the **observed death rate** on the k^{th} death day is D_k / N_k .

If the null hypothesis is true then the expected number of deaths in each group is

$$E[d_{1k} \,|\, D_k] = n_{1k}[D_k / N_k) \quad \text{and} \quad E[d_{2k} \,|\, D_k] = n_{2k}[D_k / N_k)$$

The greater the difference between d_{1k} and $E[d_{1k} \,|\, D_k]$, the greater the evidence that the null hypothesis is false.

Mantel proposed forming the 2x2 contingency tables

k^{th} death day	Treatment 1	Treatment 2	Total
Died	d_{1k}	d_{2k}	D_k
Survived	$n_{1k} - d_{1k}$	$n_{2k} - d_{2k}$	$N_k - D_k$
Total	n_{1k}	n_{2k}	N_k

on each death day and performing a Mantel-Haenszel χ^2 test.

This test was renamed the **log-rank test** by Peto who studied its mathematical properties.

If the time interval is short enough that $d_k \leq 1$ for each interval, then the test of H_0 depends only on the order in which the deaths occur and not on their time of occurrence.

It is in this sense that the test is a rank test.

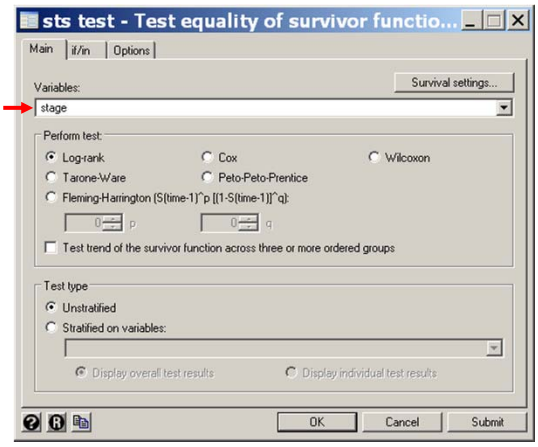
b) Example: Tumor stage in lymphoma patients

Lymphoma.log continues as follows:

```
. * Statistics > Survival... > Summary... > Test equality of survivor...
. sts test stage
                                                                        {1}
     failure _d:  fate
   analysis time _t:  time

Log-rank test for equality of survivor functions

      |  Events          Events
stage |  observed       expected
------+-------------------------
3     |         8          16.69
4     |        46          37.31
------+-------------------------
Total |        54          54.00

           chi2(1) =      6.71
           Pr>chi2 =    0.0096                                        {2}
```

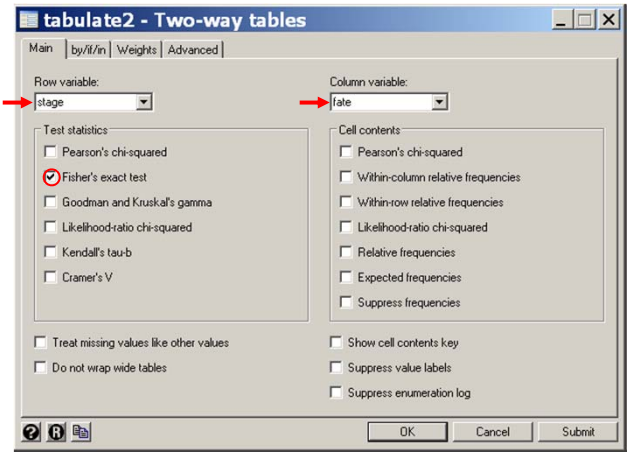
{1} Perform a **log-rank** test for equality of survivor functions in patient groups defined by different values of *stage*. In this example, stage 3 patients are compared to stage 4 patients.

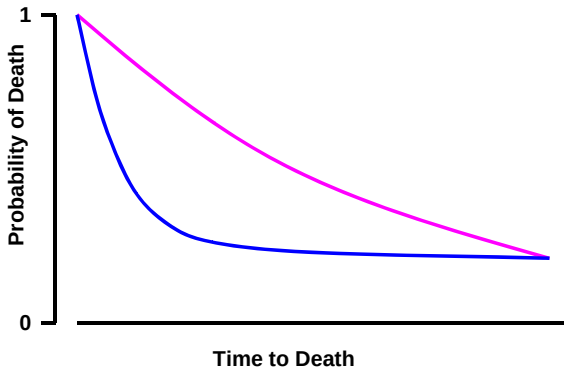
{2} In this example, the log-rank P value = **0.0096**, indicating that the marked **difference** in survivorship between stage **3** and stage **4** lymphoma patients is not likely to be due to chance.




```
. * Statistics > Summaries... > Tables > Two-way tables with measures...
. tabulate stage fate, exact                                          {3}

Lymphoma   | fate
Stage      |     Alive       Dead |     Total
-----------+----------------------+----------
         3 |        11          8 |        19
         4 |        15         46 |        61
-----------+----------------------+----------
     Total |        26         54 |        80

          Fisher's exact =                 0.011
  1-sided Fisher's exact =                 0.009
```

{3} The ***tabulate*** command cross-tabulates patients by stage and fate. The ***exact*** option calculates Fisher's exact test of the hypothesis that the proportion of deaths in the two groups are equal. Fisher's **exact** test differs from the **log-rank** test in that the latter takes into consideration **time to death** as well as numbers of deaths while the former only considers **numbers** of deaths. In this example, the two tests give very similar results. However, if the true survival curves look like this .....





…the log-rank test may be highly significant even though the observed death rates in each group are equal. Fisher's exact test, however, will not be significant if the death rates are the same.

c) Log-rank test for multiple patient groups

The log-rank test generalizes to allow the comparison of survival in several groups.

These groups are defined by the number of distinct levels taken by the variable specified in the *sts test* command. E.g. in the preceding example if there were four different lymphoma stages define by *stage* then *sts test stage* would compare the four survival curves for these groups of patients. The test statistic has an asymptotic χ^2 distribution with one degree of freedom less than the number of patient groups being compared.

8. Hazard Functions

Suppose that a patient is alive at time t and that her probability of dying in the short time interval $(t, t+\Delta t)$ is

$$\lambda[t]\Delta t$$

Then $\lambda[t]$ is said to be the hazard function for the patient at time t.

More precisely

$$\lambda[t] = \frac{\Pr\left[\begin{array}{l|l}\text{Patient dies by} & \text{Patient alive} \\ \text{time } t+\Delta t & \text{at time } t\end{array}\right]}{\Delta t} \qquad \{7.5\}$$

For a very large population

$$\lambda[t]\Delta t \cong \frac{\text{The number of deaths in the interval } (t, t+\Delta t)}{\text{Number of people alive at time } t}$$

$\lambda[t]$ is the instantaneous rate per unit time at which people are dying at time t.

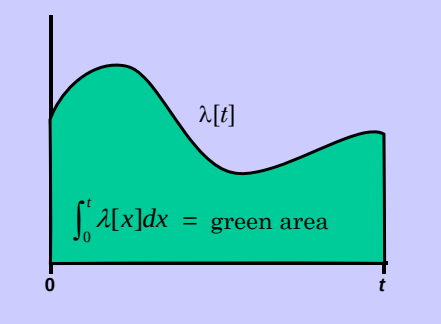
$\lambda[t] = 0$ implies that there is no risk of death at time t and $S[t]$ is flat at time t.

Large values of $\lambda[t]$ imply a rapid rate of decline in $S[t]$.

The hazard function is related to the survival function through the equation

$$S[t] = \exp\left[-\int_0^t \lambda[x]dx\right]$$

where $\int_0^t \lambda[x]dx$ is the area under the curve $\lambda[x]$ between 0 and t.



a) Proportional hazards

Suppose that $\lambda_0[t]$ and $\lambda_1[t]$ are the hazard functions for control and experimental for treatments, respectively.

Then these treatments have proportional hazards if

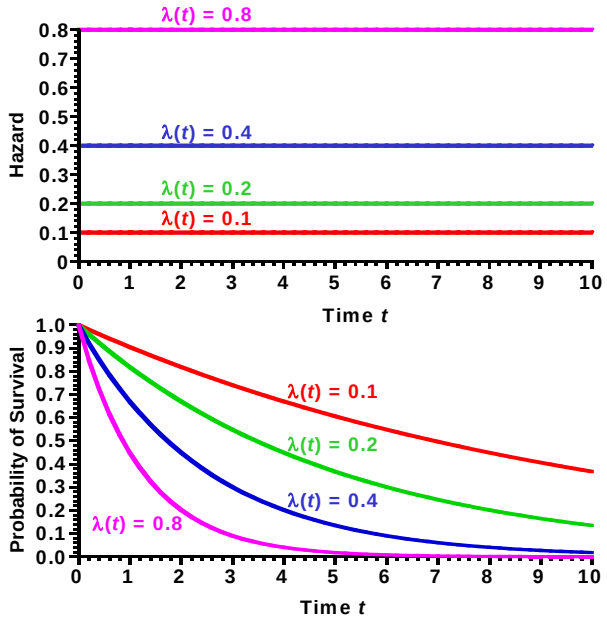
$$\lambda_1[t] = R\,\lambda_0[t]$$

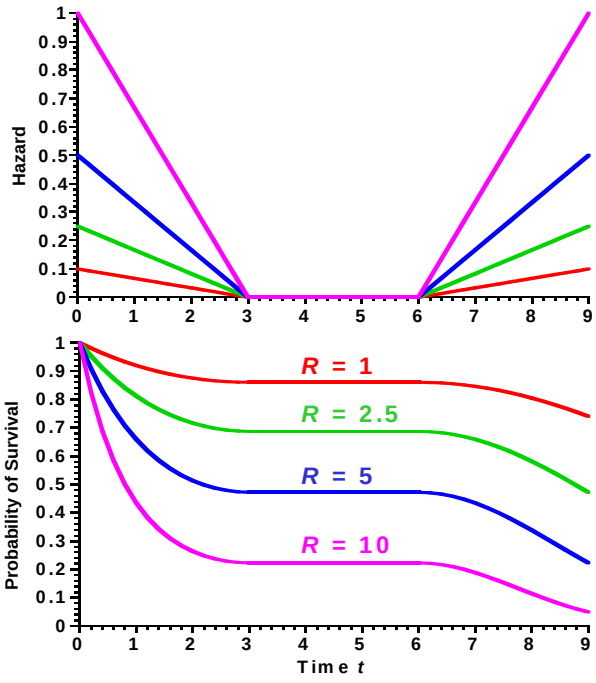
for some constant R.

The proportional hazards assumption places no restrictions on the shape of $\lambda_0(t)$ but requires that

$$\lambda_1[t] / \lambda_0[t] = R$$

Examples:





b) Relative risks and hazard ratios

Suppose that the risks of death by time $t + \Delta t$ for patients on control and experimental treatments who are alive at time t are $\lambda_0[t]\Delta t$ and, $\lambda_1[t]\Delta t$ respectively.

Then the risk of experimental subjects at time t relative to control is

$$\frac{\lambda_1[t]\Delta t}{\lambda_0[t]\Delta t} = \frac{\lambda_1[t]}{\lambda_0[t]}$$

If $\lambda_1[t] = R\lambda_0[t]$ at all times, then this relative risk is

$$\frac{\lambda_1[t]}{\lambda_0[t]} = \frac{R\lambda_0[t]}{\lambda_0[t]} = R$$

Thus the ratio of two hazard functions can be thought of as an instantaneous relative risk, or as a relative risk if this ratio is constant.

9. Proportional Hazards Regression Analysis

a) The model

Suppose that $\lambda_0[t]$ and $\lambda_1[t]$ are the hazard functions for the control and experimental therapies and β is an unknown parameter. The **proportional hazards** model assumes that

$$\lambda_1[t] = \lambda_0[t]\exp[\beta]$$

This model is said to be semi-nonparametric in that it makes no assumptions about the shape of the control hazard function.

If $\hat{\beta}$ is an estimate of β then $\exp[\hat{\beta}]$ estimates the relative risk of the experimental therapy relative to controls since

$$R = \frac{\lambda_1[t]}{\lambda_0[t]} = \frac{\exp[\beta]\lambda_0[t]}{\lambda_0[t]} = \exp[\beta]$$

b) Example: Risk of stage 3 vs. stage 4 lymphoma

In Stata proportional hazards regression analysis is performed by the *stcox* command. The *Lymphoma.log* file continues as follows.

```
. *
. *  Preform proportional hazards regression analysis of
. *  lymphoma patients by stage of tumor.
. *
. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox stage                                                                 {1}

        failure _d:  fate
  analysis time _t:  time

Iteration 0:  Log Likelihood = -207.5548
Iteration 1:  Log Likelihood =-203.86666
Iteration 2:  Log Likelihood =-203.73805
Iteration 3:  Log Likelihood =-203.73761
Refining estimates:
Iteration 0:  Log Likelihood =-203.73761

Cox regression -- Breslow method for ties

No. of subjects =           80                     Number of obs    =        80
No. of failures =           54
Time at risk    =         9718
                                                   LR chi2(1)       =      7.63
Log likelihood  =   -203.73761                     Prob > chi2      =    0.0057

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       stage |   2.614362   1.008191     2.49   0.013     1.227756    5.566976 {2}
------------------------------------------------------------------------------
```

{1} This command fits the **proportional hazards** regression model.

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

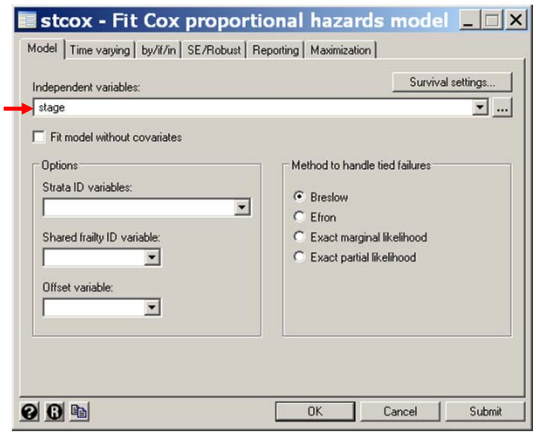
A **stset** command must precede the **stcox** command to define the fate and follow-up variables.

This model can be written $\lambda(t,3) = \lambda_0(t)e^{3\beta}$ and $\lambda(t,4) = \lambda_0(t)e^{4\beta}$ for stage 3 and 4 patients, respectively. Hence the hazard ratio for stage **4** patients **relative** to stage **3** patients is

$$\frac{\lambda(t,4)}{\lambda(t,3)} = \frac{\lambda_0(t)e^{4\beta}}{\lambda_0(t)e^{3\beta}} = e^{4\beta-3\beta} = e^{\beta}$$

which we interpret as the **relative risk** of death for stage **4** patients compared to stage **3** patients. Note that we could have redefined stage to be an indicator variable that equals 1 for stage 4 patients and 0 for stage 3 patients. Had we done that, the hazard for stage 3 and 4 patients would have been $\lambda_0(t)$ and $\lambda_0(t)e^{\beta}$ respectively. The **hazard ratio,** however, would still be e^{β}

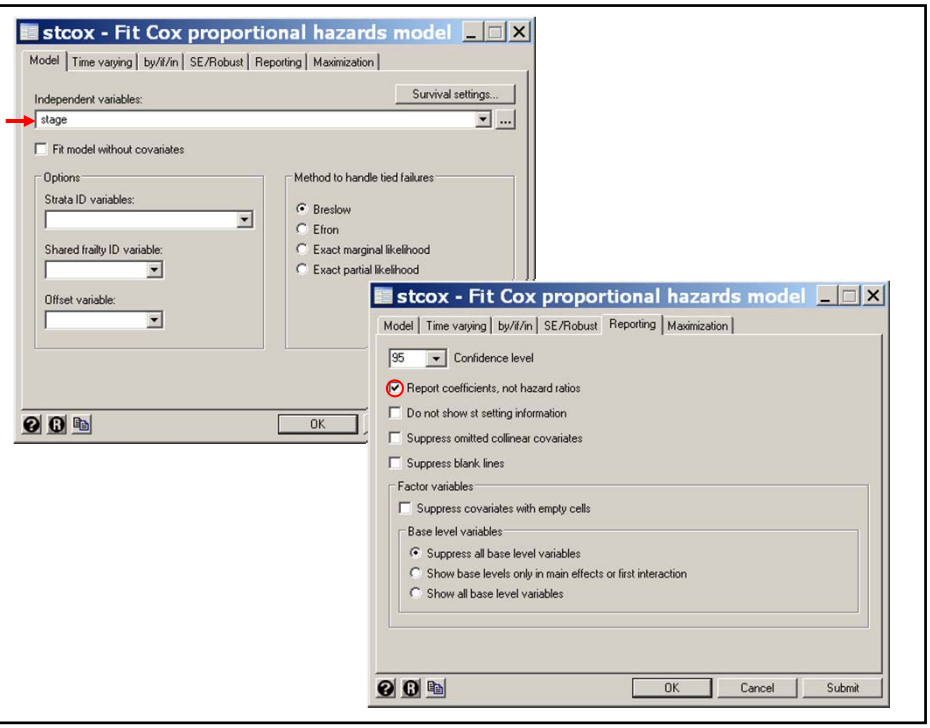
{2} This **hazard ratio** or **relative risk** equals 2.61 and is significantly different from zero (P=0.013)



```
. * Statistics > Survival... > Regression... > Cox proportional hazards model
. stcox stage,nohr                                                          {3}

         failure _d:  fate
   analysis time _t:  time

Iteration 0:  Log Likelihood = -207.5548
Iteration 1:  Log Likelihood =-203.86666
Iteration 2:  Log Likelihood =-203.73805
Iteration 3:  Log Likelihood =-203.73761
Refining estimates:
Iteration 0:  Log Likelihood =-203.73761

Cox regression -- Breslow method for ties

No. of subjects =           80                     Number of obs    =        80
No. of failures =           54
Time at risk    =         9718
                                                   LR chi2(1)       =      7.63
Log likelihood  =   -203.73761                     Prob > chi2      =    0.0057

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       stage |   .9610202   .3856356     2.49   0.013     .2051884    1.716852 {4}
------------------------------------------------------------------------------
```

{3} It is often useful to obtain direct estimates of the parameters of a hazard regression model. We do this with the ***nohr*** option, which stands for *no hazards ratios*.

{4} The estimate of β is 0.961. Note that exp(0.961) = 2.61, the hazard ratio obtained previously.



c) Estimating relative risks together with their 95% confidence intervals

The mortal risk of stage 4 lymphoma patients relative to stage 3 patients is exp(0.9610) = 2.61.

The **95% confidence interval** for this risk is

(2.61exp(-1.96*0.3856), 2.61exp(1.96*0.3856))

= (1.2, 5.6).

Note that Stata gave us this confidence interval when we did not specify the *nohr* option.

```
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       stage |   .9610202   .3856356    2.492   0.013     .2051884    1.716852
------------------------------------------------------------------------------


------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       stage |   2.614362   1.008191    2.492   0.013     1.227756    5.566976
------------------------------------------------------------------------------
```

d) Tied failure times

The most straight forward computational approach to the proportional hazards model can produce biased parameter estimates if a large proportion of the failure times are identical. For this reason it is best to record failure times as precisely as possible to avoid ties in this variable.

If there are extensive ties in the data, the *exactm, exactp,* or *efron* options of the *stcox* commands may be used to reduce this bias.

exactm and *exactp* are the most accurate, but can be computationally intensive.

An alternate approach is to use Poisson regression, which will be discussed in Chapters 7 and 8.

10. What we have covered

- Survival data: time to event
  - Right censored data
- Kaplan-Meier survival curves: the ***sts graph*** command
- Kaplan-Meier cumulative mortality curves: the ***failure*** option
  - Greenwood confidence bands for survival and mortality curves
    the ***ci*** option
  - Displaying censoring times
    the ***censored(single)*** option
  - Displaying numbers of patients at risk
    the ***risktable*** option
- Estimating survival probabilities: the ***sts list*** command
- Censoring and biased Kaplan-Meier survival curves
- Log rank test for comparing survival curves: the ***sts test*** command
- Hazard functions and cumulative mortality
  - Hazard rate ratios and relative risk
  - Estimating relative risks from proportional hazards models
- Simple proportional hazards regression model: the ***stcox*** command
- Tied failure times and biased relative risk estimates

Cited References

Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research*. Malden MA: Blackwell Science, Inc. 2002.

McKelvey EM, Gottlieb JA, Wilson HE, Haut A, Talley RW, Stephens R, Lane M, Gamble JF, Jones SE, Grozea PN, Gutterman J, Coltman C, Moon TE. Hydroxyldaunomycin (Adriamycin) combination chemotherapy in malignant lymphoma. *Cancer* 1976;38:1484-93.

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.

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

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

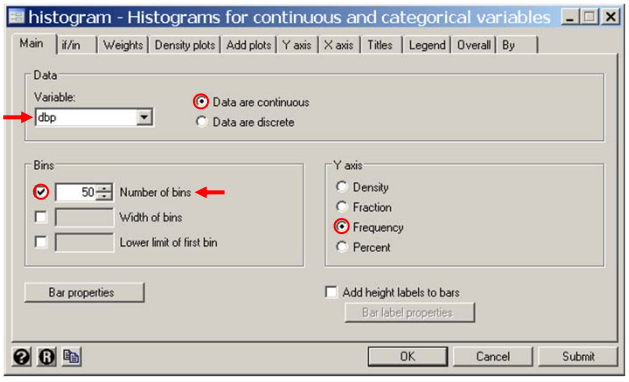
estimates an odds ratio, while in proportional hazards regression this expression estimates a relative risk.

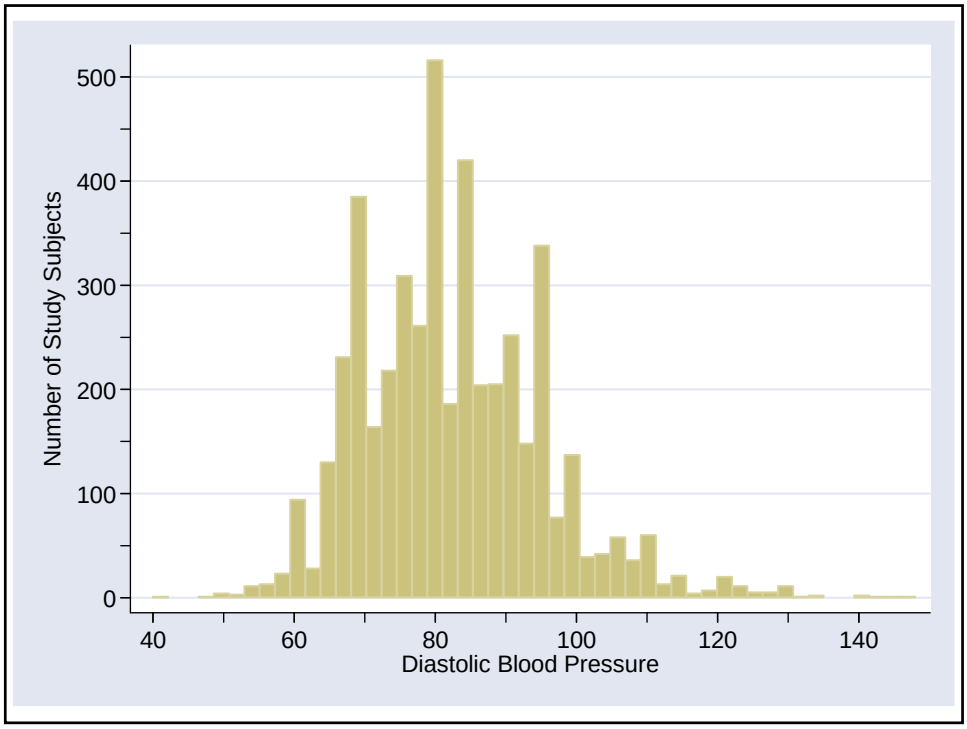
b) Example: Diastolic blood pressure and gender on risk of coronary heart disease

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 $110 < dbp$.

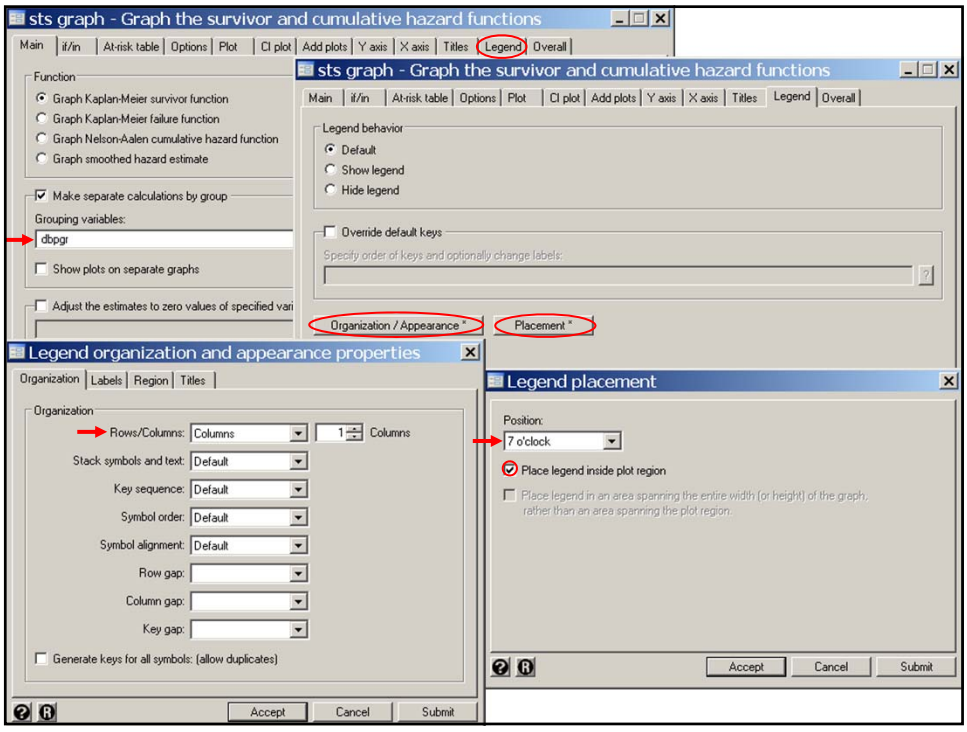
{3} This ***tabulate*** statement shows that the preceding ***recode*** statement **worked**. Subjects with DBPs less than 61 or greater than 110 are rare. However, the database is large enough to provide **255** such subjects.

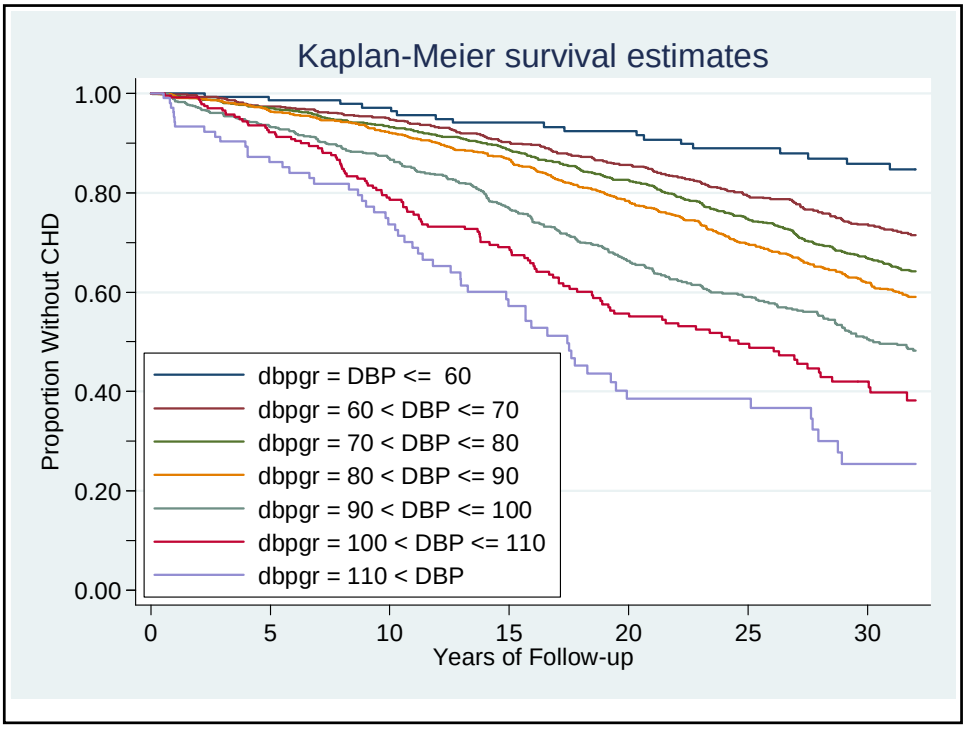
```
. * Variables Manager
. label define dbp 60  "DBP <=  60"      70  "60 < DBP <= 70"         ///
>                  90  "80 < DBP <= 90"  80  "70 < DBP <= 80"         ///
>                 100 "90 < DBP <= 100" 110 "100 < DBP <= 110" 111 "110 < DBP"

. 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.



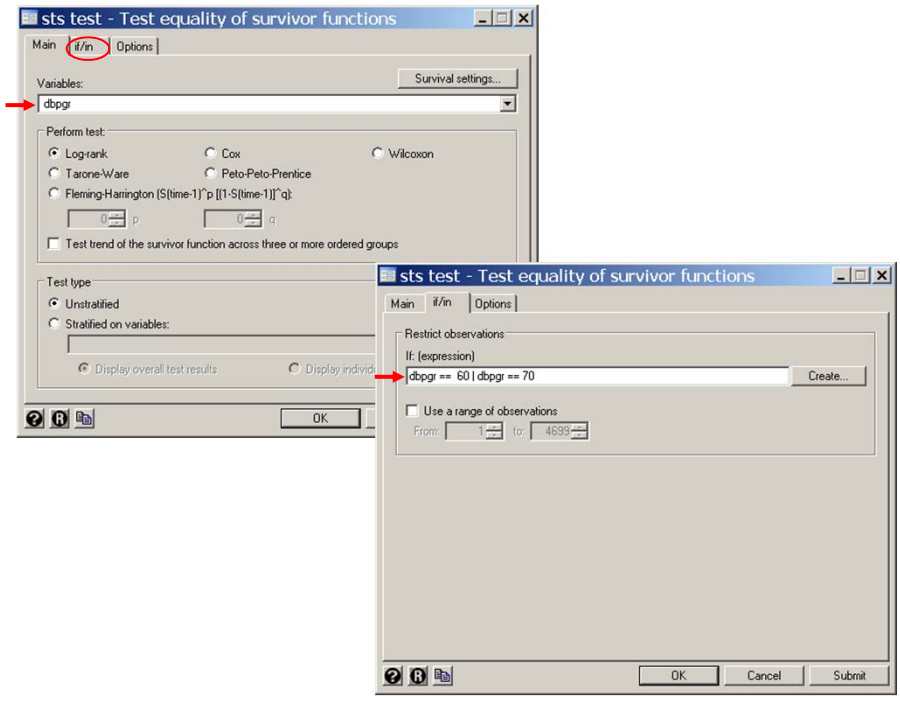



```
. * Statistics > Survival... > Summary... > Test equality of survivor...
. sts test dbpgr                                                        {6}

         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
```

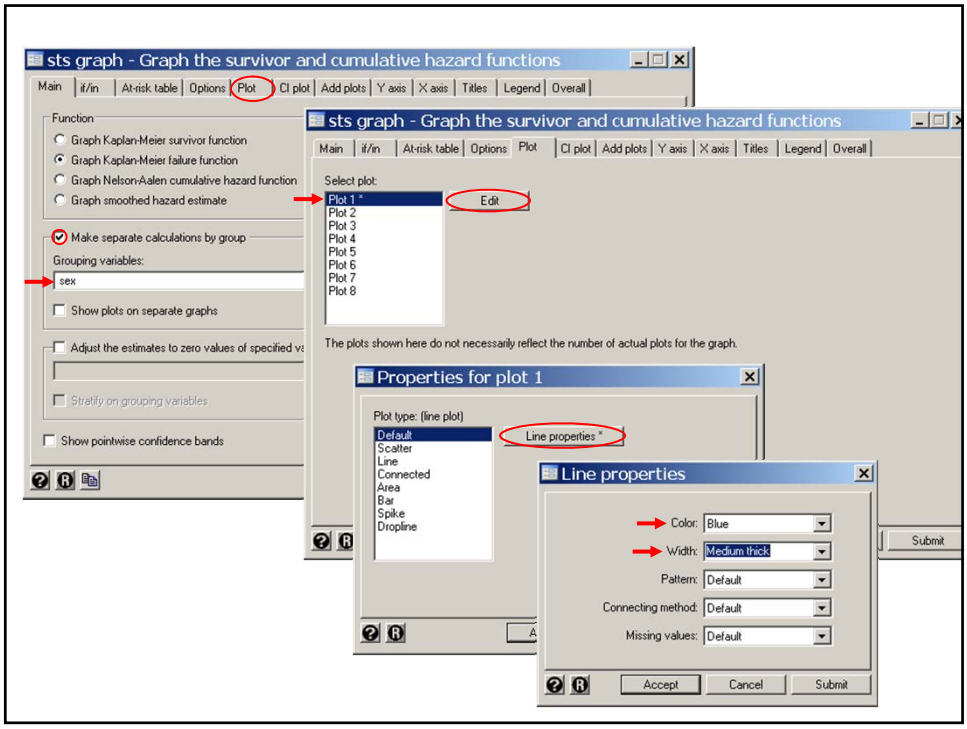
{8} All pair-wise logrank tests of adjacent DBP group levels are not statistically significant (output deleted).

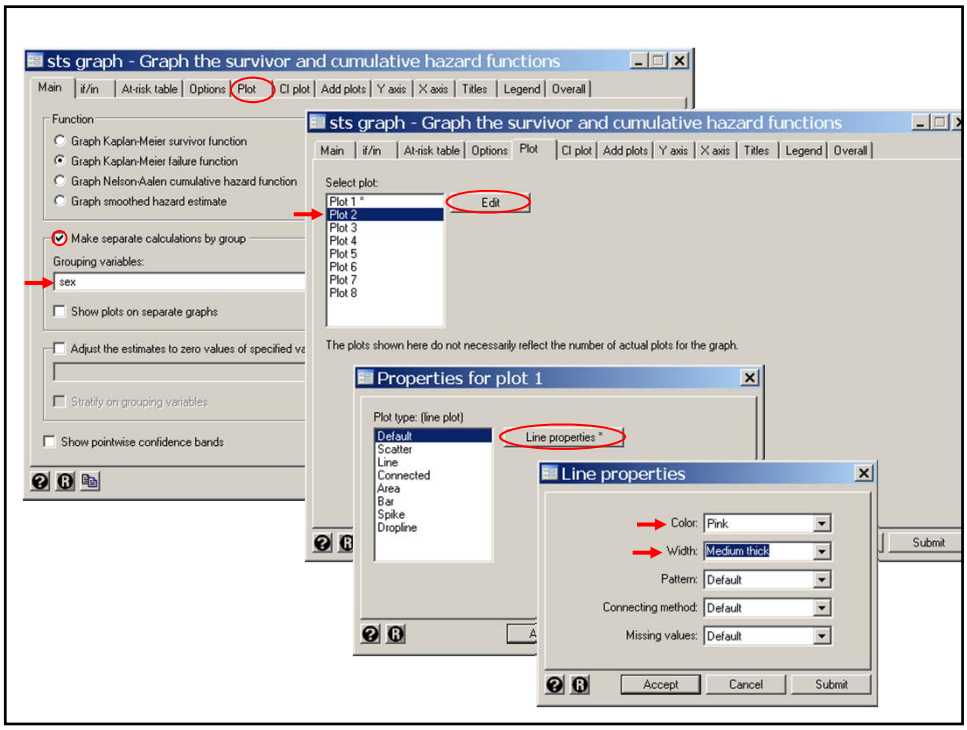
```
. *
. *  Univariate analysis of the effect of gender on CHD
. *
. * Graphics > Survival analysis graphs > Kaplan-Meier survivor function
. sts graph, by(sex) plot1opts(color(blue) lwidth(medthick) )     ///  {9}
>    plot2opts(color(pink)  lwidth(medthick))                     ///
>    ytitle(Cumulative 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
```

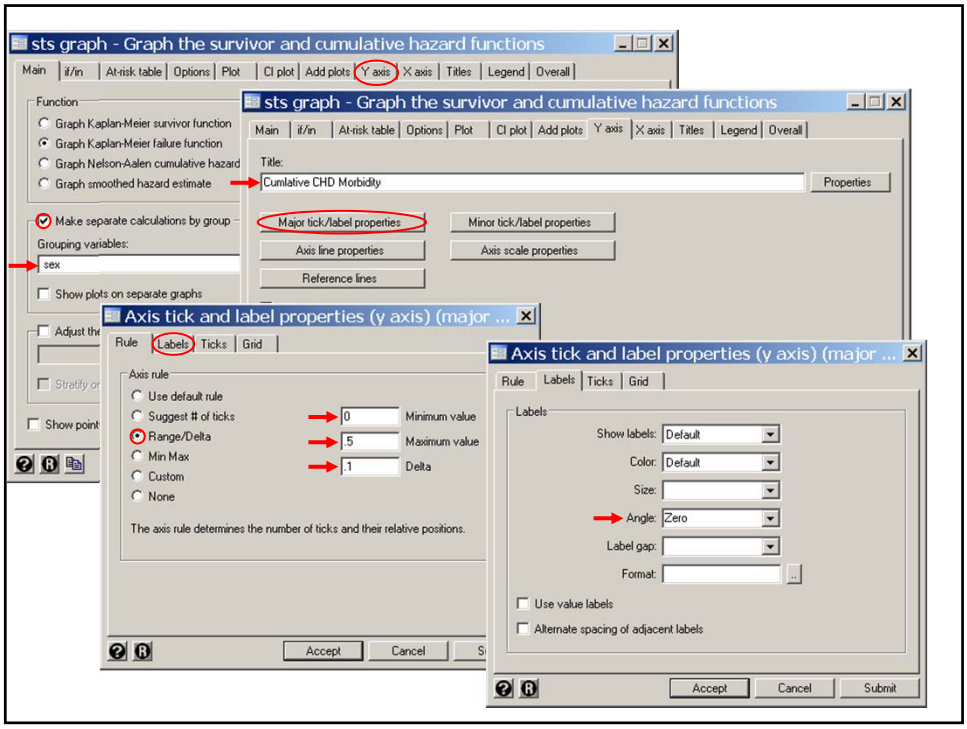
{9} The ***plot1opts*** and ***plot2opts*** options control the appearance of the first and second plot, respectively. The ***color*** and ***lwidth*** suboptions control the color and width of the plotted lines. In this example blue and pink curves of medium thickness are plotted for men and women,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

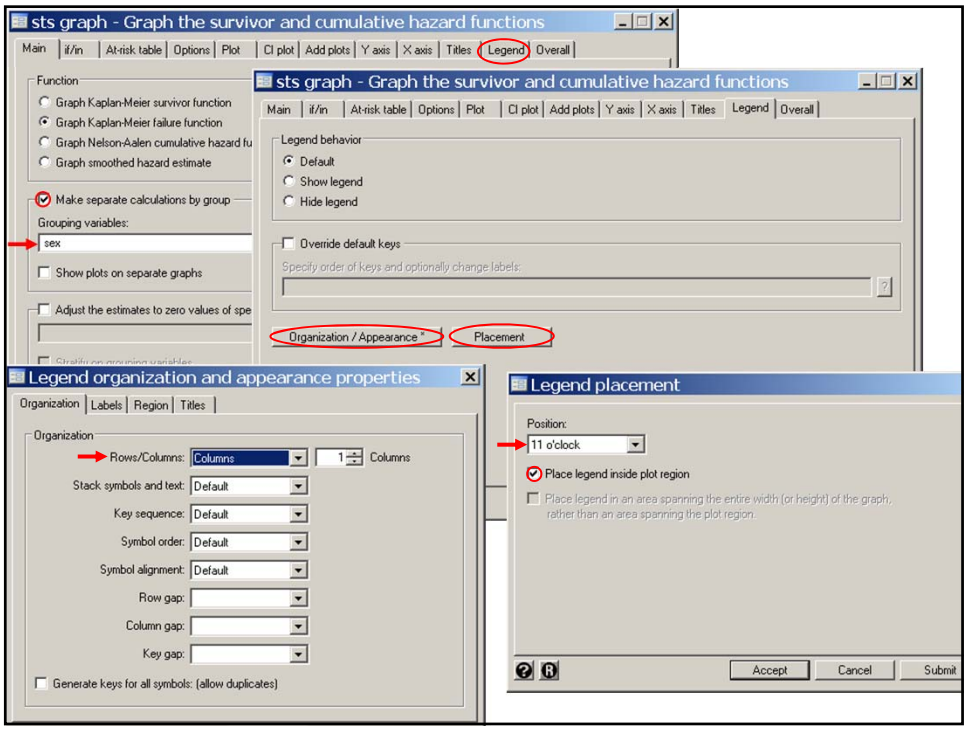
{10} The failure option converts a standard survival curve into a cumulative morbidity curve.

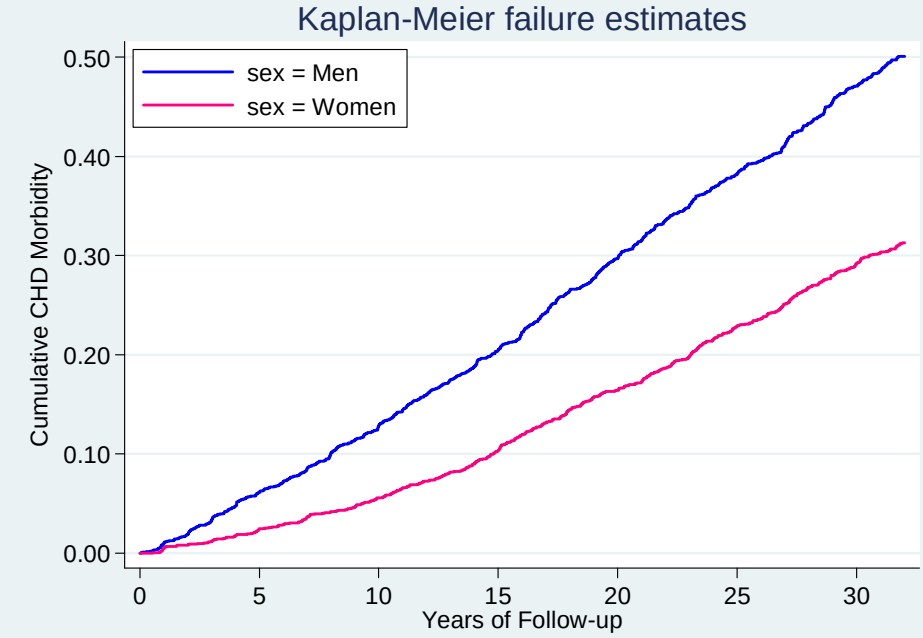
Cumulative morbidity plots are particularly **effective** when a large proportion of subjects **never** suffer the **event** of interest. Note that in this plot of CHD morbidity by sex that the *y*-axis only extends to 0.5


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:

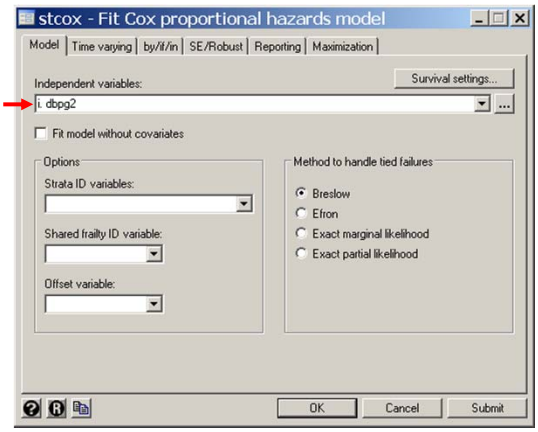
90.dbpg2 = 1 if *dbpg2* = 90, and = 0 otherwise;
110.dbpg2 = 1 if dbpg2 = 110, and = 0 otherwise;
111.dbpg2 = 1 if *dbpg2* = 111, and = 0 otherwise.

Our model is

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

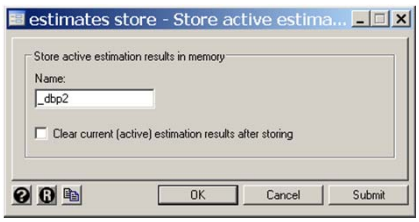
This allows us to obtain the following **relative risk** estimates for CHD compared to people with DBP≤60.

$e^{\hat{\beta}_1} = 2.58$ = 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}
```

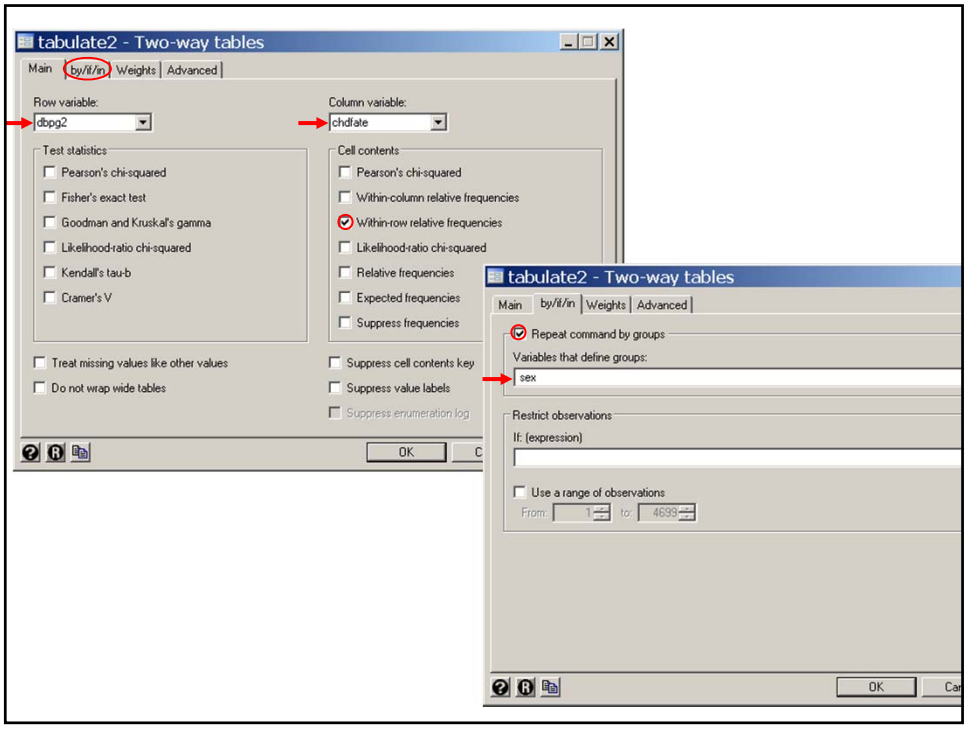
{14} The maximum value of the log likelihood function (as well as other statistics) from this model is stored under the name ***_dbpg2***



```
. 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
------------------------------------------------------------------------------
```

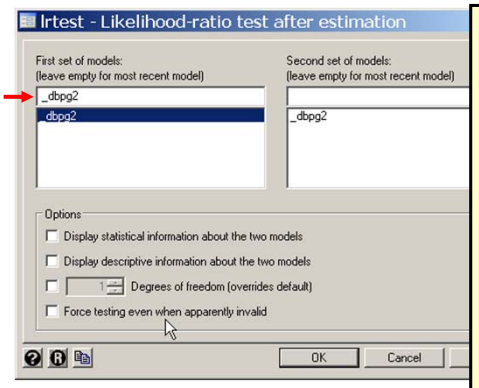
{17} We next fit a **multiplicative** model of **gender** and our four **DBP** groups. That is we fit a model without gender-DBP interaction terms.

{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
```



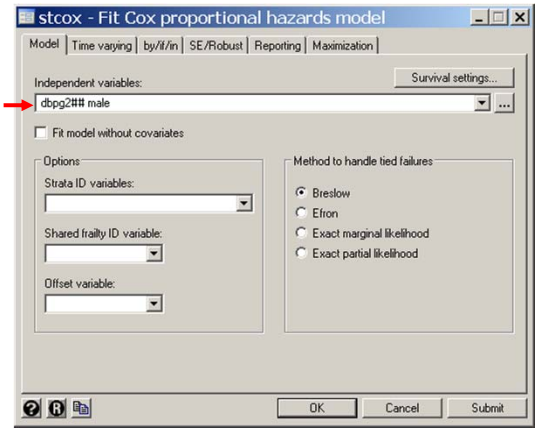
{20} The ***lrtest*** command performs the same change in model deviance calculation that we just did by hand. ***_dbpg2*** is the name that we assigned to the parameter estimates in the model with just the ***i.dbpg2*** covariates. The period refers to the most recent regression command. This command 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.

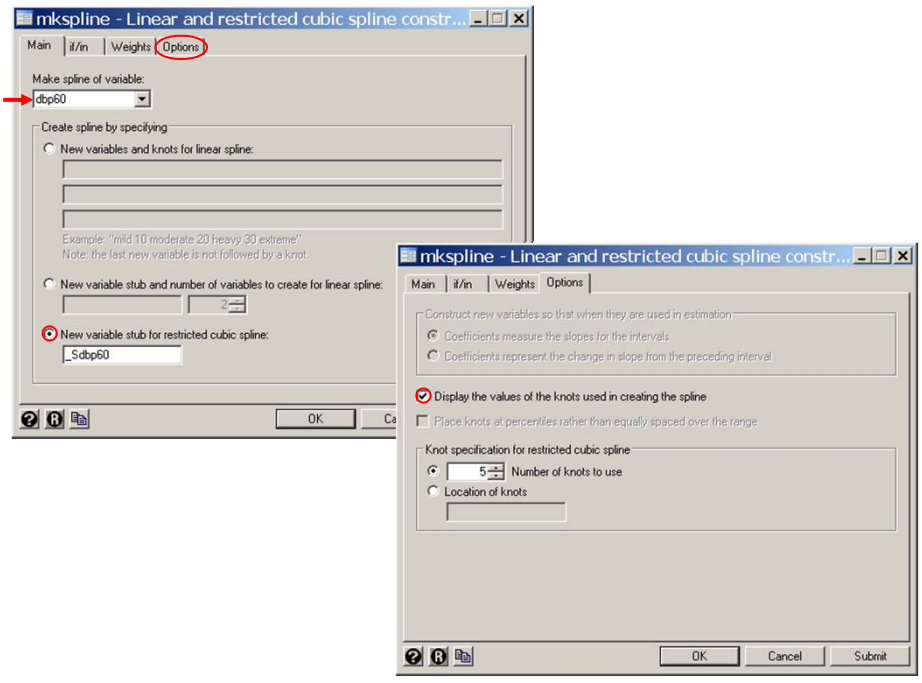
4. Fitting a Cubic Spline Model for the effect of DBP on CHD

```
. *  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.

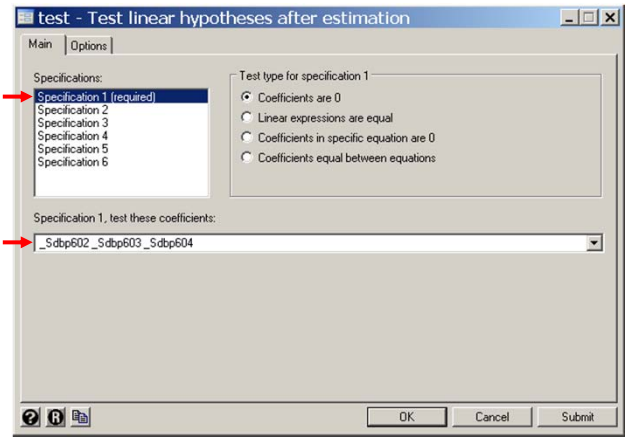
This command generates spline covariates named *_Sdbp601*, *_Sdbp602*, *_Sdbp603*, and *_Sdbp604*.



```
. * 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
```

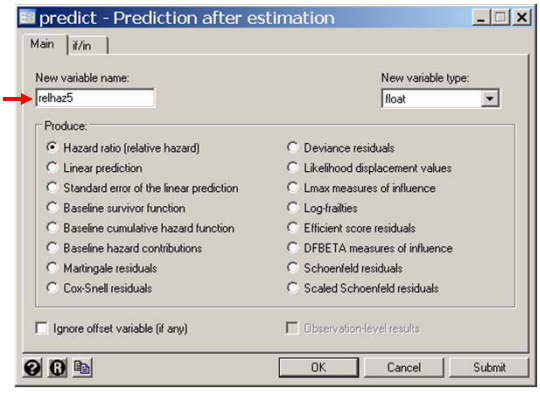
{3} Do a proportional hazards regression of CHD morbidity against the spline covariates. The ***nohr*** option causes the parameter estimates to be displayed.

{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.

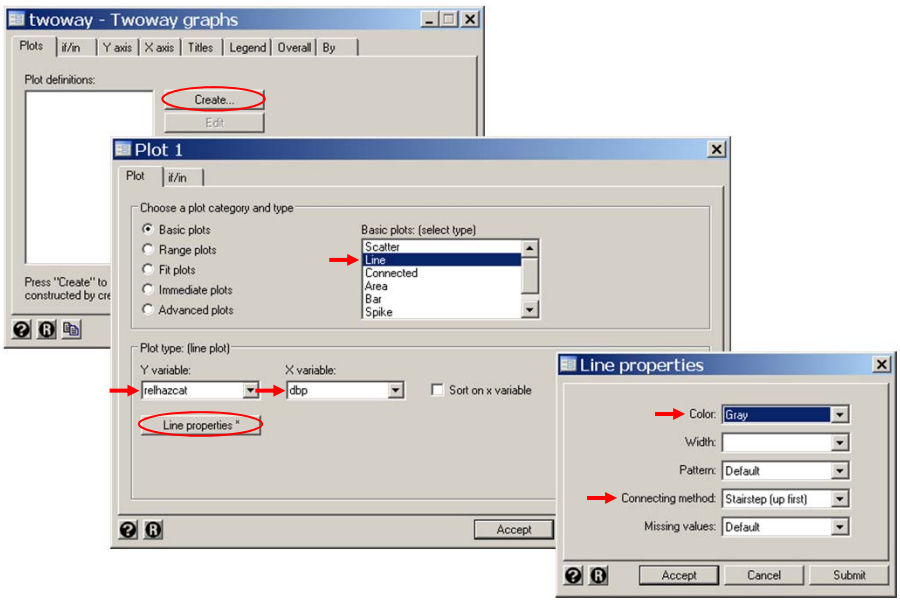
{12} The second spline covariate is not significantly different from zero.

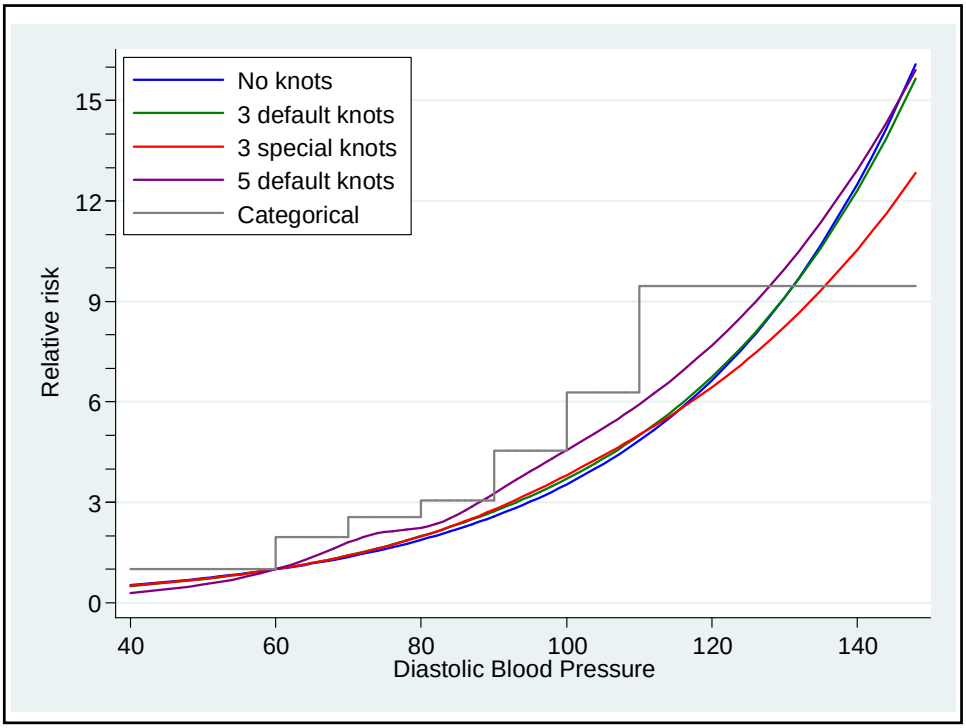
```
. *
. *  Calculate the relative hazard from model 7.12 in the text
. *
. generate relhazcat = 1

. 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



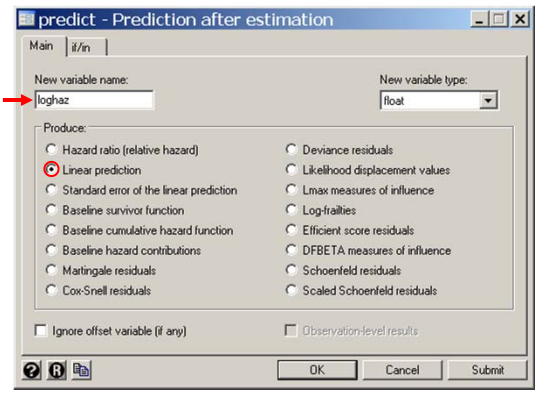


Note that the categorical model has all patients with a DBP $\leq$ 60 in the denominator of the relative risk while for the other models this denominator is patients with DBP = 60. This explains why the categorical relative risks are higher than the risks for the other models.

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.

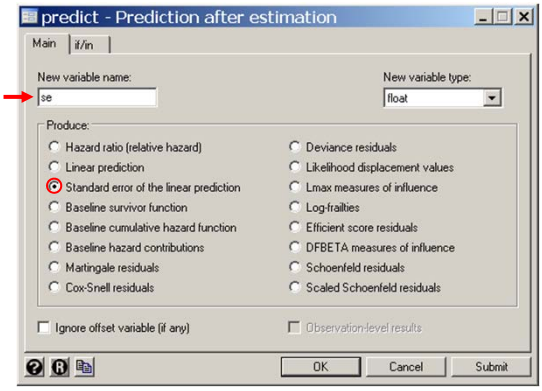
```
. predict loghaz, xb                                                {14}
```

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




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

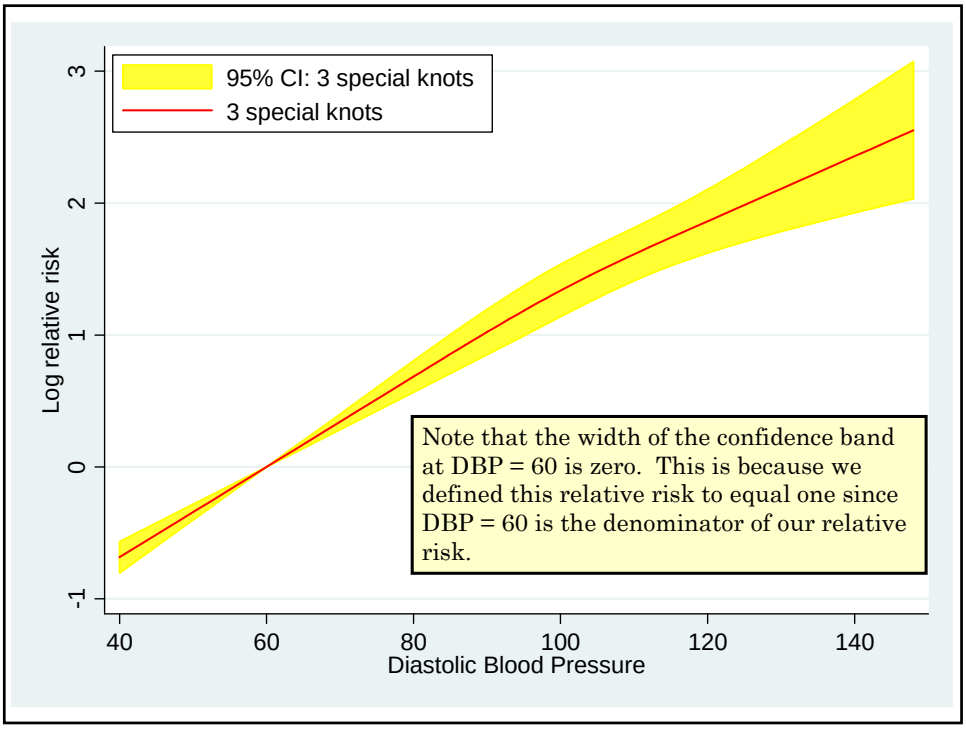
{15} *se* is the standard error of *loghaz*.



```
. 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)
```

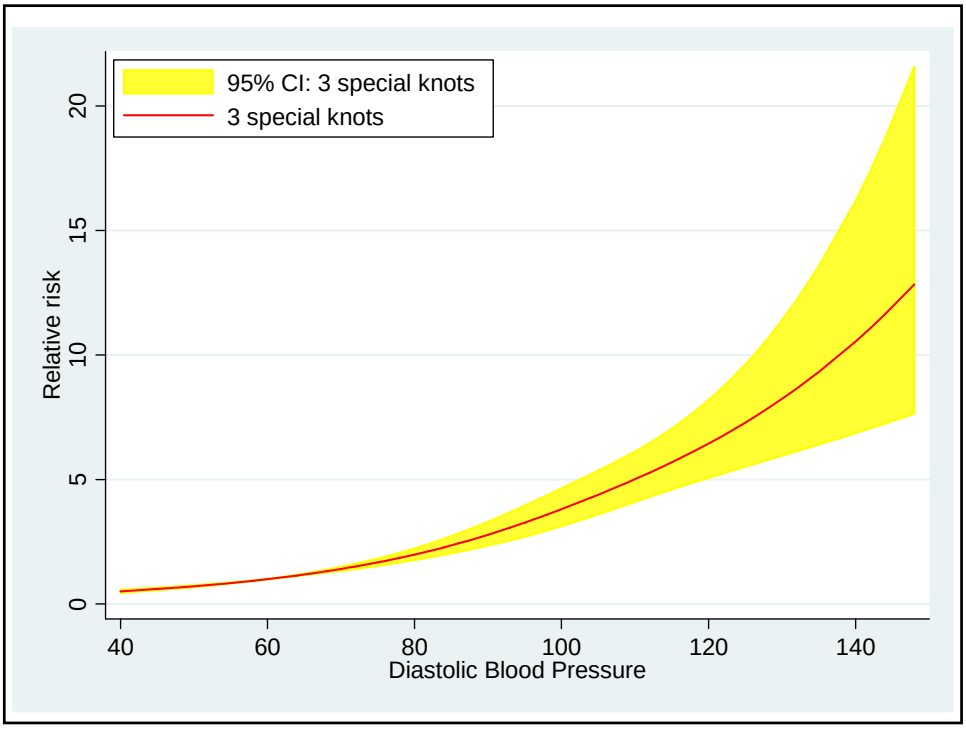
{16} *logcil* and *logciu* are the 95% confidence bands for *loghaz*.

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



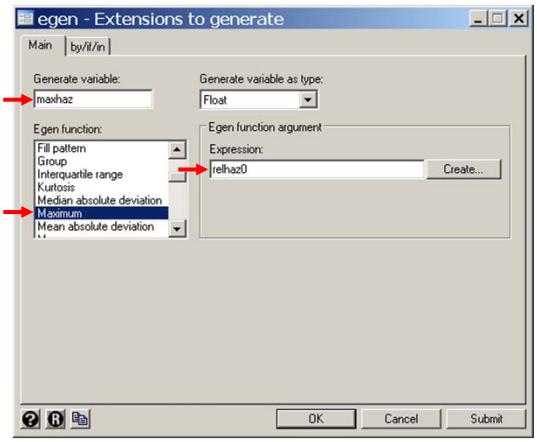
```
. generate cil3a = exp(logcil)

. 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}
```

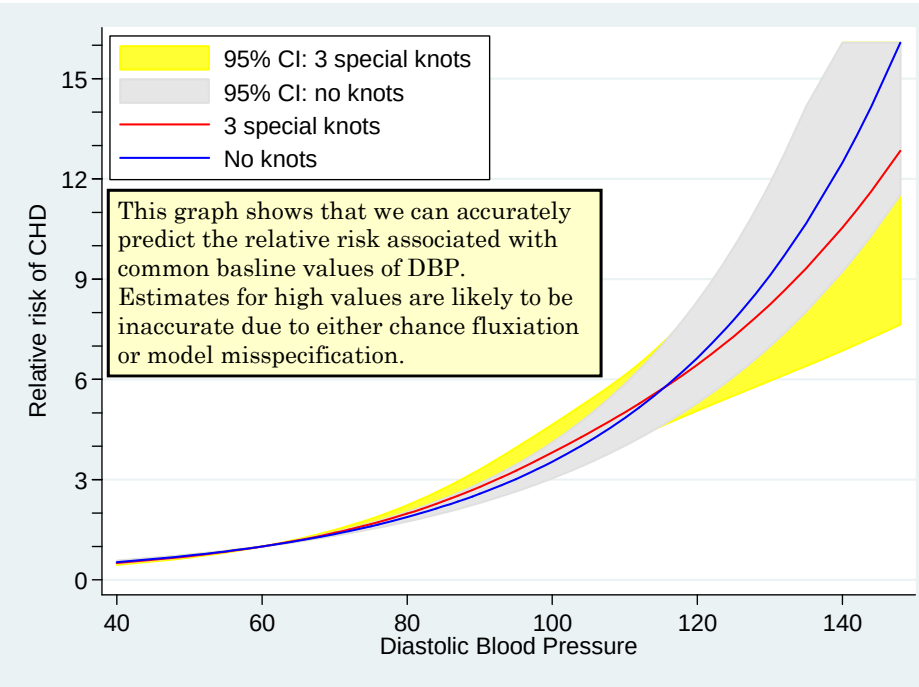
{19} This command defined *maxhaz* to equal the maximum value of *relhaz0* in the entire data set.



```
. 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)
```

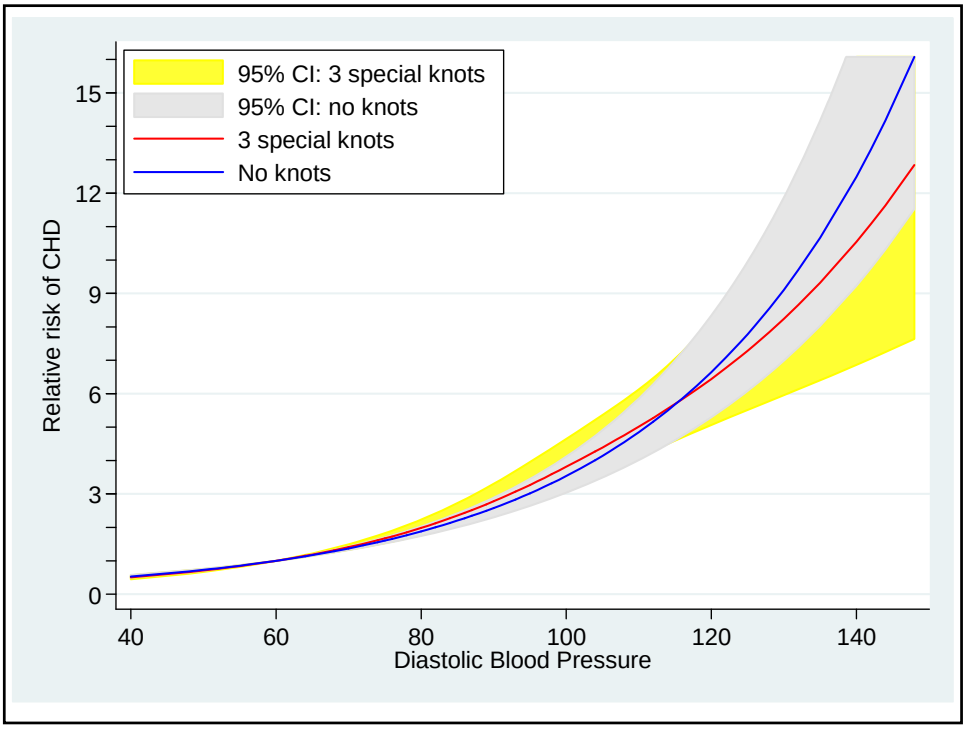
{21} Increase the number of records in the data set to 4739 by adding 40 dummy records. All of the variables in these records will be missing.

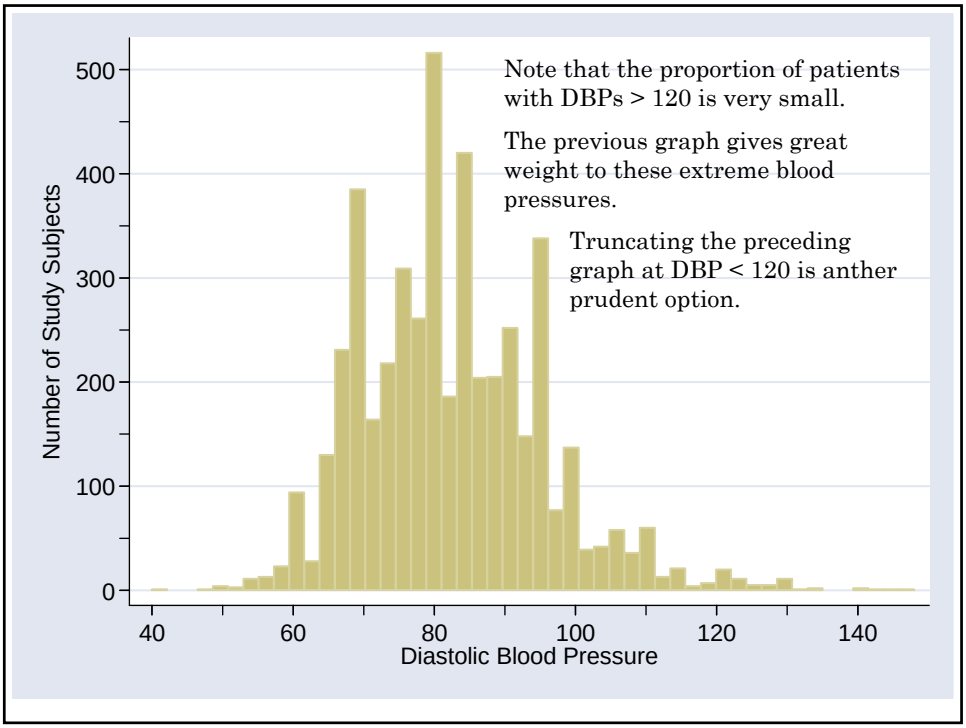
{22} There are no real blood pressures observed between 135 and 140. In these new records define *dbp* to range from 135.1 to 139 in increments of 0.1

{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.

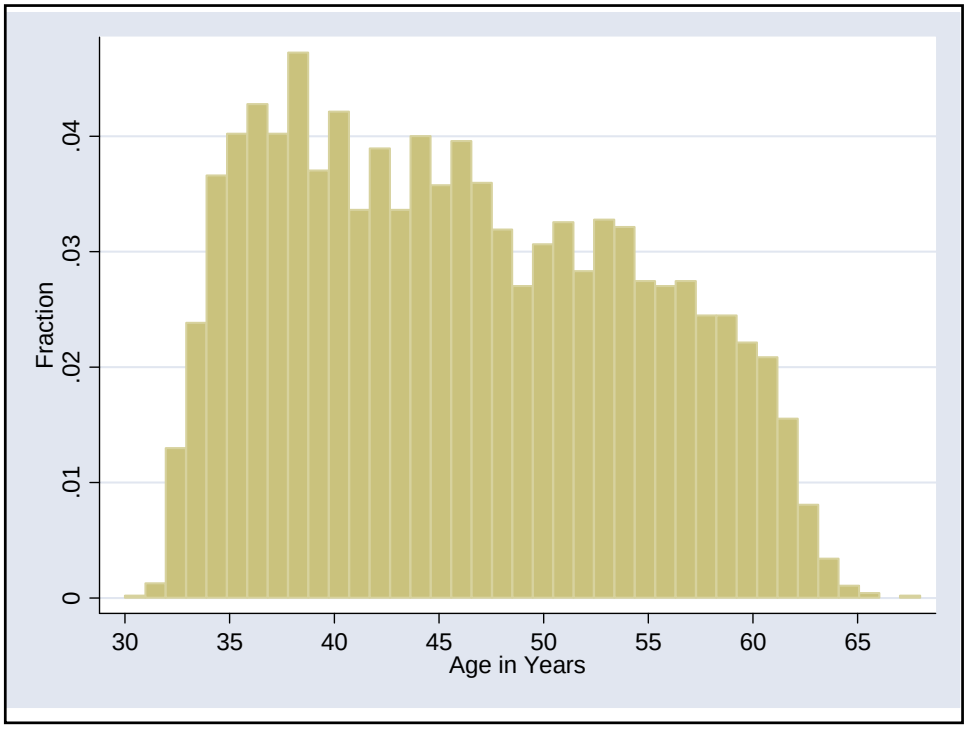
These analyses must be interpreted as the effect of age and other covariates on the risk of failure conditioned on the fact that each patient had not failed prior to her age of recruitment.

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
```

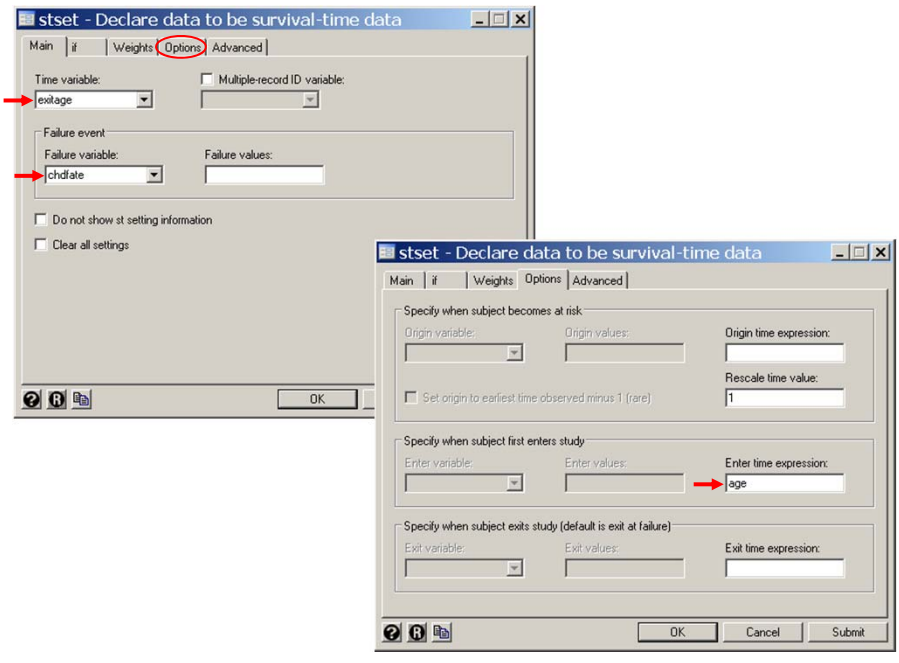
{2} We define ***exitage*** to be the patient's age at exit.

{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.

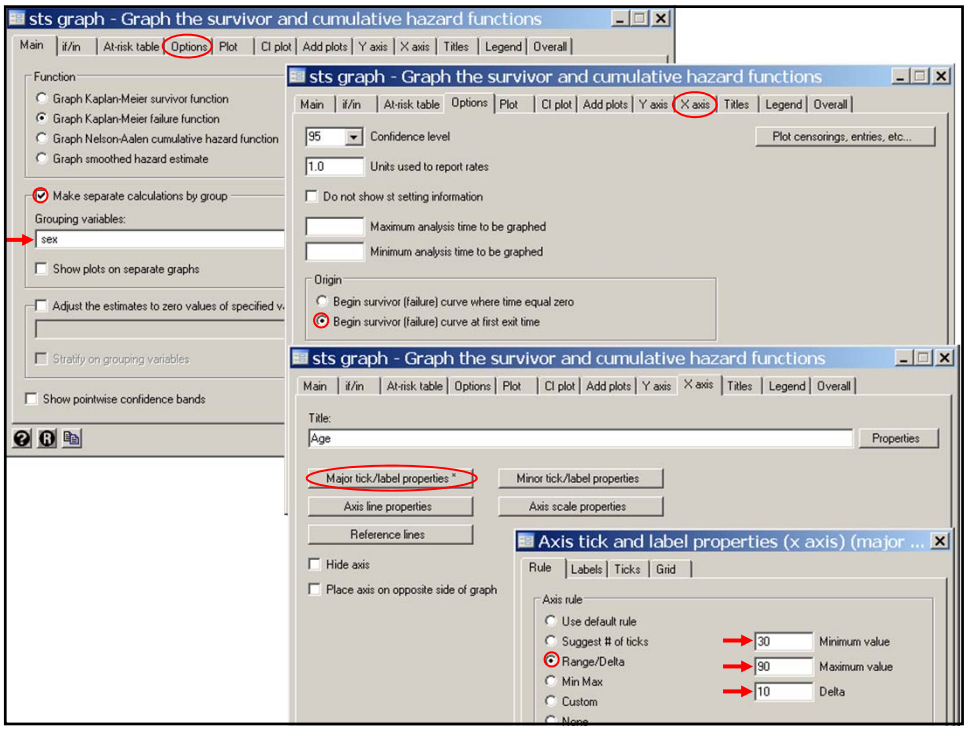
enter(time age) defines age to be the patient's entry time. That is, patients enter the analysis when they reach the age of their baseline exam. We know that all patients were free from CHD at that time.

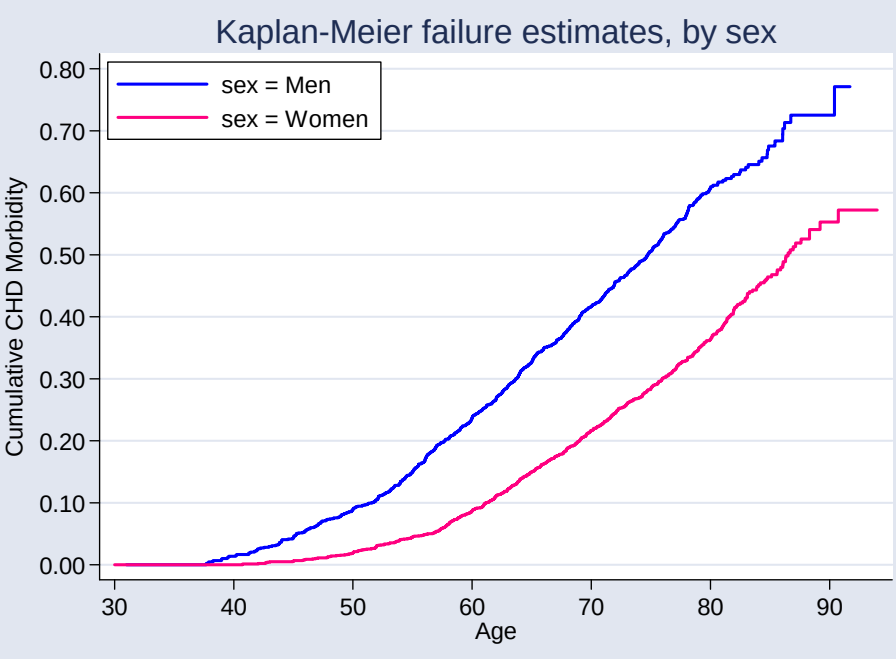


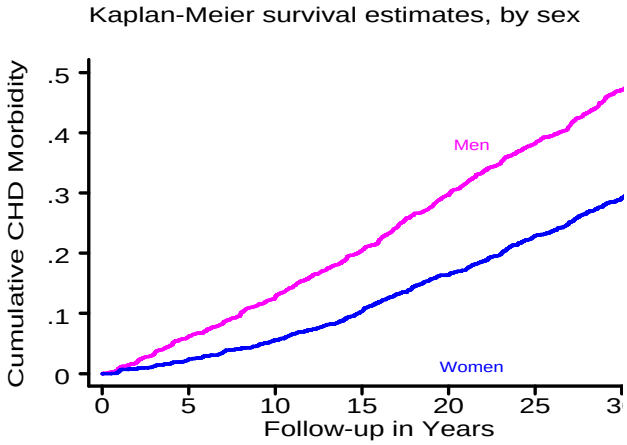
```
. * Graphics > Survival analysis graphs > Kaplan-Meier failure function
. sts graph, by(sex) failure ytitle(Cumulative CHD Morbidity) xtitle(Age) /// {4}
>     ylabel(0(.1).8, angle(0)) legend(ring(0) position(11) col(1))       ///
>     plot1opts(color(blue) lwidth(medthick))                   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

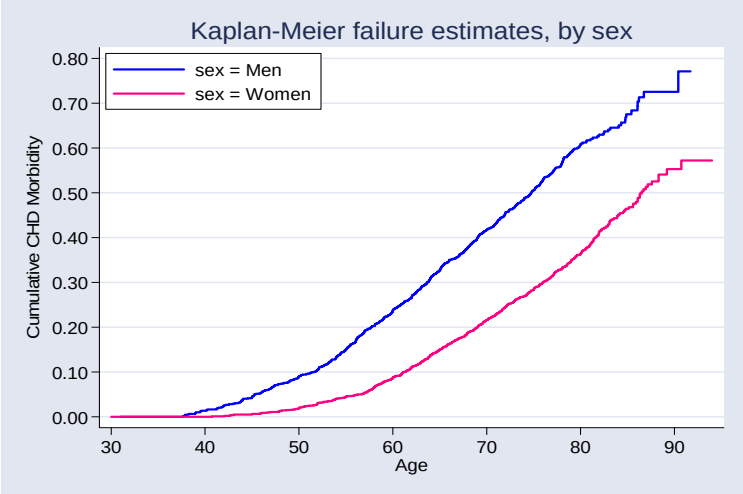






Compare the preceding graph with the analogous graph that we plotted as a function of time since recruitment. In the former graph, the morbidity curves continually widen, which indicates that men remain at greater risk than women regardless of the number of years since recruitment.

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.



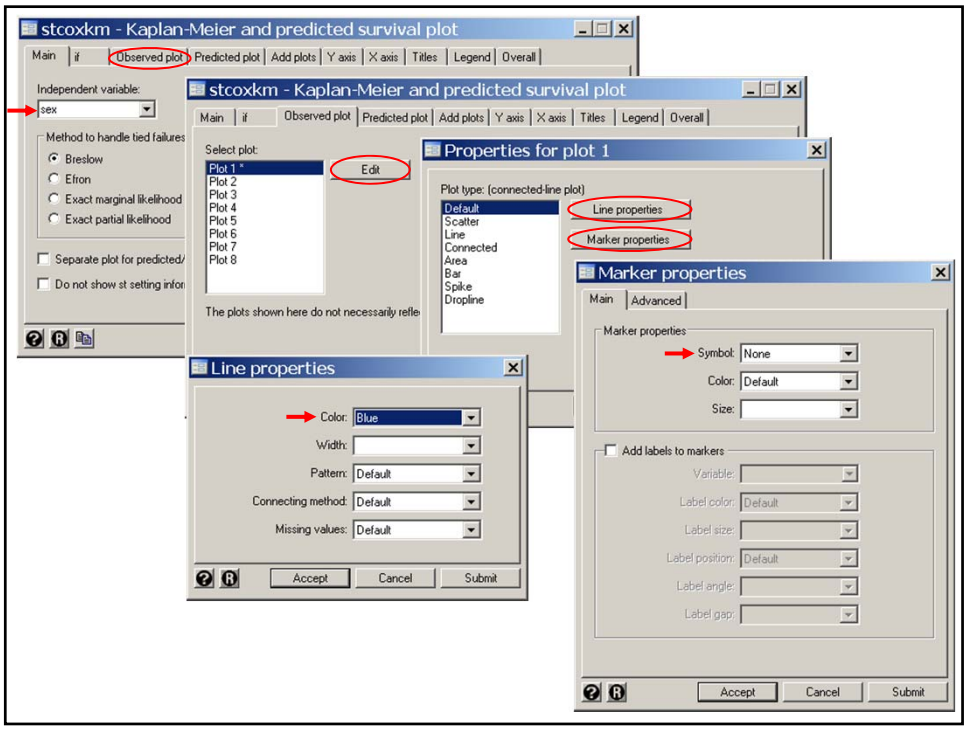
In the latter graph the curves for men and women separate rapidly as women approach the age of menopause. After age 70, however, the curves become parallel, which indicates a similar age-specific incidence for men and women. Hence this analysis is consistent with the hypothesis that the protective effect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

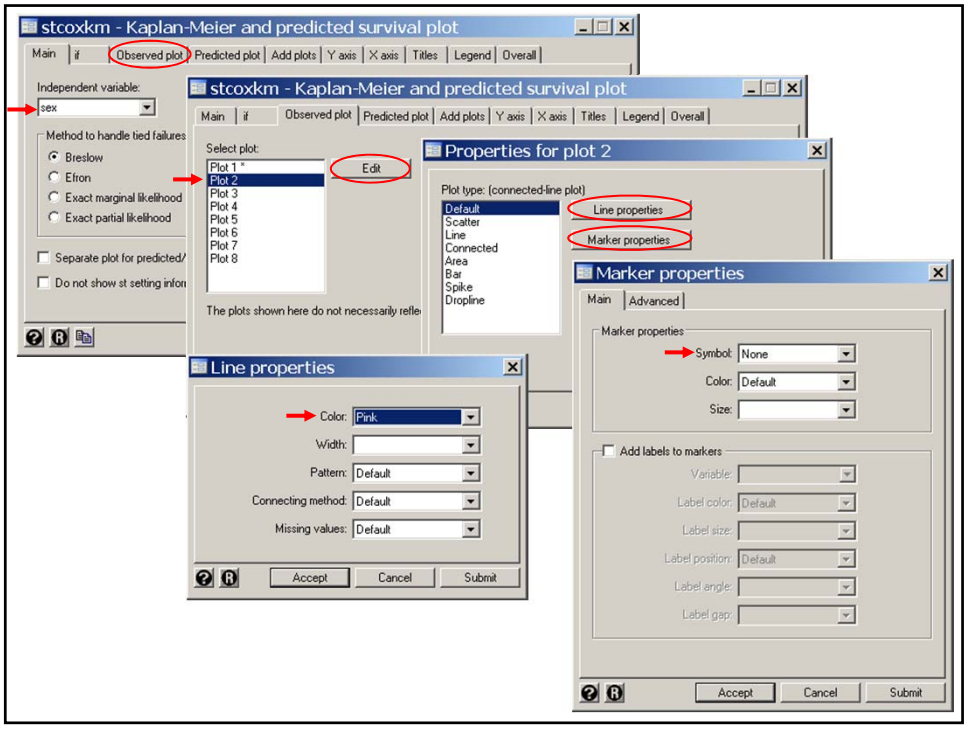
```
. *
. *  Compare Kaplan-Meier curve with best fitting survival curves under the
. *  proportional hazards model.
. *
. * Graphics > Survival analysis graphs > Compare Kaplan-Meier and Cox survival...
. stcoxkm, by(sex) obs1opts(symbol(none) color(blue))     ///        {5}
>     pred1opts(symbol(none) 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
```

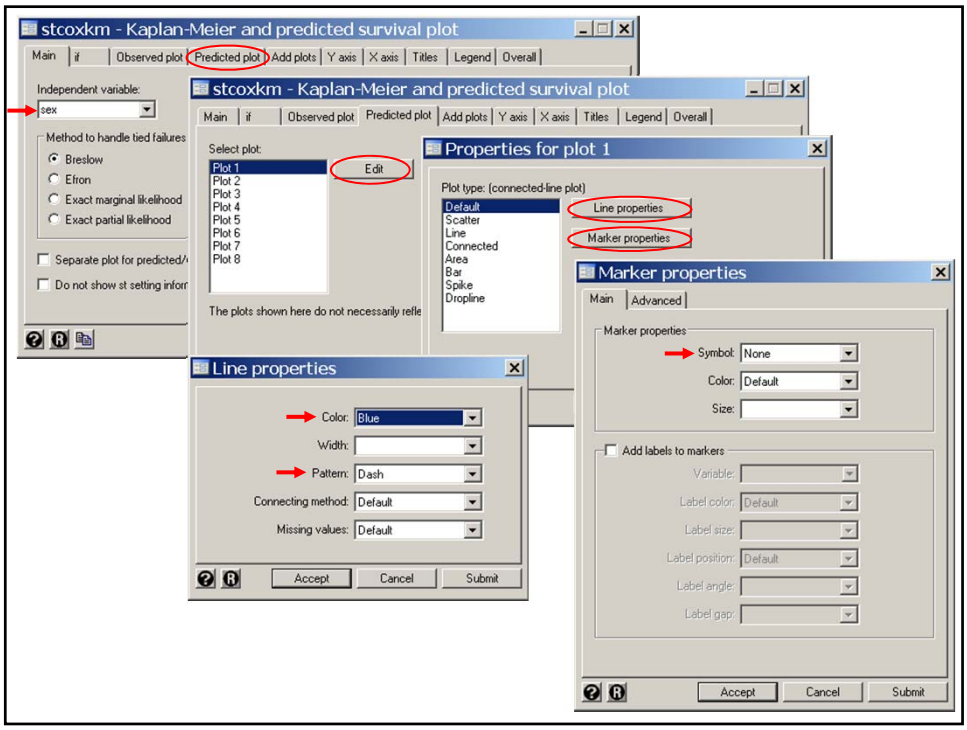
{5} This command plots the Kaplan-Meier survival curves for each sex together with the best fitting survival curves for each gender under the proportional hazards model.

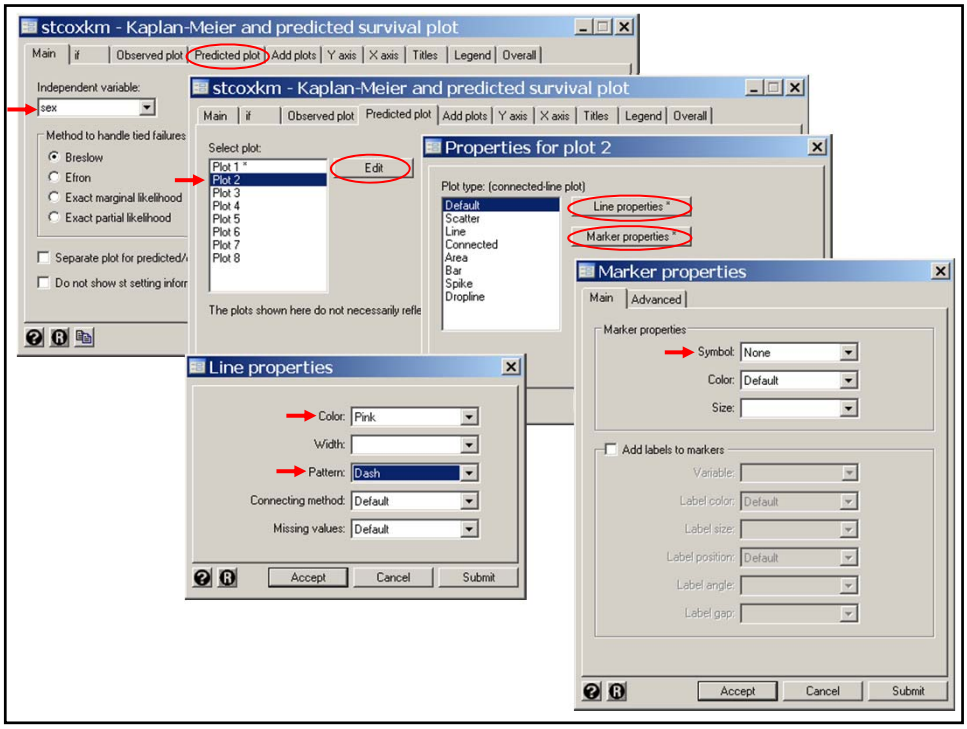
{6} The ***obs1opts*** and ***pred1opts*** options specify the characteristics of the observed and predicted male survival curves, respectively. The suboptions of these options are similar to those of the ***plot1opts*** option ***sts graph*** command. By default, ***stcoxkm*** plots a symbol at each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

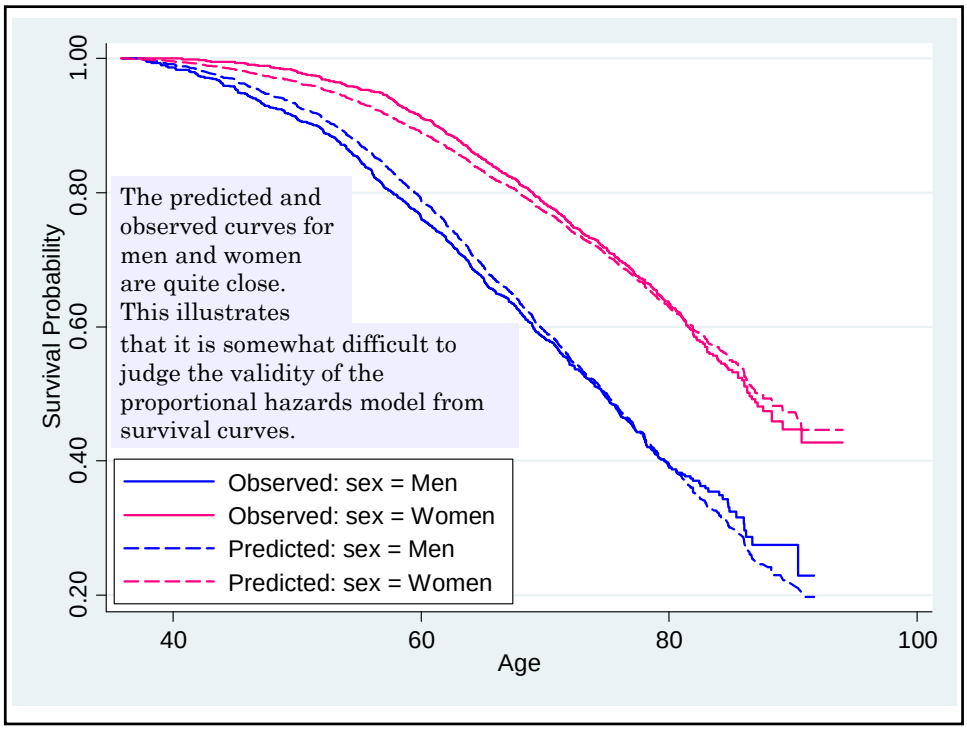
{7} The characteristics of the observed and predicted survival curves for women are similarly defined by the ***obs2opts*** and ***pred2opts*** respectively; ***obs1opts*** and ***obs2opts*** refer to men and women, respectively because the coded value of ***sex*** = 1 for men is less than that for women (***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 $\beta_1\left(x_{i1} - x_{j1}\right) + \beta_2\left(x_{i2} - x_{j2}\right) + \cdots + \beta_1\left(x_{iq} - x_{jq}\right)$.

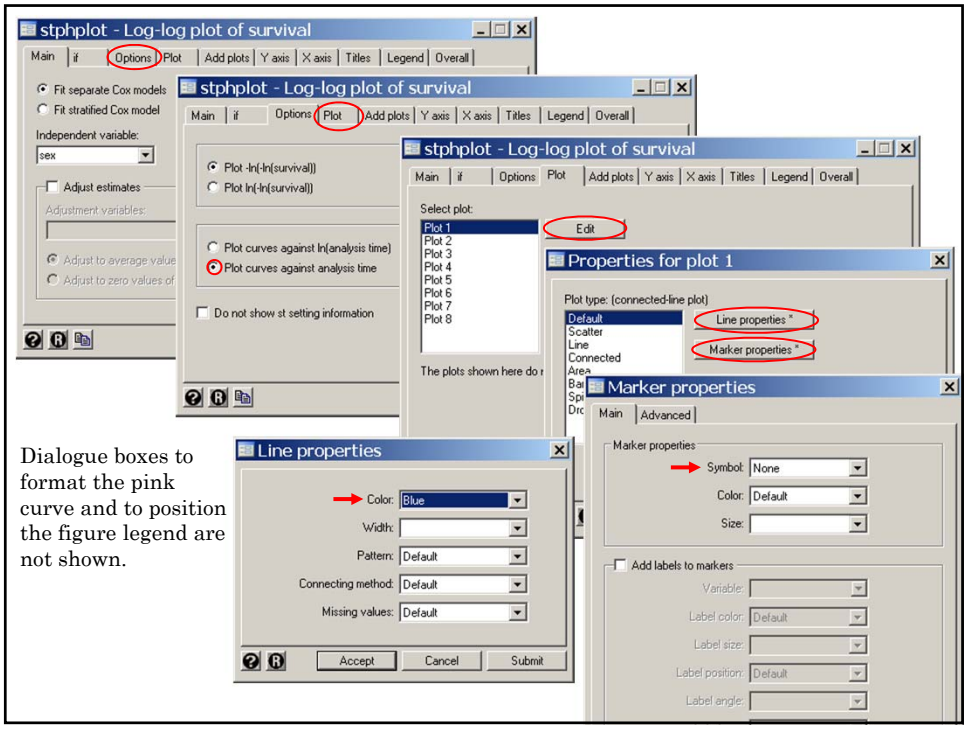
We draw such plots to visually evaluate the proportinal hazards assumption. *Framingham.age.log* continues as follows:

```
. *
. *  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
```

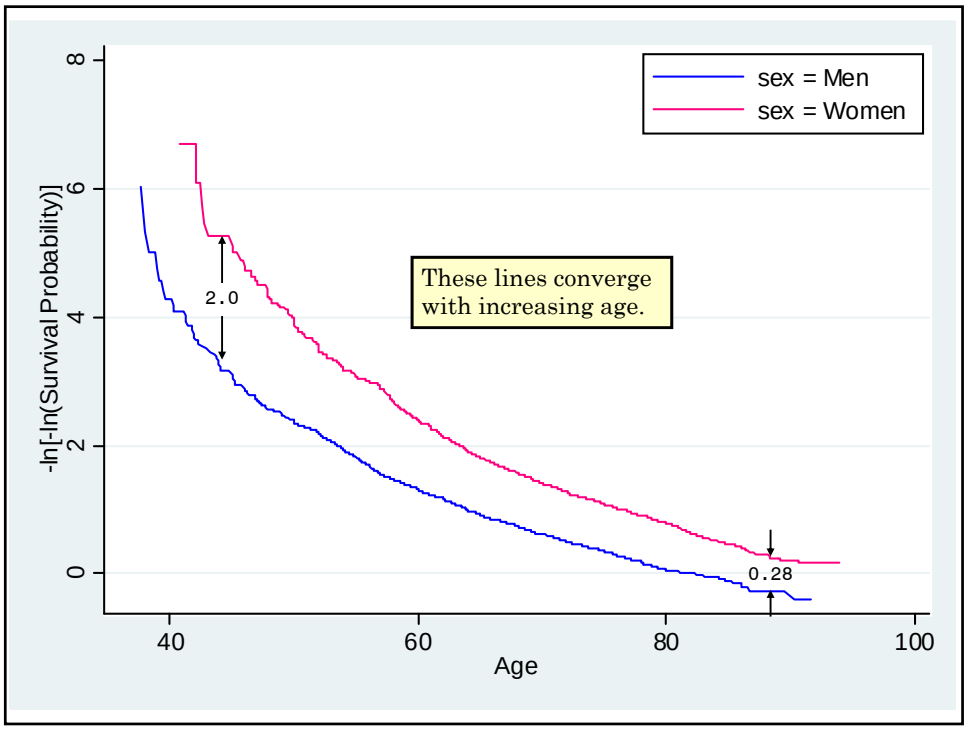
{8} The *stphplot* command draws log-log plots for each unique value of the covariate specified with the *by* option (in this example *sex*). It fits a proportional hazards model regressing *chdfate* against *sex* as defined by the previous *stset* command.

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
------------------------------------------------------------------------------
```

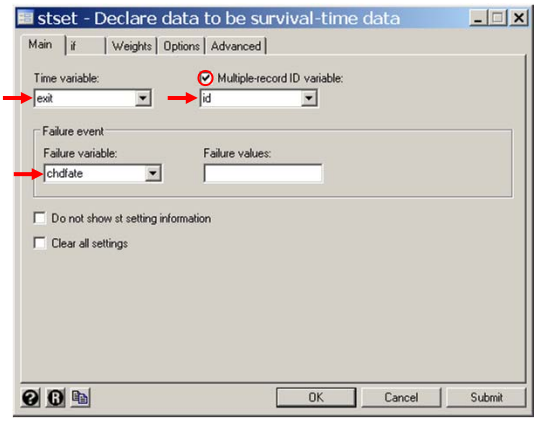
{1} First, we run the proportional hazards analysis of the effect of gender on CHD. This analysis estimates that men have **2.01** times the CHD risk of women, with overwhelming statistical significance.

```
. *
. * 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
```

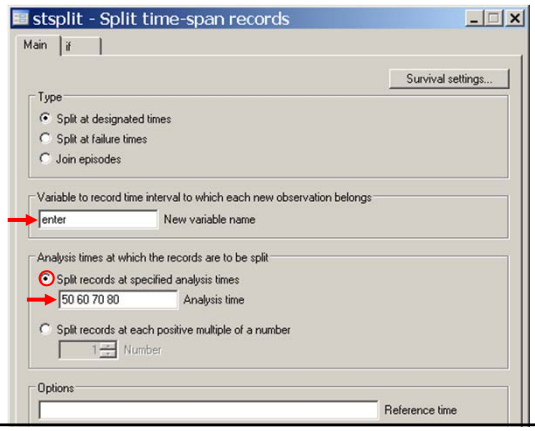
{2} The next few commands will create the multiple records that we need. It is **prudent** to be cautious doing this and to create **before** and **after tables** to confirm that we have done what we intended to do.

```
. *
. *  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.

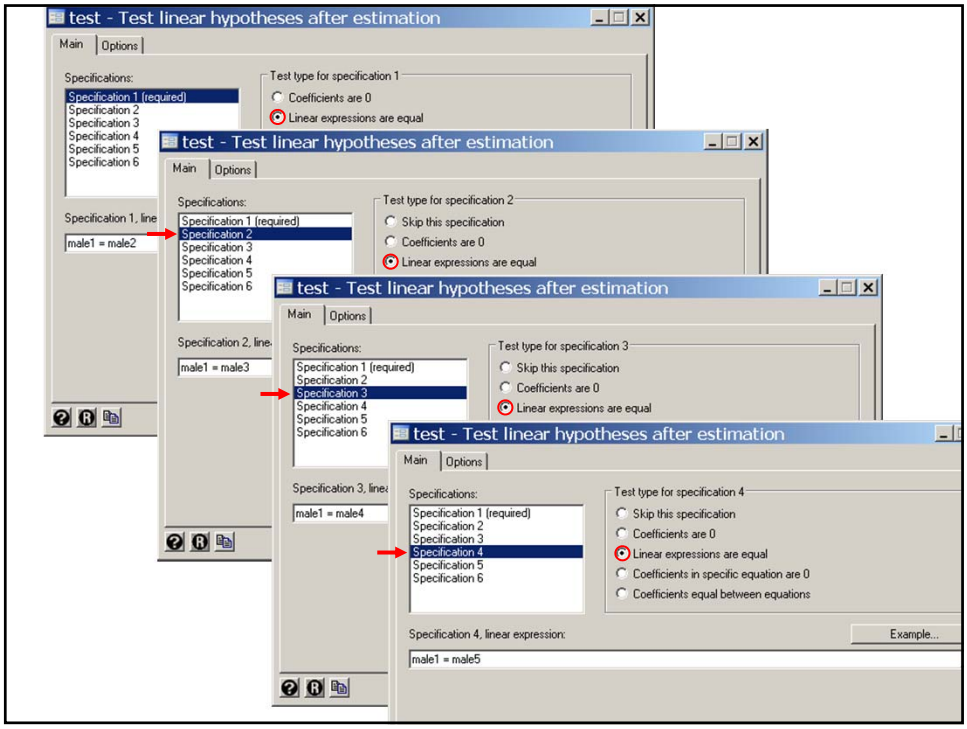
Hence, we can test the proportional hazards assumption by testing whether $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$

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).

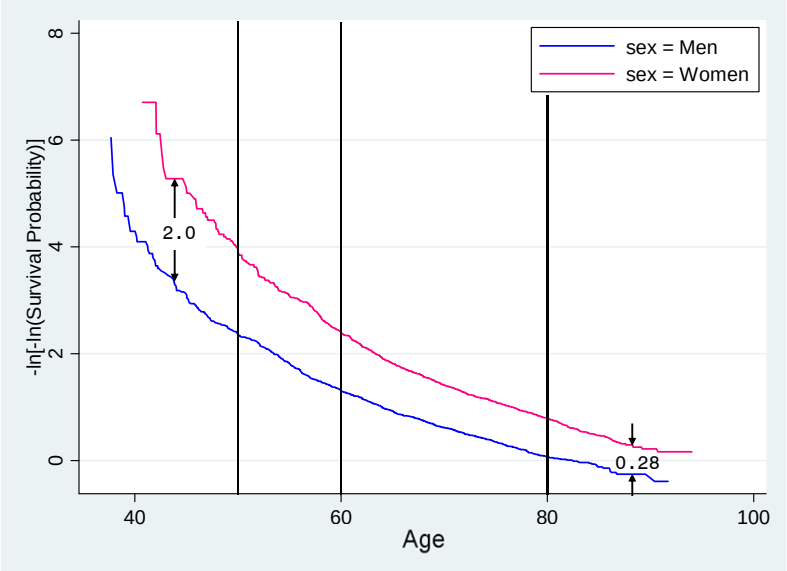
```
. lincom male3 - male4                                                  {15}

 ( 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.

VII. INTRODUCTION TO POISSON REGRESSION
Inferences on Morbidity and Mortality Rates

- Elementary statistics involving rates
  - Incidence and relative risk
- Classical methods for deriving 95% confidence intervals for relative risks
- Relationship between the binomial and Poisson distributions
- Poisson regression and 2x2 contingency tables
- Estimating relative risks from Poisson regression models
  - Offsets in Poisson regression models
- Poisson regression is an example of a generalized linear model
  - Assumptions of the Poisson regression model
  - Contrast between logistic and Poisson regression
  - 95% confidence intervals for relative risk estimates
- Poisson Regression and survival analysis
  - Converting survival records to person-year records with Stata

1. Elementary Statistics Involving Rates

The Framingham Heart Study data set contains information on 4,699 subjects with 103,710 patient-years of follow-up. We can extract the following table from this data.

	Men	Women	Total
Cases of Coronary Heart Disease	d_1 = 823	d_0 = 650	1,473
Person-years of Follow-up	n_1 = 42,259	n_0 = 61,451	103,710

a) Incidence

The incidence of CHD in men is

d_1 / n_1 = 823/42,259

= 0.01948.

The incidence of CHD in women is

d_0 / n_0 = 650/61,451

= 0.01058

b) Relative Risk

The relative risk of CHD in men compared to women is estimated by

$\hat{R} = (d_1 / n_1)/(d_0 / n_0)$ = 0.01948/0.01058 = 1.841.

c) 95% confidence interval for a relative risk

If d_i is small compared to n_i (i = 0 or 1) then

The variance of ($\log \hat{R}$) is approximated by

$$s^2_{\log(\hat{R})} = \frac{1}{d_1} + \frac{1}{d_0} \qquad \{7.1\}$$

$$= \frac{1}{823} + \frac{1}{650} = 0.002754$$

Hence a 95% confidence interval for R is

$$\hat{R} \exp\left(\pm z_{0.025} s_{\log(\hat{R})}\right) \qquad \{7.2\}$$

= [1.841 exp(-1.96× $\sqrt{0.002754}$), 1.841 exp(0.1029)]

= [1.66, 2.04]

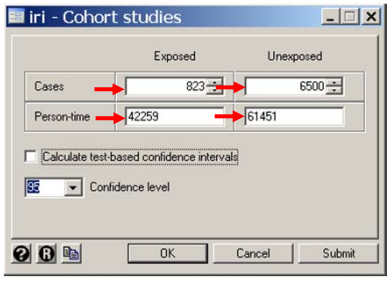
In Stata these calculations are done as follows:

```
. * 8.2.Framingham.log
. *
. *  Estimate the crude (unadjusted) relative risk of
. *  coronary heart disease in men compared to women using
. *  person-year data from the Framingham Heart Study (Levy 1999).
. *
. * Statistics > Epidemiology... > Tables... > Incidence-rate ratio calculator
. iri 823 650 42259 61451                                                {1}

                |   Exposed   Unexposed  |      Total
----------------+------------------------+------------
          Cases |       823         650  |       1473
    Person-time |     42259       61451  |     103710
----------------+------------------------+------------
                |                        |
  Incidence rate |  .0194751    .0105775  |   .0142031
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
 Inc. rate diff. |         .0088976       |    .0073383     .010457
 Inc. rate ratio |          1.84118       |    1.659204     2.043774 (exact)
 Attr. frac. ex. |            .45687      |    .3973015     .510709 (exact)
 Attr. frac. pop |         .2552641       |
                +-------------------------------------------------
                    (midp)   Pr(k>=823) =                  0.0000 (exact)
                    (midp) 2*Pr(k>=823) =                  0.0000 (exact)
```

{1} The ***ir*** command is used for incidence rate data.

Shown here is the immediate version of this command, called ***iri***, which analyses the four data values given in the command line.

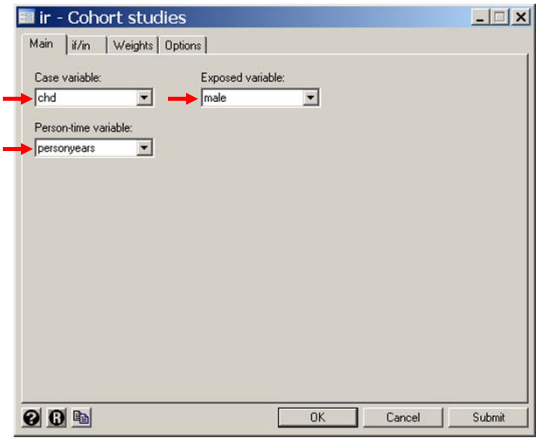
These data are the number exposed and unexposed cases together with the person-years of follow of exposed and unexposed subjects.



```
. *
. *  The equivalent ir command is illustrated below.
. *
. use 8.2.Framingham.dta, clear
. * Data > Describe data > List data
. list
     +------------------------+
     |  male   chd   per_yrs  |
     |------------------------|
  1. | Women   650      61451 |
  2. |   Men   823      42259 |
     +------------------------+
```

```
. * Statistics > Epidemiology... > Tables ... > Incidence-rate ratio
. ir chd male per_yrs                                                  {2}
                 | Male                    |
                 |   Exposed   Unexposed   |      Total
-----------------+-------------------------+------------
    CHD patients |       823         650   |       1473
 P-yrs follow-up |     42259       61451   |     103710
-----------------+-------------------------+------------
                 |                         |
  Incidence rate |  .0194751    .0105775   |   .0142031
                 |                         |
                 |      Point estimate     |    [95% Conf. Interval]
                 |-------------------------+------------------------
 Inc. rate diff. |         .0088976        |    .0073383     .010457
 Inc. rate ratio |          1.84118        |    1.659204    2.043774 (exact)
 Attr. frac. ex. |            .45687       |    .3973015     .510709 (exact)
 Attr. frac. pop |         .2552641        |
                 +-------------------------------------------------
                     (midp)   Pr(k>=823) =                  0.0000 (exact)
                     (midp) 2*Pr(k>=823) =                  0.0000 (exact)
```

{2} Here is the conventional version of this command. Person-years of follow-up may be distributed over multiple records. If there is one record per subject then
per_yrs gives each subject's years of follow-up;
chd = 1 if the subject had CHD, 0 otherwise; and
male = 1 for men, 0 for women.



We next introduce **Poisson regression** which is used for analyzing rates.

Poisson regression is used when the original data available to us is expressed as events per person-years of observation.

Poisson regression is also useful for analyzing data from large cohorts when the proportional hazards assumption is false. In this situation Poisson regression is quicker and easier to use than hazard regression with time-dependent covariates.

2. The Binomial and Poisson Distribution

Let

- n be the number of people at risk of death
- d be the number of deaths
- λ be the probability that any patient dies.

Then d has a **binomial distribution** with parameters n and λ,

mean $n\lambda$, and

variance $n\lambda(1-\lambda)$.

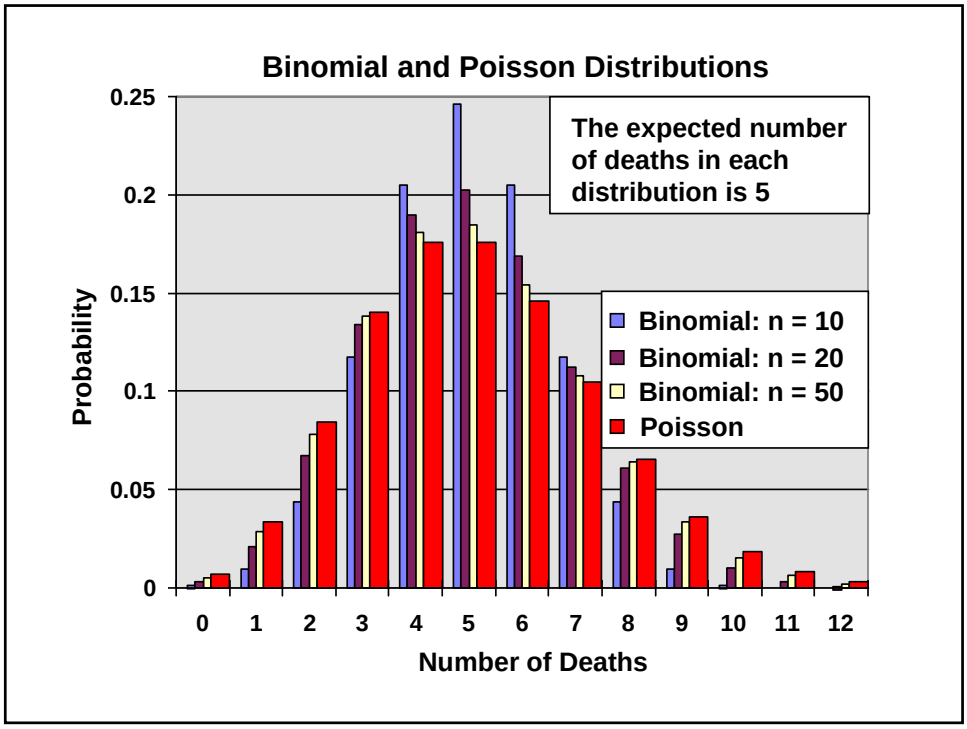
Pr[d deaths]

$$= \frac{n!}{(n-d)!\,d!}\pi^{d}(1-\pi)^{(n-d)} \qquad \{7.3\}$$

Poisson (1781–1849) showed that when n is large and π is small the distribution of d is closely approximated by the **Poisson distribution**, whose mean and variance both equal $n\pi = \lambda$.

$$\Pr[d \text{ deaths}] = \frac{e^{-\lambda}(\lambda)^{d}}{d!} \qquad \{7.4\}$$

Although it is not obvious from these formulas, the convergence of the binomial distribution to the Poisson is quite rapid.



3. Poisson Regression and the 2x2 Contingency Table

a) True and estimated death rates and relative risks

Consider a 2x2 contingency table

Died	Exposed	
	Yes	No
Yes	d_1	d_0
No	n_1-d_1	n_0-d_0
Total	n_1	n_0

Let

λ_i be the true death rate in people who are (i = 1) or are not (i = 0) exposed.

Died	Exposed	
	Yes	No
Yes	d_1	d_0
No	n_1-d_1	n_0-d_0
Total	n_1	n_0

Let

λ_i be the true death rate in people who are (i = 1) or are not (i = 0) exposed.

Then

$R = \lambda_1 / \lambda_0$ is the **relative risk** of death associated with exposure and $\lambda_1 = R\lambda_0$,

$\hat{\lambda}_i = d_i / n_i$ is the **estimated death rate** in people who are (i=1) or are not (i=0) exposed, and

$\hat{R} = \hat{\lambda}_1 / \hat{\lambda}_0$ is the **estimated relative risk** of death associated with exposure.

The expected number of deaths in group i is $E(d_i) = n_i\lambda_i$.

For any constant k and statistic d, $E(kd) = kE(d)$

Now

$$\lambda_0 = E[\hat{\lambda}_0] = E[d_0 / n_0] = E[d_0] / n_0$$

$$\log[\lambda_0] = \log[E[d_0]] - \log[n_0] \quad \text{, and}$$

$$\log[\lambda_1] = \log[E[d_1]] - \log[n_1]$$

But

$$\log[\lambda_1] = \log[R] + \log[\lambda_0]$$

Hence

$$\log[E[d_0]] = \log[n_0] + \log[\lambda_0]$$

$$\log[E[d_1]] = \log[n_1] + \log[\lambda_0] + \log[R]$$

Let $\alpha = \log[\lambda_0]$,

$\beta = \log[R]$,

$x_0 = 0$, and $x_1 = 1$.

Then

$$\log[E[d_i]] = \log[n_i] + \alpha + x_i\beta \text{ for } i = 0 \text{ or } 1, \qquad \{7.5\}$$

where d_i has a Poisson distribution whose mean and variance are estimated by d_i.

This is the simplest of all possible **Poisson regression models.**

b) Estimating relative risks from the model coefficients

Our primary interest is in β. Given an estimate of β

then $\hat{R} = e^{\hat{\beta}}$

c) Nuisance parameters

α is called a **nuisance parameter**. This is one that is required by the model but is not used to calculate interesting statistics

d) Offsets

log (n_i) is a known quantity that must be included in the model. It is called an **offset**.

4. Poisson Regression and Generalized Linear Models

Poisson regression is another example of a generalized linear model. The random component, linear predictor and link function for Poisson regression are as follows.

a) The random component

d_i is the **random component** of the model. In Poisson regression, d_i has a Poisson distribution with mean $E(d_i)$.

b) The linear predictor

$\log(n_i) + \alpha + x_i\beta$ is called the **linear predictor**.

c) Link function

$E(d_i)$ is related to the linear predictor through a logarithmic **link function.**

5. Contrast Between Simple Poisson Logistic and Linear Regression

The models:

Linear $E(y_i) = \alpha + x_i\beta$ for $i = 1, 2, \ldots, n$.

Logistic $\text{logit}(E(d_i/m_i)) = \alpha + x_i\beta$ for $i = 0$ or 1,

Poisson $\log(E(d_i)) = \log(n_i) + \alpha + x_i\beta$ for $i = 0$ or 1,

Linear Regression –

In linear regression the **random component** is y_i, which has a normal distribution with standard deviation σ. The **linear predictor** is $\alpha + x_i\beta$ and the **link function** is the identity function $I(x) = x$.

n must be fairly large since we must estimate **σ** before we can estimate α or β.

Logistic Regression –

In logistic regression we observe d_i events in m_i trials. The **random component** is d_i, which has a **binomial** distribution. The **linear predictor** is $\alpha + x_i\beta$. The model has a logit **link function**.

Poisson Regression –

In Poisson regression we observe d_i events in n_i trials. The **random component** is d_i, which has a **Poisson** distribution. The **linear predictor** is $\log(n_i) + \alpha + x_i\beta$. The model has a logarithmic **link function**.

In **Poisson and logistic** regression examples i has only **2** values. It is possible to estimate β from these equations since we have reasonable estimates of the **mean and variance** of d_i for both of these models.

Poisson regression models generalize in the usual way. For example, suppose

$x_i = i$ for i = 1 to 3 denotes three levels of a risk factor. Then a simple Poisson regression model would be

$$\log(E(d_i)) = \log(n_i) + \alpha + z_{2i}\beta_2 + z_{3i}\beta_3 \qquad \textbf{\{7.6\}}$$

where

d_i is the number of deaths observed in n_i person-years of follow-up in group i,

$$z_{2i} = \begin{cases} 1 : i = 2 \\ 0 : \text{ otherwise} \end{cases} \quad \text{and} \quad z_{3i} = \begin{cases} 1 : i = 3 \\ 0 : \text{ otherwise} \end{cases} .$$

Subtracting $\log(n_i)$ from both sides of equation {7.6} gives

$$\log\left(\mathrm{E}(d_i)/n_i\right) = \log\left(\mathrm{E}(d_i/n_i)\right) = \log(\lambda_i) = \alpha + z_{2i}\beta_2 + z_{3i}\beta_3 \qquad \textbf{\{7.7\}}$$

where λ_i is the true death rate for patients with risk level i.

$$\log\left(\mathrm{E}(d_i)/n_i\right) = \log\left(\mathrm{E}(d_i/n_i)\right) = \log(\lambda_i) = \alpha + z_{2i}\beta_2 + z_{3i}\beta_3 \qquad \textbf{\{7.7\}}$$

When i =2 {7.7} reduces to

$$\log(\lambda_2) = \alpha + \beta_2 \qquad \textbf{\{7.8\}}$$

When i =1 {7.7} reduces to

$$\log(\lambda_1) = \alpha \qquad \textbf{\{7.9\}}$$

Subtracting {7.9} from {7.8} gives

$$\log(\lambda_2/\lambda_1) = \beta_2$$

Hence β_2 equals the log relative risk of patients in group 2 relative to group 1.

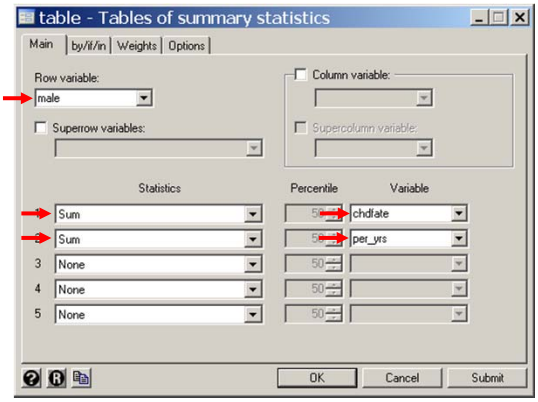
Similarly, β_3 equals the log relative risk of patients in group 3 relative to group 1.

6. Analyzing a 2x2 Contingency Table with Stata

a) Example: Gender and Coronary Heart Disease

```
. *  8.7.Framingham.log
. *
. *  Simple Poisson regression analysis of the effect of gender on
. *  Coronary heart disease in the Framingham Heart Study
. *
. use   2.20.Framingham.dta, clear
. gen male = sex==1
. gen per_yrs = followup/365.25
. * Statistics > Summaries, ... > Tables > Table of summary statistics (table)
. table male, contents(sum chdfate sum per_yrs)                    {1}

--------------------------------------
     male | sum(chdfate)  sum(per_yrs)
----------+---------------------------
        0 |          650      61451.17
        1 |          823      42258.92
--------------------------------------
```

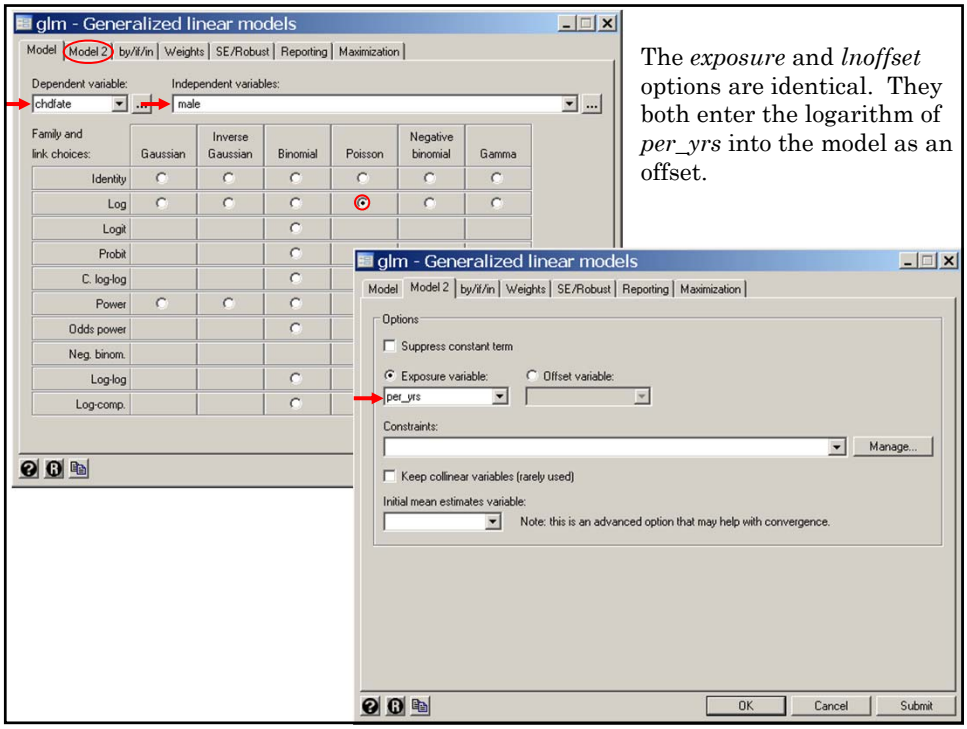
{1} Tabulate the sum of *chdfate* and *per_yrs* by gender. Recall that *2.20.Framingham.dta* contains one record per patient, with *followup* giving the number of days of follow-up for each patient.



```
. * Statistics > Generalized linear models > Generalized linear models (GLM)
. glm chdfate male , family(poisson) link(log) lnoffset(per_yrs)            {2}

Iteration 0:   log likelihood = -4240.3694
Iteration 1:   log likelihood =  -3906.885
Iteration 2:   log likelihood = -3906.5506
Iteration 3:   log likelihood = -3906.5505

Generalized linear models                         No. of obs      =      4699
Optimization     : ML                             Residual df     =      4697
                                                  Scale parameter =         1
Deviance         =  4867.101078                   (1/df) Deviance =  1.036215
Pearson          =  12820.44155                   (1/df) Pearson  =  2.729496

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =  1.663567
Log likelihood   = -3906.550539                   BIC             = -34846.53

------------------------------------------------------------------------------
             |                 OIM
     chdfate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .6104111   .0524741    11.63   0.000     .5075638    .7132584
       _cons |  -4.549026   .0392232  -115.98   0.000    -4.625902    -4.47215
     per_yrs | (exposure)
------------------------------------------------------------------------------
```

{2} Regress **chdfate** against **male**. The options **family(poisson)** and **link(log)** specify that Poisson regression is to be used. **lnoffset(per_yrs)** specifies that the logarithm of per_yrs is to be used as an offset. In short, this statement specifies model

$$\log[\mathrm{E}[chd]] = \log[per\_yrs] + \alpha + male \times \beta$$

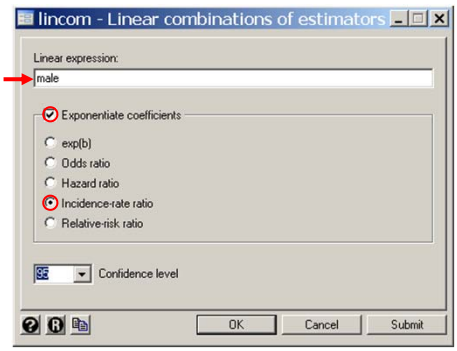


The *exposure* and *lnoffset* options are identical. They both enter the logarithm of *per_yrs* into the model as an offset.

```
. *Statistics > Postestimation > Linear combinations of estimates

. lincom male,irr                                                    {3}

 ( 1)  [chd]male = 0.0
------------------------------------------------------------------------------
         chd |        IRR   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.832227   .0961444     11.54   0.000     1.653154    2.030698
------------------------------------------------------------------------------
```

{3} The **irr** option has the same effect as the **or** option. That is, it calculates $e^{\hat{\beta}}$. The only difference is that this statistic is labeled **"IRR"** rather than **"Odds Ratio"**. **IRR** stands for **incidence rate ratio**, which is a synonym for **relative risk**. The estimate of β is 0.6055324. Hence the **relative risk** of CHD for men compared to women is $e^{\hat{\beta}} = \exp(0.6055324) = 1.832227$.

N.B. The ***or*** option of the ***lincom*** command really means "calculate $e^{\hat{\beta}}$" rather than "calculate an odds ratio" The label ***odds ratio*** in the output would be **incorrect**, since in Poisson regression $e^{\hat{\beta}}$ estimates a relative risk rather than an odds ratio.




```
. * Statistics > Epidemiology... > Tables... > Incidence-rate ratio calculator
. iri 823 650 42259 61451
                |   Exposed   Unexposed  |      Total
----------------+------------------------+-----------
          Cases |       823         650  |       1473
    Person-time |     42259       61451  |     103710
----------------+------------------------+-----------
                |                        |
 Incidence rate |   .0194751    .0105775  |   .0142031
                |                        |
                |     Point estimate     |    [95% Conf. Interval]
                |------------------------+------------------------
Inc. rate diff. |         .0088976       |     .0073383     .010457
Inc. rate ratio |         1.84118        |     1.659204     2.043774 (exact)
Attr. frac. ex. |           .45687       |     .3973015     .510709 (exact)
Attr. frac. pop |         .2552641       |
                +-------------------------------------------------
                    (midp)   Pr(k>=823) =                  0.0000 (exact)
                    (midp) 2*Pr(k>=823) =                  0.0000 (exact)
```

```
------------------------------------------------------------------------------
             |                 OIM
     chdfate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .6104111   .0524741    11.63   0.000     .5075638    .7132584
       _cons |  -4.549026   .0392232  -115.98   0.000    -4.625902    -4.47215
     per_yrs | (exposure)
------------------------------------------------------------------------------
```

c) 95% confidence intervals for relative risk estimates

$\hat{\beta}$ has an asymptotically normal distribution which allows us to estimate the 95% CI for β to be

.6104111 $\pm$ 1.96x0.05247 = (0.5075, 0.7132).

The 95% CI for the relative risk R = 1.832 is

(exp(0.5075), exp(0.7132)) = (1.661, 2.041).

d) Comparison of classical and Poisson risk estimates

The classical and Poisson relative risk estimates are in exact agreement.

The classical and Poisson 95% confidence intervals for this relative risk agree to three significant figures.

```
. lincom male,irr

  ( 1)  [chdfate]male = 0

------------------------------------------------------------------------------
     chdfate |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.841188   .0966146    11.63   0.000     1.661239     2.04063
------------------------------------------------------------------------------
```

Testing the null hypothesis that $R = 1$ is equivalent to testing the null hypothesis that $\beta = 0$.

The P value associated with this test is < 0.0005.

7. Assumptions needed for Poisson Regression

The distribution of d_i will be well approximated by a Poisson distribution if the following is true

a) Low death rates

The proportion of patients who die in each risk group should be small.

b) Independent events

Deaths in different patients are independent events.

The denominators of rates used in Poisson regressions is often patient-years rather than patients. Strictly speaking, deaths used in these rates are not independent since we can only die once. However, the independence assumption is not badly violated as long as the number of patients is large relative to the maximum number of years of follow-up per patient, and d_i is small.

8. Poisson Regression and Survival Analysis

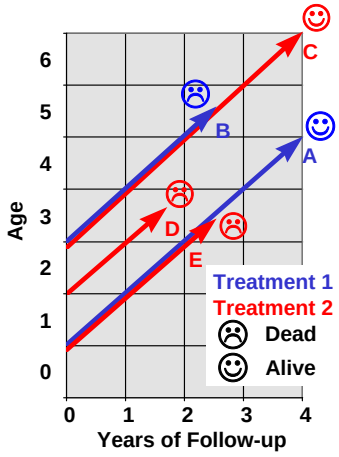
For large data sets Poisson regression is much faster than hazard regression analysis with time dependent covariates. If we have reason to believe that the proportional hazards assumption is false, it makes sense to do our exploratory analyses using Poisson regression. Before we can do this we must first convert the data from survival format to person-year format.

a) Recoding data on patients as patient-year data

Consider the following example:

Patient ID	Entry Age	Exit Age	Treatment	Fate
A	1	4	1	Alive
B	3	5	1	Dead
C	3	6	2	Alive
D	2	3	2	Dead
E	1	3	2	Dead

This data can be represented graphically as follows:



	Treatment 1		Treatment 2	
	Person-Years of Follow-up	Deaths	Person-Years of Follow-up	Deaths
A	0	0	1	0
B	1	1	1	0
C	2	0	1	0
D	2	0	3	2
E	1	0	2	0
	1	0	1	0

We need to convert the 5 patient records into 11 records of patient-years of follow-up.

9. Converting Survival Records to Person-Years of Follow-up.

The following program may be used as a template to convert survival records on individual patients into records giving person-years of follow-up.

```
. * 8.8.2.Survival_to_Person-Years.log
. *
. *   Convert survival data to person-year data.
. *   The survival data set must have the following
. *   variables
. *       id      =  patient id
. *       age_in  =  age at start of follow-up
. *       age_out =  age at end of follow-up
. *       fate    =  fate at exit: censored = 0, dead = 1
. *       treat   =  treatment variable.
. *
. *   The person-year data set created below will
. *   contain one record per unique combination of
. *   treatment and age.
. *
```

```
. *   Variables in the person-year data set that must not
. *   be in the original survival data set are
. *       age_now = an age of people in the cohort
. *       pt_yrs  = number of patient-years of observations
. *                 of people receiving therapy treat who
. *                 are age_now years old.
. *       deaths  = number of events (fate=1) occurring in
. *                 pt_yrs years of follow-up for this
. *                 group of patients.
. *
. use C:\WDDtext\8.8.2.Survival.dta, clear
. * Data > Describe data > List data
. list
```

	id	age_in	age_out	treat	fate
1.	A	1	4	1	0
2.	B	3	5	1	1
3.	C	3	6	2	0
4.	D	2	3	2	1
5.	E	1	3	2	1

```
. generate exit = age_out + 1
```
{1}

{1} A patient who is *age_out* years old at his end of follow-up will be in his *age_out* plus 1st year of life at that time. We define *exit* to be the patient's year of life at the end of follow-up.

```
. * Statistics > Survival... > Setup... > Declare data to be survival...
. stset exit, id(id) enter(time age_in) failure(fate)

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

------------------------------------------------------------------------------
        5  total obs.
        0  exclusions
------------------------------------------------------------------------------
        5  obs. remaining, representing
        5  subjects
        3  failures in single failure-per-subject data
     13.5  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         1
                                  last observed exit t =       6.5

. * Statistics > Survival... > Setup... > Split time-span records
. stsplit age_now, at(0(1)6)                                              {2}
(11 observations (episodes) created)
```

{2} This command, in combination with the preceding *stset* command expands the data set so that there is one record for each patient-year of follow-up. The effects of this command are illustrated by the following *list* command. See also Handout 6, pages 60 – 61.

```
stset exit, id(id) enter(time age_in) failure(fate)
stsplit age_now, at(0(1)6)

. * Data > Describe data > List data
. list  id age_in age_out treat fate exit age_now
```

	id	age_in	age_out	treat	fate	exit	age_now	{3,4}
1.	A	1	4	1	.	2	1	
2.	A	1	4	1	.	3	2	
3.	A	1	4	1	.	4	3	
4.	A	1	4	1	0	5	4	
5.	B	3	5	1	.	4	3	
6.	B	3	5	1	.	5	4	
7.	B	3	5	1	1	6	5	
8.	C	3	6	2	.	4	3	
9.	C	3	6	2	.	5	4	
10.	C	3	6	2	.	6	5	
11.	C	3	6	2	0	7	6	
12.	D	2	3	2	.	3	2	
13.	D	2	3	2	1	4	3	
14.	E	1	3	2	.	2	1	
15.	E	1	3	2	.	3	2	
16.	E	1	3	2	1	4	3	

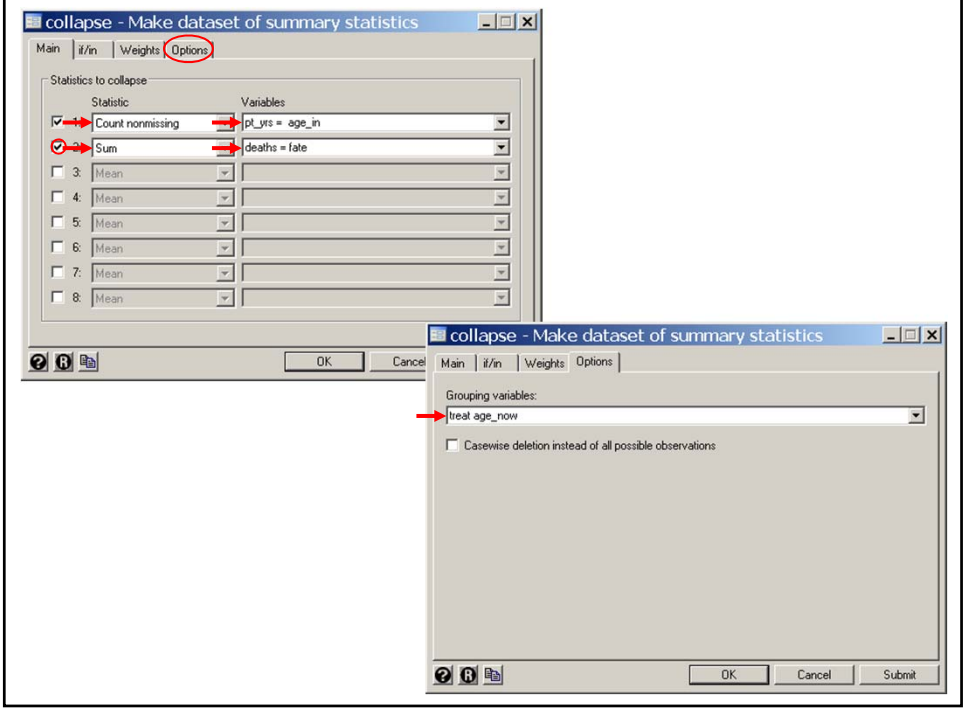
{3} There is now one record for each year of life that each patient had complete or partial follow-up. *age_now* equals *age_in* in each patient's first record and is incremented sequentially in subsequent records. It equals *age_out* at the last record.

{4} *fate* is the patient's true fate in his last record and is missing for other records. *stsplit* divides the observed follow-up into one year epochs with one record per epoch. Each epoch starts at *age_now* and ends at *exit*; *fate* gives the patient's fate at the end of the epoch.

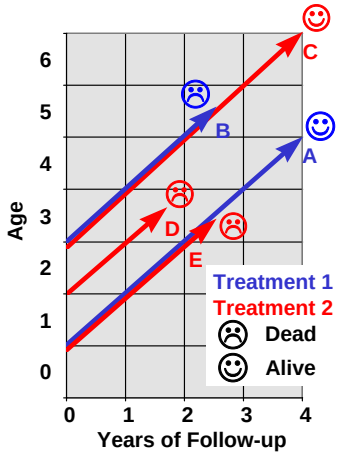
```
. sort treat age_now

. * Data > Create... > Other variable-trans... > Make dataset of means...
. collapse (count) pt_yrs=age_in (sum) deaths=fate, by(treat age_now)   {5}
```

{5} This statement **collapses** all records with **identical** values of ***treat*** and ***age_now*** into a single record. ***pt_yrs*** is set equal to the number of **records** collapsed. (More precisely, it is the count of collapsed records with non-missing values of *age_in*.)

deaths is set equal to the number of **deaths** (the sum of non-missing values of *fate* over these records). All **variables** are **deleted** from memory except ***treat age_now pt_yrs*** and ***deaths***.



```
. * Data > Describe data > List data
. list treat age_now pt_yrs deaths

     +-----------------------------------+
     | treat   age_now   pt_yrs   deaths |
     |-----------------------------------|
  1. |     1         1        1        0 |
  2. |     1         2        1        0 |
  3. |     1         3        2        0 |
  4. |     1         4        2        0 |
  5. |     1         5        1        1 |
     |-----------------------------------|
  6. |     2         1        1        0 |
  7. |     2         2        2        0 |
  8. |     2         3        3        2 |
  9. |     2         4        1        0 |
 10. |     2         5        1        0 |
     |-----------------------------------|
 11. |     2         6        1        0 |
     +-----------------------------------+

. save 8.8.2.Person-Years.dta, replace
file 8.8.2.Person-Years.dta saved
```



	Treatment 1		Treatment 2	
	Person-Years of Follow-up	Deaths	Person-Years of Follow-up	Deaths
A	0	0	1	0
B	1	1	1	0
C	2	0	1	0
D	2	0	3	2
E	1	0	2	0
	1	0	1	0

N.B.

a) If you are working on a large data set with many covariates you can reduce the computing time by only keeping the covariates that you will need in your model(s) before you start to convert to patient-year data.

b) It is a good idea to check that you have not changed the number of deaths or number of years of follow-up in your program. See the *8.9.Framingham.log* file in the next section for an example of how this can be done.

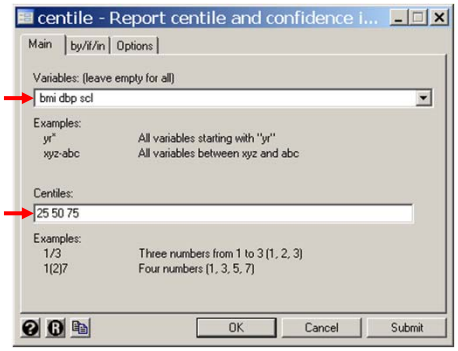
10. Converting the Framingham Survival Data to Person-time Data

The following log file shows how the Framingham Heart Study survival data set may be converted to a person-time data set that is suitable for Poisson regression analysis.

```
. * 8.9.Framingham.log
. *
. use  C:\WDDtext\2.20.Framingham.dta, clear
. *
. *  Convert bmi, scl and dbp into categorical variables
. *  that subdivide the data set into quartiles for each
. *  of these variables.
. *
. * Statistics > Summaries... > Summary and ... > Centiles with CIs
. centile bmi dbp scl, centile(25,50,75)                          {2}
```

{2} In the next chapter we will consider **body mass index**, **serum cholesterol**, and **diastolic blood pressure** as **confounding** variables in our analyses. We convert these data into **categorical** variables grouped by **quartiles**. This *centile* statement gives the 25th, 50th, and 75th quartile for *bmi, dbp* and *scl*. These are then used as arguments in the ***recode*** function to define categorical variables *bmi_gr, dbp_gr* and *scl_gr*.

```
                                                    -- Binom. Interp. --
Variable | Obs  Percentile    Centile        [95% Conf. Interval]
---------+-------------------------------------------------------------
     bmi | 4690         25       22.8            22.7          23
         |              50       25.2            25.1    25.36161
         |              75         28            27.9          28.1
     dbp | 4699         25         74              74            74
         |              50         80              80            82
         |              75         90              90            90
     scl | 4666         25        197             196           199
         |              50        225             222           225
         |              75        255             252           256
```




```
. generate bmi_gr = recode(bmi, 22.8, 25.2, 28, 29)
(9 missing values generated)

. generate dbp_gr = recode(dbp, 74,80,90,91)

. generate scl_gr = recode(scl, 197,225,255,256)
(33 missing values generated)
. *
. *  Calculate years of follow-up for each patient.
. *  Round to nearest year for censored patients.
. *  Round up to next year when patients exit with CHD
. *
. generate years=int(followup/365.25)+1 if chdfate                {3}
(3226 missing values generated)

. replace years=round(followup/365.25, 1) if ~chdfate             {4}
(3226 real changes made)
```

{3} The last follow-up interval for most patients is a fraction of a year. If the patient's follow-up was terminated because of a **CHD** event, we include the patient's **entire last year** as part of her follow-up. The ***int*** function facilitates this by truncating follow-up in years to the largest whole integer less than than *followup/365.25*. We then add 1 to this number to include the entire last year of follow-up.

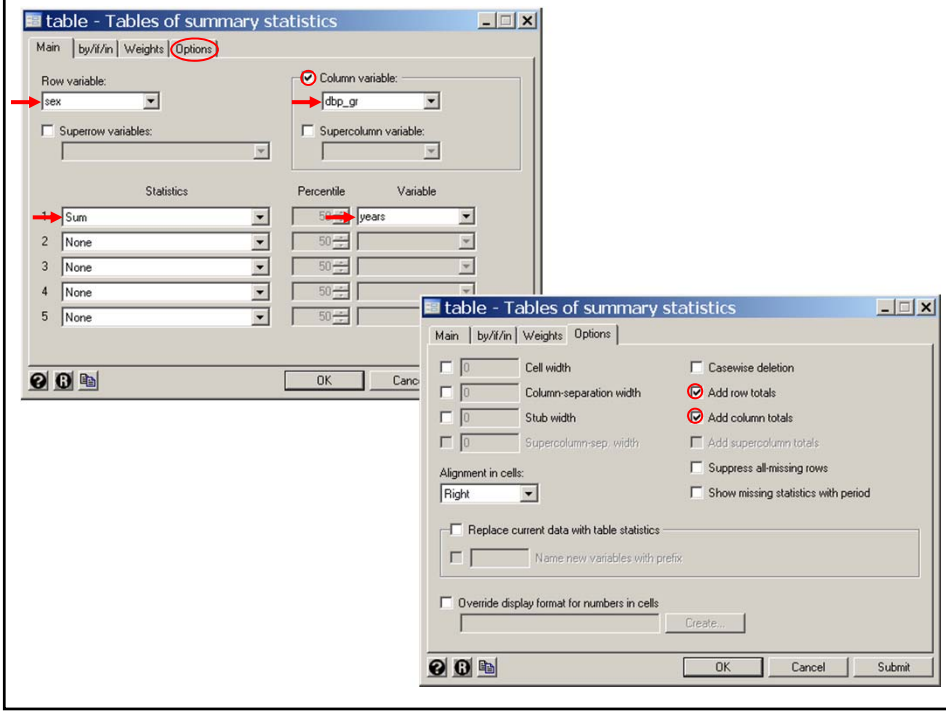
{4} If the patient is **censored** at the end of follow-up we **round** this number to the nearest integer using the *round* function. *round*(*x*,1) rounds *x* to the nearest integer.

```
. * Statistics > Summaries... > Tables > Table of summary statistics (table).
. table sex dbp_gr, contents(sum years) row col                          {5}

----------+-------------------------------------
          |                dbp_gr
      Sex |     74      80      90      91   Total
----------+-------------------------------------
      Men |  10663   10405   12795    8825   42688                      {6}
    Women |  21176   14680   15348   10569   61773
          |
    Total |  31839   25085   28143   19394  104461
----------+-------------------------------------
```

{5} So far, we haven't added any records or modified any of the original variables. Before doing this it is a good idea to **tabulate** the number of **person-years** of follow-up and CHD **events** in the data set. At the end of the transformation we can recalculate these tables to ensure that we have not lost or added any spurious years of follow-up or CHD events.

The next two tables show these data cross tabulated by *sex* and d*bp_gr*. The ***contents*(*sum years*)** option causes *years* to be summed over every **unique combination** of values of ***sex*** and ***dbp_gr*** and displayed in the table.

{6} For example, the sum of the *years* variable for men with *dbp_gr* = 90 is 12,795. This means that there are 12,795 person-years of follow-up for men with baseline diastolic blood pressures between 80 and 90.



```
. * Statistics > Summaries... > Tables > Table of summary statistics (table).
. table sex dbp_gr, contents(sum chdfate) row col                    {7}

----------+---------------------------------
          |              dbp_gr
      Sex |    74     80     90     91  Total
----------+---------------------------------
      Men |   161    194    222    246    823
    Women |   128    136    182    204    650
          |
    Total |   289    330    404    450   1473
----------+---------------------------------
```

{7} This table shows the corresponding number of **CHD** events.

```
. generate age_in = age
. generate exit = age + years
. summarize age_in exit
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      age_in |      4699    46.04107    8.504363         30         68
        exit |      4699    68.27155    10.09031         36         94

. *
. *  Transform data set so that there is one record per patient-year of
. *  follow-up.  Define age_now to be the patient's age in each record
. *
. * Statistics > Survival... > Setup... > Declare data to be survival...
. stset exit, id(id) enter(time age_in) failure(chdfate)

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

. * Statistics > Survival... > Setup... > Split time-span records
. stsplit age_now, at(30(1)94)
(99762 observations (episodes) created)
```

```
. * Data > Describe data > List data
. list id age_in years exit age_now  in 278/282                    {8}

     +---------------------------------------+
     |   id   age_in   years   exit   age_now |
     |---------------------------------------|
278. | 4075       59       3     62        61 |
279. | 4182       41       3     42        41 |
280. | 4182       41       3     43        42 |
281. | 4182       41       3     44        43 |
282. | 1730       46       3     47        46 |
     +---------------------------------------+
```

{8} The **expansion** of the data set by the *stset* and *stsplit* commands, and the definitions of *age_now*, and *exit* are done in the same way as in *8.8.2.Survival_to_Person-Years.log*. This *list* command shows the effects of these transformations. Note that patient **4182** entered the study at age **41** and exits at age **43** in his **44th** year of life. The expanded data set contains one record for each of these years.

```
. generate age_gr = recode(age_now, 45,50,55,60,65,70,75,80,81)      {9}
. label define age 45 "<= 45" 50 "45-50" 55 "50-55" 60 "55-60" 65 ///
>     "60-65" 70 "65-70" 75 "70-75" 80 "75-80" 81 "> 80"
. label values age_gr age
. sort sex bmi_gr scl_gr dbp_gr age_gr
. *
. *  Combine records with identical values of
. *  sex bmi_gr scl_gr dbp_gr and age_gr.
. *
. * Data > Create... > Other variable-trans... > Make dataset of means...
. collapse (count) pt_yrs=age_in (sum) chd_cnt=chdfate               {10}
>     , by(sex bmi_gr scl_gr dbp_gr age_gr)
. * Data > Describe data > List data
. list sex bmi_gr scl_gr dbp_gr age_gr pt_yrs chd_cnt in 310/315
>     , nodisplay

     +-------------------------------------------------------------+
     | sex   bmi_gr   scl_gr   dbp_gr   age_gr   pt_yrs   chd_cnt |
     |-------------------------------------------------------------|
310. | Men       28      197       90    45-50      124         0 |
311. | Men       28      197       90    50-55      150         1 |
312. | Men       28      197       90    55-60      158         2 |
313. | Men       28      197       90    60-65      161         4 |  {11}
314. | Men       28      197       90    65-70      100         2 |
     |-------------------------------------------------------------|
315. | Men       28      197       90    70-75       55         1 |
     +-------------------------------------------------------------+
```

{9} Recode *age_now* into 5-year age groups.

{10} **Collapse** records with identical values of *sex*, *bmi_gr*, *scl_gr*, *dbp_gr* and *age_gr*. ***pt_yrs*** records the number of **patient-years** of follow-up associated with each record while ***chd_cnt*** records the corresponding number of **CHD events**.

{11} For example, the subsequent listing shows that there were 161 patient-years of follow-up in men
aged 60 to 65 with
body mass indexes between 25.2 and 28,
serum cholesterols less than or equal to 197, and
diastolic blood pressures between 80 and 90 on their baseline exams.
Four CHD events occurred in these patients during these years of follow-up.

```
. * Statistics > Summaries... > Tables > Table of summary statistics (table).
. table sex dbp_gr, contents(sum pt_yrs) row col                        {12}
----------+-------------------------------------
          |                dbp_gr
      Sex |     74      80      90      91   Total
----------+-------------------------------------
      Men |  10663   10405   12795    8825   42688
    Women |  21176   14680   15348   10569   61773
          |
    Total |  31839   25085   28143   19394  104461
----------+-------------------------------------

. table sex dbp_gr, contents(sum chd_cnt) row col                       {13}
----------+---------------------------------
          |              dbp_gr
      Sex |    74     80     90     91  Total
----------+---------------------------------
      Men |   161    194    222    246    823
    Women |   128    136    182    204    650
          |
    Total |   289    330    404    450   1473
----------+---------------------------------
. generate male = sex == 1

. display _N
1267

. save  8.12.Framingham.dta, replace                                    {14}
(note: file 8.12.Framingham.dta not found)
file 8.12.Framingham.dta saved
```

{12} This table shows total **person-years** of follow-up cross-tabulated by *sex* and d*bp_gr*. Note that this table is identical to the one produced before the data transformation

Sex	dbp_gr 74	80	90	91	Total
Men	10663	10405	12795	8825	42688
Women	21176	14680	15348	10569	61773
Total	31839	25085	28143	19394	104461

{13} This table shows **CHD events** of follow-up cross-tabulated by *sex* and *dbp_gr*. This table is also identical to its pre-transformation version and supports the hypothesis that we have successfully transformed the data in the way we intended.

{14} The person-year data set is stored away for future analysis.

N.B. It is very important that you specify a **new** name for the transformed data set. If you use the original name you will **loose** the original data set. It is also a very good idea to always keep **back-up** copies of your original data sets in case you accidentally destroy the copy that you are working with.

11. What we have covered

- Elementary statistics involving rates
  - Incidence and relative risk
- Classical methods for deriving 95% confidence intervals for relative risks : the *iri* command
- Relationship between the binomial and Poisson distributions
- Poisson regression and 2x2 contingency tables: the *glm* command
- Estimating relative risks from Poisson regression models
  - Offsets in Poisson regression models: the *lnoffset* option
- Poisson regression is an example of a generalized linear model
  - Assumptions of the Poisson regression model
  - Contrast between logistic and Poisson regression
  - 95% confidence intervals for relative risk estimates
- Poisson Regression and survival analysis
  - Converting survival records to person-year records with Stata

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.

VIII. POISSON REGRESSION WITH MULTIPLE EXPLANATORY VARIABLES.

- Generalization of Poisson regression model to include multiple covariates
  - Deriving relative risk estimates from Poisson regression models
- Analyzing a complex survival data set with Poisson regression
  - The Framingham data set
  - Adjusting for confounding variables
  - Adding interaction terms
- Residual analysis

1. The Multiple Poisson Regression Model

Suppose that data on patients (or patient-years of follow-up) can be logically grouped into J strata based on age or other factors.

Let

$j = 1,\ldots,J$ denote the patient's strata.

Suppose that patients in strata j may be grouped into K exposure categories denoted by $k = 1,\ldots, K$.

Let $x_{jk1}, x_{jk2}, \ldots, x_{jkp}$ be explanatory variables that describe the k^{th} exposure group of patients in strata j, and

$\mathbf{x}_{jk} = \left(x_{jk1}, x_{jk2}, \ldots, x_{jkp}\right)$ denote the values of all of the covariates for patients in the j^{th} strata and k^{th} exposure category.

λ_{jk} be the probability that someone in strata j and exposure group k will die.

Then the **multiple Poisson regression** model assumes that

$$\log\left[\mathrm{E}\left[d_{jk} \mid \mathbf{x}_{jk}\right]\right] = \log\left[n_{jk}\right] + \alpha_j + \beta_1 x_{jk1} + \beta_2 x_{jk2} + \ldots + \beta_p x_{jkp} \qquad \{8.1\}$$

where

n_{jk} is the number of patients at risk in the j[th] strata who are in exposure group k

d_{jk} is the number of deaths (events) among these patients. d_{jk} is assumed to have a Poisson distribution with mean $n_{jk}\,\lambda_{jk}$,

$\alpha_1, \ldots, \alpha_J$ are unknown nuisance parameters, and

$\beta_1, \beta_2, \ldots, \beta_p$ are unknown parameters of interest.

For example, suppose that there are

$J = 5$ = five age strata.

and that patients are classified as light or heavy drinkers and light or heavy smokers in each strata. Then the are

$K = 4$ exposure categories (2 drinking categories times 2 smoking categories).

We might choose

$p = 2$ and let $\quad x_{jk1} = x_1 = \begin{cases} 1: \text{ Patient is heavy drinker} \\ 0: \text{ Patient is light drinker} \end{cases}$

$$x_{jk2} = x_2 = \begin{cases} 1: \text{ Patient is heavy smoker} \\ 0: \text{ Patient is light smoker} \end{cases}$$

Then the Poisson regression model is

$$\log(E(d_{jk})) = \log(n_{jk}) + \alpha_j + x_{jk1}\beta_1 + x_{jk2}\beta_2$$

where

$j = 1, 2, \ldots, 5;$

$k = 1, 2, 3, 4.$

		$k = 1$	$k = 2$	$k = 3$	$k = 4$
	$K = 4$ $J = 5$ $p = 2$	Light Drinker Light Smoker $x_1 = 0\ x_2 = 0$	Light Drinker Heavy Smoker $x_1 = 0\ x_2 = 1$	Heavy Drinker Light Smoker $x_1 = 1\ x_2 = 0$	Heavy Drinker Heavy Smoker $x_1 = 1\ x_2 = 1$
AGE	$j = 1$	$x_{111} = x_1 = 0$ $x_{112} = x_2 = 0$	$x_{121} = x_1 = 0$ $x_{122} = x_2 = 1$	…	$x_{141} = x_1 = 1$ $x_{142} = x_2 = 1$
	$j = 2$	$x_{211} = x_1 = 0$ $x_{212} = x_2 = 0$	$x_{221} = x_1 = 0$ $x_{222} = x_2 = 1$	…	…
	$j = 3$	$x_{311} = x_1 = 0$ $x_{312} = x_2 = 0$	$x_{321} = x_1 = 0$ $x_{322} = x_2 = 1$	…	…
	$j = 4$	$x_{411} = x_1 = 0$ $x_{412} = x_2 = 0$	$x_{421} = x_1 = 0$ $x_{422} = x_2 = 1$	$x_{431} = x_1 = 1$ $x_{432} = x_2 = 0$	…
	$j = 5$	$x_{511} = x_1 = 0$ $x_{512} = x_2 = 0$	$x_{521} = x_1 = 0$ $x_{522} = x_2 = 1$	…	$x_{541} = x_1 = 1$ $x_{542} = x_2 = 1$

Note that if we subtract $\log(n_{jk})$ from both sides of {8.1} we get

$$\log(E(d_{jk})/n_{jk}) =$$

$$\log(\lambda_{jk}) = \alpha_j + x_{jk1}\beta_1 + x_{jk2}\beta_2 + \ldots + x_{jkp}\beta_p \qquad \{8.2\}$$

Two patient groups with covariates $x_{jk'1}, x_{jk'2}, \ldots, x_{jk'p}$ and $x_{jk1}, x_{jk2}, \ldots, x_{jkp}$ will have log probabilities

$$\log(\lambda_{jk'}) = \alpha_j + x_{jk'1}\beta_1 + x_{jk'2}\beta_2 + \ldots + x_{jk'p}\beta_p$$

$$\log(\lambda_{jk}) = \alpha_j + x_{jk1}\beta_1 + x_{jk2}\beta_2 + \ldots + x_{jkp}\beta_p$$

Subtracting the latter equation from the former gives

$$\log(\lambda_{jk'}/\lambda_{jk}) =$$

$$(x_{jk'1} - x_{jk1})\beta_1 + (x_{jk'2} - x_{jk2})\beta_2 + \ldots + (x_{jk'p} - x_{jkp})\beta_p \qquad \{8.3\}$$

Thus, we can estimate log relative risks in Poisson regression models in precisely the same way that we estimated log odds ratios in logistic regression.

Indeed, the only difference is that in logistic regression weighted sums of model coefficients are interpreted as log odds ratios while in Poisson regression they are interpreted as log relative risks.

2. The 8.12.Framingham.dta Data Set

This is a person-time data set

The covariates are

BMI	grouped in quartiles
Serum cholesterol	grouped in quartiles
DBP	grouped in quartiles
gender	
age	$\leq$ 45, 46 – 50, …, 76 – 80, > 80

For each unique combination of covariate values we also have

pt_yrs	the number of patient-years of follow-up for patients with these covariate values
chd_cnt	the number of coronary heart disease events observed in these patient-years of follow-up

A patient who enters on his 44th birthday and exits at age 51 with CHD will contribute

2 patient-years of follow-up to the record for his covariate values and age 41 – 45,

5 patient-years of follow-up to the record for his covariate values and age 46 – 50, and

1 patient-year of follow-up to the record for his covariate values and age 51 – 55

He contributes

1 CHD event to the record for his covariate values and age 51 – 55

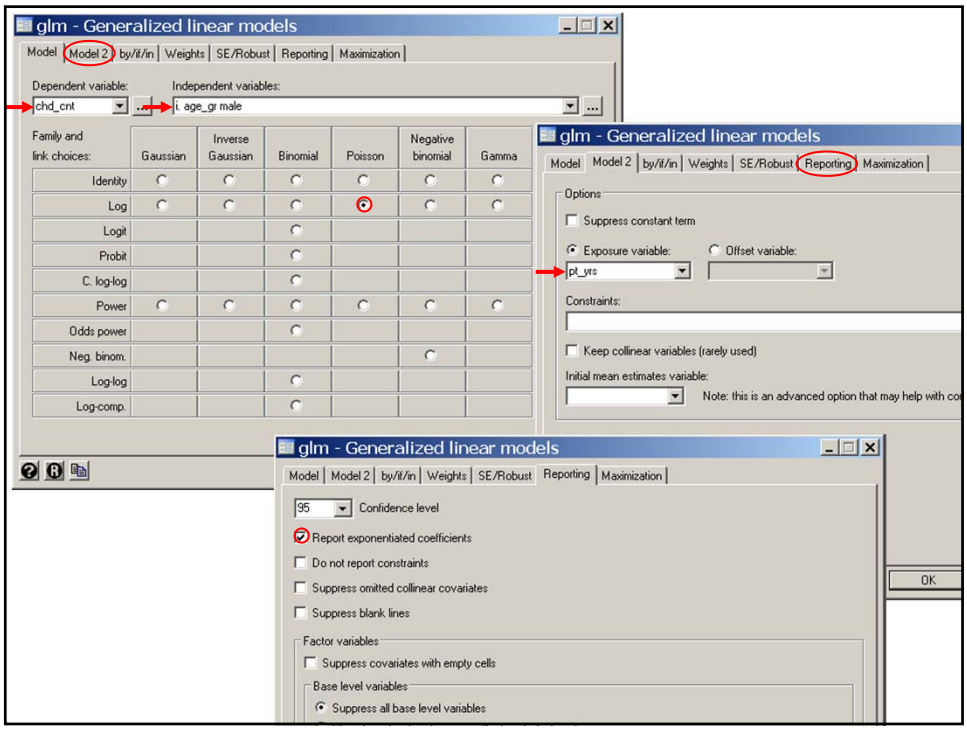
3. Gender, Age and CHD in the Framingham Heart Study

a) Analyzing the multiplicative model with Stata

```
. *  9.3.Framingham.log
. *
. *  Estimate the effect of age and gender on coronary heart disease CHD)
. *  using several Poisson regression models (Levy 1999).
. *
. use  C:\WDDtext\8.12.Framingham.dta, clear
. *
. *  Fit multiplicative model of effect of gender and age on CHD
. *
. * Statistics > Generalized linear models > Generalized linear models (GLM)
. glm chd_cnt i.age_gr male, family(poisson) link(log)                    {1}
> lnoffset(pt_yrs) eform
```

{1} We fit the model $\log(E(chd\_cnt)) = \log(pt\_yrs) + \alpha$

$$+50.age\_gr \times \alpha_2 + 55.age\_gr \times \alpha_3 + \ldots + 81.age\_gr \times \alpha_9 + male \times \beta$$



```
Generalized linear models                         No. of obs      =      1267
Optimization     : ML: Newton-Raphson             Residual df     =      1257
                                                  Scale parameter =         1
Deviance         =  1391.341888                   (1/df) Deviance =  1.106875
Pearson          =  1604.542689                   (1/df) Pearson  =  1.276486

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]
Standard errors  : OIM

Log likelihood   = -1559.206456                   AIC             =  2.477043
BIC              = -7589.177938
------------------------------------------------------------------------------
     chd_cnt |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_gr |
          50 |   1.864355   .3337745     3.48   0.001     1.312618    2.648005
          55 |   3.158729   .5058088     7.18   0.000     2.307858    4.323303
          60 |   4.885053   .7421312    10.44   0.000     3.627069    6.579347
          65 |    6.44168   .9620181    12.47   0.000     4.807047    8.632168
          70 |   6.725369   1.028591    12.46   0.000     4.983469    9.076127
          75 |   8.612712   1.354852    13.69   0.000     6.327596    11.72306
          80 |   10.37219   1.749287    13.87   0.000     7.452702    14.43534
          81 |   13.67189   2.515296    14.22   0.000     9.532967    19.60781
             |
        male |   1.996012   .1051841    13.12   0.000     1.800144    2.213192
      pt_yrs | (exposure)
------------------------------------------------------------------------------
```

The estimate of the coefficient for gender is 0.6918, which gives an age adjusted relative risk of CHD for men compared to women of

exp(0.6918) = 2.00.

This estimate is consistent with our previous estimates or this risk from other chapters.

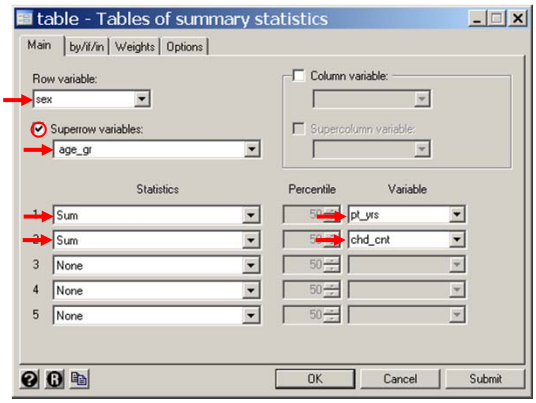
This risk is of limited interest because we know from Chapter VI that there is a powerful interaction between age and gender on coronary heart disease.

b) Age-sex specific incidence of CHD

Let us next plot the age specific incidence of CHD in men and women. *9.3.Framingham.log* continues.

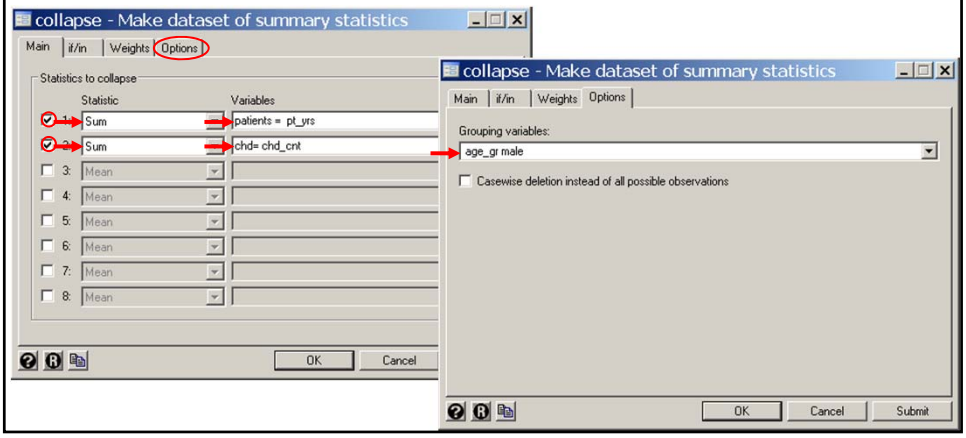
```
. *
. *  Tabulate patient-years of follow-up and number of
. *  CHD events by sex and age group.
. *
. * Statistics > Summaries... > Tables > Table of summary statistics (table)
. table sex, contents(sum pt_yrs sum chd_cnt) by(age_gr)
```

Age Group and Sex	sum(pt_yrs)	sum(chd_cnt)
<= 45		
Men	7370	43
Women	9205	9
45-50		
Men	5835	53
Women	7595	25
50-55		
Men	6814	110
Women	9113	46
55-60		
Men	7184	155
Women	10139	105
60-65		
Men	6678	178
Women	9946	148
65-70		
Men	4557	121
Women	7385	120
70-75		
Men	2575	94
Women	4579	88
75-80		
Men	1205	50
Women	2428	59
> 80		
Men	470	19
Women	1383	50




```
. *
. *  Calculate age-sex specific incidence of CHD
. *
. * Data > Create... > Other variable-trans... > Make dataset of means...
. collapse (sum) patients = pt_yrs chd = chd_cnt, by(age_gr male)            {1}
```

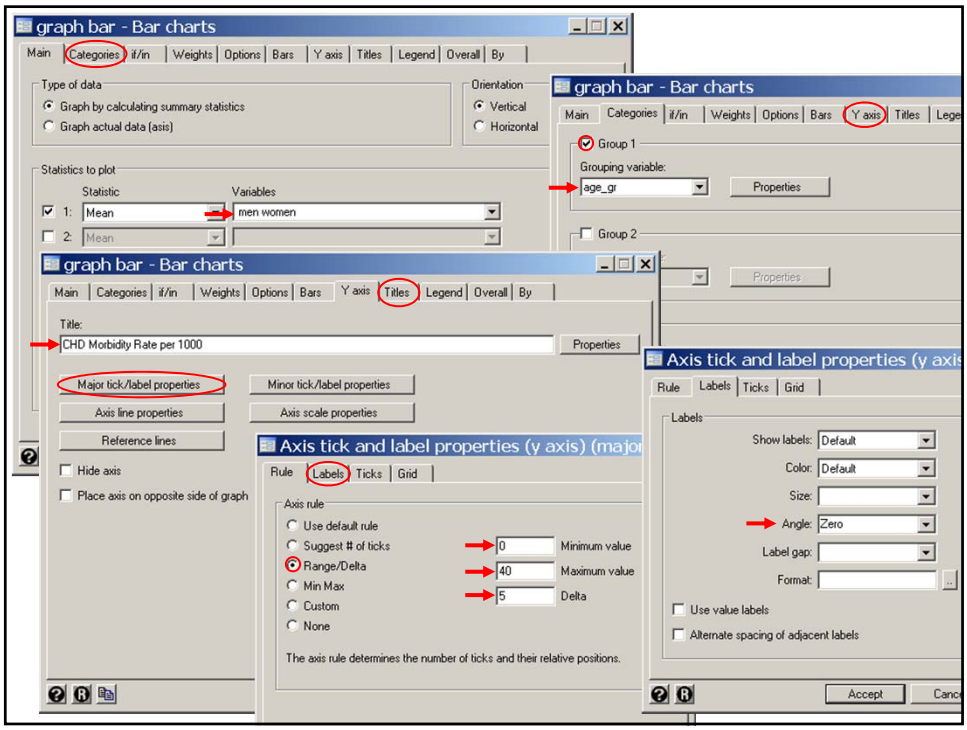
{1} Collapse the data file to one record for each combination of ***age_gr*** and ***sex***. Let ***patients*** be the total number of patient-years of follow-up and let ***chd*** be the total number CHD events in these groups.

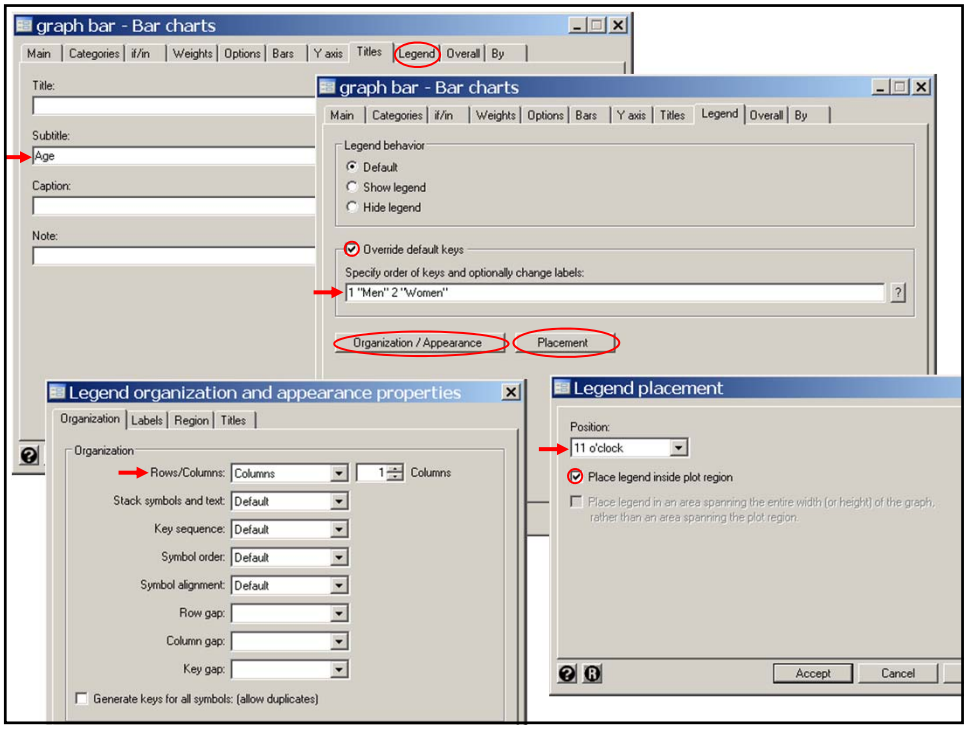


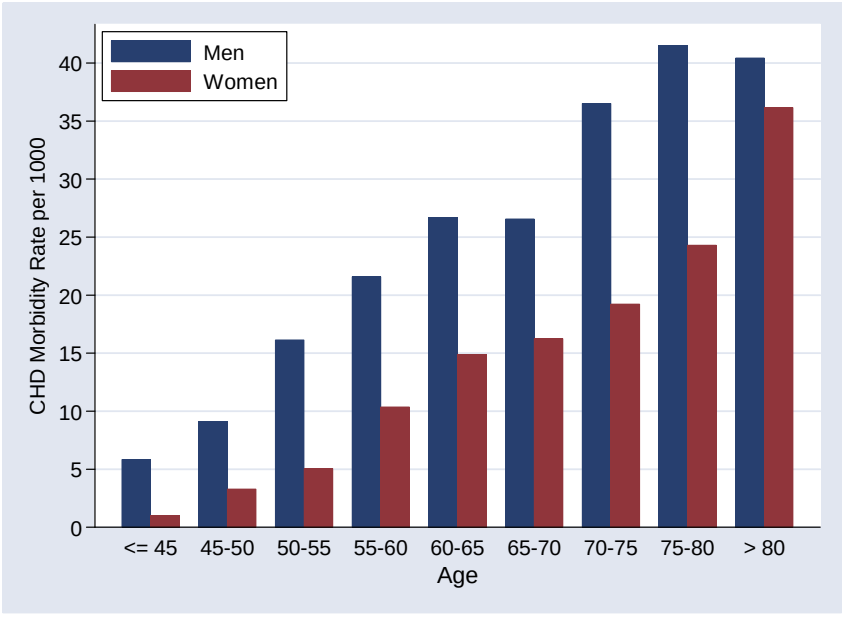
```
. generate rate = 1000*chd/patients                                          {2}
. generate men = rate if male==1
(9 missing values generated)
. generate women = rate if male==0
(9 missing values generated)

.* Graphics > Bar chart
. graph bar men women, over(age_gr) ytitle(CHD Morbidity Rate per 1000)    /// {3}
>     ylabel(0(5)40, angle(0)) subtitle(Age, position(6))                  ///
>     legend(order(1 "Men" 2 "Women") ring(0) position(11) col(1))
```

{2} ***rate*** is the age-sex specific incidence rate of CHD per year per 1,000.

{3} The ***bar*** option specifies that a **bar graph** is to be produced. The two variables ***men*** and ***women*** together with the ***over(age_gr)*** option specify that a **grouped bar graph** of ***men*** and ***women*** stratified by ***age_gr*** is to be drawn. The *y*-axis is the mean of the values of *men* and *women* in all records with identical values of *age_gr*. However, in this particular example, there is only one non-missing value of *men* and *women* for each age group.







This graph is consistent the results of the survival analysis and indicates that the relative risk of CHD in men vs. women diminishes with increasing age.

c) Using Poisson regression to model the effects of gender and age on CHD risk

Let us now model this relationship. *9.3.Framingham.log* continues.

```
. use  C:\WDDtext\8.12.Framingham.dta, clear                                {1}

. *
. *  Add interaction terms to the model
. *
. * Statistics > Generalized linear models > Generalized linear models (GLM)
. glm chd_cnt age_gr##male, family(poisson) link(log) lnoffset(pt_yrs)      {2}
```

{1} In creating the preceding bar graph we collapsed the data set. We need to restore the original data set before preceding.

{2} In this model we add 9 interaction terms of the form

50.age_gr#1.male = *50.age_gr×1.male*,
55.age_gr#1.male = *55.age_gr×1.male*,
.
.
.
80.age_gr#1.male = *80.age_gr×1.male*, and
81.age_gr#1.male = *81.age_gr×1.male*.

The syntax is identical to that used in Chapter IV.

```
Iteration 0:   log likelihood = -1621.7301
Iteration 1:   log likelihood = -1547.0628
Iteration 2:   log likelihood = -1544.3498
Iteration 3:   log likelihood = -1544.3226
Iteration 4:   log likelihood = -1544.3226

Generalized linear models                         No. of obs      =      1267
Optimization     : ML: Newton-Raphson             Residual df     =      1249
                                                  Scale parameter =         1
Deviance         =  1361.574107                   (1/df) Deviance =  1.090131
Pearson          =  1556.644381                   (1/df) Pearson  =  1.246313

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]
Standard errors  : OIM

Log likelihood   = -1544.322566                   AIC             =  2.466176
BIC              = -7561.790461
```

chd_cnt	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age_gr						
50	1.213908	.3887301	3.12	0.002	.4520112	1.975805
55	1.641462	.3644863	4.50	0.000	.9270817	2.355842
60	2.360093	.3473254	6.80	0.000	1.679348	3.040838
65	2.722564	.3433189	7.93	0.000	2.049671	3.395457
70	2.810563	.3456074	8.13	0.000	2.133185	3.487941
75	2.978378	.3499639	8.51	0.000	2.292462	3.664295
80	3.212992	.3578551	8.98	0.000	2.511609	3.914375
81	3.61029	.3620927	9.97	0.000	2.900602	4.319979
1.male	1.786305	.3665609	4.87	0.000	1.067858	2.504751
age_gr#male						
50 1	-.771273	.4395848	-1.75	0.079	-1.632843	.0902975
55 1	-.623743	.4064443	-1.53	0.125	-1.420359	.1728731
60 1	-1.052307	.3877401	-2.71	0.007	-1.812263	-.2923503
65 1	-1.203381	.3830687	-3.14	0.002	-1.954182	-.4525805
70 1	-1.295219	.3885418	-3.33	0.001	-2.056747	-.5336915
75 1	-1.144716	.395435	-2.89	0.004	-1.919754	-.3696772
80 1	-1.251231	.4139035	-3.02	0.003	-2.062467	-.4399949
81 1	-1.674611	.4549709	-3.68	0.000	-2.566338	-.7828845
_cons	-6.930278	.3333333	-20.79	0.000	-7.583599	-6.276956
pt_yrs	(exposure)					

```
. lincom 1.male, irr                                                    {3}

  ( 1)  [chd_cnt]male = 0
------------------------------------------------------------------------------
     chd_cnt |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    5.96736   2.187401     4.87   0.000     2.909143    12.24051
------------------------------------------------------------------------------
```

{3} The risk of CHD for a man ≤ 45 years of age is 5.97 times that of a woman of comparable age.

```
. lincom 1.male + 50.age_gr#1.male, irr                                 {4}

  ( 1)  [chd_cnt]1.male + [chd_cnt]50.age_gr#1.male = 0
------------------------------------------------------------------------------
     chd_cnt |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   2.759451   .6695176     4.18   0.000     1.715134    4.439635
------------------------------------------------------------------------------
.
.
.
```

{4} The log incidence of CHD for a man aged 45-50 is

$$\_cons + 1.male + 50.age\_gr + 50.age\_gr\#1.male \qquad \{8.4\}$$

For women, the corresponding log incidence is

$$\_cons + 50.age\_gr \qquad \{8.5\}$$

Subtracting {8.5} from {8.4} gives that the log relative risk for men aged 45-50 compared to women of the same age is

$$1.male + 50.age\_gr\#1.male$$

We put these terms in the *lincom* statement to estimate the relative risk for men in this age group to be 2.76.

Similar *lincom* commands permit us to complete the following table.

Table 8.1. Age-specific relative risks of CHD in men compared to women (5 year age intervals).

Age	Patient-years of follow-up		CHD Events		Relative Risk	95% Confidence Interval
	Men	Women	Men	Women		
< 45	7,370	9,205	43	9	5.97	2.9 - 12
46 - 50	5,835	7,595	53	25	2.76	1.7 - 4.4
51 - 55	6,814	9,113	110	46	3.20	2.3 - 4.5
56 - 60	7,184	10,139	155	105	2.08	1.6 - 2.7
61 - 65	6,678	9,946	178	148	1.79	1.4 - 2.2
66 - 70	4,557	7,385	121	120	1.63	1.3 - 2.1
71 - 75	2,575	4,579	94	88	1.90	1.4 - 2.5
76 - 80	1,205	2,428	50	59	1.71	1.2 - 2.5
> 80	470	1,383	19	50	1.12	0.66 - 1.9

From the preceding table it appears reasonable to collapse ages 46 - 55 into one interval, and ages 61 - 80 into another. We do this next as *9.3.Framingham.log* continues.

```
. *
. *  Refit model with interaction terms using fewer parameters.
. *
. generate age_gr2 = recode(age_gr, 45,55,60,80,81)                    {1}

. * Statistics > Generalized linear models > Generalized linear models (GLM)
. glm chd_cnt age_gr2##male                                        ///
>     , family(poisson) link(log) lnoffset(pt_yrs) eform               {2}

Iteration 0:   log likelihood = -1648.0067
Iteration 1:   log likelihood = -1566.4477
Iteration 2:   log likelihood = -1563.8475
Iteration 3:   log likelihood = -1563.8267
Iteration 4:   log likelihood = -1563.8267

Generalized linear models                         No. of obs      =      1267
Optimization     : ML: Newton-Raphson             Residual df     =      1257
                                                  Scale parameter =         1
Deviance         =  1400.582451                   (1/df) Deviance =  1.114226
Pearson          =  1656.387168                   (1/df) Pearson  =   1.31773

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]
Standard errors  : OIM

Log likelihood   = -1563.826738                   AIC             =  2.484336
                                                  BIC             = -7579.937
```

{1} This model is identical to the preceding one except that we have fewer age groups. We can generate the following table using *lincom* commands similar to those used to produce Table 8.1.

{2} ***eform*** exponentiates the coefficients in the output table

chd_cnt	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age_gr2						
55	4.346255	1.537835	4.15	0.000	2.172374	8.695524
60	10.59194	3.678849	6.80	0.000	5.362059	20.92278
80	17.43992	5.876004	8.48	0.000	9.010534	33.75503
81	36.97678	13.38902	9.97	0.000	18.18508	75.18703
1.male	5.96736	2.187401	4.87	0.000	2.909143	12.24051
age_gr2#male						
55 1	.5081773	.1998025	-1.72	0.085	.2351496	1.098212
60 1	.3491314	.1353722	-2.71	0.007	.1632841	.746507
80 1	.2899566	.1081168	-3.32	0.001	.1396186	.6021748
81 1	.1873811	.0852529	-3.68	0.000	.0768164	.4570857
pt_yrs	(exposure)					

```
. lincom 1.male + 55.age_gr2#1.male, irr

 ( 1)  [chd_cnt]1.male + [chd_cnt]55.age_gr2#1.male = 0
------------------------------------------------------------------------------
     chd_cnt |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   3.032477   .4312037     7.80   0.000     2.294884    4.007138
------------------------------------------------------------------------------

. lincom 1.male + 60.age_gr2#1.male, irr

  ( 1)  [chd_cnt]1.male + [chd_cnt]80.age_gr2#1.male = 0
------------------------------------------------------------------------------
     chd_cnt |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   2.083393   .2633282     5.81   0.000     1.626239    2.669057
------------------------------------------------------------------------------
```

Table 8.2. Age-specific relative risks of CHD in men compared to women (variable age intervals).

Age	Patient-years of follow-up		CHD Events		Relative Risk	95% Confidence Interval
	Men	Women	Men	Women		
< 45	7,370	9,205	43	9	5.97	2.9 - 12
46 - 55	12,649	16,708	163	71	3.03	2.3 - 4.0
56 - 60	7,184	10,139	155	105	2.08	1.6 - 2.7
61 - 80	15,015	24,338	443	415	1.73	1.5 - 2.0
> 80	470	1,383	19	50	1.12	0.66 - 1.9

This table suggests that **men** are at substantially **increased** risk of CHD compared to **premenopausal** women of the same age. After the menopause this risk ratio declines but remains significant until age 80. After age **80** there is **no** significant difference in CHD risk between men and women.

d) Adjusting CHD risk for confounding variables

Of course Table 8.2 is based on observational data, and may be influenced by confounding variables. We next adjust these results for possible confounding due to body mass index, serum cholesterol, and diastolic blood pressure. *9.3. Framingham.log* continues.

```
. table bmi_gr

----------------------
   bmi_gr |      Freq.
----------+-----------
     22.8 |        312
     25.2 |        290
       28 |        320
       29 |        312
----------------------

. *
. *  The i. syntax only works for integer variables.  bmi_gr gives the
. *  quartile boundaries to one decimal place.  We multiply this variable
. *  by 10 in order to be able to use this syntax.  Since indicator
. *  covariates are entered into the model, multiplying by 10 will
. *  not affect our estimates
. *
. gen bmi_gr10 = bmi_gr*10
(33 missing values generated)
```

```
. *
. *  Adjust analysis for body mass index (BMI)
. *
. * Statistics > Generalized linear models > Generalized linear models (GLM)
. glm chd_cnt age_gr2##male i.bmi_gr10                          ///
>     , family(poisson) link(log) lnoffset(pt_yrs)

Generalized linear models                         No. of obs      =      1234
Optimization     : ML: Newton-Raphson             Residual df     =      1221
                                                  Scale parameter =         1
Deviance         =   1327.64597                   (1/df) Deviance =  1.087343
Pearson          =  1569.093606                   (1/df) Pearson  =  1.285089

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]
Standard errors  : OIM

Log likelihood   = -1526.358498                   AIC             =  2.494908
                                                  BIC             = -7363.452
```

This model is **nested** within the preceding model and contains **3** more **parameters**. Therefore the reduction in model deviance will have an asymptotically χ^2 distribution with 3 degrees of freedom under the null hypothesis that the simpler model is correct.

This reduction is 1,401 - 1,328 = **73**, which is overwhelmingly significant ($P < 10^{-14}$). We will leave ***i.bmi_gr10*** in the model.

```
. *
. *  Adjust estimates for BMI and serum cholesterol
. *
. * Statistics > Generalized linear models > Generalized linear models (GLM)
. glm chd_cnt age_gr2##male i.bmi_gr10  i.scl_gr                        ///
>     , family(poisson) link(log) lnoffset(pt_yrs)

Iteration 0:   log likelihood =  -1506.494
Iteration 1:   log likelihood = -1461.0514
Iteration 2:   log likelihood = -1460.2198
Iteration 3:   log likelihood = -1460.2162
Iteration 4:   log likelihood = -1460.2162

Generalized linear models                         No. of obs      =      1134
Optimization     : ML: Newton-Raphson             Residual df     =      1118
                                                  Scale parameter =         1
Deviance         =  1207.974985                   (1/df) Deviance =  1.080479
Pearson          =  1317.922267                   (1/df) Pearson  =  1.178821

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]
Standard errors  : OIM

Log likelihood   = -1460.216152                   AIC             =  2.603556
                                                  BIC             = -6655.485
```

The model **deviance** is reduced by 1,328 - 1208 = 120, which has a χ^2 distribution with 3 degrees of freedom with P $<10^{-25}$.

```
. *
. *  Adjust estimates for BMI serum cholesterol and
. *  diastolic blood pressure
. *
. * Statistics > Generalized linear models > Generalized linear models (GLM)
. glm chd_cnt age_gr2##male i.bmi_gr10  i.scl_gr  i.dbp_gr              ///
>     , family(poisson) link(log) lnoffset(pt_yrs) eform
   .
   .
   .
Generalized linear models                         No. of obs      =      1134
Optimization     : ML: Newton-Raphson             Residual df     =      1115
                                                  Scale parameter =         1
Deviance         =  1161.091086                   (1/df) Deviance =  1.041337
Pearson          =  1228.755896                   (1/df) Pearson  =  1.102023

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]
Standard errors  : OIM

Log likelihood   = -1436.774203                   AIC             =  2.567503
                                                  BIC             = -6681.269
```

The model **deviance** is reduced by 1208 - 1161 = 47, which has a χ^2 distribution with 3 degrees of freedom with P $<10^{-9}$.

chd_cnt	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age_gr2						
55	3.757544	1.330347	3.74	0.000	1.877322	7.520891
60	8.411826	2.926018	6.12	0.000	4.254059	16.63325
80	12.78983	4.320508	7.54	0.000	6.596628	24.79748
81	23.92787	8.701246	8.73	0.000	11.73192	48.80217
1.male	4.637662	1.703034	4.18	0.000	2.257991	9.525239
age_gr2#male						
55 1	.5610101	.2207001	-1.47	0.142	.2594836	1.212918
60 1	.4230946	.1642325	-2.22	0.027	.1977092	.9054158
80 1	.3851572	.1438922	-2.55	0.011	.1851974	.8010161
81 1	.2688892	.1234925	-2.86	0.004	.1093058	.6614603
bmi_gr10						
252	1.159495	.0991218	1.73	0.083	.9806235	1.370994
280	1.298532	.1077862	3.15	0.002	1.103564	1.527944
290	1.479603	.1251218	4.63	0.000	1.253614	1.746332
scl_gr						
225	1.189835	.1004557	2.06	0.040	1.008374	1.403952
255	1.649807	.1339827	6.16	0.000	1.407039	1.934462
256	1.793581	.1466507	7.15	0.000	1.527999	2.105323
dbp_gr						
80	1.18517	.0962869	2.09	0.037	1.010709	1.389744
90	1.122983	.0892217	1.46	0.144	.9610473	1.312205
91	1.638383	.1302205	6.21	0.000	1.402041	1.914564
pt_yrs	(exposure)					

```
. lincom 1.male + 55.age_gr2#1.male, irr                                {1}

 ( 1)  [chd_cnt]1.male + [chd_cnt]55.age_gr2#1.male = 0
------------------------------------------------------------------------------
     chd_cnt |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   2.601775   .3722797     6.68   0.000     1.965505    3.444019
------------------------------------------------------------------------------

. lincom 1.male + 60.age_gr2#1.male, irr

  ( 1)  [chd_cnt]1.male + [chd_cnt]60.age_gr2#1.male = 0
------------------------------------------------------------------------------
     chd_cnt |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    1.96217   .2491985     5.31   0.000     1.529793    2.516752
------------------------------------------------------------------------------
```

{1} We next use *lincom* statements in the same way as before to construct Table 8.3.

Table 8.3. Age-specific relative risks of CHD in men compared to women. Risks are adjusted for body mass index, serum cholesterol and diastolic blood pressure.

Age	Patient-years of follow-up		CHD Events		Relative Risk	95% Confidence Interval
	Men	Women	Men	Women		
< 45	7,370	9,205	43	9	4.64	2.3 – 9.5
46 - 55	12,649	16,708	163	71	2.60	2.0 - 3.4
56 - 60	7,184	10,139	155	105	1.96	1.5 - 2.5
61 - 80	15,015	24,338	443	415	1.79	1.6 - 2.0
> 80	470	1,383	19	50	1.25	0.73 - 2.1

Compare Tables 8.3 and 8.2.

Both tables indicate a pronounced reduction in CHD risk for women that diminishes with age.

Adjusting for body mass index, serum cholesterol and diastolic blood pressure reduces but does not eliminate the magnitude of this benefit.

Age	Patient-years of follow-up		CHD Events		8.2. Unadjusted		8.2. Adjusted for BMI, SCL & DBP	
	Men	Women	Men	Women	Relative Risk	95% Confidence Interval	Relative Risk	95% Confidence Interval
< 45	7,370	9,205	43	9	5.97	2.9 - 12	4.64	2.3 – 9.5
46 - 55	12,649	16,708	163	71	3.03	2.3 - 4.0	2.60	2.0 - 3.4
56 - 60	7,184	10,139	155	105	2.08	1.6 - 2.7	1.96	1.5 - 2.5
61 - 80	15,015	24,338	443	415	1.73	1.5 - 2.0	1.79	1.6 - 2.0
> 80	470	1,383	19	50	1.12	0.66 - 1.9	1.25	0.73 - 2.1

4. Confounding versus Overmatching

It cannot be overemphasized that the correct model depends on the biologic context and cannot be ascertained solely through mathematical analysis.

One of the many ways we can go wrong is to confuse a true confounding variable with one that is on the causal pathway to the outcome of interest.

Such variables look like confounding variables in that they are correlated with both the exposure and disease outcome of interest.

Adjusting for such variables is called **overmatching** and can cause a serious underestimate of the true relative risk.

Consider the preceding example.

We know that

- Low density serum cholesterol (LDSC) is an independent risk factor for CHD.
- Exogenous estrogens reduce LDSC, and women who take hormonal replacement therapy have reduced risks of CHD.

Thus, it is plausible that the reduced CHD risk of premenopausal women results, in part, from a reduction in LDSC due to endogenous estrogens.

In this case adjusting for serum cholesterol may constitute overmatching and may falsely lower the relative risk of CHD for middle aged men.

5. Residual Analyses for Poisson Regression

Looking for outliers or poor model fit is done as follows.

a) Deviance residuals

Let

$$\log(E(d_{jk})) = \log(n_{jk}) + \alpha_j + x_{jk1}\beta_1 + x_{jk2}\beta_2 + \ldots + x_{jkp}\beta_p$$

be the standard Poisson regression model defined by equation {8.1},

$D = \sum_{jk} c_{jk}$ be the model Deviance, where c_{jk} is a non-negative value that represents the contribution to the deviance of the group of patients with identical covariate values, and

$$r_{jk} = \text{sign}\left(d_{jk} - E\left(\hat{d}_{jk}\right)\right)\sqrt{c_{jk}} \qquad \textbf{\{8.6\}}$$

where $E\left(\hat{d}_{jk}\right)$ is the estimated value of $E(d_{jk})$ under the model.

Then r_{jk} is the **deviance residual** for these patients and $D = \sum_{jk} r_{jk}^2$

As with Pearson residuals, deviance residuals are affected by varying degrees of leverage associated with the different covariate patterns. This leverage tends to shorten the residual by pulling the estimate of $\hat{\lambda}_{jk}$ in the direction of d_{jk}/n_{jk}

We can adjust for this shrinkage by calculating the **standardized deviance residual**

$$r_{jk}^s = r_{jk}/\sqrt{1-h_{jk}}$$

where h_{jk} is the leverage of the jk^{th} covariate pattern.

If the model is correct, roughly 95% of these residuals should lie between $\pm$ 2

It doesn't matter how many records have identical covariates when we are fitting a Poisson regression model.

However, many such records with residuals having the same sign may result in a poor model fit that does not show up in a residual analysis that calculates a separate residual for each identical record.

For this reason it is best to compress such records before analyzing our residuals.

b) Residual analysis of CHD model of sex, age and other variables

9.3.Framingham.log continues.

```
*
*  Compress data set for residual plot
*
. sort male bmi_gr scl_gr dbp_gr age_gr2                                {1}
. * Data > Create... > Other variable-trans... > Make dataset of means...
. collapse (sum) pt_yrs=pt_yrs chd_cnt=chd_cnt,          ///            {2}
> by (male bmi_gr10 scl_gr dbp_gr age_gr2)
```

{1} Before compressing the data file we must bring all records with identical covariates together. We do this with the *sort* command.

{2} This command combines all records with identical values of *male, bmi_*gr, *scl_gr, dbp_gr3,* and *age_gr2* together. ***pt_yrs*** and ***chd_cnt*** denote the total number of **patient-years** of observation and total number of CHD **events** in these records, respectively.

```
. *
. *  Re-analyze previous model using collapsed data set.
. *
. * Statistics > Generalized linear models > Generalized linear models (GLM)
. glm chd_cnt age_gr2##male i.bmi_gr10  i.scl_gr  i.dbp_gr        ///       {3}
> , family(poisson) link(log) lnoffset(pt_yrs)
```

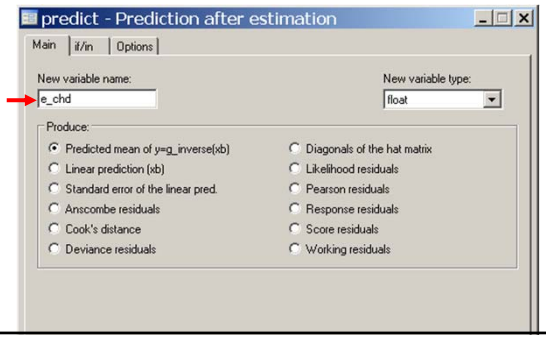
{3} This command fits the same model used for Table 8.3.

```
Generalized linear models                         No. of obs      =       623
Optimization     : ML: Newton-Raphson             Residual df     =       604
                                                  Scale parameter =         1
Deviance         =  600.7760472                   (1/df) Deviance =  .9946623 {4}
Pearson          =  633.8816072                   (1/df) Pearson  =  1.049473

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =  2.862427
Log likelihood   =  -872.645946                   BIC             =  -3285.69
```

{4} Collapsing the data set reduces the model deviance but has **no effect** on the model's parameter **estimates** or their standard errors. The table of coefficients, standard errors and confidence intervals is not shown here (see the output from the last time we ran this model in Section 2c).

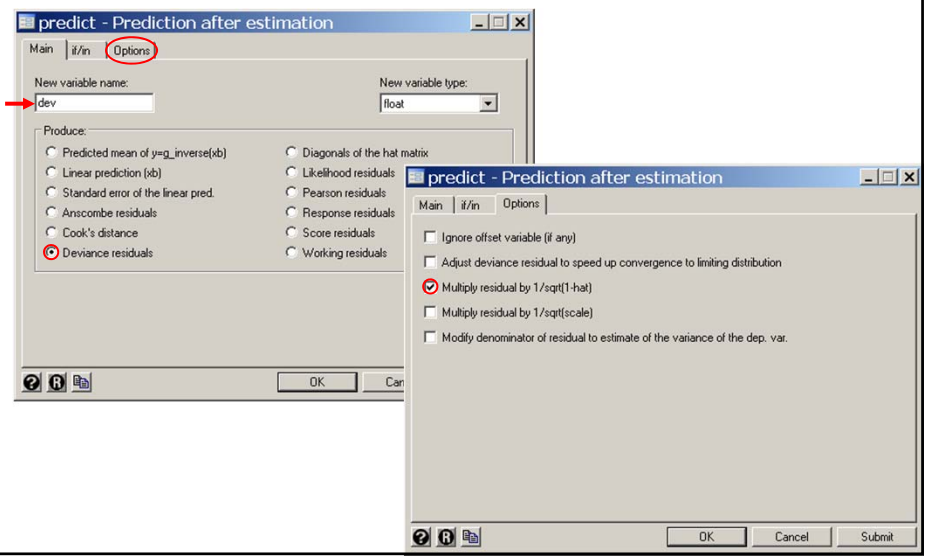
```
. *
. *  Estimate the expected number of CHD events and the
. *  standardized deviance residual for each record in the data set.
. *
. predict e_chd, mu                                                       {5}
(82 missing values generated)
```

{5} The ***mu*** option of this command defines ***e_chd*** to equal $\hat{E}(d_{jk})$, the estimated expected number of deaths for each record. More generally, it calculates the **inverse** of the **link function** evaluated at the **linear predictor** for the given record. For Poisson regression this is the exponentiated value of the linear predictor.



```
. predict dev, standardized deviance                                    {6}
(82 missing values generated)
```

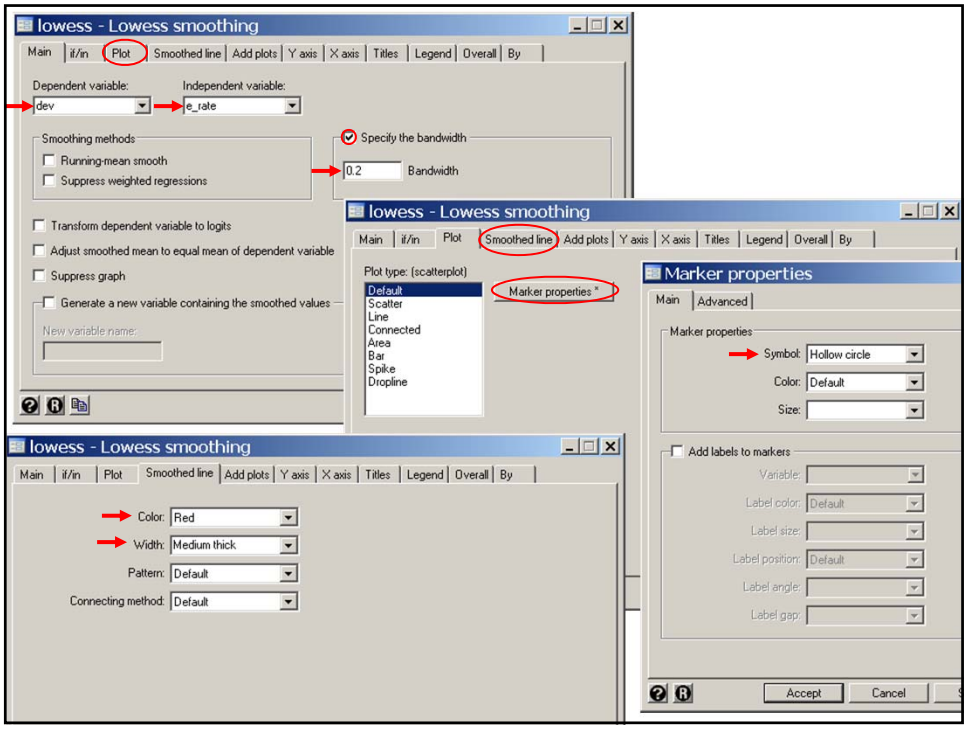
{6} This **predict** command calculates ***dev*** to equal the **standardized deviance residual.**

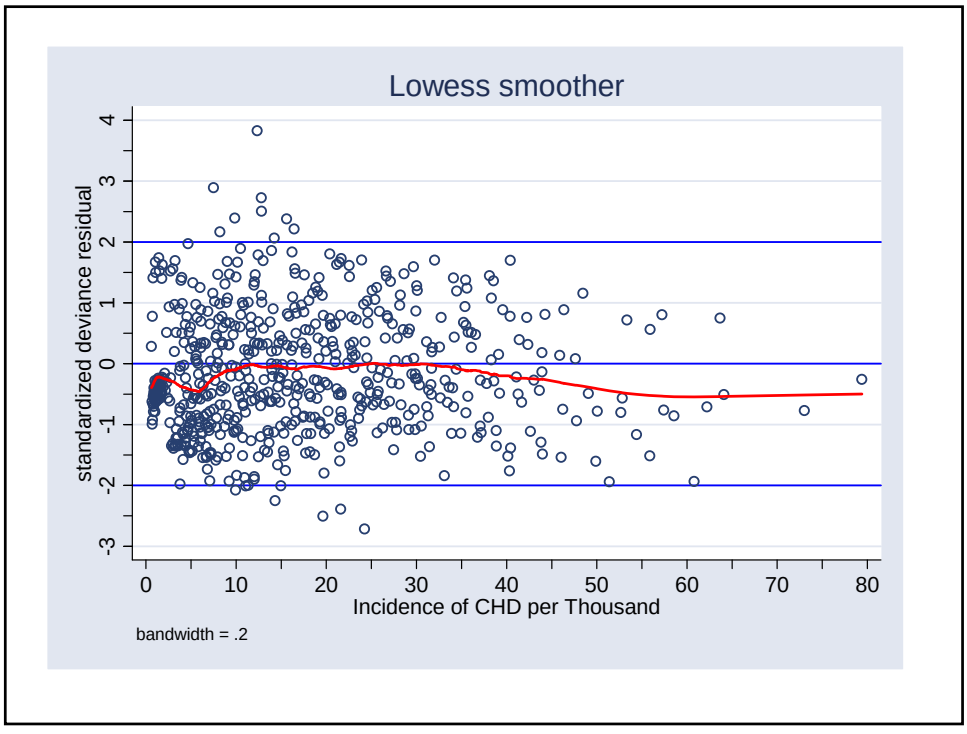



```
. generate e_rate = 1000*e_chd/pt_yrs
(82 missing values generated)

. label variable e_rate "Incidence of CHD per Thousand"

. *
. *  Draw scatterplot of the standardized deviance residual versus the
. *  estimated incidence of CHD.  Include lowess regression curve on this plot.
. *
. * Graphics > Smoothing and densities > Lowess smoothing
. lowess dev e_rate, bwidth(0.2) msymbol(Oh) ylabel(-3(1)4) ytick(-3(0.5)4) /// {7}
> lineopts(color(red) lwidth(medthick)) yline(-2 0 2 , lcolor(blue))        /// {8}
>     xlabel(0(10)80) xtick(5(10)75)
```

{7} Plot a **lowess** regression of the standardized deviance residual against the expected number of CHD events.

{8} This ***lineopts*** option specifies the color and thickness of the regression line.


lowess - Lowess smoothing
Main | if/in | Plot | Smoothed line | Add plots | Y axis | X axis | Titles | Legend | Overall | By
Dependent variable:
dev
Independent variable:
e_rate
Smoothing methods
Running-mean smooth
Suppress weighted regressions
Specify the bandwidth
0.2
Bandwidth
Transform dependent variable to logits
Adjust smoothed mean to equal mean of dependent variable
Suppress graph
Generate a new variable containing the smoothed values
New variable name:
Plot type: (scatterplot)
Default
Scatter
Line
Connected
Area
Bar
Spike
Dropline
Marker properties
Marker properties
Main | Advanced
Symbol: Hollow circle
Color: Default
Size:
Add labels to markers
Variable:
Label color: Default
Label size:
Label position: Default
Label angle:
Label gap:
Accept
Cancel
Color: Red
Width: Medium thick
Pattern: Default
Connecting method: Default


Lowess smoother
standardized deviance residual
4
3
2
1
0
-1
-2
-3
0
10
20
30
40
50
60
70
80
Incidence of CHD per Thousand
bandwidth = .2

The deviance residual plot indicates that the model fit is quite good, with most of the residuals lying between ± 2.

There is a suggestion of a negative drift for residuals associated with a large numbers of expected CDH events.

The standard deviation of these residuals may also be lower than those associated with low event rates.

6. What we have covered

- Generalization of Poisson regression model to include multiple covariates
  - Deriving relative risk estimates from Poisson regression models
- Analyzing a complex survival data set with Poisson regression
  - The *family(poisson)* and *link(log)* options of the *glm* command
  - The Framingham data set
  - Adjusting for confounding variables
  - Adding interaction terms
- Residual analysis
  - Deviance residuals
  - The *standardized deviance* option of the *predict* 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.

IX. FIXED EFFECTS ANALYSIS OF VARIANCE

- Regression analysis with categorical variables and one response measure per subject
- One-way analysis of variance
  - 95% confidence intervals for group means
  - 95% confidence intervals for the difference between group means
  - Testing for homogeneity of standard deviations across groups
- Multiple comparisons issues
  - Fisher's protected least significant difference approach
  - Bonferroni's multiple comparison adjustment
- Reformulating analysis of variance as a linear regression model
- Non-parametric one-way analysis of variance
  - Kruskal-Wallis test
  - Wilcoxon rank-sum test
- Two-Way Analysis of Variance
  - Simultaneously evaluating two categorical risk factors
- Analysis of Covariance
  - Analyzing models with both categorical and continuous covariates

1. Analysis of Variance

Traditionally, analysis of variance referred to regression analysis with categorical variables.

For example **one-way analysis of variance** involves comparing a continuous response variable in a number of groups defined by a single categorical variable.

In the middle of this century, great ingenuity was expended to devise specially balanced experimental designs that could be solved with an electric calculator.

Today, it is reasonable to consider analysis of variance as a special case of linear regression. In Stata the *xi:* prefix may be used with the *regress* command.

A critical assumption of these analyses is that the error terms for each observation are independent and have the same normal distribution. This assumption is often reasonable as long as we only have one response observation per patient.

These analyses assume that all parameters are attributes of the underlying population, and that we have obtained a representative sample of this population. These parameters measure attributes that are called **fixed-effects**.

In contrast, we often have multiple observations per patient. In this case some of the parameters measure attributes of the individual patients in the study. Such attributes are called **random effects**. A model that has both random and fixed effects is called a **mixed effects** model or a **repeated measures** model.

2. One-Way Analysis of Variance

Let n_i be the number of subjects in the i^{th} group

$n = \sum n_i$ be the total number of study subjects

y_{ij} be a continuous response variable on the j^{th} patient from the i^{th} group.

We assume for $i = 1,2,\ldots k;\ j = 1,2,\ldots,n_i$ that

$$y_{ij} = \beta_i + \varepsilon_{ij} \qquad \{9.1\}$$

where

$\beta_1,\beta_2,\ldots\beta_k$ are unknown parameters, and

ε_{ij} are mutually independent, normally distributed error terms with mean 0 and standard deviation σ.

Under this model, the expected value of y_{ij} is $\mathrm{E}\left[y_{ij} \mid i \right] = \beta_i$

Models like {9.1} are called **fixed-effects** models because the parameters $\beta_1,\beta_2,\ldots\beta_k$ are fixed constants that are attributes of the underlying population.

The response y_{ij} differs from β_i only because of the error term ε_{ij}. Let

$b_1,b_2,\ldots b_k$ be the least squares estimates of $\beta_1,\beta_2,\ldots\beta_k$, respectively,

$\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ be the sample mean for the i^{th} group,

and

$s^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2/(n-k)$ be the mean squared error (MSE) estimate of σ^2 {9.2}

We estimate σ by s, which is called the root MSE. It can be shown that $b_i = \bar{y}_i$, $E[b_i] = \beta_i$, and $E[s^2] = \sigma^2$. A 95% confidence interval for β_i is given by $\bar{y}_i \pm t_{n-k,0.025}\left(s/\sqrt{n_i}\right)$ {9.3}

Note that model {9.1} assumes that the standard deviation of ε_{ij} is the same for all groups. If it appears that there is appreciable variation in this standard deviation among groups then the 95% confidence interval for β_i should be estimated by

$$\bar{y}_i \pm t_{n_i-1,0.025}\left(s_i/\sqrt{n_i}\right) \quad \{9.4\}$$

where s_i is the sample standard deviation of y_{ij} within the i_{th} group.

We wish to test the null hypothesis that the expected response is the same in all groups. That is, we wish to test whether

$$\beta_1 = \beta_2 = \ldots = \beta_k \quad \{9.5\}$$

We can calculate a statistic that has a ***F* distribution** with k-1 and n-k degrees of freedom when this null hypothesis is true.

We reject the null hypothesis in favor of a multi-sided alternative hypothesis when the F statistic is sufficiently large.

The P value associated with this test is the probability that this statistic exceeds the observed value when this null hypothesis is true.

When there are just two groups, the F statistic will have 1 and $n-2$ degrees of freedom. In this case, the one-way analysis of variance is equivalent to an independent t test.

The square root of this F statistic equals the absolute value of the t statistic with $n-2$ degrees of freedom.

A test due to Levene (1960) can be performed to test the assumption that the standard deviation of ε_{ij} is constant within each group. If this test is significant, or if there is considerable variation in the values of s_i, then you should use equation {9.4} rather than equation {9.3} to calculate confidence intervals for the group means.

$$\bar{y}_i \pm t_{n-k,0.025}\left(s/\sqrt{n_i}\right) \qquad \{9.3\}$$

$$\bar{y}_i \pm t_{n_i-1,0.025}\left(s_i/\sqrt{n_i}\right) \qquad \{9.4\}$$

3. Multiple Comparisons

If, the analysis of variance F statistic is significant and the number of groups is not too large, we can make pair-wise comparisons of the different groups.

If the standard deviations within the k groups appears similar we can increase the power of the test that $\beta_i = \beta_j$ by using the formula

$$t_{n-k} = \left(\bar{y}_i - \bar{y}_j\right) / \left(s\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}\right) \qquad \{9.6\}$$

where s is the root MSE estimate of σ obtained from the analysis of variance.

Under the null hypothesis that $\beta_i = \beta_j$ equation {9.6} will have a t distribution with $n-k$ degrees of freedom.

This test is more powerful then the independent t test but is less robust.

A 95% confidence interval for the difference in population means between groups i and j is

$$\bar{y}_i - \bar{y}_j \pm t_{n-k,0.025}\left(s\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}\right) \qquad \{9.7\}$$

Alternately, a confidence interval based on the independent t test may be used if it appears unreasonable to assume a uniform standard deviation in all groups

$$\bar{y}_i - \bar{y}_j \pm t_{n_i+n_j-2,0.025}\left(s_p\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}\right) \qquad \{9.8\}$$

If the F test is not significant you should not report pair-wise significant differences unless they remain significant after a **Bonferroni multiple comparisons adjustment** (multiplying the P value by the number of pair wise tests.

If the number of groups is large and there is no natural ordering of the groups then a multiple comparisons adjustment may be advisable even if the F test is significant.

4. Fisher's Protected Least Significant Difference (LSD) Approach to Multiple Comparisons

The idea of only analyzing subgroup effects (e.g. differences in group means) when the main effects (e.g. F test) are significant is known as known as **Fisher's Protected Least Significant Difference (LSD) Approach to Multiple Comparisons.**

The F statistic tests the hypothesis that all of the group response means are simultaneously equal.

If we can reject this hypothesis it follows that some of the means must be different.

Fisher argued that in this situation you should be able to investigate which ones are different without having to pay a multiple comparisons penalty.

This approach is not guaranteed to preserve the experiment-wide Type I error probability, but makes sense in well structured experiments where the number of groups being examined is not too large.

5. Reformulating Analysis of Variance as a Linear Regression Model

A one-way analysis of variance is, in fact, a special case of the multiple regression model. Let

y_h denote the response from the h^{th} study subject, $h = 1,2,\ldots n$, and let

$$x_{hi} = \begin{cases} 1: \text{ if the } h^{th} \text{ patient is in the } i^{th} \text{ group} \\ 0: \text{ otherwise} \end{cases}$$

Then model (9.1) can be rewritten

$$y_h = \alpha + \beta_2 x_{h2} + \beta_3 x_{h3} + \ldots + \beta_k x_{hk} + \varepsilon_h \qquad \{9.9\}$$

where ε_h are mutually independent, normally distributed error terms with mean 0 and standard deviation σ. Note that model {9.9} is a special case of model (3.1). Thus, this analysis of variance is also a regression analysis in which all of the covariates are zero-one indicator variables.

Also,

$$\mathrm{E}\left[y_h \mid x_{h2}, x_{h3}, \cdots, x_{hk}\right] = \begin{cases} \alpha & \text{if the } h^{th} \text{ patient is from group 1} \\ \alpha + \beta_i & \text{if the } h^{th} \text{ patient is from group } i > 1 \end{cases}$$

Thus, α is the expected response of patients in the first group and β_i is the expected difference in the response of patients in the i_{th} and first groups.

The least squares estimates of α and β_i are $\bar{y}_1$ and $\bar{y}_i - \bar{y}_1$, respectively.

We can use any multiple linear regression program to perform a one-way analysis of variance, although most software packages have a separate procedure for this task.

6. Non-parametric Methods

a) Kruskal-Wallis Test

The **Kruskal-Wallis** test is the non-parametric analog of the one-way analysis of variance (Kruskal and Wallis 1952).

Model {9.1} assumes that the ε_{ij} terms are normally distributed and have the same standard deviation. If either of these assumptions is badly violated then the Kruskal-Wallis test should be used.

Suppose that patients are divided into k groups as in model {9.1} and that y_{ij} is a continuous response variable on the j^{th} patient from the i^{th} group.

The null hypothesis of this test is that the distributions of the response variables are the same in each group.

Let

n_i be the number of subjects in the i^{th} group,

$n = \sum n_i$ be the total number of study subjects.

We rank the values of y_{ij} from lowest to highest and let R_i be the sum of the ranks for the patients from the i^{th} group.

If all of the values of y_{ij} are distinct (no ties) then the Kruskal-Wallis test statistic is

$$H = \frac{12}{n(n+1)}\left(\sum \frac{R_i^2}{n_i}\right) - 3(n+1) \qquad \{9.10\}$$

When there are ties a slightly more complicated formula is used (see Steel and Torrie 1980).

Under the null hypothesis, H will have a chi-squared distribution with $k - 1$ degrees of freedom as long as the number of patients in each group is reasonably large.

Note that the value of H will be the same for any two data sets in which the data values have the same ranks. Increasing the largest observation or decreasing the smallest observation will have no effect on H. Hence, extreme outliers will not unduly affect this test.

The non-parametric analog of the independent t-test is the **Wilcoxon-Mann-Whitney rank-sum test**. This rank-sum test and the Kruskal-Wallis test are equivalent when there are only two groups of patients.

7. Example: A Polymorphism in the Estrogen Receptor Gene

The human estrogen receptor gene contains a two-allele restriction fragment length polymorphism that can be detected by Southern blots of DNA digested with the PuvII restriction endonuclease. Bands at 1.6 kb and/or 0.7 kb identify the genotype for these alleles.

Parl et al. (1989) studied the relationship between this genotype and age of diagnosis among 59 breast cancer patients.

Table 9.1

	Genotype*			Total
	1.6/1.6	**1.6/0.7**	**0.7/0.7**	
Number of Patients	14	29	16	59
Age at breast cancer diagnosis				
Mean	64.643	64.379	50.375	60.644
Standard Deviation	11.18	13.26	10.64	13.49
95% Confidence Interval				
Equation {9.3} Pooled SD estimate	(58.1 – 71.1)	(59.9 – 68.9)	(44.3 – 56.5)	
Equation {9.4} Separate SD estimates	(58.2 – 71.1)	(59.3 – 69.4)	(44.7 – 56.0)	(57.1 – 64.2)

To test the null hypothesis that the age at diagnosis does not vary with genotype, we perform a one-way analysis of variance on the ages of patients in these three groups using model {9.1}.

In this analysis, $n = 59$, $k = 3$ and β_1, β_2 and β_3 represent the expected age of breast cancer diagnosis among patients with the 1.6/1.6, 1.6/0.7, and 0.7/0.7 genotypes, respectively.

The estimates of these parameters are the average ages given in the preceding table.

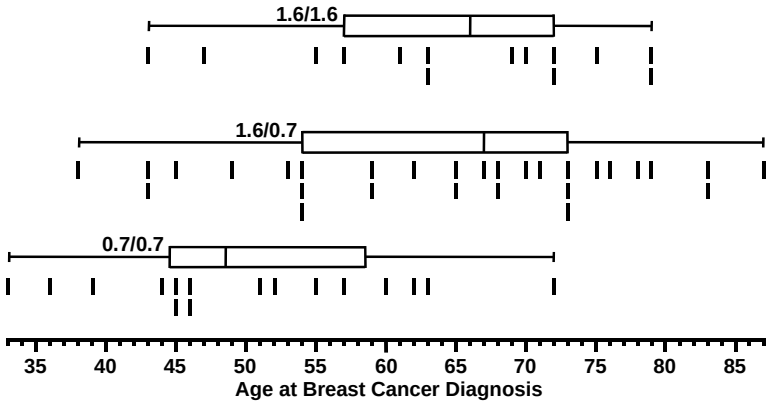
The P value form the F statistic equals 0.001.

Table 9.2

Comparison	Difference in Mean Age of Diagnosis	95% Confidence Interval	*P* Value	
			Eq. {0.7}*	Rank-sum**
1.6/0.7 vs. 1.6/1.6	-0.264	(-8.17 to 7.65)	0.95	0.96
0.7/0. 7 vs. 1.6/1.6	-14.268	(-23.2 to -5.37)	0.002	0.003
0.7/0. 7 vs. 1.6/0.7	-14.004	(-21.6 to -6.43)	< 0.0005	0.002

* Equation 7 uses the pooled estimate of s

** Wilcoxon-Mann-Whitney rank-sum test



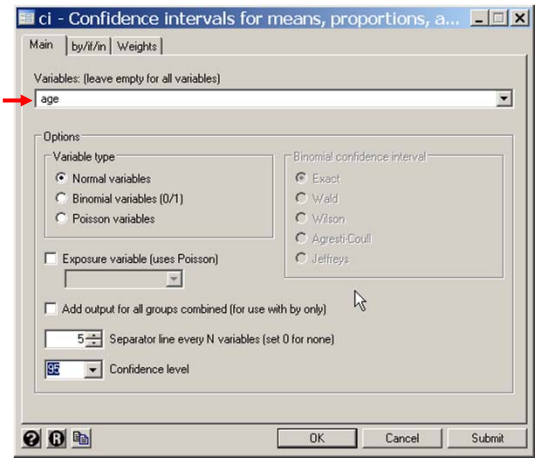
8. One-Way Analyses of Variance using Stata

The following Stata log file and comments illustrate how to perform the one-way analysis of variance discussed in the preceding section.

```
. * 10.8.ERpolymorphism.log
. *
. *  Do a one-way analysis of variance to determine whether age
. *  at breast cancer diagnosis varies with estrogen receptor (ER)
. *  genotype using the data of Parl et al. (1989).
. *
. use C:\WDDtext\10.8.ERpolymorphism.dta                                {1}
. * Statistics > Summaries, tables, ... > Summary ... > Confidence intervals
. ci age                                                                {2}

    Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+--------------------------------------------------------------
         age |      59     60.64407    1.756804        57.12744    64.16069
```

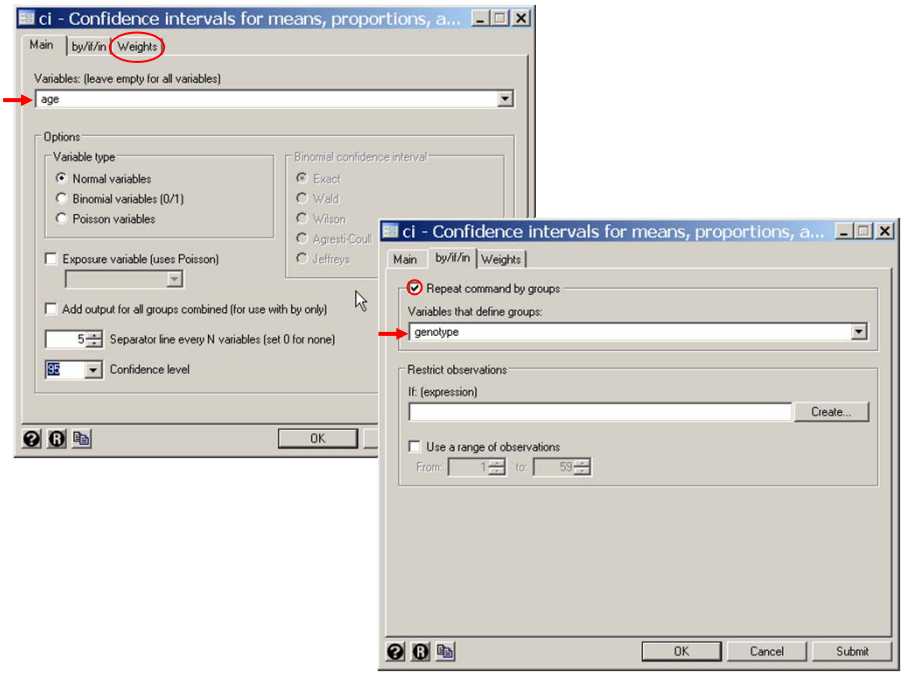
{1} This data set contains the **age of diagnosis** and **estrogen receptor genotype** of the 59 breast cancer patients studied by Parl et al. (1989). The **genotypes 1.6/1.6, 1.6/0.7 and 0.7/0.7** are coded 1, 2 and 3 in the variable *genotype,* respectively.

{2} This ***ci*** command calculates the mean age of diagnosis (*age*) together with the associated **95% confidence interval**. This confidence interval is calculated using equation {9.4}. The estimated **standard error of the mean** and the **number of patients** with non-missing ages is also given.

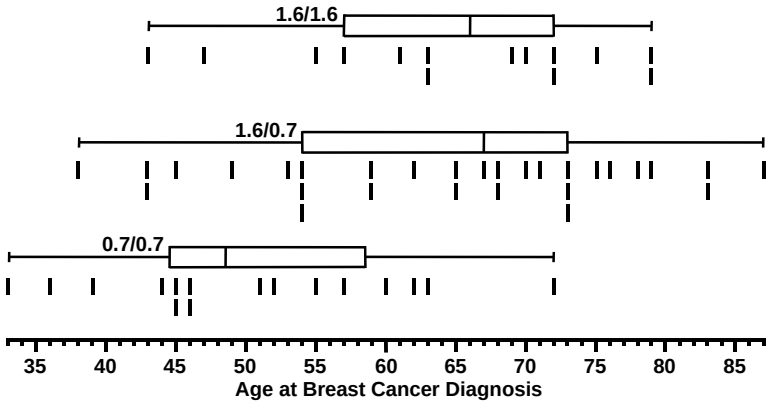



```
. * Statistics > Summaries, tables, ... > Summary ... > Confidence intervals
. by genotype: ci age                                                    {3}

-------------------------------------------------------------------------------
-> genotype = 1.6/1.6

    Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+------------------------------------------------------------------
         age |      14     64.64286    2.988269         58.1871    71.09862
-------------------------------------------------------------------------------
-> genotype = 1.6/0.7

    Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+------------------------------------------------------------------
         age |      29     64.37931    2.462234        59.33565    69.42297
-------------------------------------------------------------------------------
-> genotype = 0.7/0.7

    Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+------------------------------------------------------------------
         age |      16       50.375    2.659691          44.706      56.044
```

{3} The command prefix ***by genotype:*** specifies that means and 95% **confidence intervals** are to be calculated for **each** of the three **genotypes**.




```
. *
. *  The following graph type is not available in Stata version 8.0
. *
. graph7  age, by(genotype) box oneway                                  {4}
```



{4} The *graph7* command implements Stata Version 7 commands using version 7 syntax. The following graph is one that is not available in Version 8. The ***box*** and ***oneway*** options of this *graph* command create a graph that is similar to the Figure. See also Sections 10.7 and 10.8 of text for a prettier way of drawing this graph.

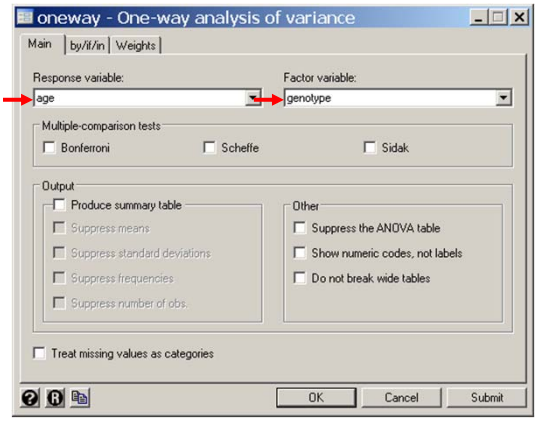
```
. * Statistics > Linear models and related > ANOVA/MANOVA > One-way ANOVA
. oneway age genotype                                                    {5}

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      2315.73355      2   1157.86678      7.86     0.0010    {6}
 Within groups      8245.79187     56   147.246283                         {7}
------------------------------------------------------------------------
    Total           10561.5254     58   182.095266

Bartlett's test for equal variances:  chi2(2) =   1.0798  Prob>chi2 =  0.583 {8}
```

{5} This ***oneway*** command performs a **one-way analysis of variance** of *age* with respect to the three distinct values of ***genotype***.

{6} The ***F*** statistic from this analysis equals **7.86**. If the mean age of diagnosis in the target population is the same for all three genotypes, this statistic will have an ***F*** distribution with $k - 1 = 3 - 1 = 2$ and $n - k = \mathbf{56}$ degrees of freedom. The probability that this statistic exceeds 7.86 is **0.001.**

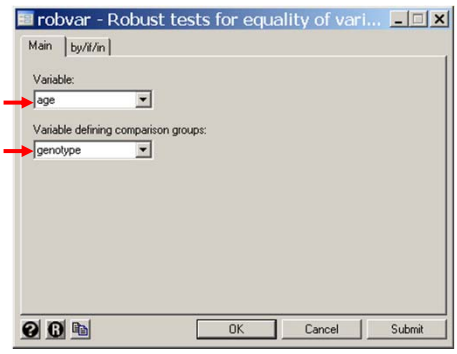
{7} The **MSE** estimate of is $= \mathbf{147.246}$.

{8} **Bartlett's test** for **equal variances** (i.e. equal standard deviations) gives a *P* value of **0.58**.



```
. *
. *  Test whether the standard deviations of age are equal in
. *  patients with different genotypes.
. *
. * Statistics > Summaries, ... > Classical ... > Robust equal variance test
. robvar age, by(genotype)


           |     Summary of Age at Diagnosis
  Genotype |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
   1.6/1.6 |   64.642857   11.181077          14
   1.6/0.7 |    64.37931   13.259535          29
   0.7/0.7 |      50.375   10.638766          16
-----------+------------------------------------
     Total |   60.644068   13.494268          59

W0  =  0.83032671   df(2, 56)     Pr > F = 0.44120161

W50 =  0.60460508   df(2, 56)     Pr > F = 0.54981692

W10 =  0.79381598   df(2, 56)     Pr > F = 0.45713722
```

This ***robvar*** command performs a test of the equality of variance among groups defined by ***genotype*** using methods of Levene (1960) and Brown and Forsythe (1974). These tests are less sensitive to departures from normality than Bartlett's test. There is no evidence of heterogeneity of variance for age in these three groups.

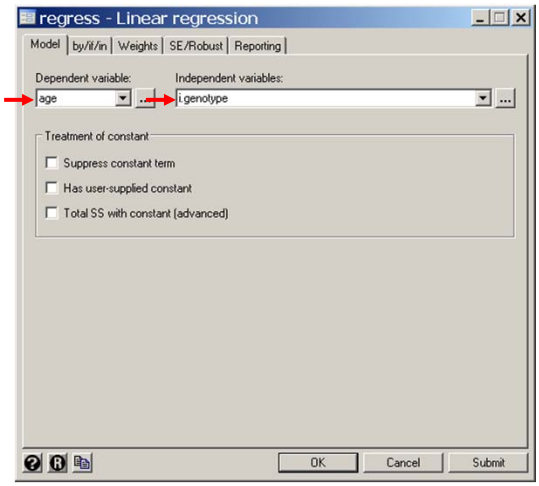


```
. *
. *  Repeat analysis using linear regression
. *
. *  Statistics > Linear models and related > Linear regression
. regress age i.genotype                                              {9}

      Source |       SS       df       MS              Number of obs =      59
-------------+------------------------------           F(  2,    56) =    7.86
       Model |  2315.73355     2  1157.86678            Prob > F      =  0.0010
    Residual |  8245.79187    56  147.246283           R-squared     =  0.2193
-------------+------------------------------           Adj R-squared =  0.1914
       Total |  10561.5254    58  182.095266           Root MSE      =  12.135

------------------------------------------------------------------------------
         age |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
genotype     |
          2  |   -.2635468   3.949057    -0.07   0.947    -8.174458    7.647365 {10}
          3  |   -14.26786   4.440775    -3.21   0.002     -23.1638   -5.371915
             |
       _cons |    64.64286   3.243084    19.93   0.000     58.14618    71.13953 {11}
------------------------------------------------------------------------------
. oneway age genotype

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      2315.73355      2   1157.86678      7.86     0.0010
Within groups       8245.79187     56   147.246283
------------------------------------------------------------------------
    Total           10561.5254     58   182.095266
```

{9} This ***regress*** command preforms exactly the same one-way analysis of variance as the ***oneway*** command given above. Note that the F statistic, the P value for this statistic and the MSE estimate of are identical to that given by the ***oneway*** command. The **syntax** of the ***xi:*** prefix is explained in Section **5.10**. The model used by this command is equation {9.9} with $k = 3$.

{10} The estimates of β_2 and β_3 in this example are $\bar{y}_2 - \bar{y}_1 = 64.379 - 64.643 = -0.264$ and $\bar{y}_3 - \bar{y}_1 = 50.375 - 64.643 = -14.268$, respectively. They are highlighted in the column labeled ***Coef.*** The 95% confidence intervals for β_2 and β_3 are calculated using equation {9.7}. The ***t*** statistics for testing the null hypotheses that $\beta_2 = 0$ and $\beta_3 = 0$ are -0.07 and -3.21, respectively. They are calculated using equation {9.6}. The highlighted values in this output are also given in **Table 9.2**.

{11} The estimate of α is $\bar{y}_1 = 64.643$. The 95% confidence interval for α is calculated using equation {9.3}. These statistics are also given in Table 10.1.




```
. lincom _cons + _Igenotype_2                                          {12}

 ( 1)  _Igenotype_2 + _cons = 0.0

------------------------------------------------------------------------------
       age |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
       (1) |   64.37931   2.253322    28.57   0.000     59.86536    68.89326 {13}
------------------------------------------------------------------------------

. lincom _cons + _Igenotype_3

 ( 1)  _Igenotype_3 + _cons = 0.0

------------------------------------------------------------------------------
       age |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
       (1) |     50.375   3.033627    16.61   0.000     44.29791    56.45209
------------------------------------------------------------------------------
```

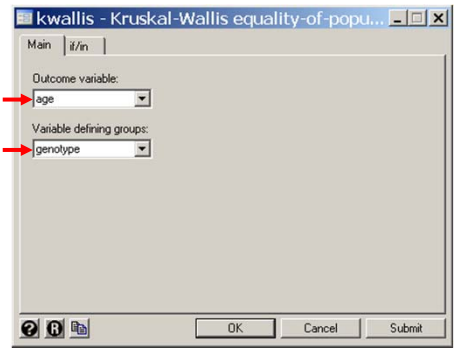
{12} This ***lincom*** command estimates $\alpha + \beta_2$ by $\hat{\alpha} + \hat{\beta}_2 = \bar{y}_2$. A 95 % confidence interval for this estimate is also given. Note that $\alpha + \beta_2$ equals the population mean age of diagnosis among women with the 1.6/0.7 genotype. Output from this and the next ***lincom*** command are also given in Table 9.1.

{13} This confidence interval is calculated using equation {9.3}.

```
. lincom 3.genotype - 2.genotype                                              {14}

 ( 1) - 2.genotype + 3.genotype = 0.0

------------------------------------------------------------------------------
         age |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -14.00431   3.778935    -3.71   0.000    -21.57443   -6.434194
------------------------------------------------------------------------------
. *
. *  Perform a Kruskal-Wallis analysis of variance
. *
. * Statistics > Nonparametric... > Tests of hypotheses > Kruskal-Wallis...
. kwallis age, by(genotype)                                                   {15}

Test: Equality of populations (Kruskal-Wallis test)

  +---------------------------+
  | genotype | Obs | Rank Sum |
  |----------+-----+----------|
  |  1.6/1.6 |  14 |   494.00 |
  |  1.6/0.7 |  29 |   999.50 |
  |  0.7/0.7 |  16 |   276.50 |
  +---------------------------+

chi-squared =    12.060 with 2 d.f.
probability =     0.0024

chi-squared with ties =    12.073 with 2 d.f.
probability =     0.0024
```

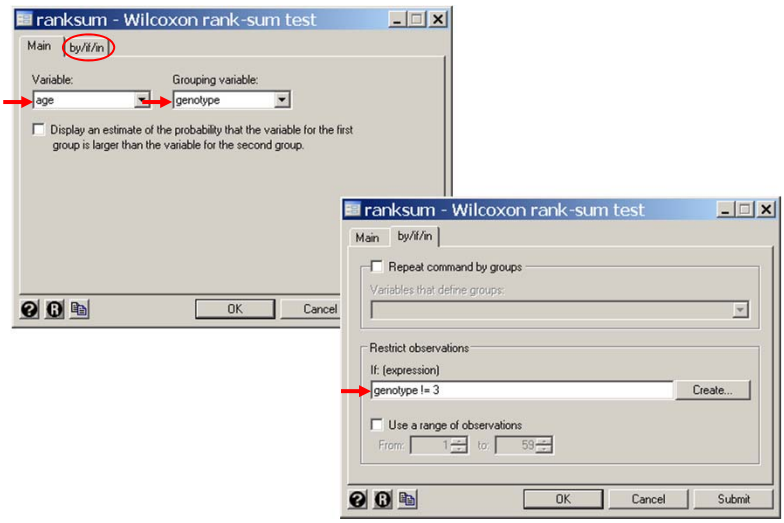
{14} This command estimates $\beta_3 - \beta_2$ by $\hat{\beta}_3 - \hat{\beta}_2 = \bar{y}_3 - \bar{y}_2 = 50.375 - 64.379 = -14.004$. The null hypothesis that $\beta_3 = \beta_2$ is the same as the hypothesis that the mean age of diagnosis in groups 2 and 3 are equal. The **confidence interval** for $\beta_3 - \beta_2$ is calculated using equation {9.7}. The highlighted values are also given in Table 9.2.

{15} This ***kwallis*** command performs a **Kruskal-Wallis** test of ***age*** by ***genotype***. The test statistic, adjusted for ties, equals 12.073. The associated *P* value equal 0.0024.




```
. * Statistics > Nonparametric... > Tests... > Wilcoxon rank-sum test
. ranksum age if genotype !=3, by(genotype)                          {16}

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

    genotype |      obs    rank sum    expected
-------------+---------------------------------
     1.6/1.6 |       14         310         308
     1.6/0.7 |       29         636         638
-------------+---------------------------------
    combined |       43         946         946

unadjusted variance     1488.67
adjustment for ties       -2.70
                        ----------
adjusted variance       1485.97

Ho: age(genotype==1.6/1.6) = age(genotype==1.6/0.7)
             z =   0.052
    Prob > |z| =   0.9586
```

{16} This command performs a **Wilcoxon-Mann-Whitney rank-sum** test on the age of diagnosis of women with the 1.6/1.6 genotype versus the 1.6/0.7 genotype. The *P* value for this test is **0.96**. The next two commands perform the other two pair-wise comparisons of age by genotype using this rank-sum test. The highlighted *P* values are included in **Table 10.2**.




```
. * Statistics > Nonparametric... > Tests... > Wilcoxon rank-sum test
. ranksum age if genotype ~=2, by(genotype)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

    genotype |      obs    rank sum    expected
-------------+---------------------------------
     1.6/1.6 |       14         289         217
     0.7/0.7 |       16         176         248
-------------+---------------------------------
    combined |       30         465         465

unadjusted variance      578.67
adjustment for ties       -1.67
                        ----------
adjusted variance        576.99

Ho: age(genotype==1.6/1.6) = age(genotype==0.7/0.7)
             z =   2.997
    Prob > |z| =   0.0027
```

```
. * Statistics > Nonparametric... > Tests... > Wilcoxon rank-sum test
. ranksum age if genotype ~=1, by(genotype)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

    genotype |      obs    rank sum    expected
-------------+---------------------------------
     1.6/0.7 |       29       798.5         667
     0.7/0.7 |       16       236.5         368
-------------+---------------------------------
    combined |       45        1035        1035

unadjusted variance     1778.67
adjustment for ties       -2.23
                        ----------
adjusted variance       1776.44

Ho: age(genotype==1.6/0.7) = age(genotype==0.7/0.7)
                z =   3.120
    Prob > |z| =   0.0018
```

```
. * Statistics > Nonparametric... > Tests of hypotheses > Kruskal-Wallis...
. kwallis age if genotype ~=1, by(genotype)                          {17}

Test: Equality of populations (Kruskal-Wallis test)

  +---------------------------+
  | genotype | Obs | Rank Sum |
  |----------+-----+----------|
  |  1.6/0.7 |  29 |   798.50 |
  |  0.7/0.7 |  16 |   236.50 |
  +---------------------------+

chi-squared =     9.722 with 1 d.f.
probability =     0.0018

chi-squared with ties =     9.734 with 1 d.f.
probability =     0.0018
```

{17} This command repeats the preceding command using the **Kruskal-Wallis test**. This test is equivalent to the rank-sum test when only two groups are being compared. Note that the *P* values from these tests both equal **0.0018**.

9. Two-Way Analysis of Variance, Analysis of Covariance, and Other Models

Fixed-effects analyses of variance generalize to a wide variety of complex models. For example, suppose that hypertensive patients were treated with either a placebo, a diuretic alone, a beta-blocker alone, or with both a diuretic and a beta-blocker. Then a model of the effect of treatment on diastolic blood pressure (DBP) might be

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \qquad \{9.11\}$$

where

α , β_1 and β_2 are unknown parameters,

$x_{i1} = \begin{cases} 1: & i^{th} \text{ patient is on a diuretic} \\ 0: & \text{otherwise} \end{cases}$

$x_{i2} = \begin{cases} 1: & i^{th} \text{ patient is on a beta-blocker} \\ 0: & \text{otherwise} \end{cases}$

y_i is the DBP of the ith patient after some standard interval therapy , and

ε_i are error terms that are independently and normally distributed with mean zero and standard deviation σ

Model {9.11} is an example of a fixed-effects, **two-way analysis of variance.**

It is called two-way because each patient is simultaneously influenced by two covariates — in this case whether she did, or did not, receive a diuretic or a beta-blocker.

A critical feature of this model is that each patient's blood pressure is only observed once.

It is this feature that makes the independence assumption for the error term reasonable and makes this a fixed-effects model. In this model,

α is the mean DBP of patients on placebo,

$\alpha + \beta_1$ is the mean DBP of patients on the diuretic alone,

$\alpha + \beta_2$ is the mean DBP of patients on the beta-blocker alone, and

$\alpha + \beta_1 + \beta_2$ is the mean DBP of patients on both treatments.

The model is additive since it assumes that the mean DBP of patients on both drugs is $\alpha + \beta_1 + \beta_2$.

If this assumption is unreasonable, we can add an interaction term as in Section 3.12.

10. Fixed Effects Analysis of Covariance

This refers to linear regression models with both categorical and continuous covariates. Inference from these models is called **analysis of covariance.**

For example, we could add the patient's age to model (9.11). This gives

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 \times age_i + \varepsilon_i \qquad \{9.12\}$$

where age_i is the i^{th} patient's age, β_3 is the parameter associated with age, and the other terms are as defined in model {9.11}. The analysis of model {9.12} would be an example of analysis of covariance.

These models no longer need the special consideration that they received in years passed and can be easily handled by the *regress* command.

11. What we have covered

- Regression analysis with categorical variables and one response measure per subject
- One-way analysis of variance: The *oneway* command
  - 95% confidence intervals for group means
  - 95% confidence intervals for the difference between group means
  - Testing for homogeneity of standard deviations across groups
    The *robvar* command
- Multiple comparisons issues
  - Fisher's protected least significant difference approach
  - Bonferroni's multiple comparison adjustment
- Reformulating analysis of variance as a linear regression model
- Non-parametric one-way analysis of variance
  - Kruskal-Wallis test: The *kwallis* command
  - Wilcoxon rank-sum test: The *ranksum* command
- Two-Way Analysis of Variance
  - Simultaneously evaluating two categorical risk factors
- Analysis of Covariance
  - Analyzing models with both categorical and continuous covariates

Cited Reference

Parl FF, Cavener DR, Dupont WD. Genomic DNA analysis of the estrogen receptor gene in breast cancer. *Breast Cancer Research and Treatment* 1989;14:57-64.

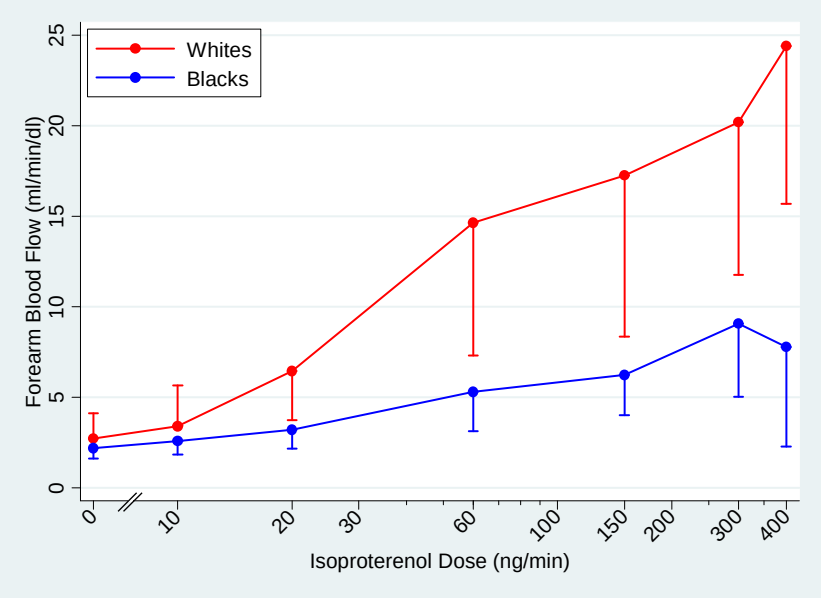
For additional references on these notes see.

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.

X. Mixed Effects Analysis of Variance

- Analysis of variance with multiple observations per patient
  - These analyses are complicated by the fact that multiple observations on the same patient are correlated with each other
- Response-feature approach to mixed effects analysis of variance
  - Reduce multiple response measures on each patient to a single statistic that captures the most biologically important aspect of the response
  - Perform a fixed effects analysis on this response feature
  - Using a regression slope as a response feature
  - Using an area under the curve as a response feature
- Generalized estimating equations (GEE) approach to mixed effects analysis of variance
  - GEE analysis with logistic or Poisson models

Lang et al. (1995) studied the effect of isoproterenol, a β-adrenergic agonist, on forearm blood flow in a group of 22 normotensive men. Nine of the study subjects were black and 13 were white. Each subject's blood flow was measured at baseline and then at escalating doses of isoproterenol.



The vertical bars indicate estimated standard deviations within each race-dose group.

There are a number of difficulties with analyzing these data.

1. Responses from the same patient are likely to be correlated. If Mr. Smith's response is 2 standard deviations above the mean day-2-treatment response on day 2, it is unlikely that he will be below the mean day-3-treatment response on day 3.

2. There is likely to be inherent variability between patients in how they respond to therapy that must be accounted for in our analysis.

3. We observe 22×7 = 154 responses. However, these observations only come from 22 patients. If we wish to make inferences about patients in general, our effective sample size is 22 rather than 154.

A common error in analyzing data like these is to use a fixed effects model. For example, a model such as

regress response race##dose

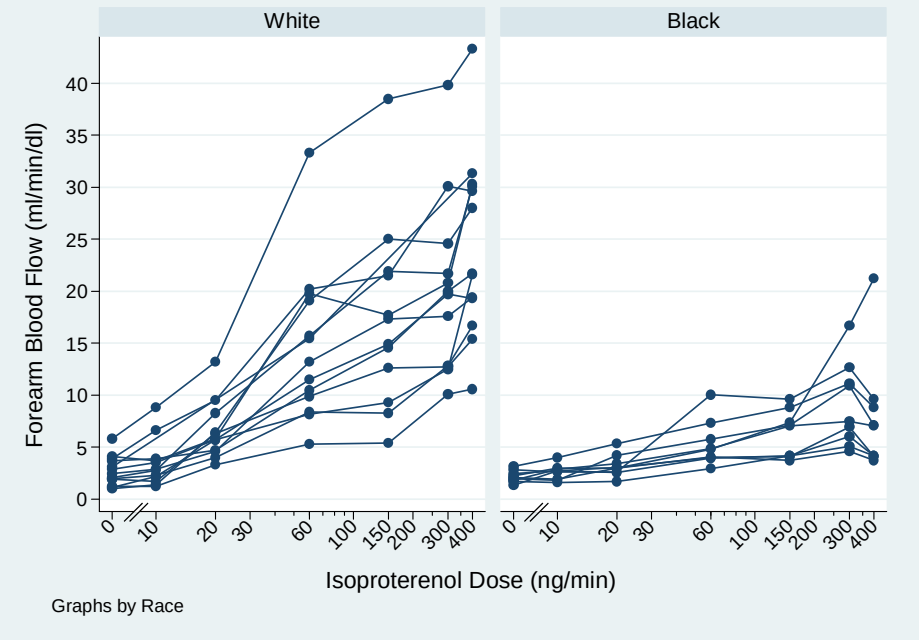
assumes that each patient's response equals
- a constant +
- an effect due to race +
- an effects due to dose +
- dose-race interaction effects +
- an independent error term.

The analysis is exactly the same as if we had had 154 distinct patients with each patient observed at a single-dose. This analysis will have 140 degrees of freedom and will seriously overestimate the significance of the dose-treatment effect.

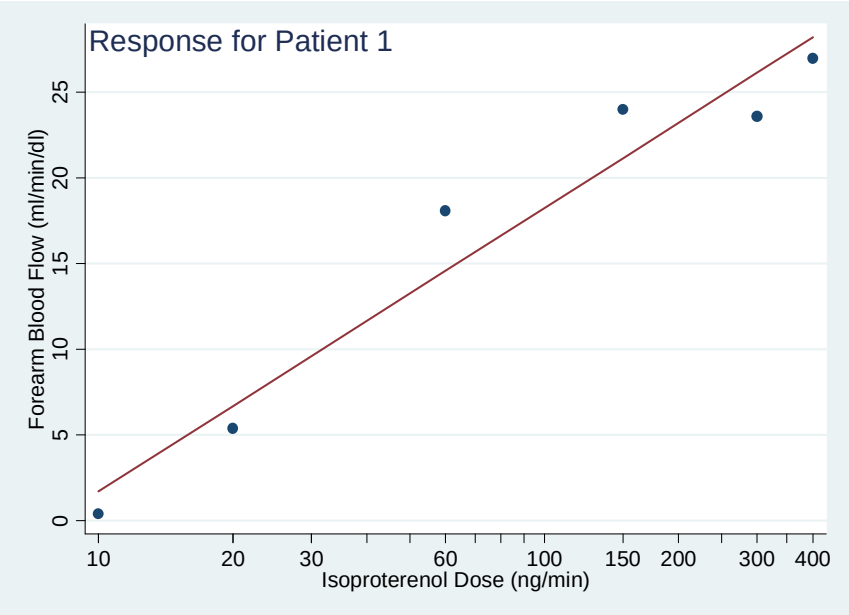
1. The Response-Feature Approach to Mixed Effects Analysis of Variance

The simplest valid way of analyzing mixed effects data is to compress each patient's response values into a single biologically sensible measure and then do an appropriate fixed effects analysis of the condensed response.

Consider the Isoproterenol-race data.

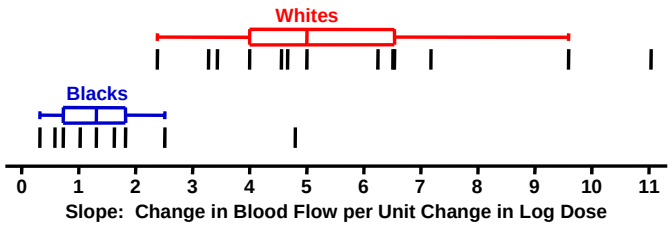


Note that there appears to be a log-linear relationship between dose and blood flow.



This suggests

1. Fit separate linear regression lines to the responses for each patient. This will give us 22 slope estimates - one for each patient.
2. Perform a Wilcoxon rank-sum test on these slopes to determine whether the slopes of black and white patients are different. It is prudent to use a non-parametric test because the individual patient slopes may have a non-normal distribution. However, you could also test these slopes with a *t*-test.



The Wilcoxon-Mann-Whiney rank sum test is significant with P=.0006.

Note that the responses between patients really are independent, so this analysis does not make any silly assumptions.

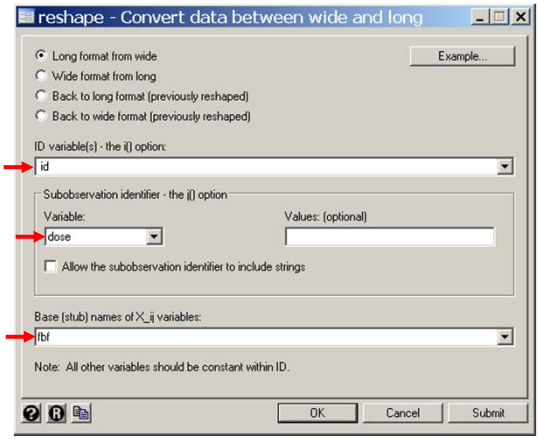
The same idea can be used in many other ways. The key idea is to compress the response data in a way that is biologically sensible. This may involve area under the curve, an average, or a weighted average.

2. Response Feature Analysis Using Stata

```
Exploratory Analysis of Repeated Measures Data Using Stata

. *  11.2.Isoproterenol.log  See Text p.364
. *
. *  Plot mean forearm blood flow by race and log dose of isoproterenol
. *  using the data of Lang et al. (1995).  Show standard deviation for
. *  each race at each drug level.
. *
. use C:\WDDtext\11.2.Isoproterenol.dta, clear
. * Statistics > Summaries... > Tables > Table of summary statistics (table).
. table race, row
 ----------------------
     Race |      Freq.
----------+-----------
    White |         13
    Black |          9
          |
    Total |         22
----------------------
. * Data > Describe data > List data
. list if id == 1| id == 22

     +-------------------------------------------------------------------------+
     | id    race   fbf0   fbf10   fbf20   fbf60   fbf150   fbf300   fbf400 |
     |-------------------------------------------------------------------------|
  1. |  1   White      1     1.4     6.4    19.1       25     24.6       28 |
 22. | 22   Black    2.1     1.9       3     4.8      7.4     16.7     21.2 |
     +-------------------------------------------------------------------------+
```

```
. generate baseline = fbf0

. *
. * Convert data from one record per patient to one record per observation.
. *
. * Data > Create > Other variable-trans... > Convert data between wide...
. reshape long fbf, i(id) j(dose)                                          {1}
(note: j = 0 10 20 60 150 300 400)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                       22   ->     154
Number of variables                  10   ->       5
j variable (7 values)                     ->   dose
xij variables:
                  fbf0 fbf10 ... fbf400   ->   fbf
-----------------------------------------------------------------------------
```

{1} The ***reshape long*** command converts data from one record per **patient** to one record per **observation**. In this command, *i*(*id*) specifies that the *id* variable identifies observations from the same subject. The variable ***fbf*** is the first three letters of variables ***fbf0***, ***fbf10***, …, ***fbf400***; ***j*(*dose*)** defines ***dose*** to be a new variable whose values are the **trailing digits** in the names of the variables ***fbf0***, ***fbf10***, …, ***fbf400***. That is, *dose* will take the values 0, 10, 20, …, 300, 400. One record will be created for each value of *fbf0*, *fbf10*, …,*fbf400*. **Other variables** in the file that are not included in this command (like *race* or *baseline*) are assumed not to vary with *dose* and are replicated in each record for each specific patient.

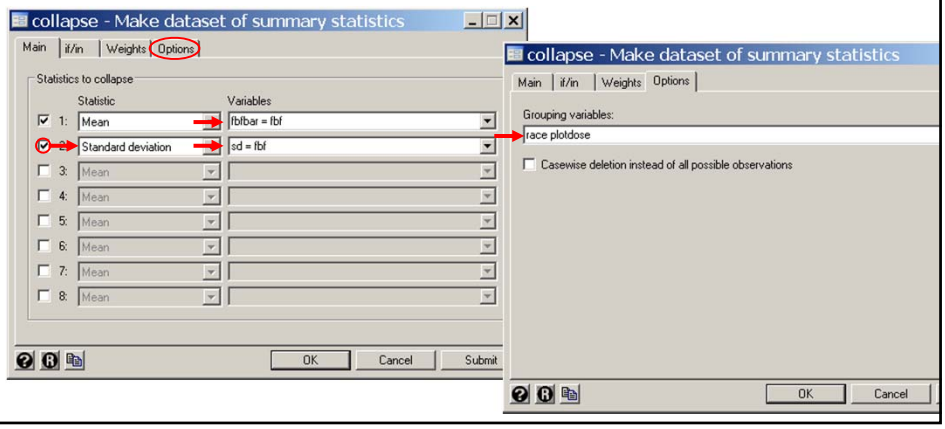



```
. * Data > Describe data > List data
. list if id == 1 | id == 22

     +-------------------------------------+
     | id   dose    race    fbf   baseline |
     |-------------------------------------|
  1. |  1      0   White      1          1 |
  2. |  1     10   White    1.4          1 |
  3. |  1     20   White    6.4          1 |
  4. |  1     60   White   19.1          1 |
  5. |  1    150   White     25          1 |
     |-------------------------------------|
  6. |  1    300   White   24.6          1 |
  7. |  1    400   White     28          1 |
148. | 22      0   Black    2.1        2.1 |
149. | 22     10   Black    1.9        2.1 |
150. | 22     20   Black      3        2.1 |
     |-------------------------------------|
151. | 22     60   Black    4.8        2.1 |
152. | 22    150   Black    7.4        2.1 |
153. | 22    300   Black   16.7        2.1 |
154. | 22    400   Black   21.2        2.1 |
     +-------------------------------------+
```

```
. generate delta_fbf = fbf - baseline
(4 missing values generated)

. label variable delta_fbf "Change in Forearm Blood Flow"

. label variable dose "Isoproterenol Dose (ng/min)"

. generate plotdose = dose

. replace plotdose = 6    if dose == 0                                {2}
(22 real changes made)

. label variable plotdose "Isoproterenol Dose (ng/min)"

. generate logdose = log(dose)
(22 missing values generated)

. label variable logdose "Log Isoproterenol Dose"
```

{2} We want to create Figures 10.1 and 10.2 that plot dose on a logarithmic scale. We also want to include the baseline dose of zero on these figures. Since the logarithm of zero is undefined, we create a new variable called *plotdose* that equals *dose* for all values greater than zero and equals 6 when *dose* = 0. We will use a graphics editor to relabel this value zero with a break in the *x*-axis when we create these figures.

```
. *
. *  Save long format of data for subsequent analyses
. *
. save C:\WDDtext\11.2.Long.Isoproterenol.dta, replace
file C:\WDDtext\11.2.Long.Isoproterenol.dta saved

. *
. *  Generate Figure 11.1
. *
. * Data > Create... > Other variable-trans... > Make dataset of means...
. collapse (mean) fbfbar = fbf (sd) sd = fbf, by(race plotdose)
```



```
. generate blackfbf = .
(14 missing values generated)

. generate whitefbf = .
(14 missing values generated)

. generate whitesd = .
(14 missing values generated)

. generate blacksd = .
(14 missing values generated)

. replace whitefbf = fbfbar if race == 1                                {3}
(7 real changes made)

. replace blackfbf = fbfbar if race == 2
(7 real changes made)
```

{3} The variable ***whitefbf*** equals the mean forearm blood flow for **white** subjects and is missing for black subjects; ***blackfbf*** is similarly defined for **black** subjects. The variables ***blacksd*** and ***whitesd*** give the standard deviations for black and white subjects, respectively.

```
. replace blacksd = sd if race == 2
(7 real changes made)

. replace whitesd = sd if race == 1
(7 real changes made)

. label variable whitefbf "Forearm Blood Flow (ml/min/dl)"

. label variable blackfbf "Forearm Blood Flow (ml/min/dl)"

. generate wsdbar = whitefbf - whitesd                                  {4}
(7 missing values generated)

. generate bsdbar = blackfbf - blacksd
(7 missing values generated)

. replace wsdbar = whitefbf + whitesd if plotdose < 20                  {5}
(2 real changes made)

. twoway connected whitefbf plotdose, color(red)                      /// {6}
>     || rcap whitefbf wsdbar plotdose, color(red)                    ///
>     || connected blackfbf plotdose, color(blue)                     ///
>     || rcap blackfbf bsdbar plotdose, color(blue)                   ///
>     ||, ytitle(Forearm Blood Flow (ml/min/dl))                       ///
>         legend(ring(0) position(11) col(1) order(1 "Whites" 3 "Blacks")) /// {7}
>         xtitle(Isoproterenol Dose (ng/min)) xscale(log)              /// {8}
>         xlabel(6 "0" 10 20 30 60 100 150 200 300 400, angle(45))     /// {9}
>         xmtick(40(10)90 250 350)
```

{4} The **distance** between ***whitefbf*** and ***wsdbar*** equals the standard deviation of the forearm blood flow for **white** subjects at each dose; ***bsdbar*** is similarly defined for black patients.

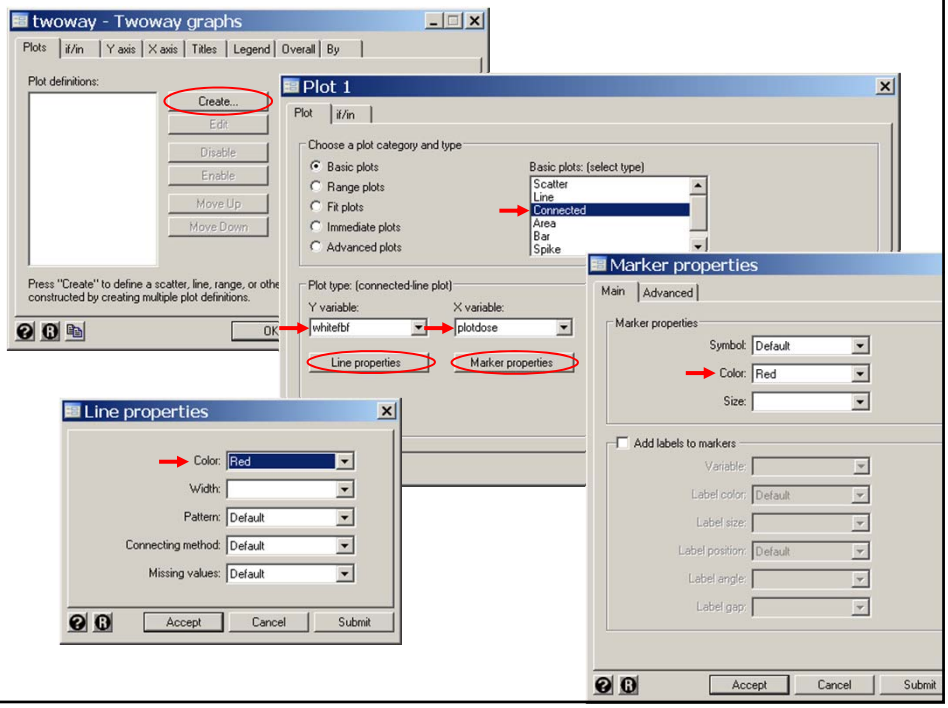
{5} This command will result in drawing the first two error bars for whites above the line to avoid collisions between the standard deviation bars for the two races.

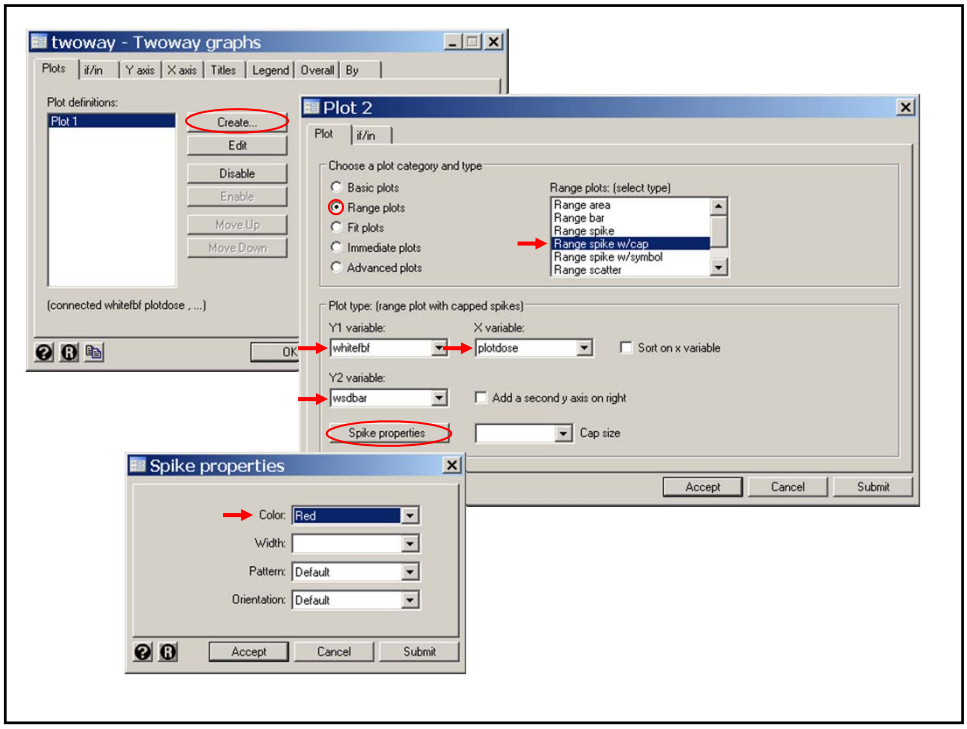
{6} This ***twoway connected*** command draws a scatter-plot of *whitefbf* by *plotdose* and connects the observations with straight lines.

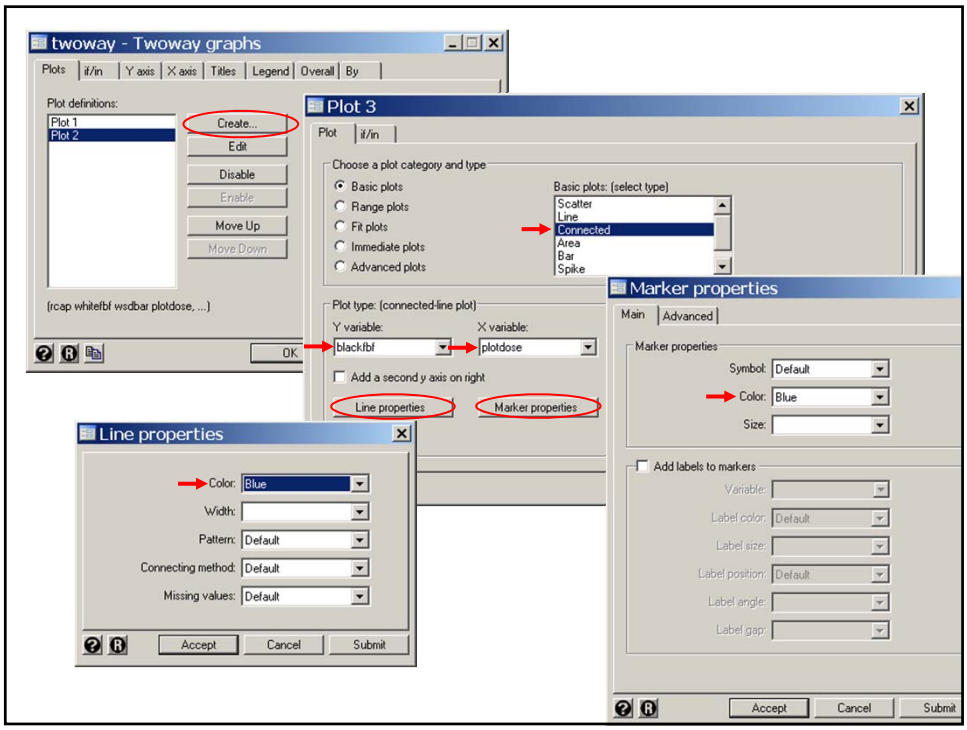
{7} The first and third variables (keys) of this plot are the mean blood flows for whites and blacks, respectively. This order option restricts the legend to these two variables and labels them "Whites" and "Blacks" respectively.

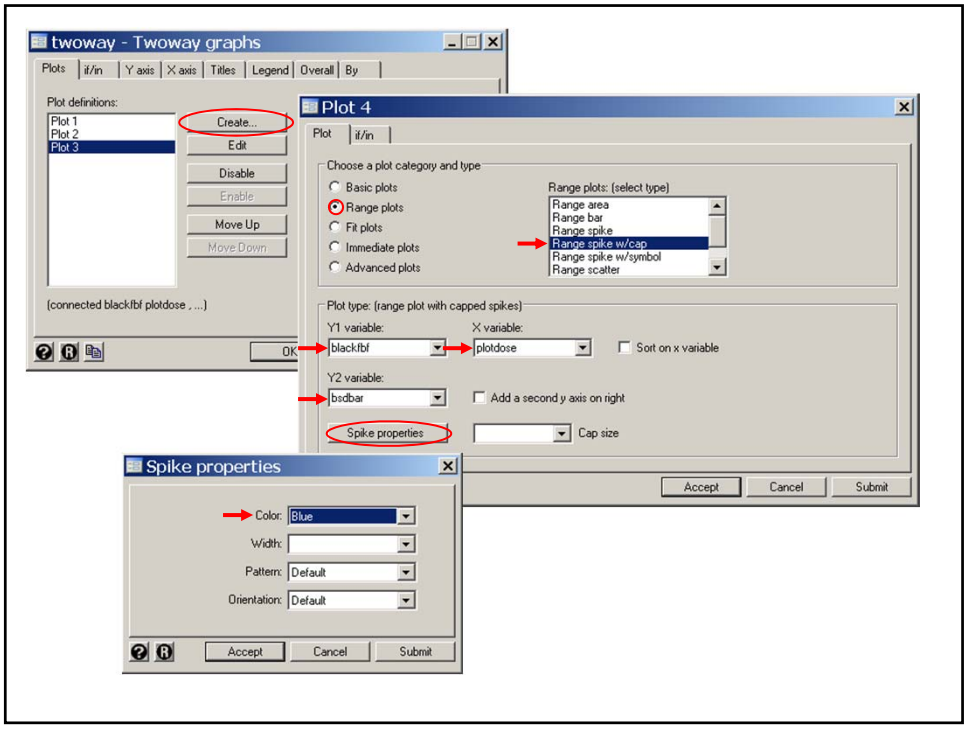
{8} The *xscale(log)* option of the *graph* command causes the *x*-axis to be drawn on a **logarithmic scale**.

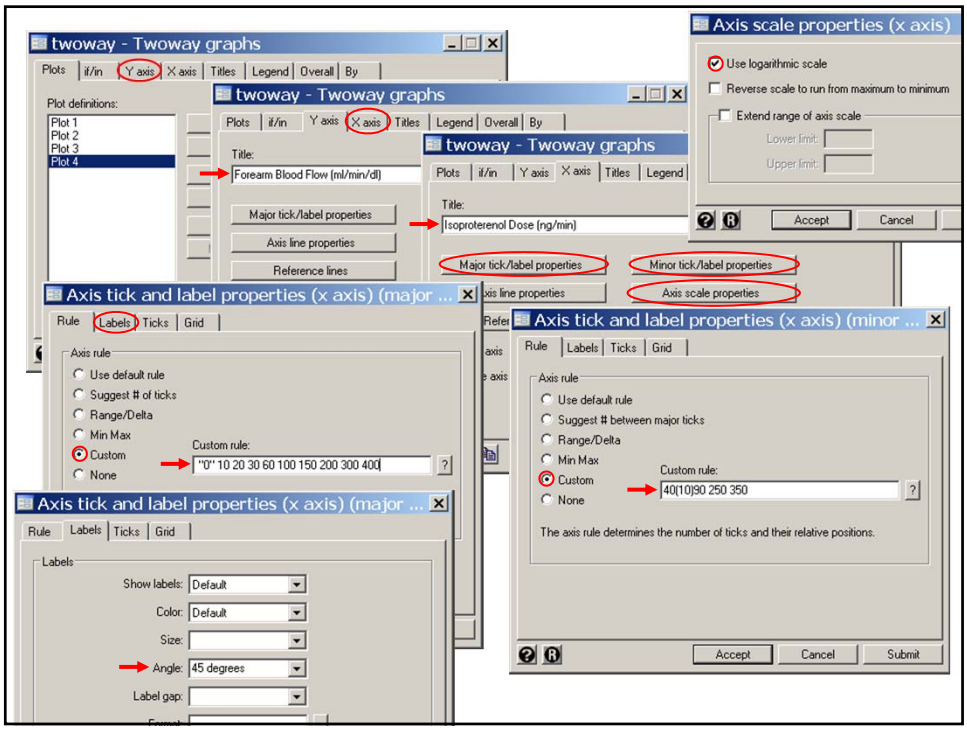
{9} The value 6 is assigned the label "0"

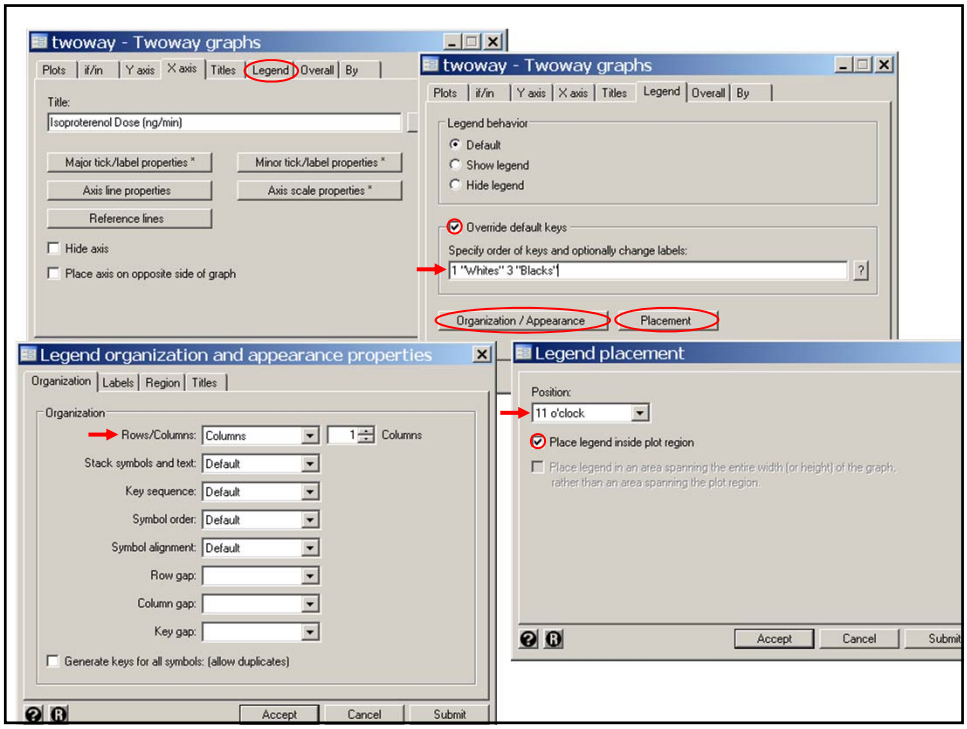


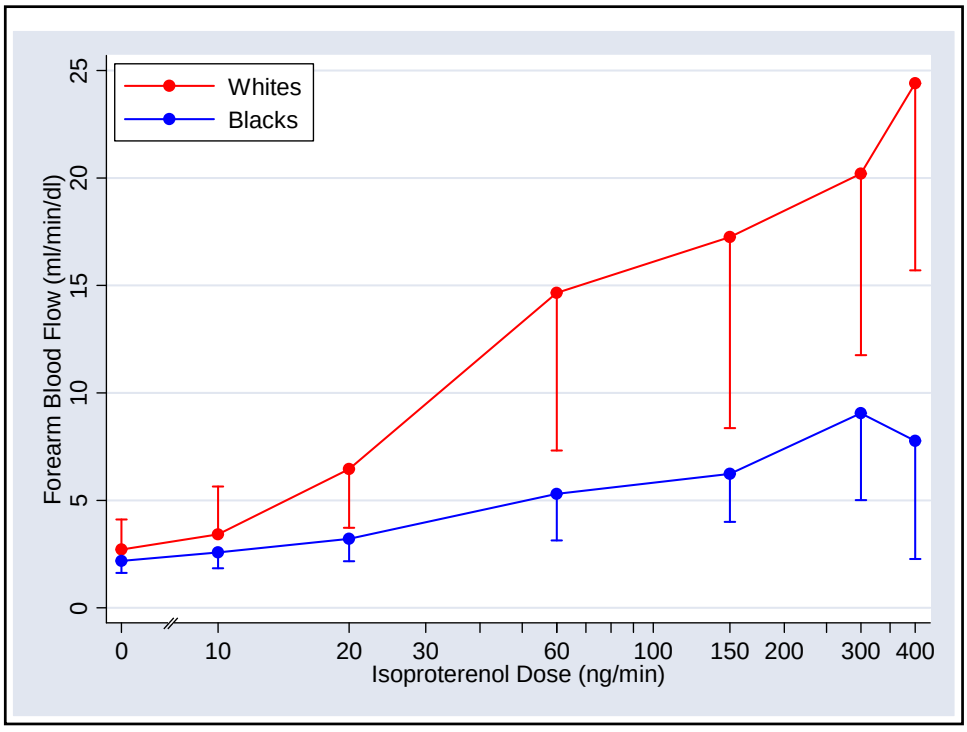

twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Create...
Edit
Disable
Enable
Move Up
Move Down
(connected whitefbf plotdose , ...)
OK
Plot 2
Plot
if/in
Choose a plot category and type
Basic plots
Range plots
Fit plots
Immediate plots
Advanced plots
Range plots: (select type)
Range area
Range bar
Range spike
Range spike w/cap
Range spike w/symbol
Range scatter
Plot type: (range plot with capped spikes)
Y1 variable:
X variable:
whitefbf
plotdose
Sort on x variable
Y2 variable:
wsdbar
Add a second y axis on right
Spike properties
Cap size
Accept
Cancel
Submit
Spike properties
Color:
Red
Width:
Pattern:
Default
Orientation:
Default
Accept
Cancel
Submit


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
Create...
Edit
Disable
Enable
Move Up
Move Down
(rcap whitefbf wsdbar plotdose, ...)
OK
Plot 3
Plot
if/in
Choose a plot category and type
Basic plots
Range plots
Fit plots
Immediate plots
Advanced plots
Basic plots: (select type)
Scatter
Line
Connected
Area
Bar
Spike
Plot type: (connected-line plot)
Y variable:
X variable:
blackfbf
plotdose
Add a second y axis on right
Line properties
Marker properties
Line properties
Color:
Blue
Width:
Pattern:
Default
Connecting method:
Default
Missing values:
Default
Accept
Cancel
Submit
Marker properties
Main
Advanced
Marker properties
Symbol:
Default
Color:
Blue
Size:
Add labels to markers
Variable:
Label color:
Default
Label size:
Label position:
Default
Label angle:
Label gap:
Accept
Cancel
Submit


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
Plot 3
Create...
Edit
Disable
Enable
Move Up
Move Down
(connected blackfbf plotdose , ...)
OK
Plot 4
Plot
if/in
Choose a plot category and type
Basic plots
Range plots
Fit plots
Immediate plots
Advanced plots
Range plots: (select type)
Range area
Range bar
Range spike
Range spike w/cap
Range spike w/symbol
Range scatter
Plot type: (range plot with capped spikes)
Y1 variable:
blackfbf
X variable:
plotdose
Sort on x variable
Y2 variable:
bsdbar
Add a second y axis on right
Spike properties
Cap size
Accept
Cancel
Submit
Spike properties
Color:
Blue
Width:
Pattern:
Default
Orientation:
Default
Accept
Cancel
Submit


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
Plot 3
Plot 4
Title:
Forearm Blood Flow (ml/min/dl)
Major tick/label properties
Axis line properties
Reference lines
Title:
Isoproterenol Dose (ng/min)
Major tick/label properties
Minor tick/label properties
Axis scale properties
Axis scale properties (x axis)
Use logarithmic scale
Reverse scale to run from maximum to minimum
Extend range of axis scale
Lower limit:
Upper limit:
Accept
Cancel
Axis tick and label properties (x axis) (major ...
Rule
Labels
Ticks
Grid
Axis rule
Use default rule
Suggest # of ticks
Range/Delta
Min Max
Custom
None
Custom rule:
"0" 10 20 30 60 100 150 200 300 400
Axis tick and label properties (x axis) (minor ...
Rule
Labels
Ticks
Grid
Axis rule
Use default rule
Suggest # between major ticks
Range/Delta
Min Max
Custom
None
Custom rule:
40(10)90 250 350
The axis rule determines the number of ticks and their relative positions.
Accept
Cancel
Submit
Axis tick and label properties (x axis) (major ...
Rule
Labels
Ticks
Grid
Labels
Show labels:
Default
Color:
Default
Size:
Angle:
45 degrees
Label gap:

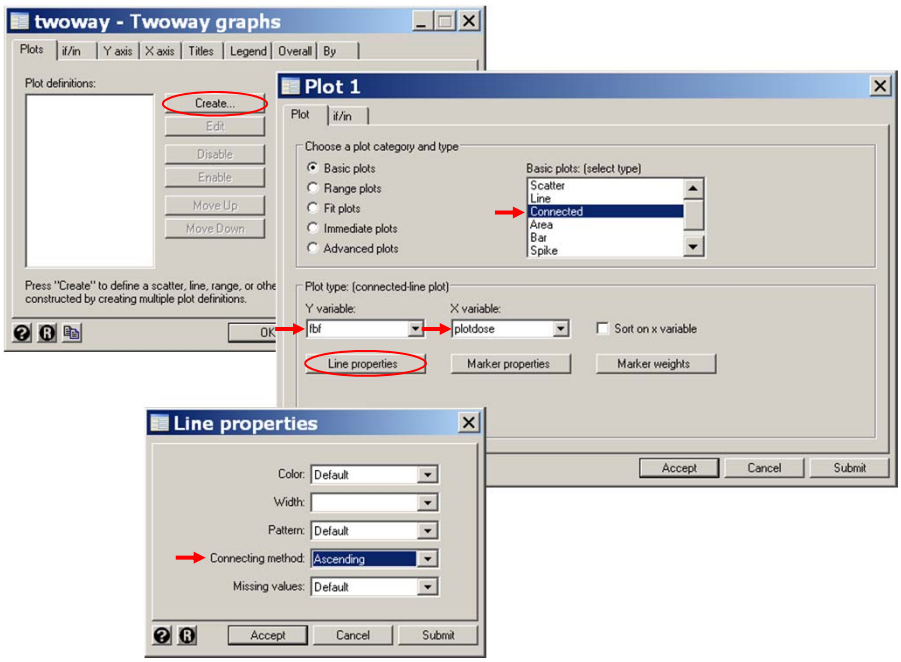

twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Title:
Isoproterenol Dose (ng/min)
Major tick/label properties
Minor tick/label properties
Axis line properties
Axis scale properties
Reference lines
Hide axis
Place axis on opposite side of graph
twoway - Twoway graphs
Legend behavior
Default
Show legend
Hide legend
Override default keys
Specify order of keys and optionally change labels:
1 "Whites" 3 "Blacks"
Organization / Appearance
Placement
Legend organization and appearance properties
Organization
Labels
Region
Titles
Rows/Columns:
Columns
Stack symbols and text:
Key sequence:
Symbol order:
Symbol alignment:
Row gap:
Column gap:
Key gap:
Generate keys for all symbols: (allow duplicates)
Accept
Cancel
Submit
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

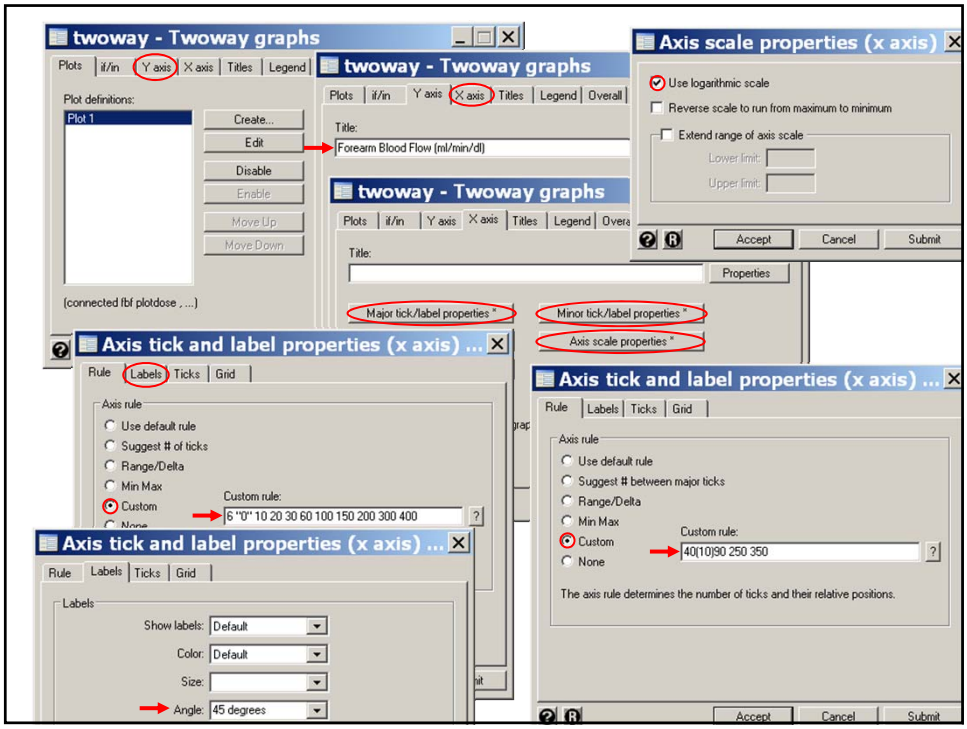

Whites
Blacks
Forearm Blood Flow (ml/min/dl)
0
5
10
15
20
25
0
10
20
30
60
100
150
200
300
400
Isoproterenol Dose (ng/min)

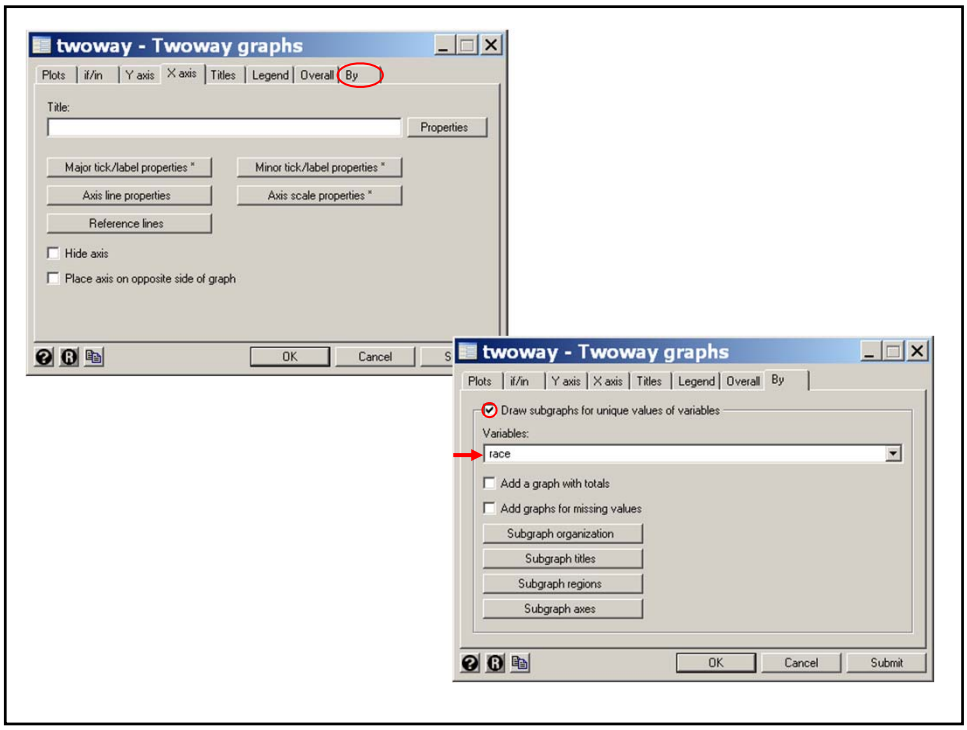
```
. *
. *  Plot individual responses for white and black patients
. *
. use C:\WDDtext\11.2.Long.Isoproterenol.dta, clear                          {10}

. sort id plotdose

. twoway connect fbf plotdose, connect(L)  xscale(log)                  /// {11}
>     xlabel(6 "0" 10 20 30 60 100 150 200 300 400, angle(45))          ///
>     xtick(40(10)90 250 350) ylabel(0(5)40, angle(0))                  ///
>     ytitle(Forearm Blood Flow (ml/min/dl)) by(race)
```

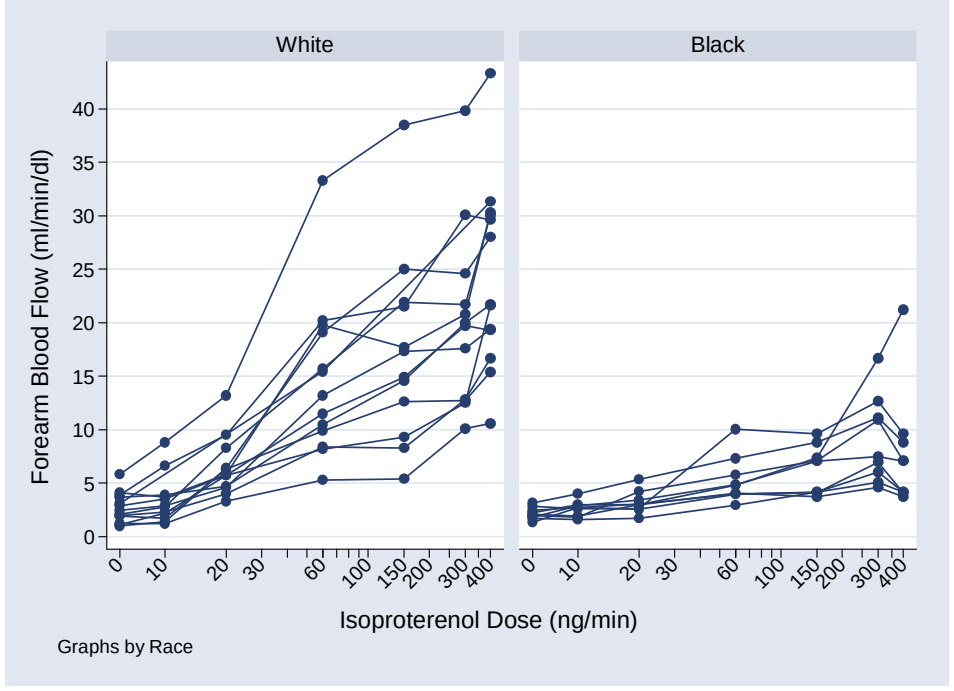
{10} We restore the long form of the data set. Note that this data set was destroyed in memory by the preceding *collapse* command.

{11} The ***connect(L)*** option specifies that straight lines are to **connect** consecutive points as long as the values of the *x*-variable, *plotdose*, are **increasing**. Otherwise the points are not connected. Note that in the preceding command we sorted the data set by *id* and *plotdose*. This has the effect of grouping all observations on the same patient together and of ordering the values on each patient by increasing values of *plotdose*. Hence, ***connect(L)*** will connect the values for **each patient** but will not **connect** the last value of one patient with the first value of the next. ***by(race)*** causes seprate graphs to be made for each race.




twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Plot definitions:
Plot 1
Create...
Edit
Disable
Enable
Move Up
Move Down
(connected fbf plotdose , ...)
Title:
Forearm Blood Flow (ml/min/dl)
Overall
Major tick/label properties "
Minor tick/label properties "
Axis scale properties "
Axis scale properties (x axis)
Use logarithmic scale
Reverse scale to run from maximum to minimum
Extend range of axis scale
Lower limit:
Upper limit:
Accept
Cancel
Submit
Properties
Axis tick and label properties (x axis) ...
Rule
Labels
Ticks
Grid
Axis rule
Use default rule
Suggest # of ticks
Range/Delta
Min Max
Custom
None
Custom rule:
6 "0" 10 20 30 60 100 150 200 300 400
Suggest # between major ticks
40(10)90 250 350
The axis rule determines the number of ticks and their relative positions.
Labels
Show labels:
Default
Color:
Size:
Angle:
45 degrees


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Title:
Properties
Major tick/label properties "
Minor tick/label properties "
Axis line properties
Axis scale properties "
Reference lines
Hide axis
Place axis on opposite side of graph
OK
Cancel
Draw subgraphs for unique values of variables
Variables:
race
Add a graph with totals
Add graphs for missing values
Subgraph organization
Subgraph titles
Subgraph regions
Subgraph axes
Submit



The following log file and comments illustrates how to perform the response feature analysis described in the preceding section.

```
. *  11.5.Isoproterenol.log
. *
. *  Perform a response feature analysis of the effect of race and dose of
. *  isoproterenol on blood flow using the data of Lang et al. (1995).
. *  For each patient, we will perform separate linear regressions of change in
. *  blood flow against log dose of isoproterenol.  The response feature that we
. *  will use is the slope of each individual regression curve.
. *
. use C:\WDDtext\11.2.Long.Isoproterenol.dta, clear
. *
. *  Calculate the regression slope for the first patient
. *
```

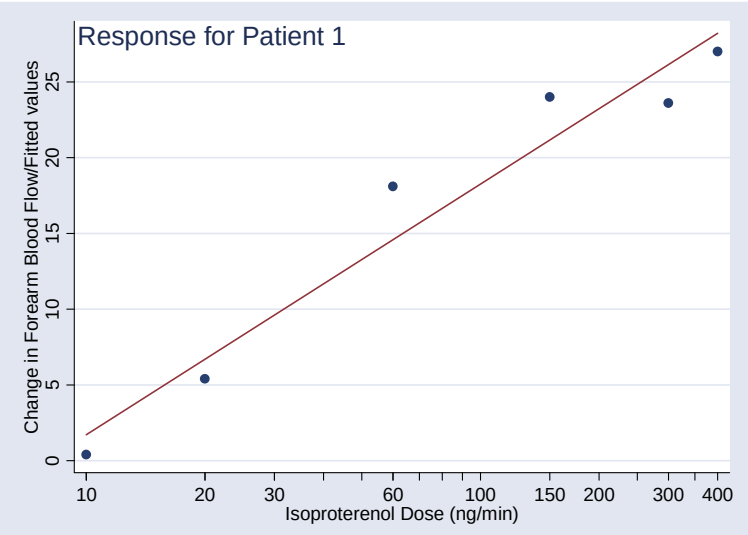
```
. regress delta_fbf logdose if id == 1                                    {1}

   Source | SS                    df         MS         Number of obs =     6
----------+-------------------------------------       F( 1, 4)      = 71.86
    Model |  570.114431   1       570.114431            Prob > F      =0.0011
 Residual |  31.7339077   4       7.93347694            R-squared     =0.9473
----------+-------------------------------------       Adj R-squared =0.9341
    Total |  601.848339   5       120.369668            Root MSE      =2.8166
------------------------------------------------------------------------------
 delta_fbf |  Coef.    Std. Err.   t          P>|t|      [95% Conf. Interval]
-----------+------------------------------------------------------------------
   logdose |  7.181315  .8471392    8.48       0.001     4.82928  9.533351
     _cons |-14.82031  3.860099    -3.84      0.018      -25.53767 -4.10296
------------------------------------------------------------------------------

. predict yhat
(option xb assumed; fitted values)
(22 missing values generated)
```

{1} We **regress** change in blood **flow** against **log dose** of isoproterenol for the observations from the **first** patient. Note that *logdose* is missing when *dose* = 0. Hence, only the six positive doses are included in this analysis. The regression slope for this patient is **7.18**. We could obtain the slopes for all 22 patients with the command

```
by id: regress delta_fbf logdose
```

However, this would require extracting the slope estimates by hand and re-entering them into Stata. This is somewhat tedious to do and is prone to transcription error. Alternately, we can use the ***statsby*** command as explained below.

```
. scatter delta_fbf dose if dose !=0 & id==1                    ///
>     || line yhat  dose if dose !=0 & id==1                    /// {2}
>     ||, ylabel(0 5 10 15 20 25) xscale(log)                   ///
>        xlabel(10 20 30 60 100 150 200 300 400)                ///
>        xtick(10(10)90 250 350) legend(off)                    ///
>        title("Response for Patient 1", position(11) ring(0))      {3}
```



{3} The position of a graph title is controled in the same way as the graph legend.

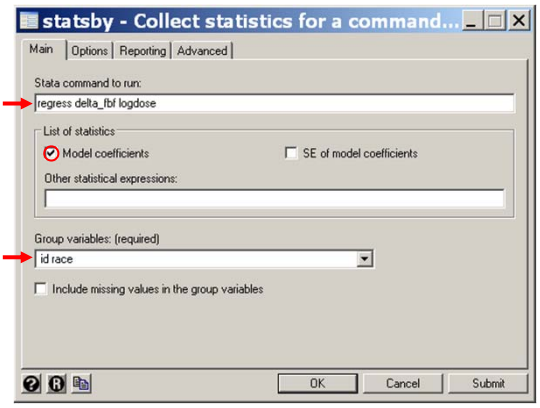
{2} Note that

```
lfit yhat dose
```

will not give the desired results since we are regressing *delta_fbf* against *logdose*.

```
. *
. * Calculate regression slopes for each patient.
. * Reduce data set to one record per patient.
. * The variable slope contains the regression slopes.
. * Race is include in the following by statement to keep this
. * variable in the data file.
. *
. * Statistics > Other > Collect statistics for a command across a by list
. statsby  slope = _b[logdose], by(id race) clear:                  ///     {3}
>     regress delta_fbf logdose
(running regress on estimation sample)

      command:  regress delta_fbf logdose
        slope:  _b[logdose]
           by:  id race

Statsby groups
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.....................
```

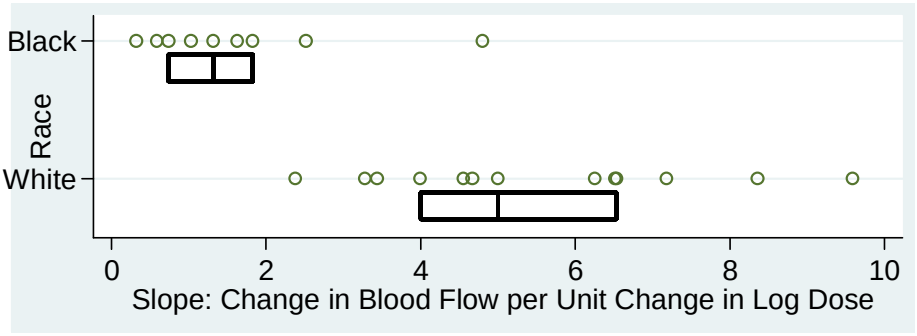
{3} This ***statsby*** command performs a **separate regression** of ***delta_fbf*** against ***logdose*** for each unique combination of values of the variables given by the ***by*** option. In this example, these variables are ***id*** **and** ***race***. The original data set is discarded and is replaced by a new data set with one record per patient. The term ***slope = _b[logdose]*** creates a new variable called ***slope*** that contains the **slope coefficient** of each individual regression. The variables that remain in the data set are ***slope*** and the ***by*** option variables (***id*** and ***race***). Note that, since *id* uniquely specifies each patient, it is not necessary to specify *race* in the *by* option to generate these regressions. However, we include *race* in the *by* option in order to keep this variable in the data set. The *clear* option allows the original data set to be replaced even if it has not been saved.




```
. * Data > Describe data > List data
. list id _b_logdose race                                          {4}
     +-----------------------+
     | id   _b_log~e    race |
     |-----------------------|
  1. |  1   7.181315   White |
  2. |  2   6.539237   White |
  3. |  3   3.999704   White |
  4. |  4   4.665485   White |
  5. |  5   4.557809   White |
     |-----------------------|
  6. |  6   6.252436   White |
  7. |  7   2.385183   White |
  8. |  8   8.354769   White |
  9. |  9   9.590916   White |
 10. | 10   6.515281   White |
     |-----------------------|
 11. | 11   3.280572   White |
 12. | 12   3.434072   White |
 13. | 13   5.004545   White |
 14. | 14   .5887727   Black |
 15. | 15   1.828892   Black |
     |-----------------------|
 16. | 16   .3241574   Black |
 17. | 17    1.31807   Black |
 18. | 18   1.630882   Black |
 19. | 19   .7392463   Black |
 20. | 20   2.513615   Black |
     |-----------------------|
 21. | 21   1.031773   Black |
 22. | 22   4.805952   Black |
     +-----------------------+
```

{4} We **list** the individual **slope estimates** for each patient. Note that the highlighted slope estimate for the first patient is identical to the estimate obtained earlier with the *regress* command.

```
. stripplot slope, over(race)  boffset(-0.1)                          /// {5}
    box(lwidth(medthick)   barwidth(0.1)) xsize(5)
                                                    ///
    xtitle(Slope: Change in Blood Flow per Unit Change in Log Dose)
```



{5} This graph, which is similar to the box plot figure, highlights the difference in the distribution of slope estimates between blacks and whites.

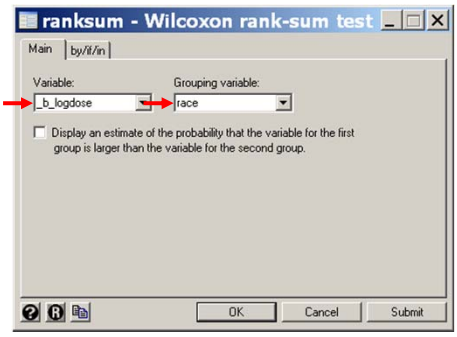
The *stripplot* command is a user-contributed command that must be downloaded using the *findit* command before use. The command

```
. graph7 slope,  by(race) oneway box
```

produces a similar graph.

```
. *
. *  Do ranksum test on slopes.
. *
. * Statistics > Nonparametric... > Tests... > Wilcoxon rank-sum test
. ranksum slope, by(race)                                                {6}
Two-sample Wilcoxon rank-sum (Mann-Whitney) test

        race |      obs    rank sum    expected
-------------+---------------------------------
       White |       13         201       149.5
       Black |        9          52       103.5
-------------+---------------------------------
    combined |       22         253         253

unadjusted variance      224.25
adjustment for ties       -0.00
                        ----------
adjusted variance        224.25

Ho: slope(race==White) = slope(race==Black)
             z =   3.439
    Prob > |z| =   0.0006
```

{6} This ***ranksum*** command performs a Wilcoxon-Mann-Whitney rank sum test of the **null hypothesis** that the **distribution of slopes** is the same for both races. The test is highly significant giving a *P* value of **0.0006**.




```
. *
. *  Do t tests comparing change in blood flow in blacks
. *  and whites at different doses

. *
. use C:\WDDtext\11.2.Long.Isoproterenol.dta, clear
                                                                {7}
 . sort dose
. *  Data > Create or change data > Keep or drop observations
. drop if dose == 0
(22 observations deleted)

. * Statistics > Summaries... > Classical... > Two-group mean-comparison test
. by dose: ttest delta_fbf , by(race) unequal                   {8}
```

{7} The preceding ***statsby*** command **deleted** most of the data. We must read in the **data** set before performing *t* tests at the different doses.

{8} This *ttest* command performs independent ***t* tests** of ***delta_fbf*** in blacks and whites at **each dose** of isoproterenol. The output for doses 60, 150 and 300 have been omitted. The highlighted output from this command is also given in the following Table 10.1.

```
-> dose = 10

Two-sample t test with unequal variances

------------------------------------------------------------------------------
   Group |   Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   White |    12   .7341667    .3088259    1.069804    .0544455    1.413888
   Black |     9   .3966667    .2071634    .6214902   -.081053    .8743863
---------+--------------------------------------------------------------------
combined |    21   .5895238    .1967903    .9018064    .1790265    1.000021
---------+--------------------------------------------------------------------
    diff |          .3375    .3718737               -.4434982    1.118498
------------------------------------------------------------------------------
Satterthwaite's degrees of freedom:  18.0903

                 Ho: mean(White) - mean(Black) = diff = 0

    Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
      t =   0.9076               t =   0.9076                  t =   0.9076
 P < t =   0.8120           P > |t| =   0.3760          P > t =   0.1880
```

```
-> dose = 20

Two-sample t test with unequal variances

------------------------------------------------------------------------------
   Group |   Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   White |    12   3.775833    .6011875    2.082575    2.452628    5.099038
   Black |     9       1.03    .3130229    .9390686     .308168    1.751832
---------+--------------------------------------------------------------------
combined |    21   2.599048    .4719216    2.162616    1.614636    3.583459
---------+--------------------------------------------------------------------
    diff |          2.745833    .6777977               1.309989    4.181677
------------------------------------------------------------------------------
Satterthwaite's degrees of freedom:  16.1415

                 Ho: mean(White) - mean(Black) = diff = 0

    Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
      t =   4.0511               t =   4.0511                  t =   4.0511
 P < t =   0.9995           P > |t| =   0.0009          P > t =   0.0005
```

{Output omitted. See Table 10.1}

```
-> dose = 400

Two-sample t test with unequal variances

------------------------------------------------------------------------------
Group     |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   White |      13    21.69308    2.163637    7.80110    16.97892    26.40724
   Black |       9     5.58666      1.803   5.410649    1.427673     9.74566
---------+--------------------------------------------------------------------
combined |      22    15.10409    2.252517   10.56524    10.41972    19.78846
---------+--------------------------------------------------------------------
    diff |            16.10641   2.816756               10.2306    21.98222
------------------------------------------------------------------------------
Satterthwaite's degrees of freedom:  19.9917

                 Ho: mean(White) - mean(Black) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
     t =   5.7181                 t =   5.7181                 t =   5.7181
 P < t =   1.0000             P > |t| =   0.0000             P > t =   0.0000
```

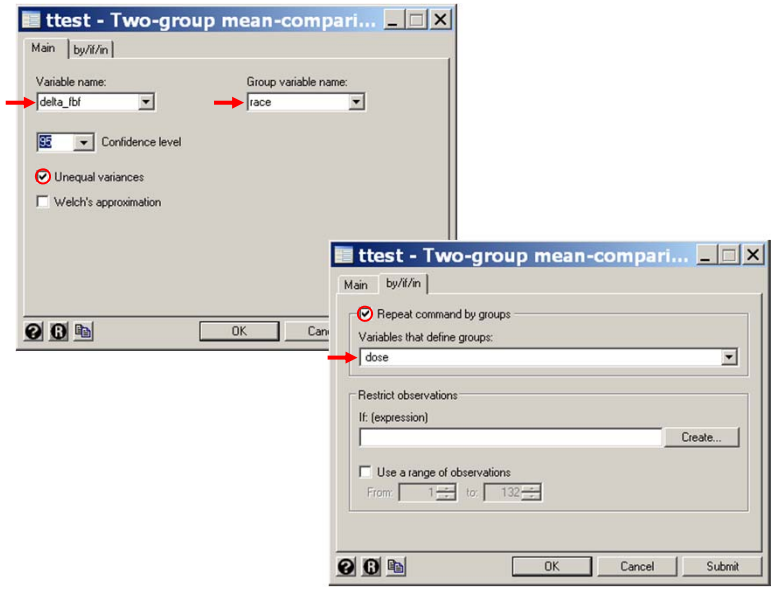


Table 10.1

	Dose of Isoproterenol (ng/min)					
	10	**20**	**60**	**150**	**300**	**400**
White Subjects						
Mean Change from Baseline	0.734	3.78	11.9	14.6	17.5	21.7
Standard Error	0.309	0.601	1.77	2.32	2.13	2.16
95% Confidence Interval	0.054 to 1.4	2.5 to 5.1	8.1 to 16	9.5 to 20	13 to 22	17 to 26
Black Subjects						
Mean Change from Baseline	0.397	1.03	3.12	4.05	6.88	5.59
Standard Error	0.207	0.313	0.607	0.651	1.30	1.80
95% Confidence Interval	-0.081 to 0.87	0.31 to 1.8	1.7 to 4.5	2.6 to 5.6	3.9 to 9.9	1.4 to 9.7
Mean Difference						
White – Black	0.338	2.75	8.82	10.5	10.6	16.1
95% Confidence Interval	-0.44 to 1.1	1.3 to 4.2	4.8 to 13	5.3 to 16	5.4 to 16	10 to 22
P value	0.38	0.0009	0.0003	0.0008	0.0005	<0.0001

3. The Area-Under-the-Curve Response Feature

A response feature that is often useful in response feature analysis is the **area under the curve.**

Let $y_i(t)$ be the response from the i^{th} patient at the time t.
$y_{ij} = y_i(t_j)$ at times $t_1, t_2, \ldots t_j, \ldots, t_n$

We can estimate the area under the curve $y_i(t)$ between t_1 and t_n as follows:

Draw a scatterplot of y_{ij} against t_j for $j = 1,2,\ldots,n$. Then draw straight lines connecting the points (t_1,y_{i1}), (t_2,y_{i2}), …, (t_n,y_{in}).

We estimate the area under the curve to be the area under these lines. Specifically, the area under the line from (t_j,y_{ij}) to $(t_j{+}_1,y_{i,j+1})$ is

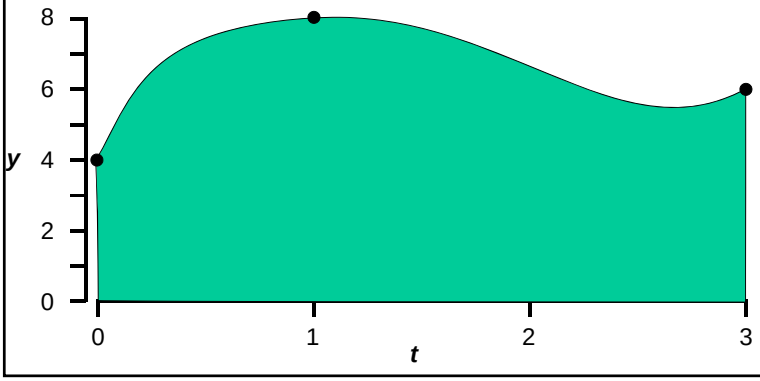
$$\left(\frac{y_{ij} + y_{i,j+1}}{2}\right)\left(t_{j+1} - t_j\right)$$

Hence, the area under the entire curve is estimated by

$$\sum_{j=1}^{n-1}\left(\frac{y_{ij}+y_{i,j+1}}{2}\right)\left(t_{j+1}-t_j\right) \qquad \{10.1\}$$

For example, if $n = 3$, $t_1 = 0$, $t_2 = 1$, $t_3 = 3$, $y_{i1} = 4$, $y_{i2} = 8$, and $y_{i3} = 6$ then equation (10.1) reduces to

$$\left(\frac{4+8}{2}\right)(1-0)+\left(\frac{8+6}{2}\right)(3-1)=20.$$

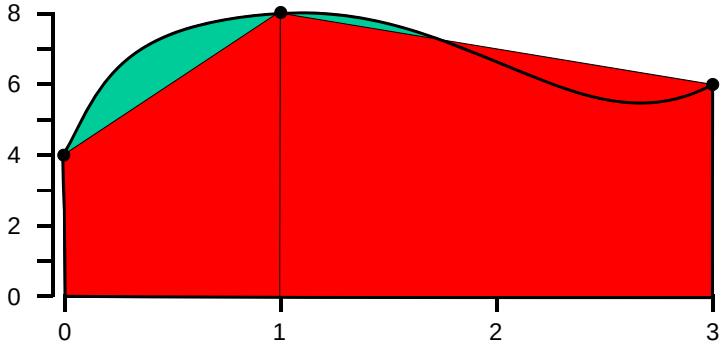


Hence, the area under the entire curve is estimated by

$$\sum_{j=1}^{n-1}\left(\frac{y_{ij}+y_{i,j+1}}{2}\right)\left(t_{j+1}-t_j\right) \qquad \{10.1\}$$

For example, if $n = 3$, $t_1 = 0$, $t_2 = 1$, $t_3 = 3$, $y_{i1} = 4$, $y_{i2} = 8$, and $y_{i3} = 6$ then equation (0.13) reduces to

$$\left(\frac{4+8}{2}\right)(1-0)+\left(\frac{8+6}{2}\right)(3-1)=20.$$

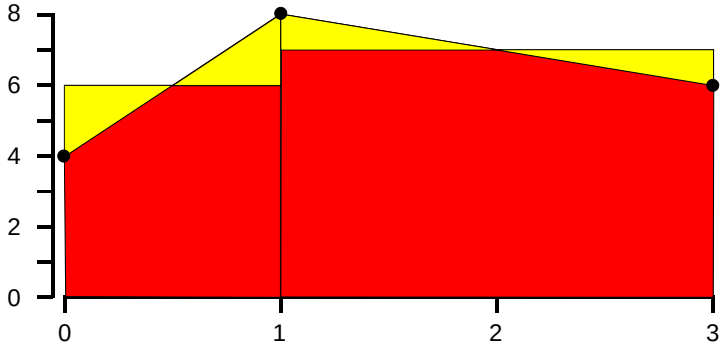


Hence, the area under the entire curve is estimated by

$$\sum_{j=1}^{n-1}\left(\frac{y_{ij}+y_{i,j+1}}{2}\right)\left(t_{j+1}-t_j\right) \qquad \{10.1\}$$

For example, if $n = 3$, $t_1 = 0$, $t_2 = 1$, $t_3 = 3$, $y_{i1} = 4$, $y_{i2} = 8$, and $y_{i3} = 6$ then equation (0.13) reduces to

$$\left(\frac{4+8}{2}\right)(1-0)+\left(\frac{8+6}{2}\right)(3-1)=20.$$



In a response feature analysis based on area under the curve, we use equation {10.1} to calculate this area for each patient and then perform a one-way analysis of variance on these areas.

Equation {10.1} can be implemented in Stata as follows. Let

id be the patient's identification number *i,*
time be the patient's time of observation t_j,
response be the patient's response $y_i\left(t_j\right)$.

Then the area under the response curve for study subjects can be calculated by using the following Stata code

```
sort id time
*
* Delete records with missing values for time or response
*
*  Data > Create or change data > Keep or drop observations
drop if time == . | response == .
generate area=(response+response[_n+1])*(time[_n+1]-time)/2 if id==id[_n+1]
collapse (sum) area = area , by(id)
*
* The variable area is now the area under the curve for
* each patient defined by equation {10.1}.  The data file
* contains one record per patient.
```

4. Generalized Estimating Equations (GEE)

This is a popular and more sophisticated approach to modeling mixed effects response data.

It is basically a generalization of the generalized linear model to allow repeated measures per subject. An appropriate correlation structure for the responses from each patient is built into the model.

Let n be the number of patients studied,
n_i, number of observations on the i^{th} patient,
y_{ij} be the response of the i^{th} patient at her j^{th} observation,
$x_{ij1}, x_{ij2}, \ldots, x_{ijq}$ be q covariates that are measured on her at this time,
$\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijq})$ denote the values of all of the covariates for the i^{th} patient at her j^{th} observation.

Then the model used by GEE analysis assumes that:

1. The distribution of y_{ij} belongs to the exponential family of distributions.

2. The expected value of y_{ij} given the patient's covariates $x_{ij1}, x_{ij2}, \ldots, x_{ijq}$ is related to the model parameters through an equation of the form

$$g\left[\mathrm{E}\left[y_{ij} \mid \mathbf{x}_{ij}\right]\right] = \alpha + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \ldots + \beta_q x_{ijq} \qquad \{10.2\}$$

g is the link function

$\alpha + \beta_1 x_{ij_1} + \beta_2 x_{ij2} + \ldots + \beta_q x_{ijq}$ is the linear predictor.

3. Responses from different patients are mutually independent.

When there is only one observation per patient (for all i), model {10.2} is, in fact, the generalized linear model. In this case,

when g is the identity function ($g[y] = y$), and y_{ij} is normally distributed, {10.2} reduces to multiple linear regression;

when g is the logit function and y_{ij} has a binomial distribution, {10.2} describes logistic regression;

when g is the logarithmic function and y_{ij} has a Poisson distribution, this model becomes Poisson regression.

Model {10.2} differs from the generalized linear model in that it does not make any assumptions about how observations on the same patient are correlated.

5. Common Correlation Structures

Let ρ_{jk} denote the population correlation coefficient between j^{th} the and k^{th} observations on the same patient. If all patients have n observations, then

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2n} \\ \rho_{31} & \rho_{32} & 1 & \cdots & \rho_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \rho_{n3} & \cdots & 1 \end{bmatrix} \qquad \{10.3\}$$

R is called the **correlation matrix** for repeated observations on study subjects. In this matrix, the coefficient in the j^{th} row and k^{th} column is the correlation coefficient between the j^{th} and k^{th} observations.

{10.3} is called an **unstructured correlation** matrix. It

- makes no assumptions about the correlation structure
- requires $n(n-1) / 2$ correlation parameters.

An **exchangeable correlation** structure assumes that

$$\mathbf{R} = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix} \qquad \{10.4\}$$

Any two distinct observations from the same patient have the same correlation coefficient ρ.

Many data sets have much more complicated correlation structures.

Observations on a patient taken closer in time are often more correlated than observations taken far apart.

Correlation structure may vary among patients.

6. GEE Analysis and the Huber-White Sandwich Estimator

GEE analysis is computationally and methodologically complex. The basic idea of the analysis can be summarize as follows:

1. We select a working correlation matrix $\mathbf{R}_i$ for each patient. $\mathbf{R}_i$, – usually with an exchangeable correlation structure.
2. We estimate the working variance-covariance matrix for the i^{th} patient.
3. Using the working variance-covariance structure we obtain estimates of the model parameters
4. We estimate the variance-covariance matrix of our model parameters using a technique called the **Huber-White sandwich estimator**.
5. We use our parameter estimates and the Huber-White variance-covariance matrix to test hypotheses or construct confidence intervals from relevant weighted sums of the parameter estimates (see Sections 5.14 through 5.16).

7. Example: Analyzing the Isoproterenol Data with GEE

Suppose that in model {10.2} , y_{ij} is a normally distributed random component and $g[y] = y$ is the identity link function. Then model {10.2} reduces to

$$\mathrm{E}\left[y_{ij} \mid \mathbf{x}_{ij} \right] = \alpha + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \ldots + \beta_q x_{ijq} \qquad \{10.5\}$$

Model {10.5} is a special case of the GEE model {10.2}.

Let

y_{ij} be the change from baseline in forearm blood flow for the i^{th} patient at the j^{th} dose of isoproterenol,

$$white_i = \begin{cases} 1: \text{ if the } i^{th} \text{ patient is white} \\ 0: \text{ if he is black} \end{cases}$$

and

$$dose_{jk} = \begin{cases} 1: \text{ if } j = k \\ 0: \text{ otherwise} \end{cases}$$

We will assume that y_{ij} is normally distributed and

$$\mathrm{E}\left[y_{ij} \mid white_i, j \right] = \alpha + \beta \times white_i + \sum_{k=2}^{6} \left(\gamma_k dose_{jk} + \delta_k \times white_i \times dose_{jk} \right) \qquad \{10.6\}$$

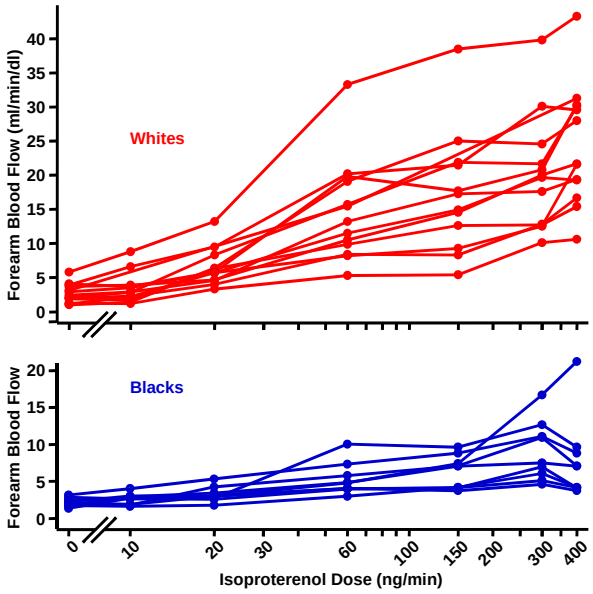
where α, β, $\{\gamma_k,\ \delta_k : k = 2, \ldots, 6\}$ are the model parameters. Model {10.6} is a special case of model {10.5). Note that this model implies that the expected change in blood flow is

α	for a black man on the first dose,	{10.7}
$\alpha + \beta$	for a white man on the first dose,	{10.8}
$\alpha + \gamma_j$	for a black man on the j^{th} dose with $j > 1$, and	{10.9}
$\alpha + \beta + \gamma_j + \delta_j$	for a white man on the j^{th} dose with $j > 1$.	{10.10}

It must be noted that patient 8 in this study has four missing blood flow measurements. This concentration of missing values in one patient causes the choice of the working correlation matrix to have an appreciable effect on our model estimates.

Regardless of the working correlation matrix, the working variance for y_{ij} in model {10.5} is constant.

Figure 10.2 suggests that this variance is greater for whites than blacks and increases with increasing dose.



Hence, it is troubling to have our parameter estimates affected by a working correlation matrix that we know is wrong.

Also, the Huber-White variance-covariance estimate is only valid when the missing values are few and randomly distributed.

For these reasons, we delete patient 8 from our analysis. Without patient 8, the Huber-White variance-covariance matrix is unaffected by the choice of $\mathbf{R}_i$.

Let $\hat{\alpha}$, $\hat{\beta}$, $\left\{\hat{\gamma}_k,\ \hat{\delta}_k : k = 2, ..., 6\right\}$ denote the GEE parameter estimates from the model. Then our estimates of the mean change in blood flow in blacks and whites at the different doses are given by equations {10.7} through {10.10} with the parameter estimates substituting for the true parameter values. Subtracting the estimate of equation {10.7} from that for equation {10.8} gives the estimated mean difference in change in flow between whites and blacks at dose 1, which is

$$\left(\hat{\alpha} + \hat{\beta}\right) - \hat{\alpha} = \hat{\beta} \qquad \{10.11\}$$

Subtracting the estimate of equation {10.9}) from that for equation {10.10} gives the estimated mean difference in change in flow between whites and blacks at dose $j > 1$, which is

$$\left(\hat{\alpha} + \hat{\beta} + \hat{\gamma}_j + \hat{\delta}_j\right) - \left(\hat{\alpha} + \hat{\gamma}_j\right) = \left(\hat{\beta} + \hat{\delta}_j\right) \qquad \{10.12\}$$

Tests of significance and 95% confidence intervals can be calculated for these estimates using the Huber-White variance-covariance matrix.

This is done in the same way as was illustrated in logistic regression. These estimates, standard errors, confidence intervals and *P* values are given in the next table.

Figure 10.2

	Dose of Isoproterenol (ng/min)					
	10	**20**	**60**	**150**	**300**	**400**
White Subjects						
Mean Change from Baseline	0.734	3.78	11.9	14.6	17.5	21.2
Standard Error	0.303	0.590	1.88	2.27	32.09	2.23
95% Confidence Interval	0.14 to 1.3	2.6 to 4.9	8.2 to 16	10 to 19	13 to 22	17 to 26
Black Subjects						
Mean Change from Baseline	0.397	1.03	3.12	4.05	6.88	5.59
Standard Error	0.200	0.302	0.586	0.629	1.26	1.74
95% Confidence Interval	0.0044 to 0.79	0.44 to 1.6	2.0 to 4.3	2.8 to 5.3	4.4 to 9.3	2.2 to 9.0
Mean Difference						
White – Black	0.338	2.75	8.79	10.5	10.6	15.6
95% Confidence Interval	-0.37 to 1.0	1.4 to 4.0	4.9 to 13	5.9 to 15	5.9 to 15	10 to 21
P value	0.35	<0.0005	<0.0005	<0.0005	<0.0005	<0.0001

The null hypothesis that there is no interaction between race and dose on blood flow is

$H_0: \delta_2 = \delta_3 = \delta_4 = \delta_5 = \delta_6 = 0$

Under this null hypothesis a chi-squared statistic can be calculated that has as many degrees of freedom as there are interaction parameters (in this case five).

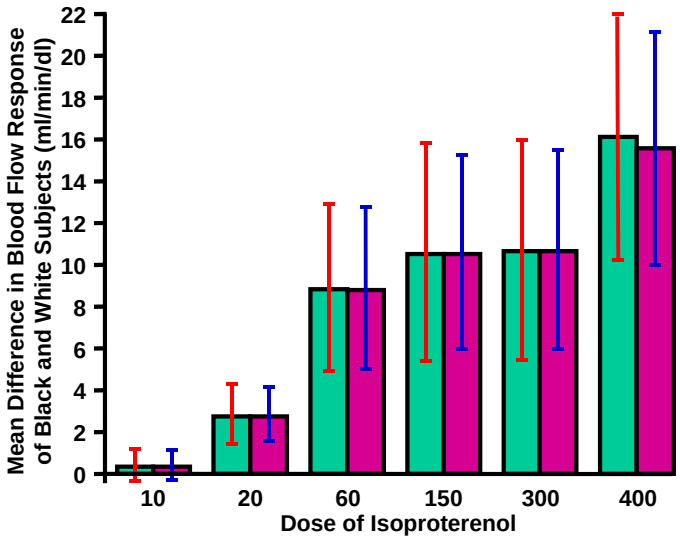
This statistic equals 40.41, which is highly significant ($P < 0.00005$). Hence, we can conclude that the observed interaction is certainly not due to chance.

The GEE and response feature analysis (RFA) in Tables 10.2 and 10.1 should be compared. Note that the mean changes in blood flow in the two races and six dose levels are very similar. They would be identical were if not for the fact that patient 8 is excluded from the GEE analysis but is included in the RFA.

This is a challenging data set to analyze in view of the fact that the standard deviation of the response variable

- increases with dose and
- differs between the races.

The following figure compares the mean difference between blacks and whites at the six different doses. The **green** and **magenta** bars are from the RFA and GEE analyses, respectively.



In this example, response feature analysis and GEE give virtually identical results.

8. Using Stata to Analyze the Isoproterenol Data Set Using GEE

The following log file and comments illustrate how to perform the GEE analysis for the isoproterenol data

```
. *  11.11.Isoproterenol.log
. *
. *  Perform a GEE analyses of the effect of race and dose
. *  of isoproterenol
. *  on blood flow using the data of Lang et al. (1995).
. *
. use C:\WDDtext\11.2.Long.Isoproterenol.dta, clear

. *  Data > Create or change data > Keep or drop observations
. drop if dose == 0 | id == 8                                          {1}
(28 observations deleted)

. generate white = race == 1
```

{1} We **drop** all records with *dose* = **0** or *id* = **8**. When ***dose* = 0**, the change from baseline, ***delta_fbf***, is by definition, **zero**. We eliminate these records as they provide no useful information to our analyses. Patient **8** has **four** missing **values**. These missing values have an adverse effect on our analysis. For this reason we eliminate all observations on this patient (see Section 7).

```
. *
. *  Analyze data using classification variables with
. *  interaction
. *
. * Statistics > Longitudinal... > Generalized est... > Generalized...(GEE)
. xtgee delta_fbf dose##white, i(id) robust                          {2}
Iteration 1: tolerance = 2.061e-13

GEE population-averaged model                   Number of obs      =       126
Group variable:                         id      Number of groups   =        21
Link:                             identity      Obs per group: min =         6
Family:                           Gaussian                     avg =       6.0
Correlation:                  exchangeable                     max =         6
                                                Wald chi2(11)      =    506.86
Scale parameter:                  23.50629      Prob > chi2        =    0.0000

                              (standard errors adjusted for clustering on id)
```

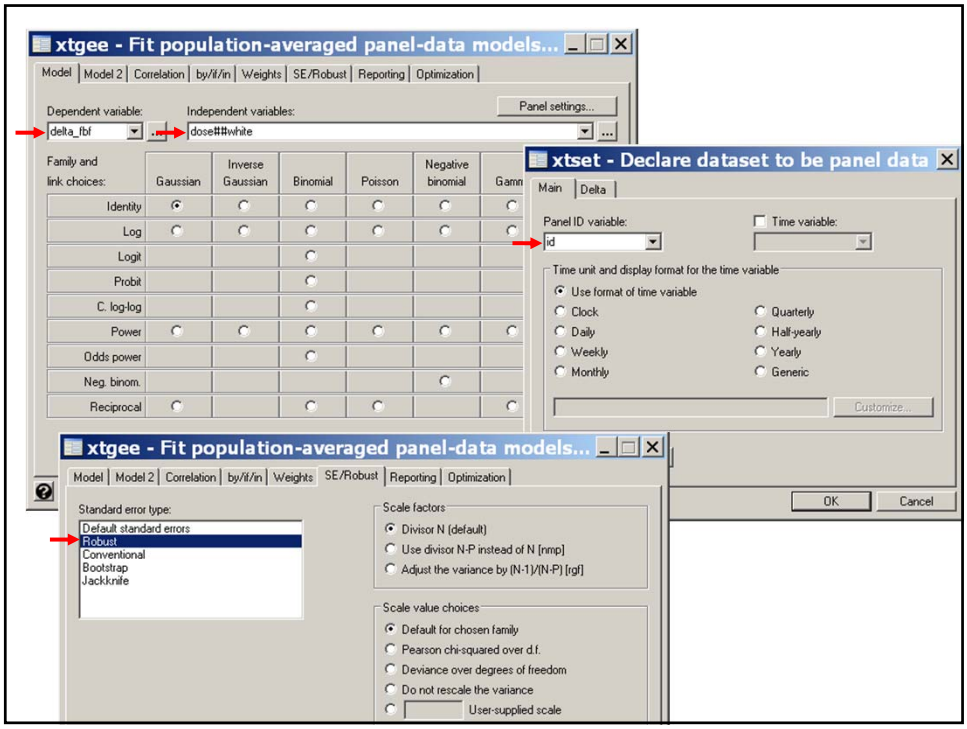
{2} This ***xtgee*** command analyzes model {10.6}. The syntax of *i.dose*white* is analogous to that used for the logistic command in Chapter 4. The **default link function** is the **identity** function. For the **identity link** function the **default random component** is the **normal distribution**. Hence, we do not need to specify either of these aspects of our model explicitly in this command. The ***i(id)*** option specifies ***id*** to be the variable that **identifies** all observations made on the same **patient**. The **exchangeable correlation structure** is the **default** working correlation structure, which we use here. The ***robust*** option specifies that the Huber-White sandwich estimator is to be used. The **table of coefficients** generated by this command is similar to that produced by other **Stata regression commands.**

Note that if we had **not** used the ***robust*** option the model would have assumed that the **exchangeable** correlation structure was true. This would have led to **inaccurate confidence intervals** for our estimates. I strongly recommend that this option always be used in any GEE analysis.

delta_fbf	Coef.	Semi-robust Std. Err.	z	P>|z|	[95% Conf.	Interval]	
dose							
20	.6333333	.2706638	2.34	0.019	.1028421	1.163825	
60	2.724445	.6585882	4.14	0.000	1.433635	4.015254	
150	3.656667	.7054437	5.18	0.000	2.274022	5.039311	
300	6.478889	1.360126	4.76	0.000	3.813091	9.144687	
400	5.19	1.830717	2.83	0.005	1.601861	8.77814	
1.white	.3375	.363115	0.93	0.353	-.3741922	1.049192	{3}
dose#white							
20 1	2.408333	.5090358	4.73	0.000	1.410642	3.406025	
60 1	8.450556	1.823352	4.63	0.000	4.876852	12.02426	
150 1	10.17667	2.20775	4.61	0.000	5.849557	14.50378	
300 1	10.30444	2.305474	4.47	0.000	5.785798	14.82309	
400 1	15.22667	2.748106	5.54	0.000	9.840479	20.61285	
_cons	.3966667	.2001388	1.98	0.047	.0044017	.7889316	{4}

{3} The highlighted term are the estimated **mean**, ***P* value** and 95% **confidence interval** for the difference in response between **white** and **black** men on the **first** dose of isoproterenol (10 ng/min). The parameter estimate associated with the *white* covariate is $\hat{\beta} = 0.3375$ in model {10.6}. The highlighted values in this and in subsequent lines of output are entered into Table 10.2.

{4} The highlighted terms are the estimated **mean**, **standard error** and 95% **confidence interval** for **black** men on the **first** dose of isoproterenol. The parameter estimate associated with *_cons* is $\hat{\alpha} = 0.3967$.




```
. lincom _cons + 1.white                                                    {5}

 ( 1)  1.white + _cons = 0.0

------------------------------------------------------------------------------
   delta_fbf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .7341667    .30298     2.42   0.015     .1403367    1.327997
------------------------------------------------------------------------------

. lincom _cons+ 20.dose                                                     {6}

 ( 1) 20.dose + _cons = 0.0

------------------------------------------------------------------------------
   delta_fbf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |       1.03   .3024088     3.41   0.001     .4372896     1.62271
------------------------------------------------------------------------------

. lincom _cons+ 20.dose +  1.white + 20.dose#1.white                        {7}

 ( 1) 20.dose + 1.white + 20.dose#1.white + _cons = 0.0

------------------------------------------------------------------------------
   delta_fbf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   3.775833   .5898076     6.40   0.000     2.619832    4.931835
------------------------------------------------------------------------------
```

{5} This command calculates $\hat{\alpha}+\hat{\beta}$, the **mean** response for **white** men at the **first** dose of isoproterenol, together with related statistics.

{6} This command calculates $\hat{\alpha}+\hat{\gamma}_2$ the **mean** response for **black** men at the **second** dose of isoproterenol, together with related statistics.

{7} This command calculates $\hat{\alpha}+\hat{\beta}+\hat{\gamma}_2+\hat{\delta}_2$, the **mean** response for **white** men at the **second** dose of isoproterenol, together with related statistics.

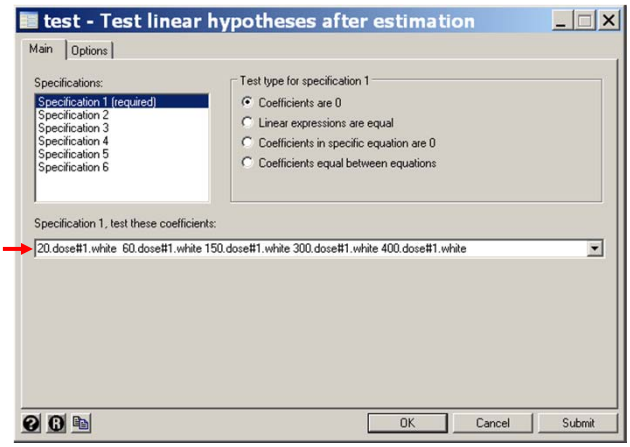
```
. lincom                1.white + 20.dose#1.white                          {8}

 ( 1) 1.white + 20.dose#1.white = 0.0

------------------------------------------------------------------------------
   delta_fbf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   2.745833   .6628153     4.14   0.000     1.446739    4.044927
------------------------------------------------------------------------------

. lincom _cons + 60.dose
                                                {output omitted. See Table 10.2}
. lincom _cons + 60.dose +  1.white +  60.dose#1.white
                                                {output omitted. See Table 10.2}
. lincom                    1.white +  60.dose#1.white
                                                {output omitted. See Table 10.2}
. lincom _cons + 150.dose
                                                {output omitted. See Table 10.2}
. lincom _cons + 150.dose + 1.white + 150.dose#1.white
                                                {output omitted. See Table 10.2}
. lincom                    1.white + 150.dose#1.white
                                                {output omitted. See Table 10.2}
. lincom _cons + 300.dose
                                                {output omitted. See Table 10.2}
. lincom _cons + 300.dose + 1.white + 300.dose#1.white
                                                {output omitted. See Table 10.2}
. lincom                    1.white + 300.dose#1.white
                                                {output omitted. See Table 10.2}
. lincom _cons + 400.dose
```

{8} This calculates $\hat{\beta} + \hat{\delta}_2$, the **mean difference** in response between **white and black** men at the **second** dose of isoproterenol, together with related statistics. Analogous *lincom* commands are also given for dose 3, 4, 5, and 6.

```
( 1) 400.dose + _cons = 0

------------------------------------------------------------------------------
   delta_fbf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   5.586667   1.742395     3.21   0.001     2.171636    9.001698
------------------------------------------------------------------------------

. lincom _cons + 400.dose + 1.white + 400.dose#1.white

 ( 1) 400.dose + 1.white + 400.dose#1.white + _cons = 0.0

------------------------------------------------------------------------------
   delta_fbf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   21.15083   2.233954     9.47   0.000     16.77236    25.5293
------------------------------------------------------------------------------

. lincom                    1.white + 400.dose#1.white

 ( 1) 1.white + 400.dose#1.white = 0.0

------------------------------------------------------------------------------
   delta_fbf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   15.56417   2.833106     5.49   0.000     10.01138    21.11695
------------------------------------------------------------------------------
```

```
. * Statistics > Postestimation > Tests > Test linear hypotheses
. test 20.dose#1.white  60.dose#1.white 150.dose#1.white    ///
>     300.dose#1.white 400.dose#1.white                                    {9}

 ( 1)  20.dose#1.white = 0
 ( 2)  60.dose#1.white = 0
 ( 3)  150.dose#1.white = 0
 ( 4)  300.dose#1.white = 0
 ( 5)  400.dose#1.white = 0

           chi2(  5) =   40.41
         Prob > chi2 =    0.0000
```

{9} This command test the **null hypothesis** that the **interaction** parameters δ_2, δ_3, δ_4, δ_5, and δ_6 are **simultaneously** equal to zero. That is, it tests the null hypothesis that the effects of race and dose on change in blood flow are additive. This test, which has five degrees of freedom, gives $P < 0.00005$, which allows us to reject the null hypothesis with overwhelming statistical significance.



9. GEE Analyses with Logistic or Poisson Models

GEE analyses can be applied to any generalized linear model with repeated measures data.

For logistic regression we use the logit link function and a binomial random component.

For Poisson regression we use the logarithmic link function and a Poisson random component.

In Stata, the syntax for specifying these terms is the same as in the *glm* command.

For logistic regression, we use the *link(logit)* and *family(binomial)* options to specify the link function and random component, respectively.

For Poisson regression, these options are *link(log)* and *family(poisson)*.

10. What we have covered

- Analysis of variance with multiple observations per patient
  - These analyses are complicated by the fact that multiple observations on the same patient are correlated with each other
- Response-feature approach to mixed effects analysis of variance
  - Reduce multiple response measures on each patient to a single statistic that captures the most biologically important aspect of the response: the *statsby* command
  - Perform a fixed effects analysis on this response feature
  - Using a regression slope as a response feature
  - Using an area under the curve as a response feature
- Generalized estimating equations (GEE) approach to mixed effects analysis of variance: the *xtgee* command
  - GEE analysis with logistic or Poisson models

Cited Reference

Lang CC, Stein CM, Brown RM, Deegan R, Nelson R, He HB, Wood M, Wood AJ. Attenuation of isoproterenol-mediated vasodilatation in blacks. N Engl J Med 1995;333:155-60.

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.

XI. OTHER TOPICS

Complicated Statistics with Nasty Properties

Bootstrap Analysis

- Treat the sample as if it were the target population
- Sample repeatedly without replacement to obtain many samples of the same size as the real sample
- Calculate the test statistic for each sample
- Examine the variation of the test statistic among bootstrapped samples to assess its dispersion.

Multiple imputation of missing values

- Most statistical packages, including Stata do complete case analyses. That is they discard the data on any patient who is missing any model covariate.
- Multiple imputation is a method that adjusts for missing data by predicting missing values from non-missing covariates.
- Lead to unbiased results if the probability of the outcome of interest is not affected by whether a specific covariate is missing.
- Stata has a very comprehensive package for doing multiple imputation
- Particularly useful to adjust for missing values in confounding variables.

1. Discriminatory Analysis

We often wish to place patients into two or more groups on the basis of a set of explanatory variables with a minimum of misclassification error.

For example, we might wish to classify patients as

- having or not having cancer,
- benefiting or not benefiting from aggressive therapy.

We typically start of with a learning set of patients whose true classification is known. We then use these patients for developing rules to classify other patients. The three most common ways of doing this are as follows.

- **Logistic Regression**

The linear predictor from a multiple logistic regression can be used to develop a classification rule. Patients whose linear predictor is greater than some value are assigned to one group; all other patients are assigned to the other.

The advantages of this approach are

⇒ It can lead to a simple rule based on a weighted sum of covariates.

⇒ By adjusting the cutoff point we can control the sensitivity and specificity of the rule. It is easy to generate **receiver operating characteristic curves** for this method.

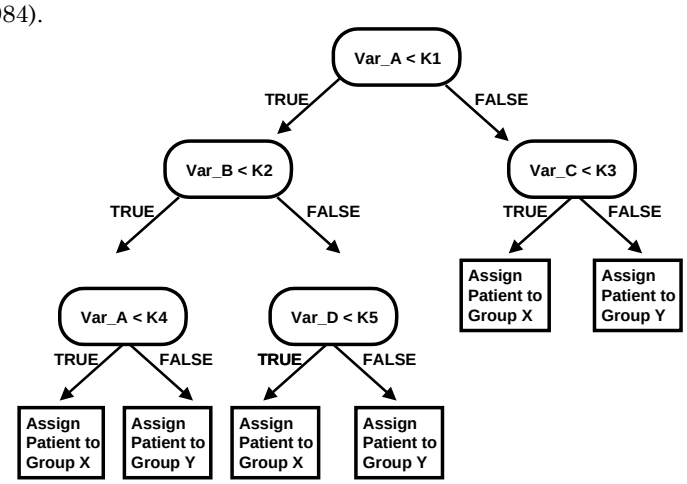
⇒ Particularly effective when used with restricted cubic splines

The disadvantage is that the rule may be less than optimal if the model is mis-specified.

- **Classification and Regression Trees**
- **Neural Networks**

2. Classification and Regression Trees (CART)

The basic idea here is to derive a tree that consists of a series of binary decisions that lead to patient classification (Breiman et al. 1984).



The CART graphic indicates the degree of increased homogeneity induced by each split. Trees can then be pruned back to produce a classification rule that makes clinical sense and is fairly easy to remember.

The advantages of this method are

- It often does better than logistic regression when the model for the latter is poorly specified.
- It gives a rule that is intelligible to clinicians and can be judged by its clinical criteria.

A disadvantage is that, when applied to continuous covariates it looses information due to the fact that it dichotomizes the selected variable at each split.

3. Neural Networks

This method attempts to outperforms the logistic regression approach by adopting models that varies from complex to extremely complex (Hinton 1992).

Advantages

- Great name.
- Sometimes does better than logistic regression.

Disadvantages

- Method is essentially a black box. You need a computer to apply it and it is very difficult to gain intuitive insight into what it is doing.
- Method usually performs only as well as the CART method or logistic regression models with restricted cubic splines.

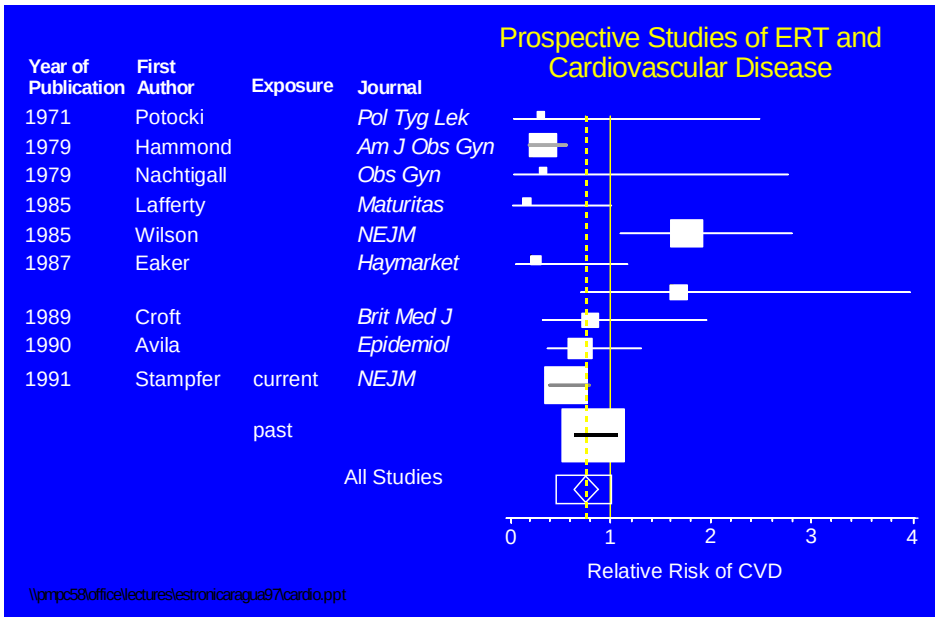
4. Meta-Analyses

One of the strengths of this approach is the meta-analysis graphic.

This is a rather pretentious term for doing quantitative reviews of the medical literature. The English refer to these techniques as **quantitative overviews,** which is a far more reasonable description. However, in this country we appear to be stuck with the term meta-analysis.

The basic steps in performing a meta-analysis are as follows:

- Systematically identify all publications that may be germane to the topic of interest.
- Review these publications. Eliminate those that are irreverent or misleading using explicitly defined criteria.
- Present the results of the individual studies graphically to show the extent to which they agree or disagree.
- Use clinical judgment and statistical methods to determine whether it is reasonable to combine some or all of the studies into a single analysis. In this case present the relative risk derived from the combined data, together with its 95% confidence interval.



- In these graphs the relative risk from each study is displayed on a single line.
- Each relative risk or odds ratio is plotted as a square.
- The size of this square is proportional to the reciprocal of the variance of the log relative risk (often referred to as the study **information**).
- The 95% confidence interval for each study is depicted as a horizontal line.
- A vertical line depicts a weighted geometric mean of the studies. This mean is weighted by the information content of each study.
- One, or preferably two, 95% confidence intervals are drawn for this combined geometric mean. These confidence intervals are usually drawn as diamonds or squares. They are calculated using either a fixed effects or random effects model.

a) Fixed effects model for meta-analysis

This approach assumes that all studies are measuring the same risk in a comparable way, and that the only variation between studies is due to chance.

If this assumption is false it will overestimate the precision of the combined estimate.

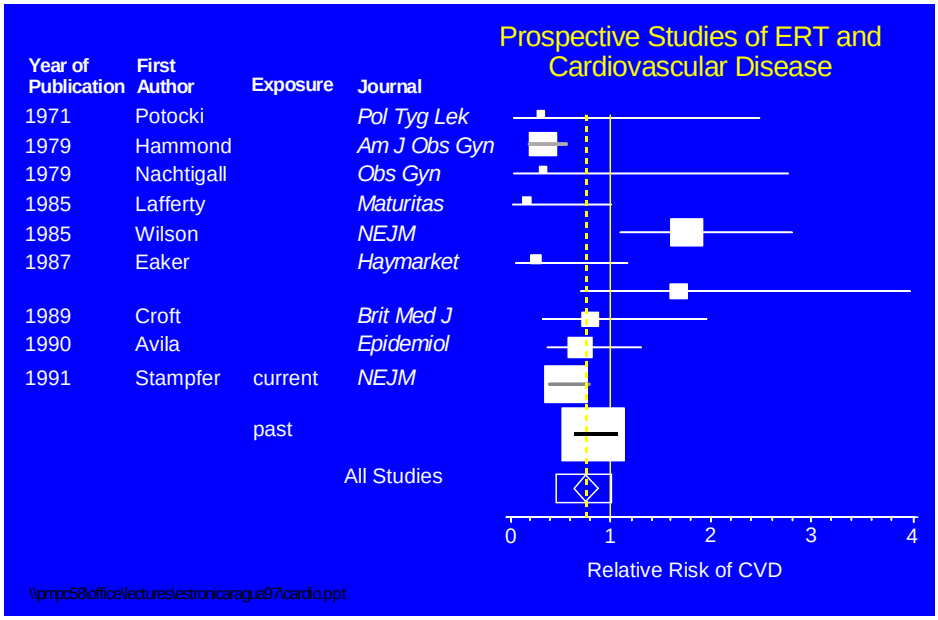
b) Random effects model for meta-analysis

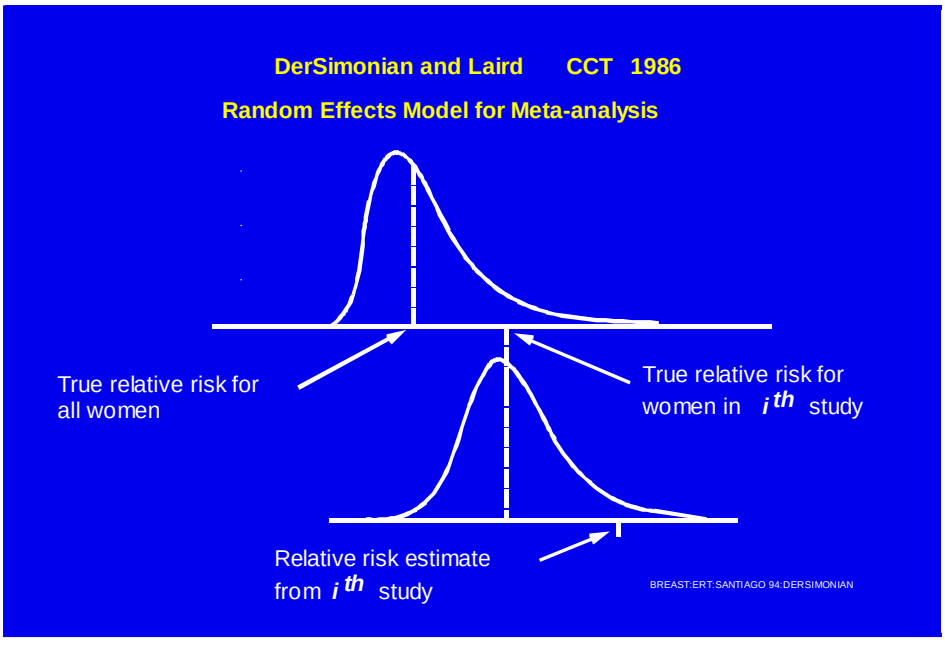
This model assumes that that each study is estimating a different unknown relative risk that is specific to that study. These risks differ from one study to the next due to differences in study populations, study designs, or biases of one kind or another.

It assumes that these study-specific relative risks follow a log-normal distribution, and that the variation in the estimated relative risks is due both to variation in the study specific risk as well as intra-study variation of study subjects.

DerSimonian and Laird (1986) devised a way to estimate the confidence interval for the combined relative risk for this model.

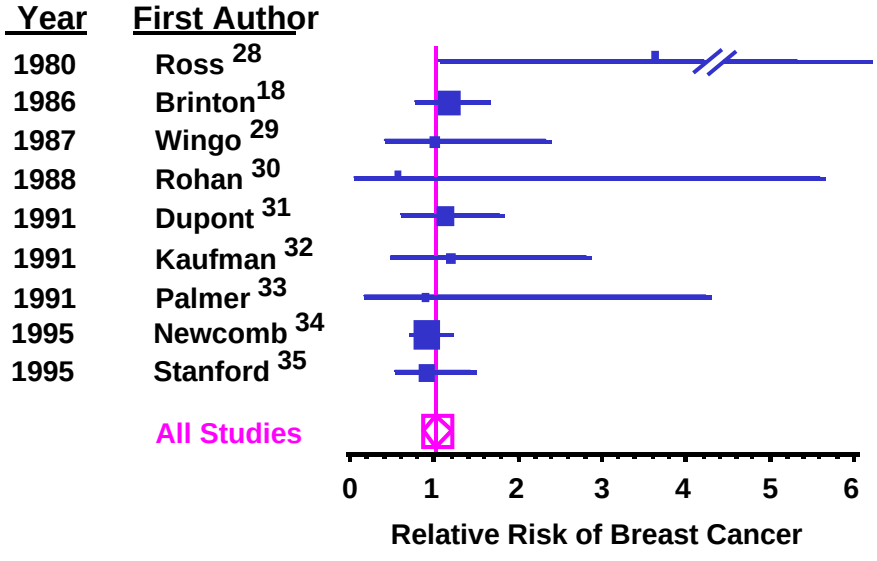
It is a good idea to plot both the fixed effects and random effects confidence intervals for the combined relative risk estimate. If these intervals disagree then the inter-study variation is greater than we would expect by chance and the studies are most likely estimating different risks. In this case we need to be very cautious about combining the results of these studies.


Prospective Studies of ERT and Cardiovascular Disease
Year of Publication
First Author
Exposure
Journal
1971 Potocki Pol Tyg Lek
1979 Hammond Am J Obs Gyn
1979 Nachtigall Obs Gyn
1985 Lafferty Maturitas
1985 Wilson NEJM
1987 Eaker Haymarket
1989 Croft Brit Med J
1990 Avila Epidemiol
1991 Stampfer current NEJM
past
All Studies
0
1
2
3
4
Relative Risk of CVD
\\pmpc58\office\lectures\estronicaragua97\cardio.ppt


DerSimonian and Laird CCT 1986
Random Effects Model for Meta-analysis
True relative risk for all women
True relative risk for women in i th study
Relative risk estimate from i th study
BREAST:ERT:SANTIAGO 94:DERSIMONIAN

On the other hand, if these estimates agree then the studies are mutually consistent and there is no statistical reason not to combine them.

Women with a History of Benign Breast Disease
Breast cancer risk among ERT users compared to non-users



5. Publication bias

One of the ways that meta-analyses can be misleading is through publication bias. That is, papers may be more likely to be published if they show that a risk factor either increases or reduces some risk than if they find a relative risk near one.

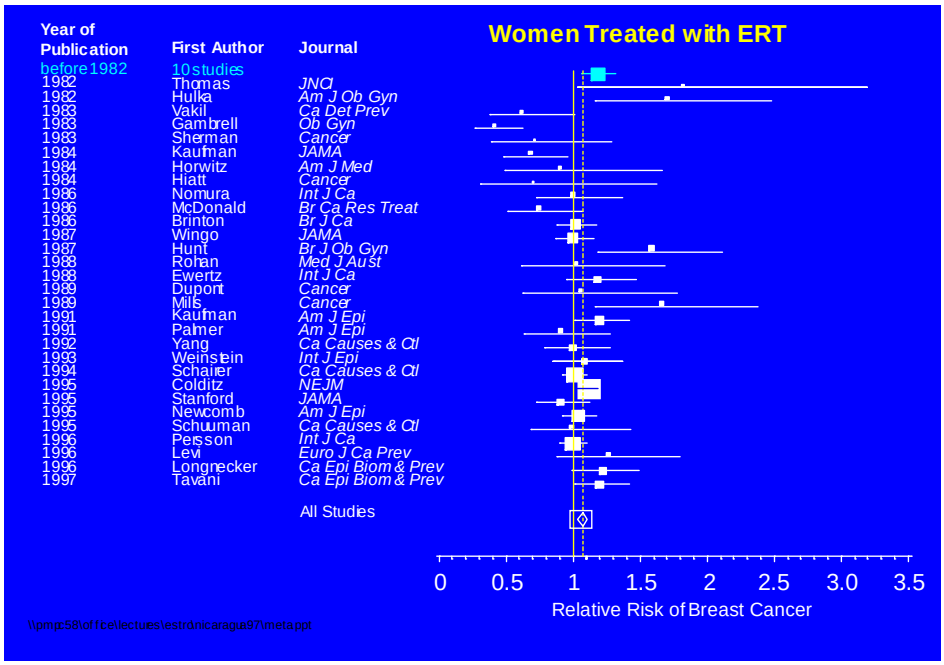
Small studies are more likely to be affected by publication bias than large ones.

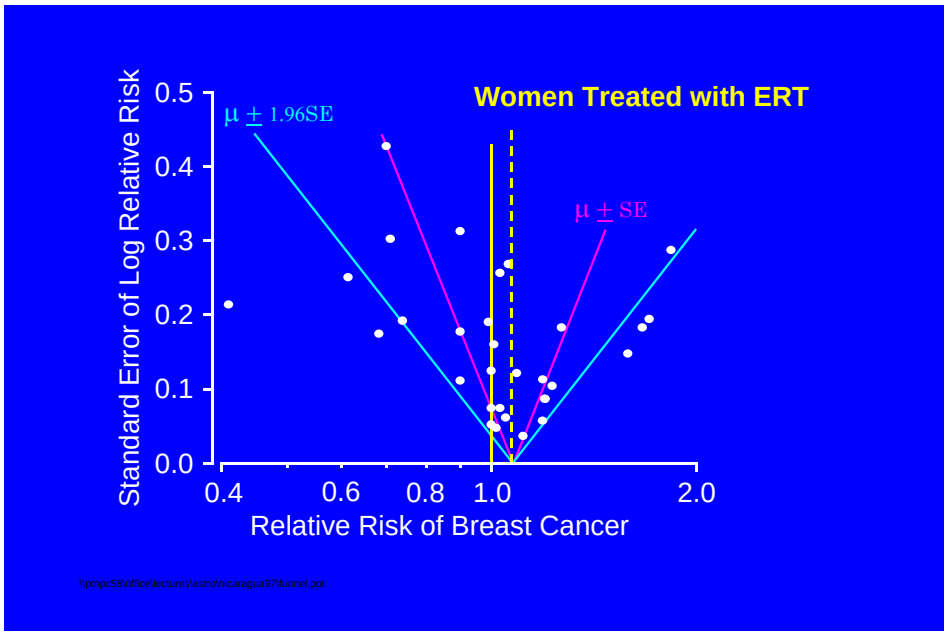
6. Funnel graphs

One way to check for publication bias is to plot funnel graphs (Light & Pillemer 1984).

In these graphs we plot the standard error of the log relative risk against log relative risk. If this plot has a funnel shape we have evidence of publication bias

When this happens it may make sense to exclude studies with a standard error of the log relative risk that is greater than some value.





Approaches to Extreme Multiple Comparisons Problems

- Permutation Tests
- Cross validation Methods
- False Discovery Rates
- Shrinkage Analysis
- Learning set Test Set Analyses

This course has been concerned with methods that are appropriate when the number of patients far exceeds the number of model parameters.

Diversity Among Statisticians

We all want to

Minimize probabilities of Type I errors

Minimize probabilities of Type II errors

All other things being equal, simple explanations are better than complex ones.

Science may be described as the art of systematic over-simplification — the art of discerning what we may with advantage omit.

Karl Popper

Today, reputable statisticians may disagree to some extent about the relative emphasis that should be placed on these three goals

XII. SUMMARY OF MULTIPLE REGRESSION METHODS

Table 1.1. Classification of Response Variables and Regression Models

Nature of Response Variable(s)	Model	Table in Appendix A	Chapters
One response per patient			
Continuous	Linear regression	A.1	2, 3, 10
Dichotomous	Logistic regression	A.2	4, 5
Categorical	Proportional odds and polytomous logistic regression	A.2	5
Survival	Hazard regression	A.3	6, 7
Rates	Poisson regression	A.4	8, 9
Multiple responses per patient			
Continuous	Response feature and generalized estimating equation analysis	A.5	11
Dichotomous	Response feature and generalized estimating equation analysis	A.5	11

Table A.1. Models for continuous response variables with one response per patient.

Model Attributes	Method of Analysis	Pages
Normally distributed response variable.		
Single continuous independent variable.		
Linear relationship between response and independent variable.	Simple linear regression.	47 – 99
Non-linear relationship between response and independent variable.	Multiple linear regression using restricted cubic splines.	138 – 159
	Transform response or independent variables and use simple linear regression.	75 – 84
	Convert continuous independent variable to dichotomous variables and use multiple linear regression.	222 – 231, 100 – 163
Single dichotomous independent variable.	Independent t-test.	36 – 41
Single categorical variable.	Convert categorical variable to dichotomous variables and use multiple linear regression.	222 – 231, 100 – 163
	One-way analysis of variance.	439 – 457

Table A1. Continued: continuous response, fixed effects

Model Attributes	Method of Analysis	Pages
Normally distributed response variable.		
Multiple independent variables.		
Independent variables have additive effects on response variable.	Multiple linear regression model without interaction terms.	100 – 124
Independent variables have non-additive effects on response variable.	Include interaction terms in multiple linear regression model.	111 – 114
Independent variables are categorical or have non-linear effects on the response variable.	Multiple linear regression: see above for single independent variable.	100 – 163
Two independent categorical variables.	Two-way analysis of variance.	457 – 458
Multiple categorical and continuous independent variables.	Analysis of covariance. This is another name for multiple linear regression.	100 – 163

Table A1. Continued: continuous response, fixed effects

Model Attributes	Method of Analysis	Pages
Skewed response variable.		
Single dichotomous independent variable.	Wilcoxon-Mann-Whitney rank-sum test.	446
Single categorical independent variable.	Kruskal-Wallis test.	445 – 446
Any combination of independent variables.	Apply normalizing transformation to response variable. Then see methods for linear regression noted above.	75 – 84

Table A.2. Models for dichotomous or categorical response variables with one response per patient.

Model Attributes	Method of Analysis	Pages
Dichotomous response variable.		
Single continuous independent variable.		
Linear relationship between log-odds of response and independent variable.	Simple logistic regression.	164 – 206
Non-linear relationship between log-odds of response and independent variable.	Multiple logistic regression using restricted cubic splines.	271 – 285
	Transform independent variable. Then use simple logistic regression.	75 – 84, 164 – 206
	Convert continuous variable to dichotomous variables and use multiple logistic regression.	222 – 230
Single dichotomous independent variable.	2×2 contingency table analysis. Calculate crude odds ratio.	193 – 197
	Simple logistic regression.	197 – 203
Single categorical variable.	Convert categorical variable to dichotomous variables and use multiple logistic regression.	222 – 231

Table A2. Continued: dichotomous response, fixed effects

Model Attributes	Method of Analysis	Pages
Dichotomous response variable.		
Multiple independent variables.		
Two dichotomous independent variables with multiplicative effects on the odds ratios.	Mantel-Haenszel odds-ratio and test for multiple 2×2 tables.	207 – 216
	Multiple logistic regression.	218 – 224
Independent variables have multiplicative effects on the odds-ratios.	Multiple logistic regression model without interaction terms.	216 – 238
Independent variables have non-multiplicative effects on the odds-ratios.	Include interaction terms in multiple logistic regression model.	238 – 244
Independent variables are categorical or have non- linear effects on the log odds.	Multiple logistic regression. See above for single independent variable.	222 – 231, 271 – 285, 75 – 84
Matched cases and controls.	Conditional logistic regression.	264 – 265

Table A2. Continued: categorical response, fixed effects

Model Attributes	Method of Analysis	Pages
Dichotomous response variable.		
Categorical response variable.		
Response categories are ordered and proportional odds assumption is valid.	Proportional odds logistic regression.	285 – 287
Response categories not ordered or proportional odds assumption invalid.	Polytomous logistic regression.	287 – 289
Independent variables have non-multiplicative effects on the odds-ratios, are categorical or have non-linear effects on the log odds.	See above for logistic regression.	238 – 244, 222 – 231, 271 – 285, 75 – 84

Table A.3. Models for survival data (follow-up time plus fate at exit observed on each patient).

Model Attributes	Method of Analysis	Pages
Categorical independent variable.	Kaplan-Meier survival curve.	298 – 305
	Log-rank test.	305 – 314
Proportional hazards assumption valid.		
Single continuous independent variable.		
Linear relationship between log-hazard and independent variable.	Simple proportional hazards regression model.	315 – 321
Non-linear relationship between log-hazard and independent variable.	Multiple proportional hazards model using restricted cubic splines.	329 – 332, 341 – 357
	Transform independent variable. Then use simple proportional hazards model.	75 – 84, 315 – 321
	Convert continuous variable to dichotomous variables. Then use multiple proportional hazards regression model.	332 – 333, 341 – 357
Time denotes age rather than time since recruitment.	Proportional hazards regression analysis with ragged entry.	358 – 363

Table A3. Continued: survival data

Model Attributes	Method of Analysis	Pages
Proportional hazards assumption valid.		
Single categorical independent variable.	Convert categorical variable to dichotomous variables and use multiple proportional hazards regression model.	222 – 224, 332 – 333, 341 – 357
Multiple independent variables.		324 – 368
Independent variables have non-multiplicative effects on the hazard ratios.	Include interaction terms in multiple proportional hazards regression.	336 – 337, 341 – 357
Independent variables are categorical or have non-linear effects on the log-hazard.	Multiple proportional hazards regression. See above for single independent variable.	324 – 368

Table A3. Continued

Model Attributes	Method of Analysis	Pages
Proportional hazards assumption invalid.	Stratified proportional-hazards regression analysis.	357 – 358
	Hazard regression analysis with time-dependent covariates.	368 – 379
Events are rare and sample size is large.	Poisson regression.	393 – 436
Independent variables have non-multiplicative effects on the hazard ratios.	Include interaction terms in time-dependent hazard regression model.	336 – 337, 368 – 379
Independent variables have non-linear effects on the log-hazard.	See above for a single continuous independent variable. Use a time-dependent hazard regression model.	329 – 332, 75 – 84, 368 – 379
Independent variables are categorical.	Convert categorical variables to dichotomous variables in time-dependent model.	332 – 333, 368 – 379
Time denotes age rather than time since recruitment	Hazards regression analysis with time-dependent covariates and ragged entry.	358 – 363, 368 – 379

Table A.4. Models for response variables that are event rates or the number of events during a specified number of patient-years of follow-up. The event must be rare.

Model Attributes	Method of Analysis	Pages
Single dichotomous independent variable.	Incident rate ratios.	383 – 386
	Simple Poisson regression.	387 – 391
Single categorical independent variable.	Convert categorical variable to dichotomous variables and use multiple Poisson regression.	222 – 224, 414 – 432
Multiple independent variables.		
Independent variables have multiplicative effects on the event rates.	Multiple Poisson regression models without interaction terms.	411 – 417
Independent variables have non-multiplicative effects on the event rates.	Multiple Poisson regression models with interaction terms.	417 – 432
Independent variables are categorical.	Multiple Poisson regression. See above for single independent variable	222 – 224, 414 – 432

Table A.5. Models with multiple observations per patient or matched or clustered patients.

Model Attributes	Method of Analysis	Pages
Continuous response measures.		
Dichotomous independent variable.	Paired *t*-test.	33 – 36
Multiple independent variables.		
	Response feature analysis: consider slopes of individual patient regressions or areas under individual patient curves.	469 – 479
	GEE analysis with identity link function and normal random component.	479 – 491
Dichotomous response measure.		
Multiple independent variables	Response feature analysis: consider within-patient event rate.	470
	GEE analysis with logit link function and binomial random component.	491

Problem	**Method**
Cross-sectional Study	
Continuous outcome	
Normally distributed	
Linear model ok	Linear regression Fixed-effects analysis of variance
Non-linear model	Linear model of transformed data Linear model with restricted cubic splines
Skewed response data	Linear model of transformed data
Dichotomous outcome	Logistic regression
Rare response	Poisson regression
Longitudinal Data	
	Response feature analysis Repeated measures analysis of variance Generalized estimating equation analysis

Problem	Method
Cohort Study	
Proportional hazards assumption ok	Hazard regression
Rare events	Poisson regression
Ragged entry	Proportional hazard regression with ragged entry times
Expensive data collection	Logistic regression (Nested case-control study)
Complete follow-up with time. to failure not important	Logistic regression
Proportional hazards invalid	
Entry uniform or ragged	Stratified hazard regression Time dependent hazard regression Poisson regression
Large study: proportional hazards assumption invalid	Poisson regression
Case-Control Study	
Unstratified or large strata	Unconditional logistic regression
Small strata	Conditional logistic regression

Additional Reading

A good reference for the response-compression approach to mixed-effects analysis of variance is Matthews et al. (1990).

Classic although rather mathematical references for generalized estimating equations are Liang and Zeger (1986) and Zeger and Liang (1986). Diggle et al. (2002) is an authoritative text on the analysis of longitudinal data.

Armitage and Berry (1994) discuss receiver operating characteristic curves.

Classification and regression trees are discussed by Breiman et al. (1984).

An introduction to neural networks is given by Hinton (1992). A comparison of neural nets with classification and regression trees is given by Reibnegger et al. (1991)

An introduction to meta-analysis is given by Greenland (1987). This paper also describes the fixed effects method of calculating a confidence interval for the combined relative risk estimate. The random effects method is given by DerSimonian and Laird (1986).

Harrell (2001) is an advanced text on modern regression methods.

References

Armitage P and Berry G: *Statistical Methods in Medical Research*, Third ed. Cambridge, MA: Blackwell Science, Inc., 1994.

Bernard GR, Wheeler AP, Russell JA, Schein R, Summer WR, Steinberg KP, Fulkerson WJ, Wright PE, Christman BW, Dupont WD, Higgins SB, Swindell BB: "The Effects of Ibuprofen on the Physiology and Survival of Patients with Sepsis". *New England Journal of Medicine,* 1997; 336: 912-918

Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Belmont CA: Wadsworth, 1984.

Breslow NE and Day NE: *Statistical Methods in Cancer Research: Vol. I The Analysis of Case-Control Studies*. Lyon: IARC Scientific Publications, 1980.

Breslow NE and Day NE: *Statistical Methods in Cancer Research: Vol. II. The Design and Analysis of Cohort Studies*. Lyon: IARC Scientific Publications, 1987.

Cleveland WS. *The Elements of Graphing Data*: Monterey, CA: Wadsworth Advanced Books and Software, Bell Telephone Laboratories, Inc., 1985.

DerSimonian R. Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; 7:177-188.

Diggle PJ, Heagerty P, Liang K-Y, Zeger SL: *Analysis of Longitudinal Data 2nd Ed.* Oxford: Oxford University Press, 2002

Fleiss JL: *Statistical Methods for Rates and Proportions, Second ed.:* New York: John Wiley & Sons, Inc., 1981.

Greene J and Touchstone J. "Urinary Tract Estriol: An Index of Placental Function. *American Journal of Obstetrics and Gynecology*, 1963; 85: 1-9.

Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiologic Reviews* 1987; 9:1-30.

Harrell, FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* New York: Springer, 2001

Hinton GE. How neural networks learn from experience. *Scientific American* September 1992; p.145-151.

Kalbfleisch JD and Prentice RL: *The Statistical Anallysis of Failure Time Data*, New York: John Wiley and Sons, 1980.

Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73:13-22.

Liang K-Y, Zeger SL. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; 42:121-130.

Light RJ, Pillemer DB. *Summing Up: The Science of Reviewing Research.* Cambridge, MA: Harvard University Press. 1984.

Matthews JNS, Altman DG, Campbell MJ, Royston P. Analysis of serial measurements in medical research. *British Medical Journal* 1990;300:230-235.

McCullagh P and Nelder JA: *Generalized Linear Models, Second ed.*: New York: Chapman and Hall, 1989.

McKelvey EM, Gottlieb JA, Wilson HE, Haut A, Talley RW, et al.: "Hydroxyldaunomycin (Adriamycin) Combination Chemotherapy in malignant lymphoma. *Cancer,* 1976; 38: 1484-1493.

Pagano M and Gauvreau K, *Principles of Biostatistics*, Belmont, CA: Duxbury Press, 1993.

Reibnegger G, Weiss G, Werner-Felmayer G, Judmaier G, Wachter H. Neural networks as a tool for utilizing laboratory information: Comparison with linear discriminant analysis and with classification and regression trees. *Proc. Natl. Acad. Sci. USA* 1991; 88:11426-11430.

Rosner B: *Fundamentals of Biostatistics, Fourth ed.*: Belmont, CA: Duxbury Press, 1995.

Schottenfeld D and Fraumeneni JF: *Cancer Epidemiology and Prevention:* Philadelphia, PA: W.B. Saunders Company, 1982.

Tuyns AJ, Pequignot G, and Jensen OM. "Le Cancer de l'oesophage en Ille-et-Villaine en fonction des niveaux de consommation d'alcool et de tabac. *Bulletin du Cancer*, 1977: 64: 45-60.

For additional references see

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