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

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

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

Hence a 95% confidence interval for R is

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

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

= [1.66, 2.04]

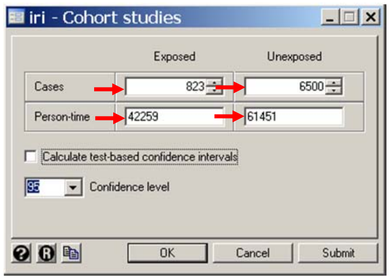
In Stata these calculations are done as follows:

```
. * 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.

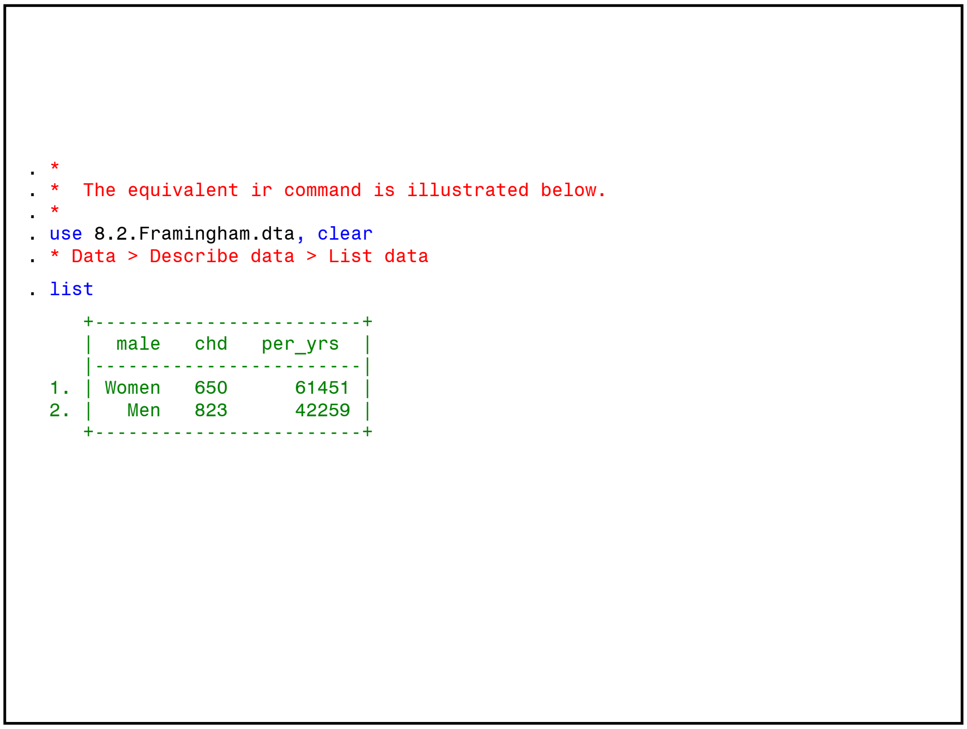
These data are the number exposed and unexposed cases together with the person-years of follow of exposed and unexposed subjects.


iri - Cohort studies

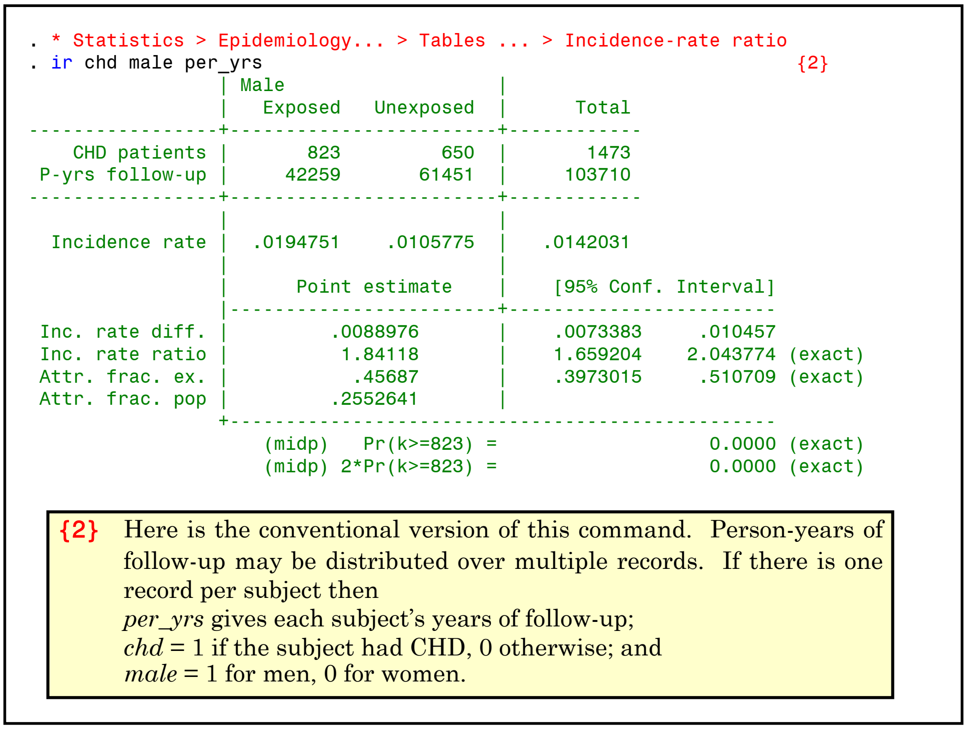
	Exposed	Unexposed
Cases	823	6500
Person-time	42259	61451

□ Calculate test-based confidence intervals

95 Confidence level

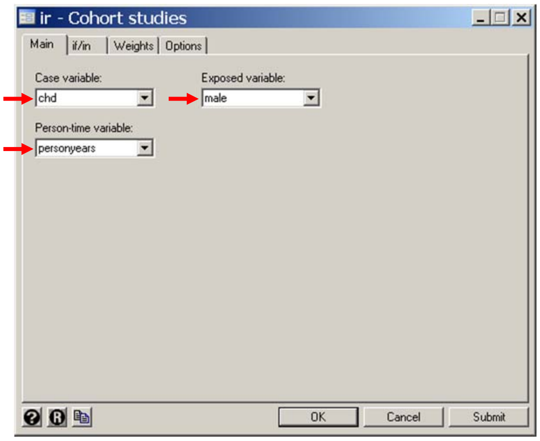


```
. *
. *  The equivalent ir command is illustrated below.
. *
. use 8.2.Framingham.dta, clear
. * Data > Describe data > List data

. 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.



Case variable: chd

Exposed variable: male

Person-time variable: personyears

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

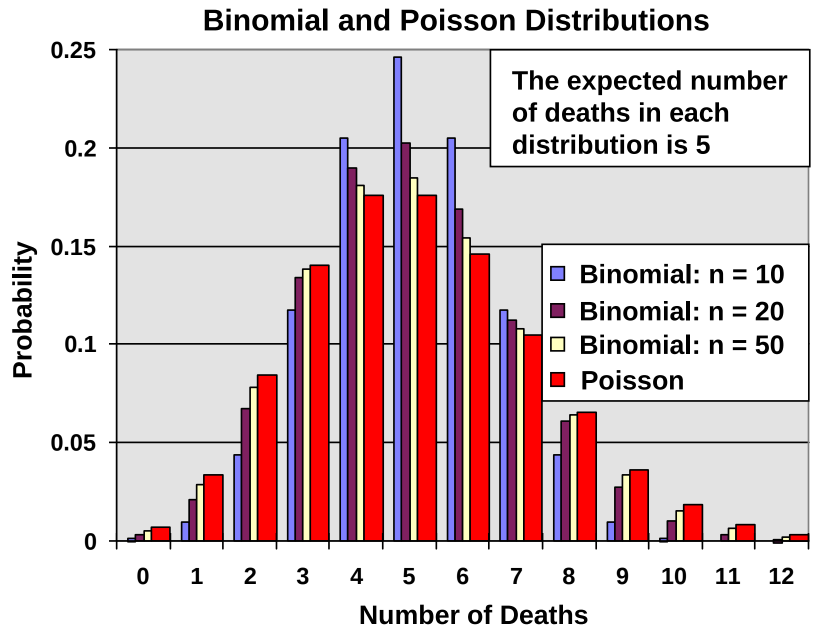
Pr[d deaths]

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

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 $n\pi = \lambda$.

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

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

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

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

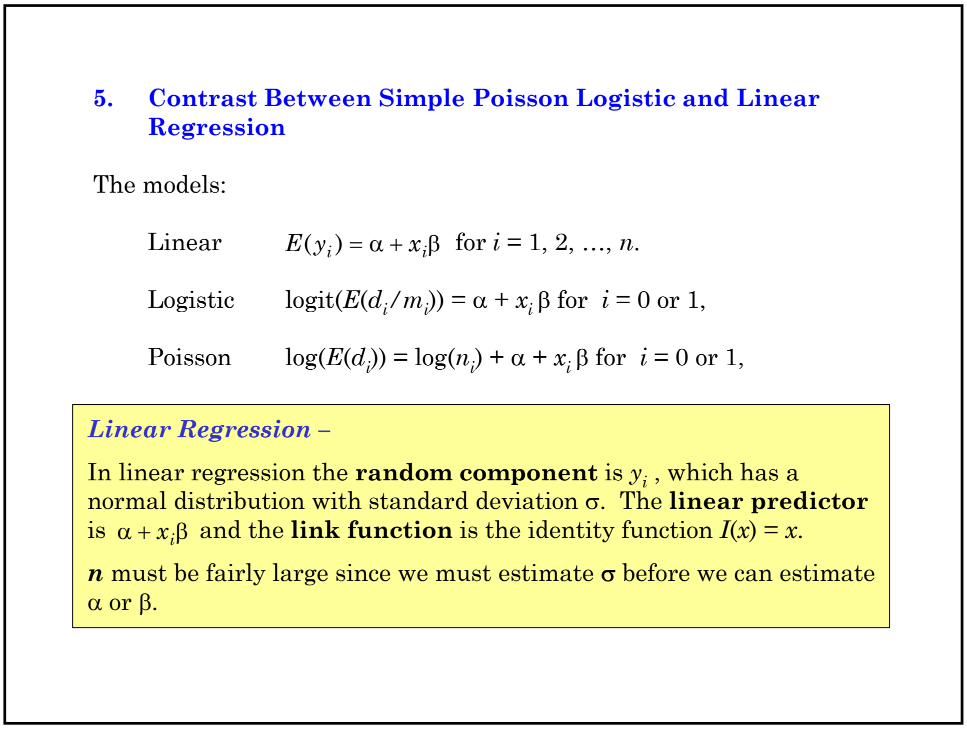
d_i is the **random component** of the model. In Poisson regression, d_i has a Poisson distribution with mean $E(d_i)$.

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 $i = 1, 2, \ldots, n$.

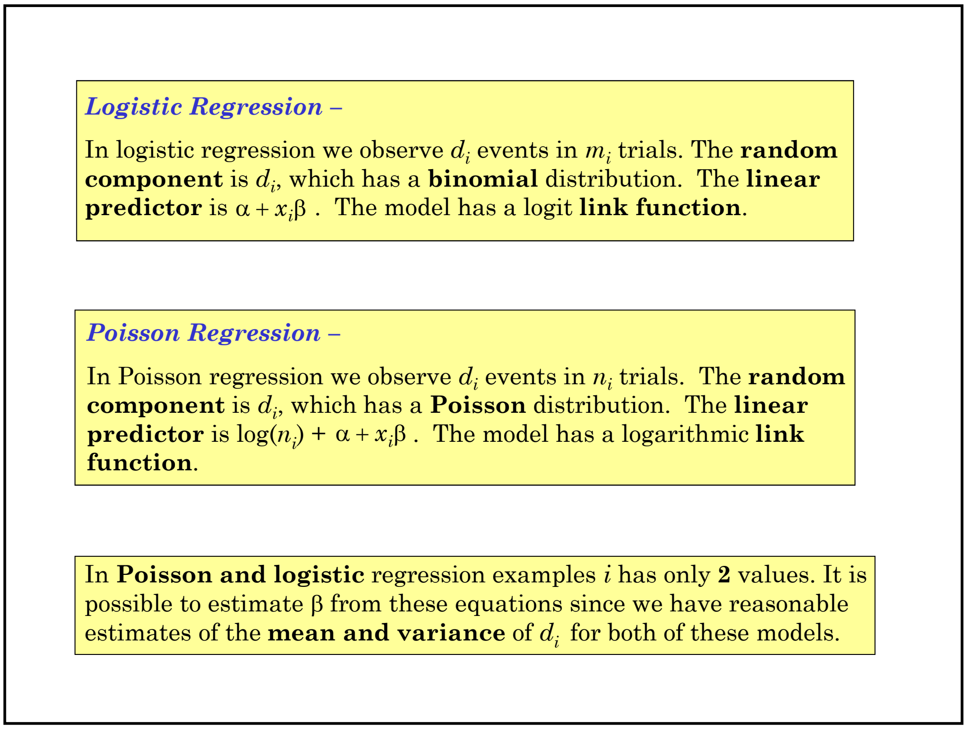
Logistic $\text{logit}(E(d_i/m_i)) = \alpha + x_i\beta$ for $i = 0$ or 1,

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 $\boldsymbol{\sigma}$ 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.

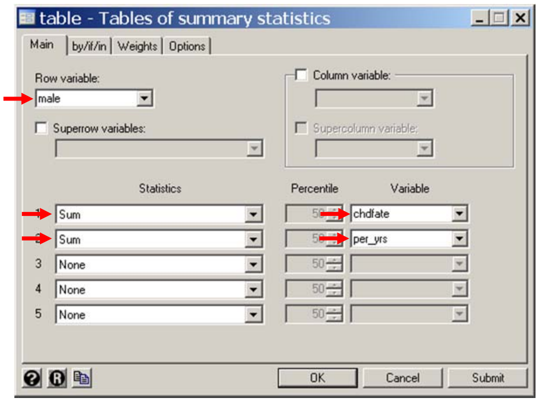
Similarly, β_3 equals the log relative risk of patients in group 3 relative to group 1.

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.



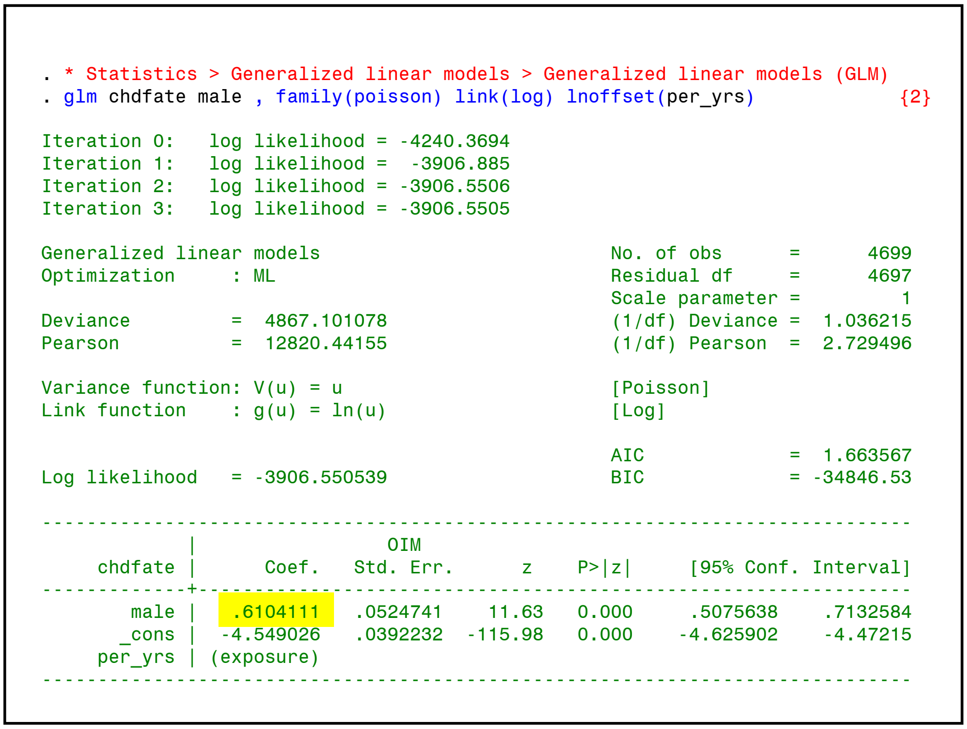
Row variable: male

□ Superrow variables:

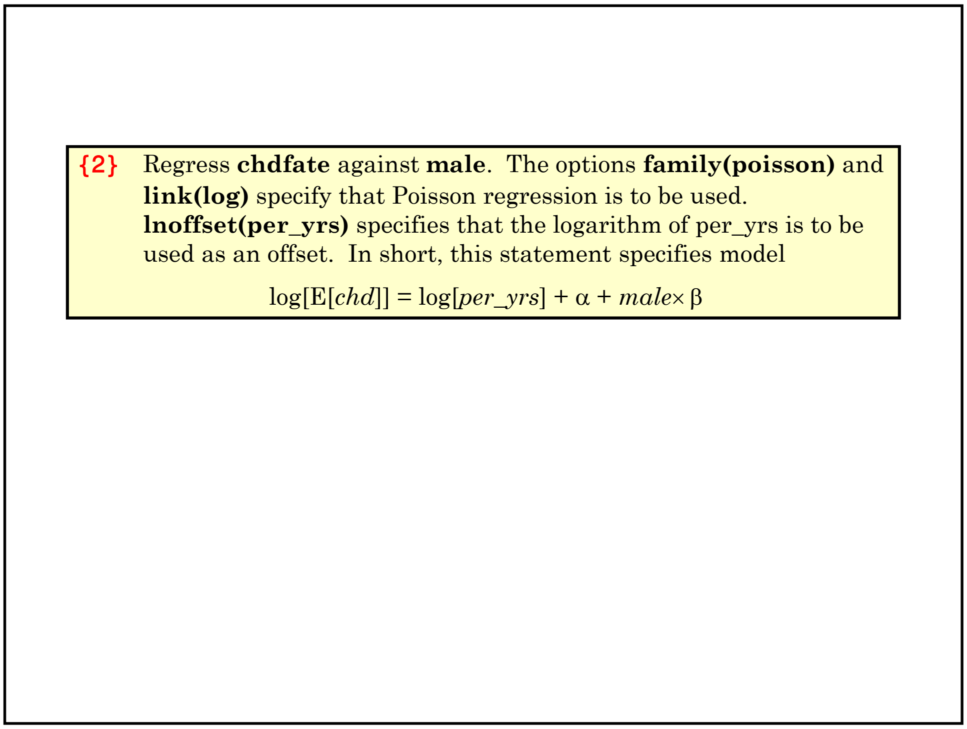
□ Column variable:

□ Supercolumn variable:

	Statistics	Percentile	Variable
1	Sum	50	chdfate
2	Sum	50	per_yrs
3	None	50	
4	None	50	
5	None	50	

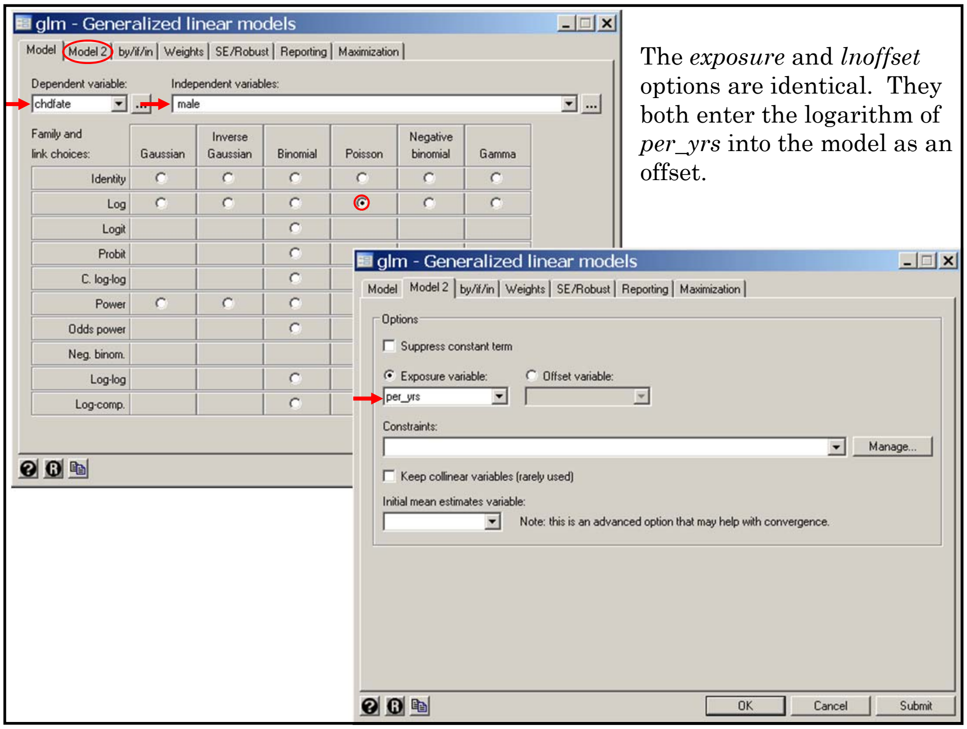


```
. * Statistics > Generalized linear models > Generalized linear models (GLM)
. glm chdfate male , family(poisson) link(log) lnoffset(per_yrs)           {2}

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

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



Dependent variable: chdfate

Independent variables: male

Family and link choices: ● Poisson, Log

● Exposure variable: per_yrs

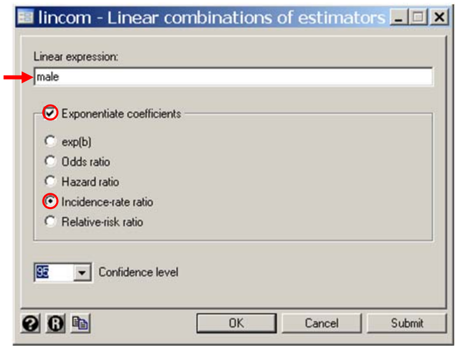
○ Offset variable:

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.



lincom - Linear combinations of estimators

Linear expression: male

☑ Exponentiate coefficients

○ exp(b)
○ Odds ratio
○ Hazard ratio
● Incidence-rate ratio
○ Relative-risk ratio

95 Confidence level

```
. * 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 ± 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

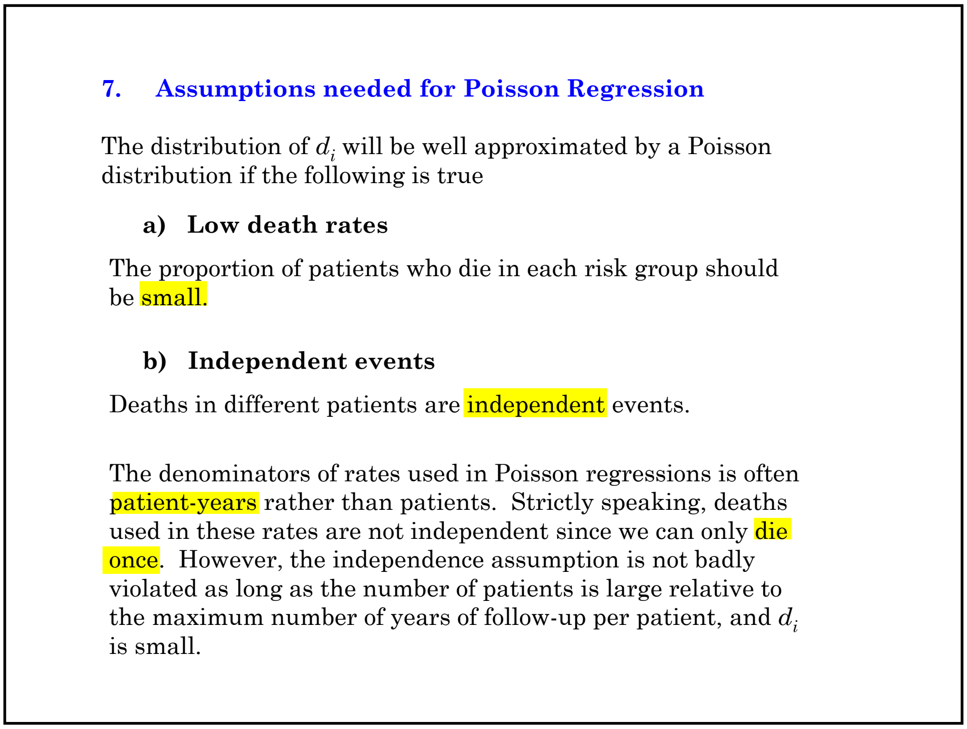
The classical and Poisson relative risk estimates are in exact agreement.

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

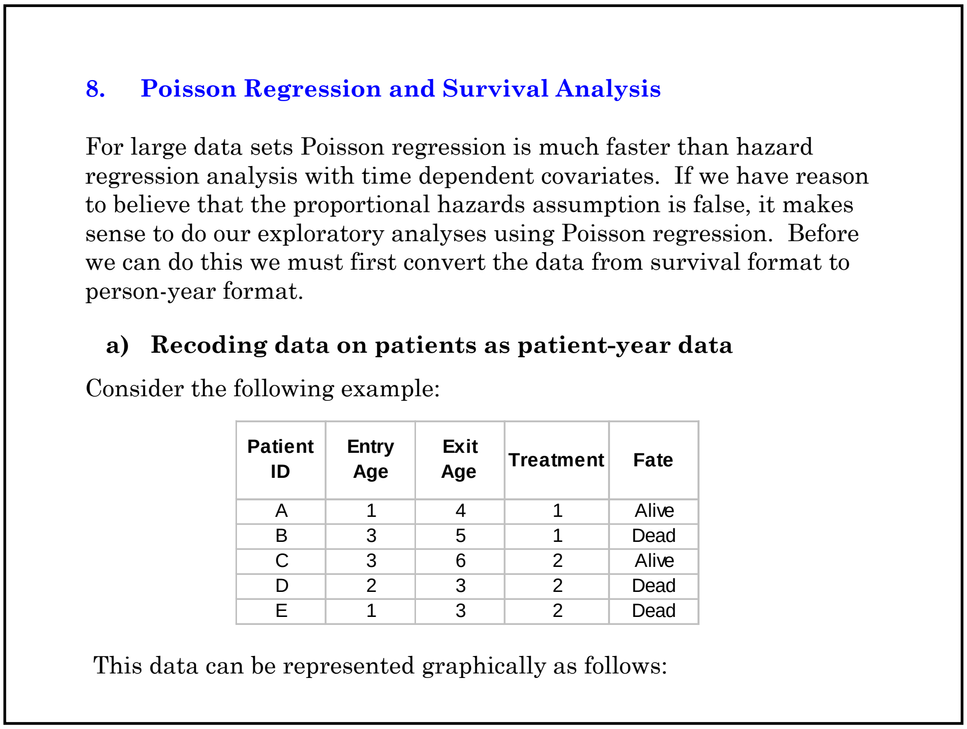
a) Low death rates

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

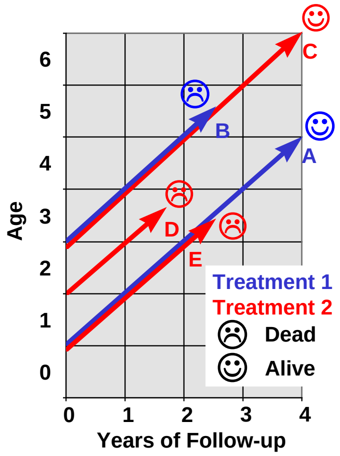
For large data sets Poisson regression is much faster than hazard regression analysis with time dependent covariates. If we have reason to believe that the proportional hazards assumption is false, it makes sense to do our exploratory analyses using Poisson regression. Before we can do this we must first 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
0	0	1	0
1	1	1	0
2	0	1	0
2	0	3	2
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 |
     +-------------------------------------------------------+
```

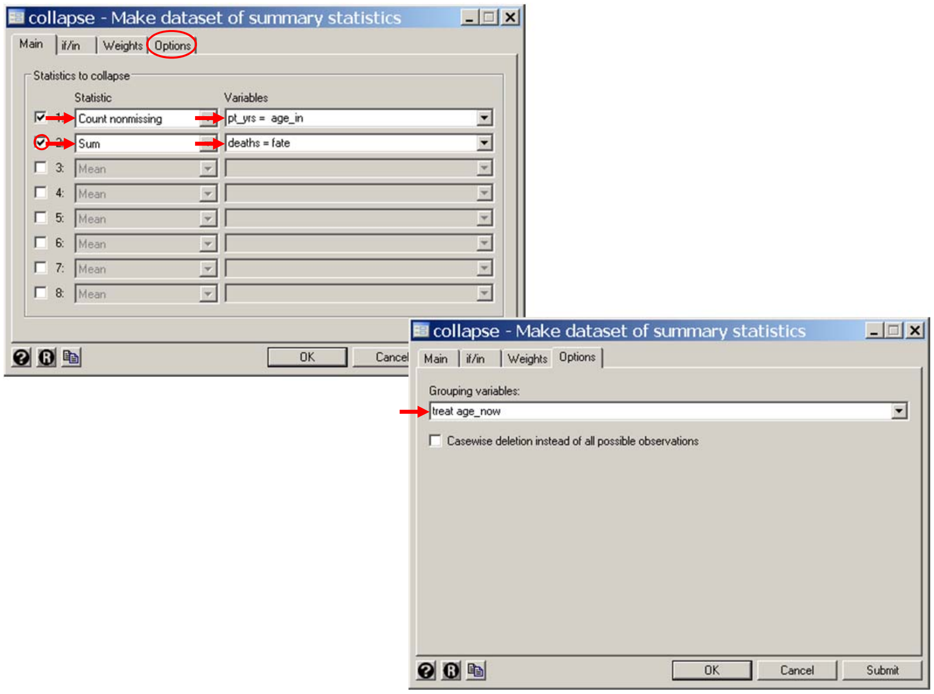
{3} There is now one record for each year of life that each patient had complete or partial follow-up. *age_now* equals *age_in* in each patient's first record and is incremented sequentially in subsequent records. It equals *age_out* at the last record.

{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***.

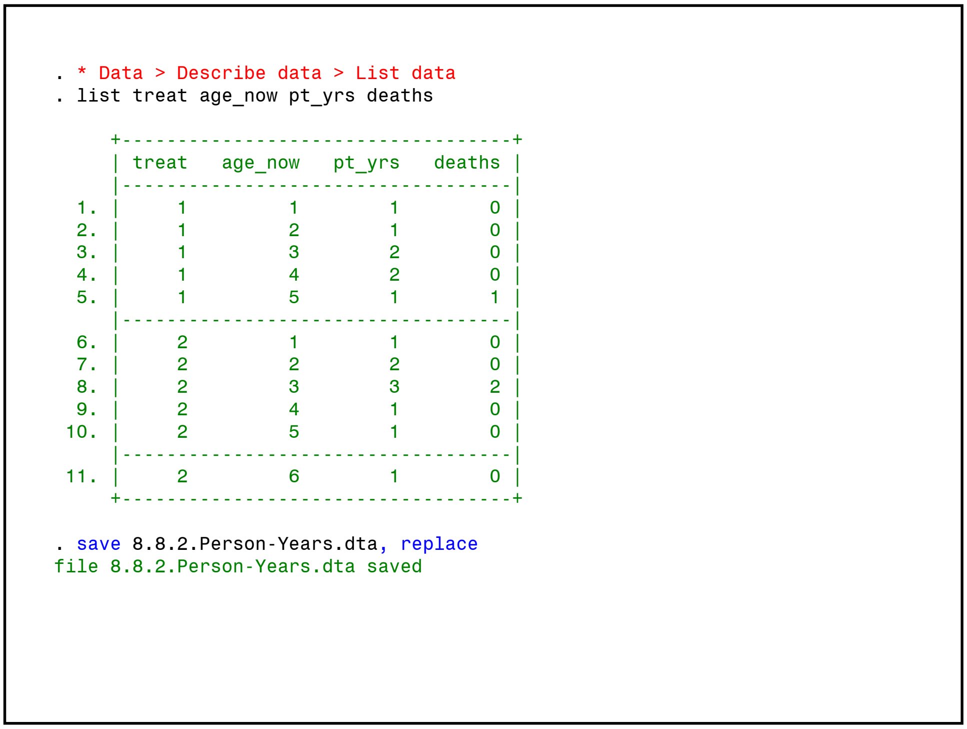


☑ Count nonmissing — pt_yrs = age_in

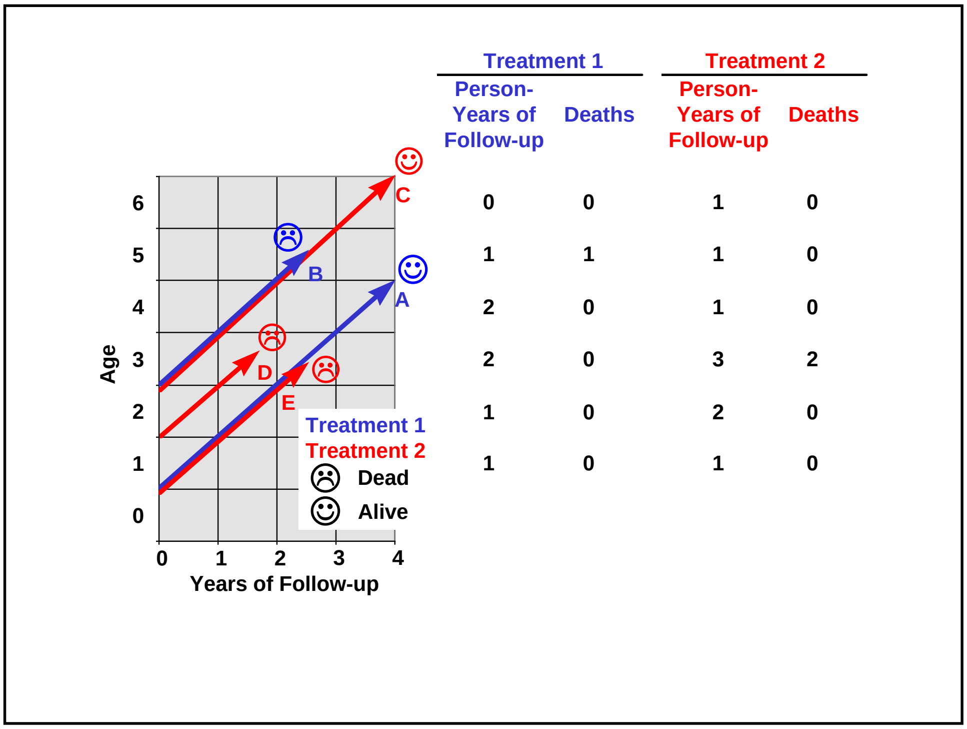
☑ 2: Sum — deaths = fate

Grouping variables: treat age_now

□ Casewise deletion instead of all possible observations



```
. * 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
0	0	1	0
1	1	1	0
2	0	1	0
2	0	3	2
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.

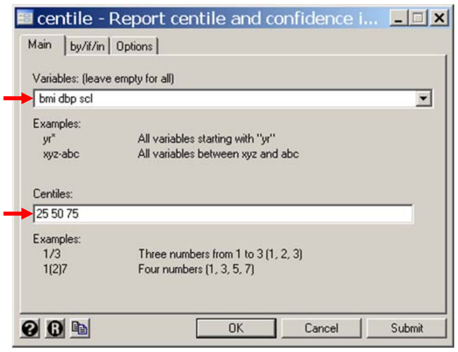
10. Converting the Framingham Survival Data to Person-time Data

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
```


centile - Report centile and confidence i...
Main
by/if/in
Options
Examples:
yr*
All variables starting with "yr"
xyz-abc
All variables between xyz and abc
Examples:
1/3
Three numbers from 1 to 3 (1, 2, 3)
1(2)7
Four numbers (1, 3, 5, 7)
OK
Cancel
Submit

Variables: (leave empty for all) bmi dbp scl

Centiles: 25 50 75

```
. 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.

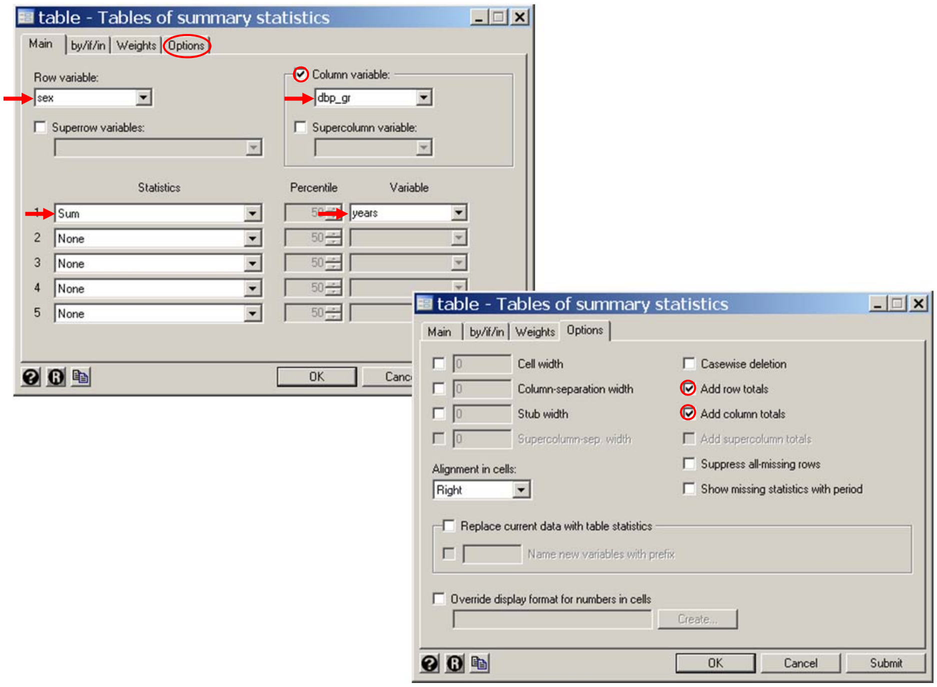


table - Tables of summary statistics

Row variable: sex

☑ Column variable: dbp_gr

□ Superrow variables:

□ Supercolumn variable:

	Statistics	Percentile	Variable
1	Sum	50	years
2	None	50	
3	None	50	
4	None	50	
5	None	50	

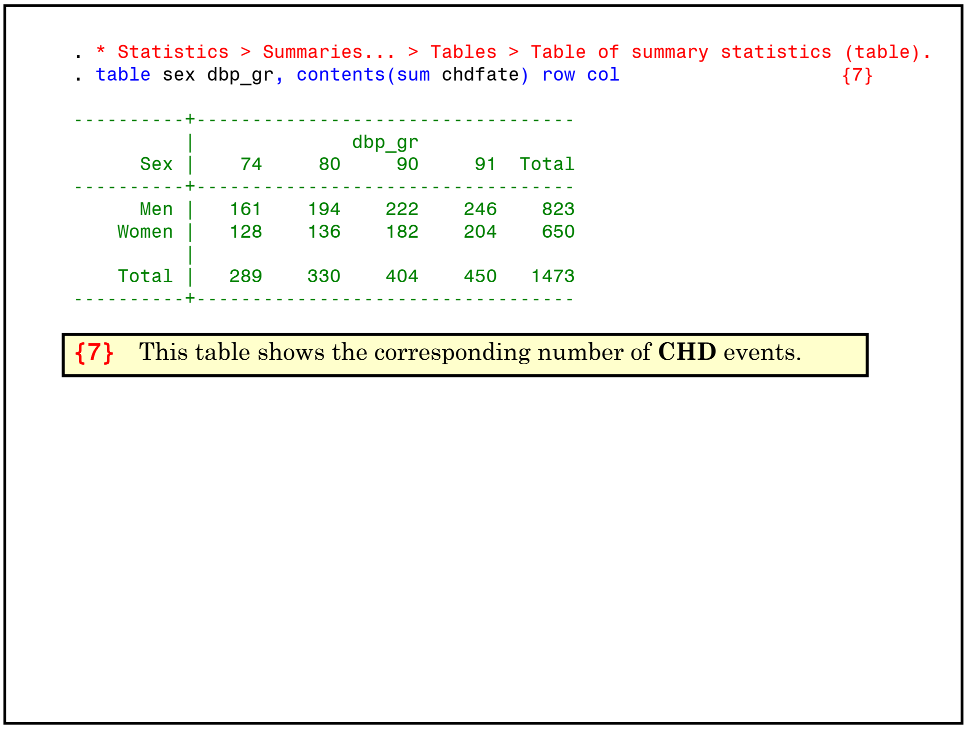
table - Tables of summary statistics (Options)

□ 0 Cell width
□ 0 Column-separation width
□ 0 Stub width
□ 0 Supercolumn-sep. width
Alignment in cells: Right

□ Casewise deletion
☑ Add row totals
☑ Add column totals
□ Add supercolumn totals
□ Suppress all-missing rows
□ Show missing statistics with period

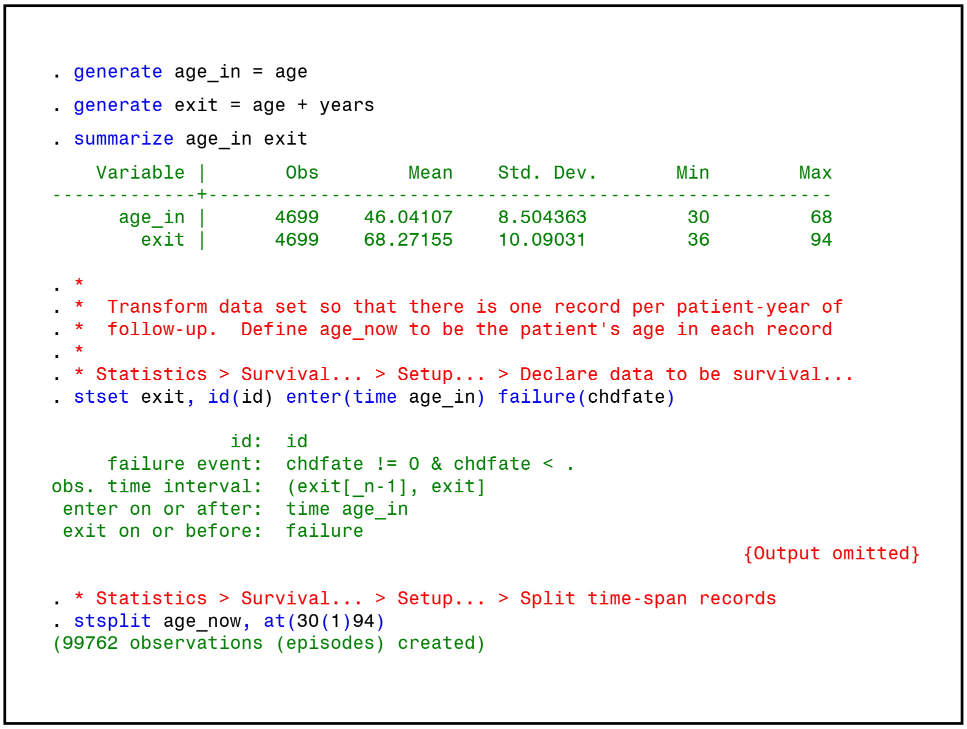
□ Replace current data with table statistics
□ Name new variables with prefix

□ Override display format for numbers in cells

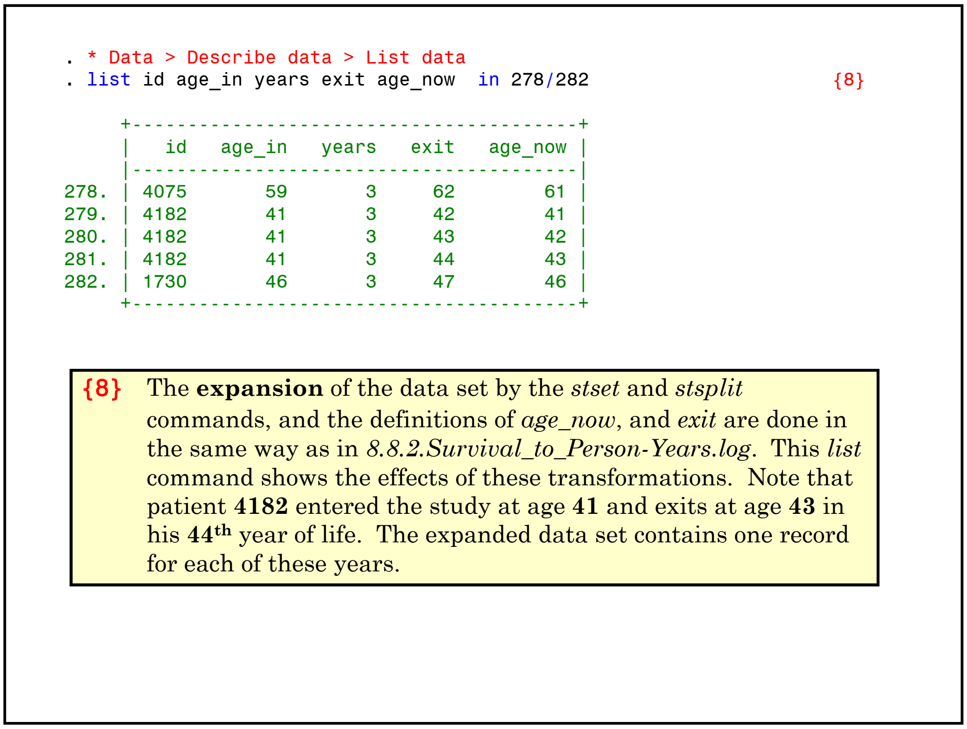


```
. * Statistics > Summaries... > Tables > Table of summary statistics (table).
. table sex dbp_gr, contents(sum chdfate) row col                     {7}

----------+--------------------------------
          |             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.

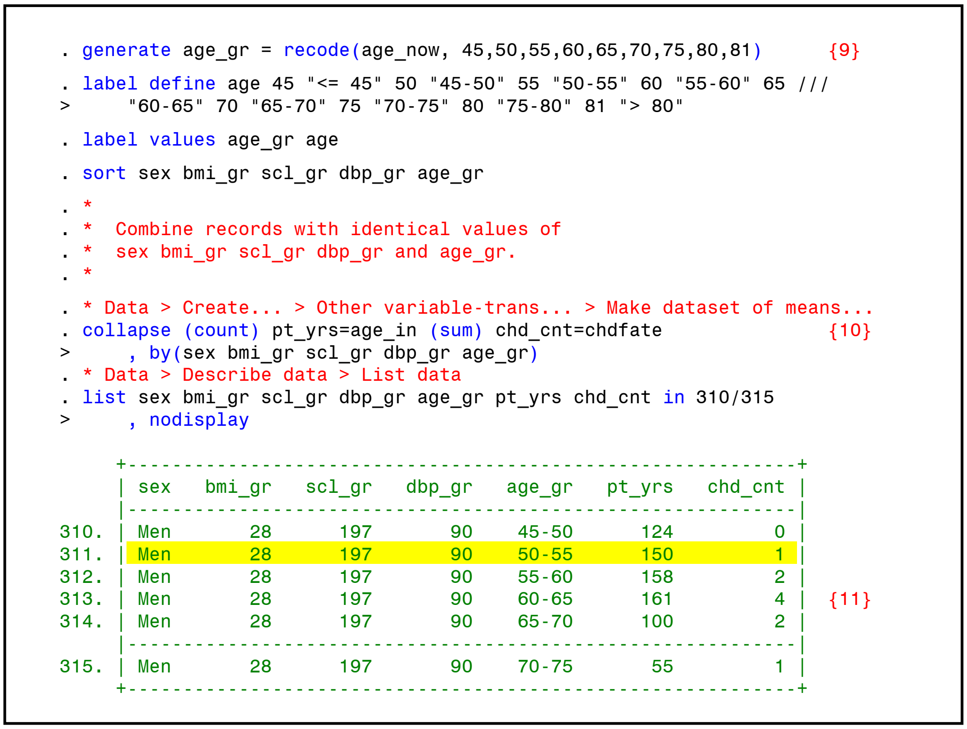


```
. generate age_in = age

. 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
```

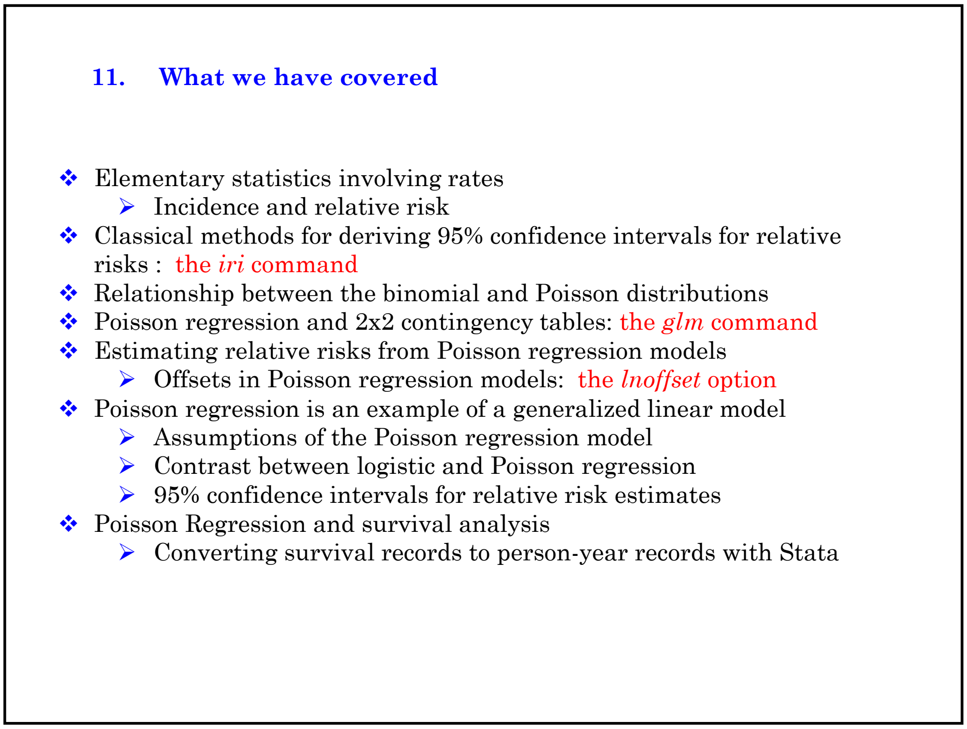
{12} This table shows total **person-years** of follow-up cross-tabulated by *sex* and *dbp_gr*. Note that this table is identical to the one produced before the data transformation

```
----------+--------------------------------------
          |                 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
----------+--------------------------------------
```

{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.

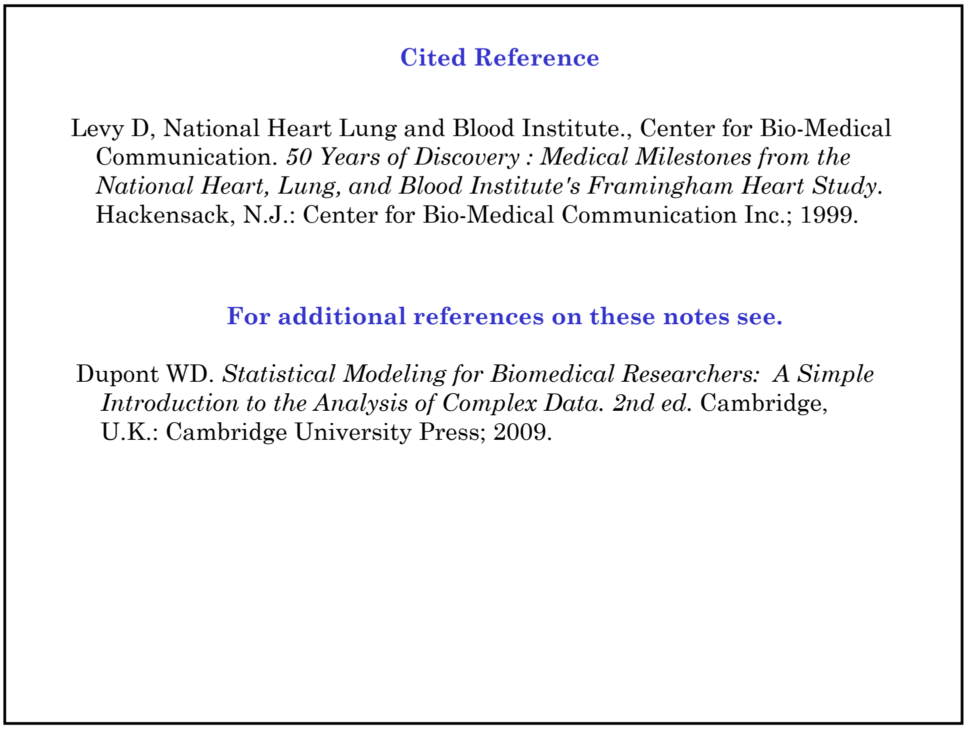
{14} The person-year data set is stored away for future analysis.

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.