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 $\qquad x_{jk1} = x_1 = \begin{cases} 1: \text{ Patient is heavy drinker} \\ 0: \text{ Patient is light drinker} \end{cases}$

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

Then the Poisson regression model is

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

where

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

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

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

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

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

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

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

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

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

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	≤ 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

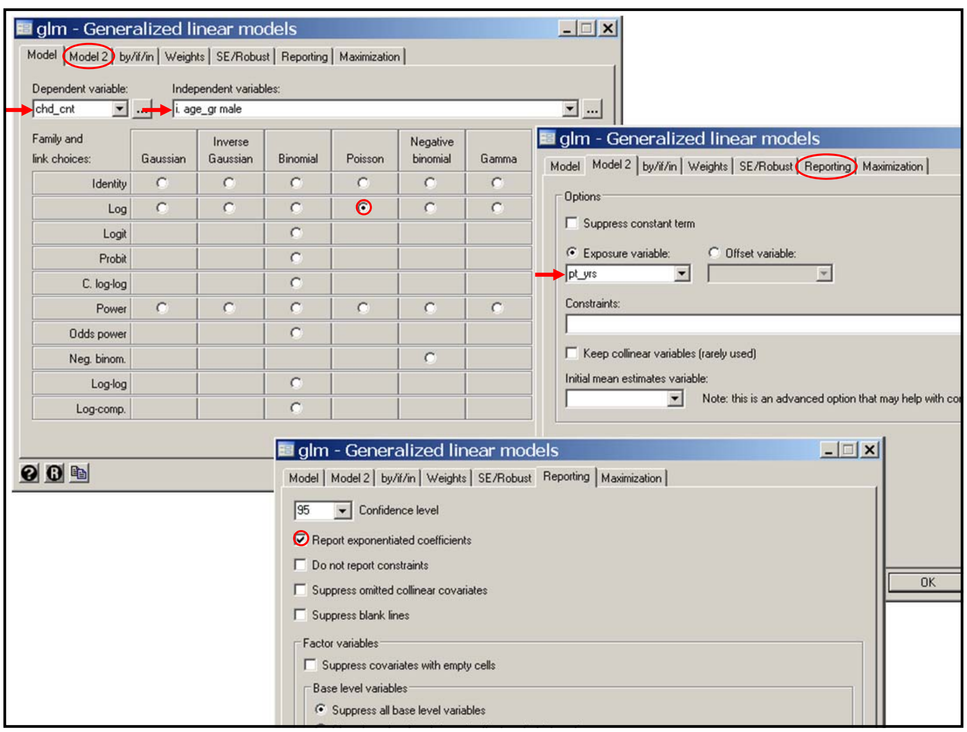
1 CHD event to the record for his covariate values and age 51 – 55

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.

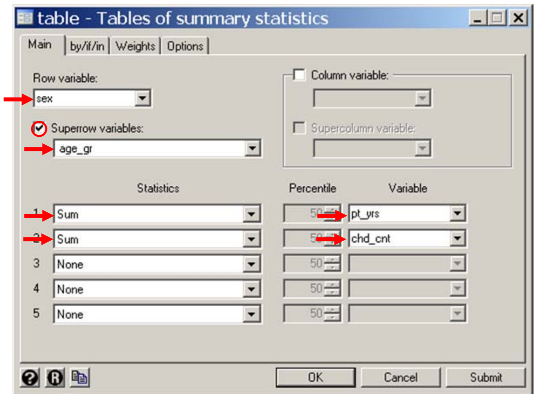
This risk is of limited interest because we know from Chapter VI that there is a powerful interaction between age and gender on coronary heart disease.

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.

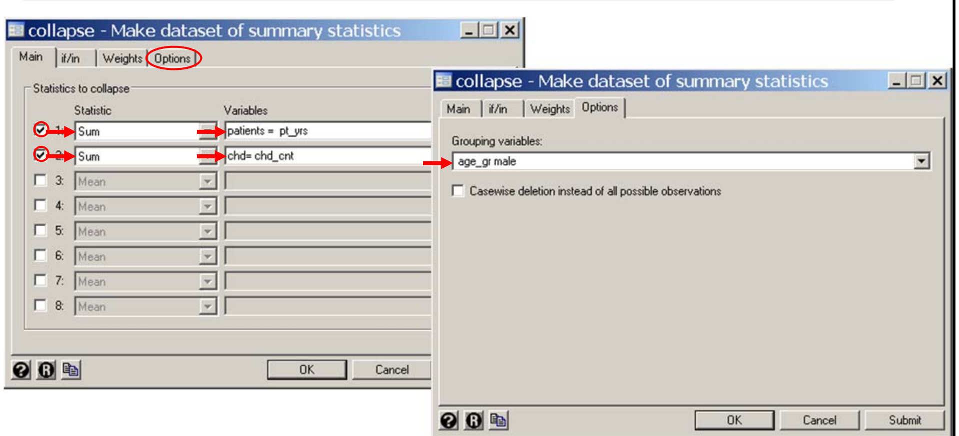
```
. *
. *  Tabulate patient-years of follow-up and number of
. *  CHD events by sex and age group.
. *
. * Statistics > Summaries... > Tables > Table of summary statistics (table)
. table sex, contents(sum pt_yrs sum chd_cnt) by(age_gr)
          ----------+
          Age Group |
          and Sex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

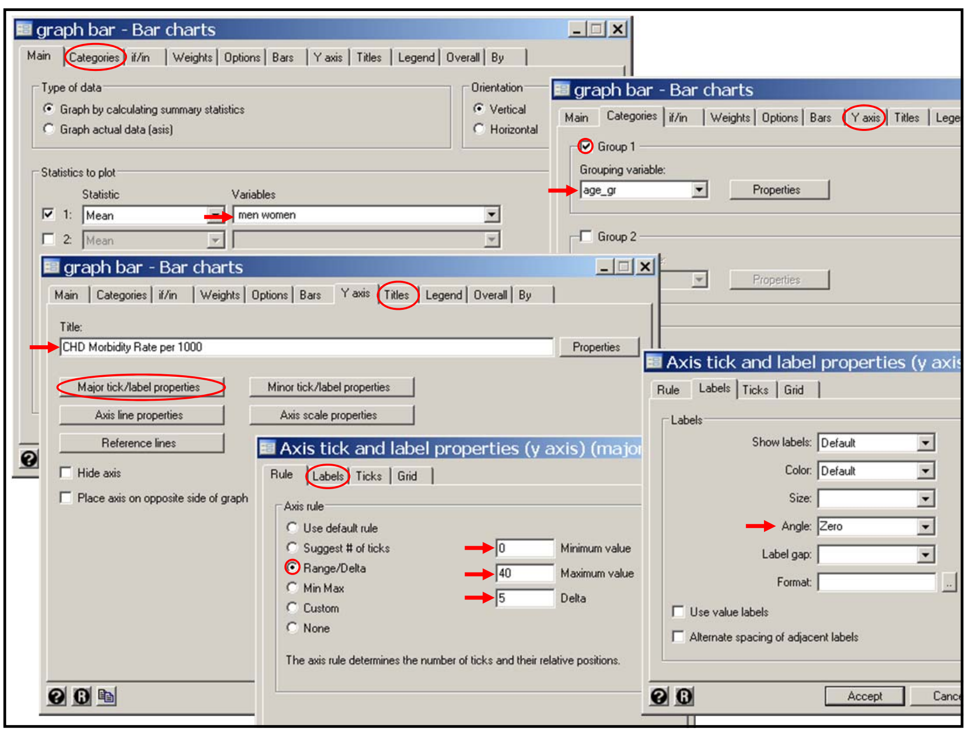
{1} Collapse the data file to one record for each combination of ***age_gr*** and ***sex***. Let ***patients*** be the total number of patient-years of follow-up and let ***chd*** be the total number CHD events in these groups.

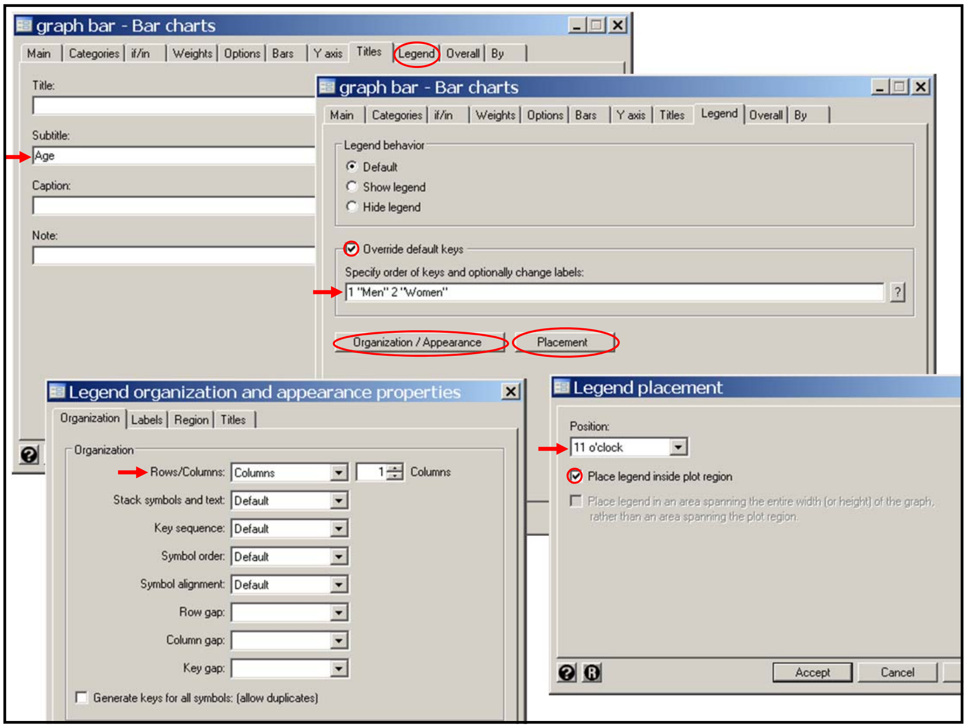


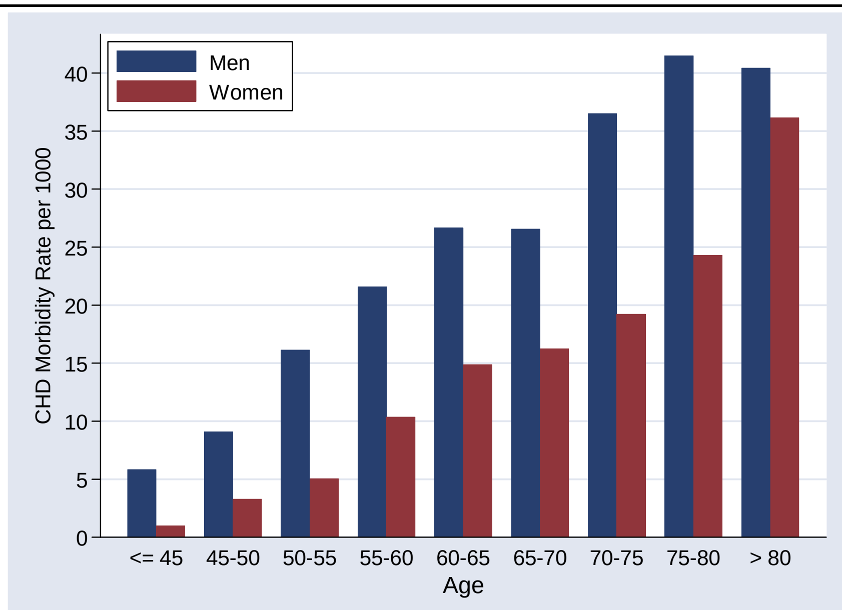
```
. generate rate = 1000*chd/patients                                    {2}

. 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 {8.4}

For women, the corresponding log incidence is

_cons + 50.age_gr {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 $E\left(d_{jk}\right)$ 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.

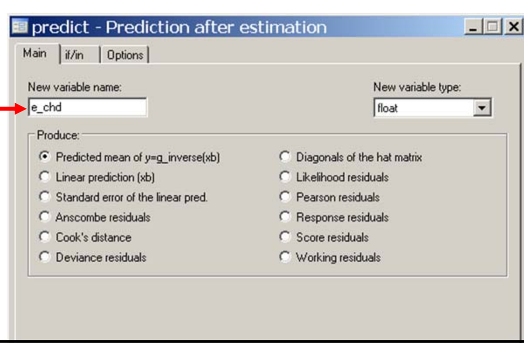
```
. *
. *  Re-analyze previous model using collapsed data set.
. *
. * Statistics > Generalized linear models > Generalized linear models (GLM)
. glm chd_cnt age_gr2##male i.bmi_gr10  i.scl_gr  i.dbp_gr       ///       {3}
> , family(poisson) link(log) lnoffset(pt_yrs)
   .
   .
   .
Generalized linear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

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

{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).

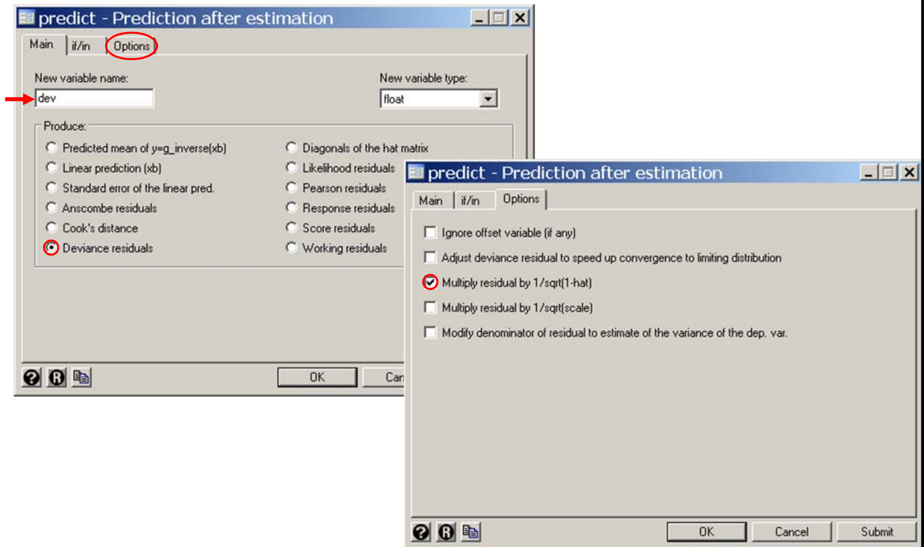
```
. *
. *  Estimate the expected number of CHD events and the
. *  standardized deviance residual for each record in the data set.
. *
. predict e_chd, mu                                                        {5}
(82 missing values generated)
```

{5} The ***mu*** option of this command defines ***e_chd*** to equal $\hat{\mathrm{E}}(d_{jk})$, 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)
```

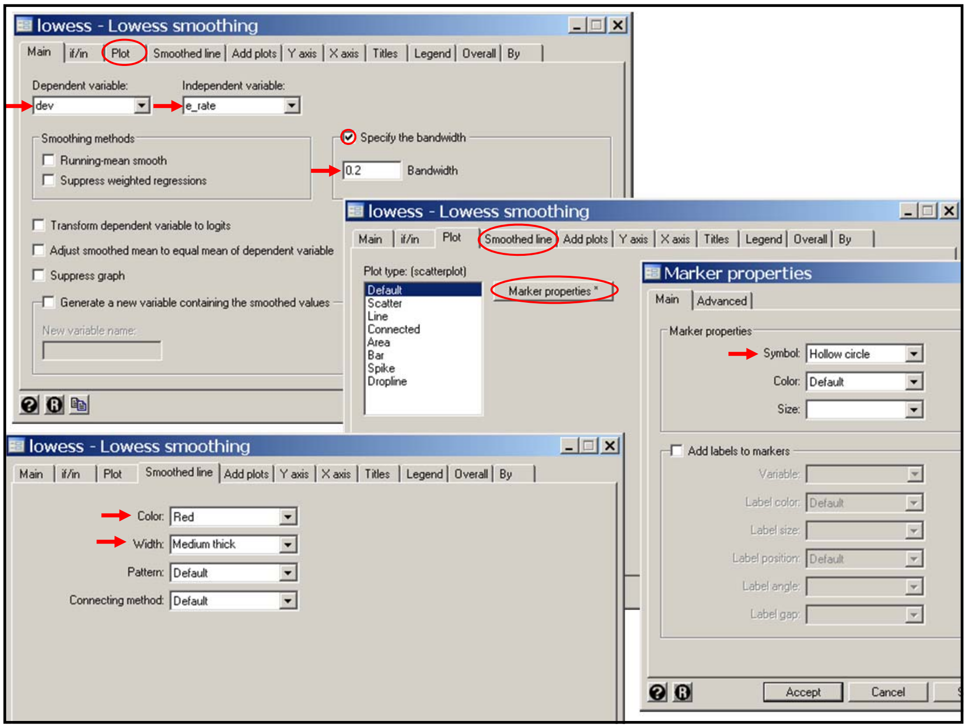
{6} This **predict** command calculates ***dev*** to equal the **standardized deviance residual.**

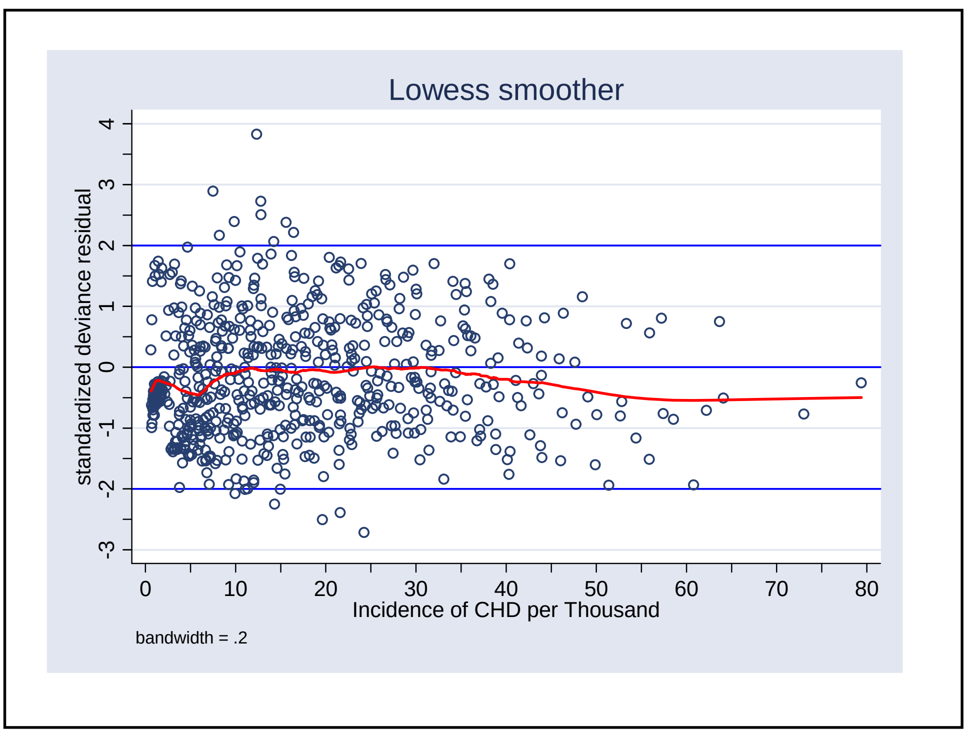



```
. generate e_rate = 1000*e_chd/pt_yrs
(82 missing values generated)

. 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
Marker properties
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
-3
-2
-1
0
1
2
3
4
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 $\pm$ 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.