Statistical Methods in Epidemiology
Lab2 - Solutions.
Rates in Follow-up Studies

I. The Diet Data Set.

1.

```
. desc

Contains data from C:\diet.dta
  obs:           337                          Diet data with dates
 vars:            13                          25 Sep 2006 16:31
 size:        17,187 (99.8% of memory free)
------------------------------------------------------------------------------
              storage  display    value
variable name   type   format     label      variable label
------------------------------------------------------------------------------
id              float  %9.0g                 Subject identity number
doe             long   %dDmCY                Date of entry
dox             long   %dDmCY                Date of exit
chd             float  %9.0g                 Outcome: 1= chd, 0 otherwise
dob             long   %dDmCY                Date of birth
job             int    %8.0g                 Occupation
month           byte   %8.0g                 month of survey
energy          float  %9.0g                 Total energy (1000kcals/day)
height          float  %9.0g                 Height (cm)
weight          float  %9.0g                 Weight (kg)
fat             float  %9.0g                 Total fat (g/day)
fibre           float  %9.0g                 Total fibre (g/day)
hieng           float  %9.0g                 Indicator for energy > 2.75
------------------------------------------------------------------------------
Sorted by:


. tab chd

Outcome: 1= |
     chd, 0 |
  otherwise |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        291       86.35       86.35
          1 |         46       13.65      100.00
------------+-----------------------------------
      Total |        337      100.00
```

This table shows the distribution of the outcome of interest, i.e. CHD, within the cohort. There are 13.65% cases with CHD among the 337 subjects.

2. `. list id doe dox chd in 1/20`

```
     +---------------------------------+
     | id         doe         dox   chd |
     |---------------------------------|
  1. |  1   16Aug1964   01Dec1976     0 |
  2. |  2   16Dec1964   01Dec1976     0 |
  3. |  3   16Nov1965   01Dec1976     0 |
  4. |  4   16Sep1965   01Dec1976     0 |
  5. |  5   16Sep1965   31Mar1976     0 |
     |---------------------------------|
  6. |  6   16Mar1965   31Aug1968     0 |
  7. |  7   16Nov1958   01Dec1976     0 |
  8. |  8   16May1965   01Dec1976     0 |
  9. |  9   16Feb1959   10Jan1962     0 |
 10. | 10   16Jul1964   16May1974     0 |
     |---------------------------------|
 11. | 11   16Oct1964   08Apr1974     0 |
 12. | 12   16Jul1964   03Aug1974     0 |
 13. | 13   16Sep1964   16Feb1974     0 |
 14. | 14   16Dec1959   01Dec1976     0 |
 15. | 15   16May1962   20Aug1976     0 |
     |---------------------------------|
 16. | 16   16May1959   31Dec1959     1 |
 17. | 17   16Feb1959   14Jan1965     0 |
 18. | 18   16Feb1959   08Mar1968     0 |
 19. | 19   16Feb1959   12Mar1966     0 |
 20. | 20   16Feb1959   26Dec1969     0 |
     +---------------------------------+
```

3. `. stset dox,fail(chd) origin(doe) scale(365.25)`

```
     failure event:  chd != 0 & chd < .
obs. time interval:  (origin, dox]
 exit on or before:  failure
    t for analysis:  (time-origin)/365.25
            origin:  time doe

--------------------------------------------------------------------------
      337  total obs.
        0  exclusions
--------------------------------------------------------------------------
      337  obs. remaining, representing
       46  failures in single record/single failure data
 4603.669  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =  20.04107
```

With this command we specify that the data are survival data, i.e. `stset` declares data to be survival-time (st) data.
Whenever you type `stset` or `streset`, Stata runs or reruns checks on your data making sure that what you are now declaring (or declared in the past) makes sense.
There are 337 total observations with 46 failures. The failures are declared in the `fail` option. `failure(varname[==numlist])` specifies the failure event. `failure()` must be specified with multiple-record data and is optional with single-record data. If `failure()` is not specified, every record is assumed to end in a failure. `failure(varname)` specifies that a failure occurs whenever varname is not zero and not missing. `failure(varname[==numlist])` specifies that a failure occurs whenever `varname` takes on any of the values of `numlist`.

`enter([varname==numlist] time exp)` specifies when a subject first comes under observation.

`exit(failure|[varname==numlist] time exp)` specifies the latest time under which the subject is both under observation and at risk of the failure event.

`origin([varname==numlist] time exp|min)` and `scale(#)` define analysis time. `origin()` defines when a person becomes at risk and `scale()` can be handy for making t units more readable (such as converting days in years).

4. `. li id _t0 _t _d _st in 1/20`

```
     +---------------------------------+
     | id   _t0          _t   _d   _st |
     |---------------------------------|
  1. |  1     0    12.29295    0     1 |
  2. |  2     0   11.958932    0     1 |
  3. |  3     0   11.041752    0     1 |
  4. |  4     0   11.208761    0     1 |
  5. |  5     0   10.537988    0     1 |
     |---------------------------------|
  6. |  6     0   3.4606434    0     1 |
  7. |  7     0   18.042437    0     1 |
  8. |  8     0   11.545517    0     1 |
  9. |  9     0    2.899384    0     1 |
 10. | 10     0   9.8316222    0     1 |
     |---------------------------------|
 11. | 11     0   9.4757016    0     1 |
 12. | 12     0   10.047912    0     1 |
 13. | 13     0   9.4182067    0     1 |
 14. | 14     0   16.960986    0     1 |
 15. | 15     0   14.264203    0     1 |
     |---------------------------------|
 16. | 16     0   .62696783    1     1 |
 17. | 17     0   5.9110198    0     1 |
 18. | 18     0   9.0568104    0     1 |
 19. | 19     0   7.0663929    0     1 |
 20. | 20     0   10.858316    0     1 |
     +---------------------------------+
```

`_t0`: time of entry on the analysis time scale
`_t`: time of exit on the analysis time scale
`_d`: reason for exit
`_st`: indicator for whether the record is included in the analysis.

5. `. stset dox,fail(chd) origin(dob) enter(doe) scale(365.25)`

```
     failure event:  chd != 0 & chd < .
obs. time interval:  (origin, dox]
 enter on or after:  time doe
 exit on or before:  failure
    t for analysis:  (time-origin)/365.25
            origin:  time dob
--------------------------------------------------------------------------
      337  total obs.
        0  exclusions
--------------------------------------------------------------------------
      337  obs. remaining, representing
       46  failures in single record/single failure data
 4603.669  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =  30.07529
```

```
                                    last observed exit t =  69.99863
```

```
. li id _t0 _t _d _st in 1/20
```

	id	_t0	_t	_d	_st
1.	1	49.615332	61.908282	0	1
2.	2	50.537988	62.49692	0	1
3.	3	58.784394	69.826146	0	1
4.	4	58.726899	69.935661	0	1
5.	5	59.460643	69.998631	0	1
6.	6	50.98152	54.442163	0	1
7.	7	45.138946	63.181383	0	1
8.	8	50.428474	61.97399	0	1
9.	9	67.099247	69.998631	0	1
10.	10	60.167009	69.998631	0	1
11.	11	60.52293	69.998631	0	1
12.	12	59.950719	69.998631	0	1
13.	13	60.580424	69.998631	0	1
14.	14	43.950719	60.911704	0	1
15.	15	55.734428	69.998631	0	1
16.	16	62.661191	63.288159	1	1
17.	17	59.780972	65.691992	0	1
18.	18	60.941821	69.998631	0	1
19.	19	60.960986	68.027379	0	1
20.	20	59.140315	69.998631	0	1

Note that here the subjects are supposed to start participating in the study at `_t0` which is the age of entering into the study. We are the ones who specified this with the option `origin(dob)`. The time of exit is the age at exit `_t`. Note however that the person time at risk is the same as before since it is `_t-_t0`.

6.
```
. tab hieng
```

Indicator for energy > 2.75	Freq.	Percent	Cum.
0	155	45.99	45.99
1	182	54.01	100.00
Total	337	100.00	

This is our exposure. 0 denotes subjects with energy intake <=2.75 Mcal and 1 those with energy intake >2.75 Mcal. The table shows the distribution of the exposure among the subjects of the cohort.

7.
```
. strate,per(1000)

        failure _d:  chd
  analysis time _t:  (dox-origin)/365.25
            origin:  time dob
 enter on or after:  time doe

Estimated rates (per 1000) and lower/upper bounds of 95% confidence
intervals
(337 records included in the analysis)
```

```
  +---------------------------------------+
  |  D        Y     Rate    Lower     Upper |
  |---------------------------------------|
  | 46   4.6037   9.9920   7.4843   13.3400 |
  +---------------------------------------+
```

`strate` tabulates the rate, estimated as the number of failures divided by the person-years, by different levels of one or more categorical explanatory variables (declared in the `varlist` of the command – see the following command). Confidence intervals for the rate are also calculated.

Here we have 46 chd cases and 4.6 per 1000 person-years in total. The rate is 9.992.

8. We generate a new variable for categorizing height:

```
. gen htgrp=height
(5 missing values generated)

. recode htgrp min/169.999=0 170/174.999=1 175/179.999=2 180/195=3
(htgrp: 332 changes made)

. tab htgrp

      htgrp |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         92       27.71       27.71
          1 |        102       30.72       58.43
          2 |         83       25.00       83.43
          3 |         55       16.57      100.00
------------+-----------------------------------
      Total |        332      100.00
```

You could check if the recording was successful by trying the following command:

```
. sort htgrp

. by htgrp: sum height, de

-------------------------------------------------------------------------------
-> htgrp = 0
                          Height (cm)
-------------------------------------------------------------
      Percentiles      Smallest
 1%         152.4          152.4
 5%       157.988         153.67
10%        160.02         157.48       Obs                  92
25%        163.83        157.734       Sum of Wgt.          92

50%        166.37                      Mean           165.5602
                          Largest       Std. Dev.      3.530403
75%      167.7924        169.3926
90%        168.91         169.545       Variance       12.46375
95%      169.1894         169.799       Skewness      -1.392987
99%       169.926         169.926       Kurtosis       5.127077

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

```
-> htgrp = 1

                          Height (cm)
-------------------------------------------------------------
      Percentiles      Smallest
 1%     170.0022       170.0022
 5%       170.18       170.0022
10%       170.18         170.18       Obs                 102
25%       171.45         170.18       Sum of Wgt.         102

50%       172.72                      Mean            172.277
                         Largest       Std. Dev.      1.338051
75%      173.482         173.99
90%       173.99        174.1932      Variance       1.790381
95%       173.99         174.498      Skewness      -.1539058
99%      174.498         174.625      Kurtosis       1.892441

------------------------------------------------------------------------
-> htgrp = 2

                          Height (cm)
-------------------------------------------------------------
      Percentiles      Smallest
 1%      175.006        175.006
 5%       175.26        175.006
10%       175.26         175.26       Obs                  83
25%      175.895         175.26       Sum of Wgt.          83

50%       176.53                      Mean           177.0533
                         Largest       Std. Dev.      1.355743
75%        177.8         179.07
90%       179.07         179.07       Variance       1.838039
95%       179.07         179.07       Skewness       .1356182
99%       179.07         179.07       Kurtosis        1.75807

------------------------------------------------------------------------
-> htgrp = 3

                          Height (cm)
-------------------------------------------------------------
      Percentiles      Smallest
 1%       180.34         180.34
 5%       180.34         180.34
10%       180.34         180.34       Obs                  55
25%       180.34         180.34       Sum of Wgt.          55

50%       182.88                      Mean           182.8292
                         Largest       Std. Dev.      2.570328
75%       184.15         187.96
90%       186.69         187.96       Variance       6.606587
95%       187.96         189.23       Skewness       1.049664
99%        190.5          190.5       Kurtosis       3.490751

------------------------------------------------------------------------
-> htgrp = .

                          Height (cm)
-------------------------------------------------------------
no observations
```

II. Rate Ratios.

```
. stmh hieng

        failure _d:  chd
  analysis time _t:  (dox-origin)/365.25
            origin:  time dob
 enter on or after:  time doe

Maximum likelihood estimate of the rate ratio
  comparing hieng==1 vs. hieng==0

RR estimate, and lower and upper 95% confidence limits

        ------------------------------------------------------------
             RR       chi2        P>chi2      [95% Conf. Interval]
        ------------------------------------------------------------
          0.520       4.84        0.0277         0.288      0.941
        ------------------------------------------------------------

  . stmh hieng,c(0,1)

        failure _d:  chd
  analysis time _t:  (dox-origin)/365.25
            origin:  time dob
 enter on or after:  time doe

Maximum likelihood estimate of the rate ratio
  comparing hieng==0 vs. hieng==1

RR estimate, and lower and upper 95% confidence limits

        ------------------------------------------------------------
             RR       chi2        P>chi2      [95% Conf. Interval]
        ------------------------------------------------------------
          1.922       4.84        0.0277         1.063      3.474
        ------------------------------------------------------------
```

In its simplest use, `stmh` estimates the ratio of the rates of failure for two categories of the explanatory variable (the first argument). Categories to be compared may be defined, as recode rules, in the compare option.

III. Exposure with more than two levels.

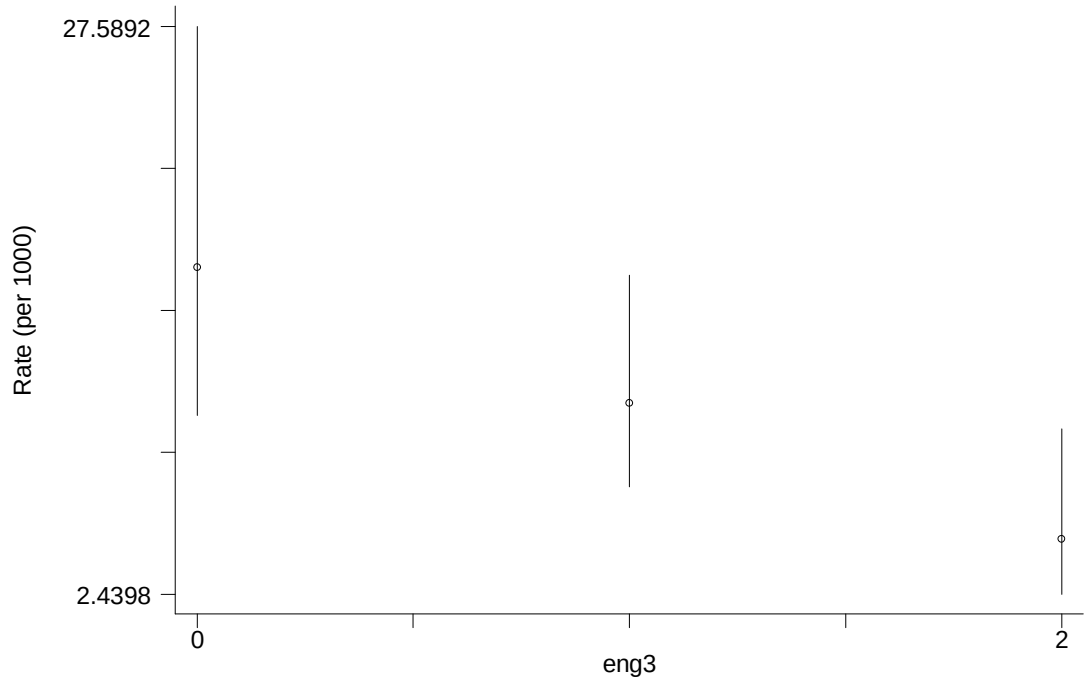
1.
```
. egen eng3=cut(energy), at(1.5, 2.5, 3.0, 4.5) icodes

. tab eng3

       eng3 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         75       22.26       22.26
          1 |        150       44.51       66.77
          2 |        112       33.23      100.00
------------+-----------------------------------
      Total |        337      100.00
```

2.
```
. strate eng3, per(1000)

        failure _d:  chd
  analysis time _t:  (dox-origin)/365.25
            origin:  time dob
 enter on or after:  time doe
```

```
Estimated rates (per 1000) and lower/upper bounds of 95% confidence
intervals
(337 records included in the analysis)

  +-------------------------------------------------+
  | eng3    D        Y      Rate     Lower     Upper |
  |-------------------------------------------------|
  |    0   16   0.9466   16.9020   10.3547   27.5892 |
  |    1   22   2.0173   10.9059    7.1810   16.5629 |
  |    2    8   1.6398    4.8787    2.4398    9.7555 |
  +-------------------------------------------------+
```

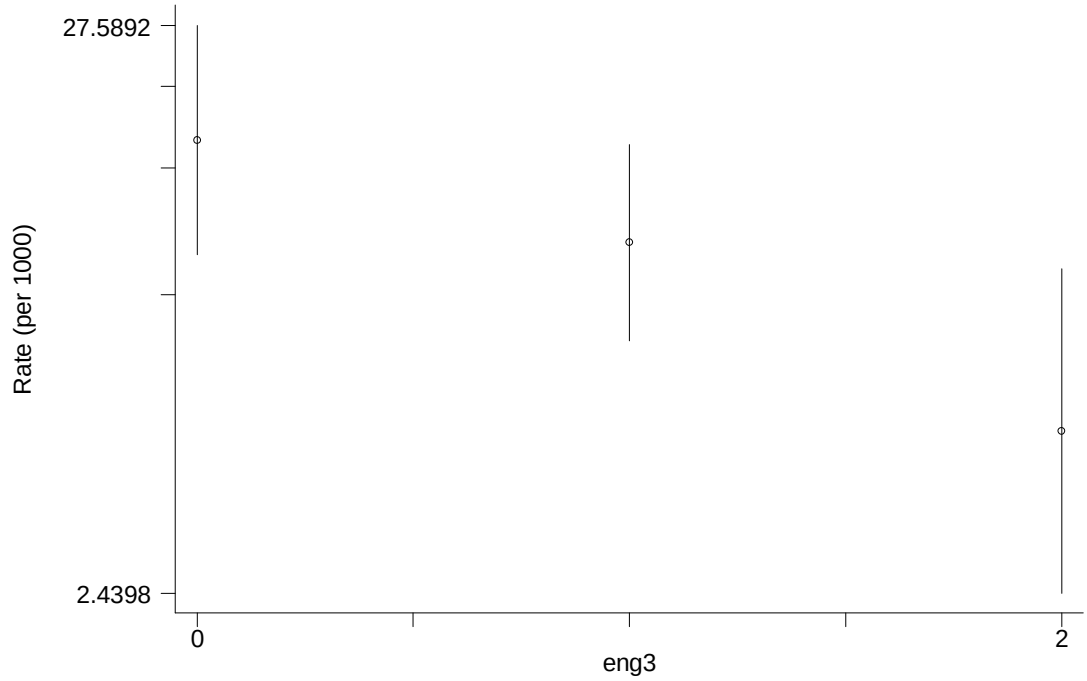
Here the rates of chd among the three levels of exposure are presented. Note that there is a decline in the actual rates as we go from the first to the last level.

```
. strate eng3, per(1000) graph
```



Note that with the option `graph` we can plot the actual rates. Therefore, this graph offers a graphical inspection of the actual difference in the rates.

```
. strate eng3, per(1000) graph ylog
```



With the additional option `ylog` we plot the log(rate). Note here that the differences we see are differences between the log(rates) and therefore we can inspect the log of the rate ratio (log(R1)-log(R2)=log(R1/R2)).

3.

```
. stmh eng3, c(1,0)

         failure _d:  chd
   analysis time _t:  (dox-origin)/365.25
             origin:  time dob
  enter on or after:  time doe

Maximum likelihood estimate of the rate ratio
  comparing eng3==1 vs. eng3==0
RR estimate, and lower and upper 95% confidence limits

        ------------------------------------------------------------
             RR       chi2        P>chi2      [95% Conf. Interval]
        ------------------------------------------------------------
          0.645       1.81        0.1789         0.339      1.229
        ------------------------------------------------------------

. stmh eng3, c(2,0)

         failure _d:  chd
   analysis time _t:  (dox-origin)/365.25
             origin:  time dob
  enter on or after:  time doe

Maximum likelihood estimate of the rate ratio
  comparing eng3==2 vs. eng3==0
RR estimate, and lower and upper 95% confidence limits

        ------------------------------------------------------------
             RR       chi2        P>chi2      [95% Conf. Interval]
        ------------------------------------------------------------
          0.289       9.35        0.0022         0.124      0.674
        ------------------------------------------------------------
```

4. `. stmh eng3`

```
         failure _d:  chd
   analysis time _t:  (dox-origin)/365.25
             origin:  time dob
  enter on or after:  time doe

Score test for trend of rates with eng3
  with an approximate estimate of the
  rate ratio for a one unit increase in eng3

RR estimate, and lower and upper 95% confidence limits

        ------------------------------------------------------------
             RR       chi2        P>chi2      [95% Conf. Interval]
        ------------------------------------------------------------
          0.548       8.98        0.0027         0.370      0.812
        ------------------------------------------------------------
```

Here we deal with the exposure in its metric form. Therefore we will estimate an average effect of changing from the one level to the other. This is given by the RR = 0.548. It seems that the rate of CHD declines as we go from the lowest to the highest level of energy intake. This reduction is statistically significant.

5. `. stmh energy`

```
         failure _d:  chd
   analysis time _t:  (dox-origin)/365.25
             origin:  time dob
  enter on or after:  time doe

Score test for trend of rates with energy
  with an approximate estimate of the
  rate ratio for a one unit increase in energy

RR estimate, and lower and upper 95% confidence limits

        ------------------------------------------------------------
             RR       chi2        P>chi2      [95% Conf. Interval]
        ------------------------------------------------------------
          0.351       9.89        0.0017         0.183      0.674
        ------------------------------------------------------------
```

Here we assess the effect of 1 unit (1 Mcal) increase in the actual value of energy intake. The rate ratio of 0.351 indicates a reduction with 1 Mcal increase in the energy intake. As expected, this is statistically significant also.

IV. Controlling for Confounding.

1. `. stmh hieng job`

```
         failure _d:  chd
   analysis time _t:  (dox-origin)/365.25
             origin:  time dob
  enter on or after:  time doe

Mantel-Haenszel estimate of the rate ratio
  comparing hieng==1 vs. hieng==0
  controlling for job
```

```
RR estimate, and lower and upper 95% confidence limits

        ------------------------------------------------------------
             RR       chi2        P>chi2      [95% Conf. Interval]
        ------------------------------------------------------------
          0.525       4.71        0.0299         0.291      0.949
        ------------------------------------------------------------
```

Here the aim is to explore the effect of energy intake on the rates of CHD, adjusting for any effect from the variable job. However, the previous form of the command `stmh` does not show the RRs within each category of job. We obtain this information with the following form of the command:

```
. stmh hieng, by(job)

         failure _d:  chd
   analysis time _t:  (dox-origin)/365.25
             origin:  time dob
  enter on or after:  time doe

Maximum likelihood estimate of the rate ratio
  comparing hieng==1 vs. hieng==0
  by job

RR estimate, and lower and upper 95% confidence limits

  +----------------------------+
  | job     RR   Lower   Upper |
  |----------------------------|
  |   0   0.41    0.12    1.36 |
  |   1   0.66    0.23    1.89 |
  |   2   0.52    0.21    1.27 |
  +----------------------------+

Overall estimate controlling for job

        ------------------------------------------------------------
             RR       chi2        P>chi2      [95% Conf. Interval]
        ------------------------------------------------------------
          0.525       4.71        0.0299         0.291      0.949
        ------------------------------------------------------------

Approx test for unequal RRs (effect modification): chi2(2) =     0.33
                                                   Pr>chi2 =   0.8468
```

What we get now is the effect of high vs. low energy intake on the rates of CHD within each category of job. We can see that these rates do not seem very different in between. We get a test for effect modification, i.e. a test for an interaction between `job` and `hieng` in the last lines of this output. The test is not statistically significant ($p = 0.8468$) but don't forget that it lacks power.

The adjusted estimated risk ratio for the effect of `hieng` on the rates of CHD is now 0.525, whereas the crude one was 0.520. Practically the adjustment did not affect the rate ratio. This is so because job does not affect the outcome (try this with `stmh job`).

2.

```
. stmh hieng, by(job htgrp)

        failure _d:  chd
  analysis time _t:  (dox-origin)/365.25
            origin:  time dob
  enter on or after:  time doe

Maximum likelihood estimate of the rate ratio
  comparing hieng==1 vs. hieng==0
  by job htgrp

RR estimate, and lower and upper 95% confidence limits

  +-----------------------------------+
  | job   htgrp     RR   Lower   Upper |
  |-----------------------------------|
  |   0     150   0.69    0.10    4.88 |
  |   0     170   0.41    0.04    3.95 |
  |   0     175   0.22    0.02    2.15 |
  |   1     150   0.59    0.15    2.37 |
  |   1     170   1.09    0.18    6.52 |
  |-----------------------------------|
  |   2     150   0.41    0.08    2.26 |
  |   2     170   0.91    0.18    4.51 |
  |   2     175   0.50    0.11    2.22 |
  +-----------------------------------+

Overall estimate controlling for job htgrp

        ------------------------------------------------------------
             RR       chi2        P>chi2      [95% Conf. Interval]
        ------------------------------------------------------------
          0.569       3.48        0.0620         0.313      1.037
        ------------------------------------------------------------

Approx test for unequal RRs (effect modification): chi2(7) =     1.85
                                                   Pr>chi2 =   0.9677
```

Here we explore the effect of `hieng` whilst controlling for two confounders, `job` and `htgrp`. The strata are now defined by the cross-classification of the two possible confounders (i.e., `job` and `htgrp`). The rate ratios are estimated within each of these strata. As you can see, the high energy group does not affect statistically significantly the rate of CHD in any of the strata (all CIs include 1).

The MH estimate of 0.569 is relevant if there is not any interaction between the three variables. In terms of modelling, this is a model which fits the main effects of `job`, `htgrp` and `hieng` as explanatory variables and the rate of `chd` as the dependent variable. What we observe in the rates within strata is that the trend in the rates between the categories of `htgrp` within each of the levels of job is not the same. Perhaps the assumption of no interaction is not relevant…