Statistical Methods in Epidemiology
Lab 2.
Rates in Follow-up Studies

I. The Diet Data Set.

The `diet` data set contains data from a pilot study of 337 men who kept a record of their fully weighted diet over two weeks in the file `diet.dta`. The variable `chd` is coded 1 for subjects with CHD and 0 otherwise.

1. Read in the data with the commands

```
. use diet, clear
. desc
. tab chd
```

2. How many cases of CHD are there? Make a note of this number. The time variables are dates of entry and exit to the study. These are stored in stata format that is, as days since 1/1/1960. For calculations dates are treated as numbers of days since 1/1/1960 but for output they are printed in standard date format. For each subject, the date of entry to the study is in `doe` and the date of exit in the `dox`. To see the dates try,

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

3. Set the `st` variables with the command

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

This command states that the time of exit is in `dox` and that the origin of the time scale to be used in the analysis is the date of entry to the study, stored in `doe`. The analysis time is then,

$$\frac{dox - doe}{scale}$$

which is time-since-entry in years. The failure variable is in `chd`, coded 1 for CHD and 0 otherwise.

4. The command `stset` creates 4 new variables called `_t0`, `_t`, `_d`, `_st`. These refer to the times of entry and exit (both on the analysis time scale), the reason for exit and whether the record is included. Try,

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

to see these new variables. Note that `_t0` is always zero because time is measured since date of entry, so zero corresponds to entry. The value of `_t` is the follow-up time in years. The value of `_d` is the same as chd in this case, but will be different when certain codes of the failure variable are selected for analysis. The value of `_st` is 1 because all observations are included in the analysis.

5. Try `stset` again, this time with `dob` as origin:

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

The time scale is now age. It is now essential to specify the date of entry otherwise the default is 0, i.e. birth. Try

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

The value of `_t0` is now the age at entry.

6. The explanatory variable we shall concentrate on is `energy`, and for a preliminary look at the data we shall use the variable `hieng` which is coded 1 for subjects with total daily energy intake greater than 2.75 Mcals and 0 otherwise:

```
. tab hieng
```

7. Use strate to tabulate the CHD rates. The command

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

will produce the overall rate per 1000 years. Try,

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

to see how the rate varies with the two levels of `hieng`. Does a high intake of energy have a protective or an adverse effect on the rate of CHD?

8. Generate a new variable called `htgrp` with

```
. egen htgrp=cut(height), at(150,170,175,180,195)
```

Use `strate` to study the way the CHD rate changes with `htgrp`. Use the option `graph` in `strate` to look at this graphically.

II. Rate Ratios.

Use the command stmh to find the rate ratio for the high energy group compared to the low energy group:

```
. stmh hieng
```

To find the rate ratio the other way round, try

`. stmh hieng, c(0,1)` *(where* `c` *means "compare")*

III. Exposure with more than two levels.

Grouping the values of total energy into just two groups does not tell us much about how the CHD rate changes with total energy. It is a useful exploratory device, but to look more closely we need to group the total energy into perhaps 3 or 4 groups. In this example, we shall use the cut points 1.5, 2.5, 3.0, 4.5.

1. Use the command

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

to create a new variable `eng3` coded 0 for values of `energy` in the range 1.5-2.499, 1 for values in the range 2.5-2.999, and 2 for values in the range 3.0-4.499. The codes 0, 1, 2 are called the levels of the variable.

2. To find the rate for different levels of `eng3` try

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

The option `graph` will show a graph of rate against levels of exposure and the option `ylog` will plot the rates on a log scale thereby concentrating on the *rate ratio* with changing level of energy, not the *rate difference*. Try

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

3. For an exposure variable with more than two levels, `stmh` can be used to compare any two levels, *i* and *j* say, using the option `c(i,j)`. For example, to compare level 1 with level 0 and level 2 with level 0. For `eng3` use,

```
. stmh eng3, c(1,0)
. stmh eng3, c(2,0)
```

4. Another way of studying metric exposures like energy is to group their values and assess the effect of changing *from one level of exposure to the subsequent one*. To find the effect of changing from one level of energy to the other you can use `stmh` in the usual way, i.e.,

```
. stmh eng3
```

What do you think about the effect of `eng3` on the rate of CHD?

5. In the same way, a metric exposure can be considered as continuous thus its effect can be assessed via the change on the rate of interest with 1 unit in the exposure (linear effect). The command is the same as above.
Investigate the effect of height in the rate of CHD. What is your conclusion? What is the interpretation of this effect?

IV. Controlling for Confounding.

1. To find the effect of high energy controlled for job the previous command can be used in a slightly different way:

```
. stmh hieng job
```

The remaining variables (i.e. those different from the exposure) are categorical variables which are to be controlled for using stratification. Strata are defined by cross-classification by all these variables and the rate ratio estimate is combined

over strata using the Mantel-Haenszel method. Using the `by` option, the variation of the rate ratio with further categorical variables may be explored.
Try now,

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

to see the additional information which you get. Do you think that the assumption underlying the MH estimate is relevant?

2. Investigate the relationship of `hieng` with the rate of CHD controlling for `job` and `htgrp`. What are your conclusions?