1 Analysis of IHD data in C&H Table 22.6

BACKGROUND: this dataset is first introduced, without the age-stratification, on page of the of Clayton & Hills (C&H) chapter 13:

Table 13.1. Incidence of ischaemic heart disease by energy intake

	Energy intake	
	< 2750 kcals (exposed)	≥ 2750 kcals (unexposed)
Person years	1857.5 (Y_1)	2768.9 (Y_0)
New cases	28 (D_1)	17 (D_0)
Estimated rate	15.1	6.1
90% interval	(11.1 → 20.6)	(4.1 → 9.1)

Table 13.1 shows a preliminary tabulation of some data which will be analysed in detail in this and the following chapter.* The data relate subsequent incidence of ischaemic heart disease (IHD) to dietary energy intake. The study cohort consisted of 337 men whose energy intake was assessed by a seven-day weighed dietary survey. The subsequent follow-up was for an average of 13.7 years and yielded 45 new cases of IHD. The table divides this cohort into an exposed group consisting of men whose energy intake was less than 2750 kcals per day, the remaining men being regarded as unexposed. Although it might seem odd to denote the low energy intake group as exposed, this is because low energy intake is a surrogate measure for physical inactivity. Table 13.1 also introduces some algebraic notation: D_0, D_1 for the number of disease events observed in the unexposed and exposed cohorts respectively, and Y_0, Y_1 for the corresponding person-years observation.
*Unpublished data. The study is described by Morris, J.N. et al. (1977) British Medical Journal, 19 November 1977, 2, 1307-1314.

The full citation, *Morris JN, Marr JW, Clayton DG. Diet and heart: a postscript. Br Med J. 1977 Nov 19;2(6098):1307-14,* shows that Clayton was a co-author. The abstract reads:

> During 1956-66, 337 healthy middle-aged men in London and south-east England participated in a seven-day individual weighed dietary survey. By the end of 1976, 45 of them had developed clinical coronary heart disease (CHD) which showed two main relationships with diet. Men with a high energy intake had a lower rate of disease than the rest, and, independently of this, so did men with a high intake of dietary fibre from cereals. Energy intake reflects physical activity, but the advantage of a diet high in cereal fibre cannot be explained; there was no evidence that the disease was associated with consumption of refined carbohydrates. Fewer cases of CHD developed among men with a relatively high ratio of polyunsaturated to saturated fatty acids in their diet, but the difference was not statistically significant.

Morris was an influential epidemiologist. The headline of the (2009) obituary in the Financial Times describes him as "The man who invented exercise". See the link `http://www.ft.com/cms/s/2/e6ff90ea-9da2-11de-9f4a-00144feabdc0.html`.
He also wrote a classic textbook, *Uses of Epidemiology,* now hard to find S. Harper – no, not the PM, the other S Harper – has a copy. For more, see the 'Jerry Morris (physician)' entry in Wikipedia, or the appreciation in `http://ije.oxfordjournals.org/cgi/content/full/36/6/1184`.

EXERCISE – like gardening, this type of exercise may not measurably improve c-v health

`R` code, with additional notes interspersed with the code, is available under the resources for 'Regression models for (incidence) rates.'.

i. Fit an *'additive rates'* model[1] to the (age-stratified) data in Table 22.6, and present the results in the same format as Table 22.7 of Clayton and Hills – but with + signs instead of × signs (Ch 22 was handed out earlier, and is also available in the resources).

ii. Fit Clayton and Hills' multiplicative model and verify that the fitted model is the same as that given in their Table 22.7.

iii. Fit a multiplicative model but with age used as an interval ('continuous') rather than a categorical variable. Use two versions (a) and (b) of this 'age' variable. Comment on the differences between the fitted coefficients in these two models and those in (ii), and also on the differences in interpretation of the coefficients between versions (a) and (b).[2]

(a) `age=c( 0, 0, 10,10, 20,20)`

(b) `age=c(45,45, 55,55, 65,65)`

2 Do Oscar Winners Live Longer than Less Successful Peers? A Reanalysis of the Evidence

The aims are to carry out (1) the 'P-Y' analysis described in the 2006 'McGill' re-analysis, and (2) calculate the 'fewer-assumptions involved' Mantel-Haenszel summary ID ratio that the McGill authors calculated but – not to confuse the reader with yet another analysis – omitted from their article. Later on in the course, we will analyze the data with the same (time-dependent Cox PH) model that was reported on in the 2006 article.

[1] You will need to fill in a few blanks in the `R` code.

[2] It is a good idea, both for interpretation and for remembering, to code continuous X's so that resulting values are on both sides of zero ('centered') or mostly (or entirely) immediately to the right of the starting point of the data. For example, which formula for *ideal weight* – the weight below such that the health risks balance those of being above it – is easier to remember
F: 100 lbs. + 5 lbs for every inch above 5 feet, or ... -300 lbs. + 5 lbs * height in inches ?
M: 110 lbs. + 6 lbs for every inch above 5 feet, or ... -360 lbs. + 6 lbs * height in inches ?

Under the Resources for Regression Models for (incidence) Rates, you will find (a) the Oscar data set[3] with one data-record per performer (b) a dataset (with approx. 20,000 records) in which each the performer's post-1st-nomination data-record has been converted (split) into 1-year data-records, and classified according to age, period, AND Oscar-status, (c) a smaller dataset in which the individual performer-years (and numbers of deaths) in (b) have been aggregated into 'sex-age-period-Oscar' cells, with 5-year age-bands and 10 year calendar-year-bands,[4] and (d) a file similar to (c), but where *all* of a performer's post-nomination performer-time is allocated to the 'winners' category if that performer *ever* won an Oscar, or to the 'nominated' category if (s)he was nominated but never won.[5]

In the *description* of (b) and (c) below, the name of the Oscar-status indicator is shortened to O, with $O = 0$ indicating performer-time lived as a nominee, and $O = 1$ indicating performer-time lived as an Oscar winner. In the *actual dataset to be analyzed, i.e. in* (c), $O = 0$ corresponds to `w.cat=0` and $O = 1$ to `w.cat=1`.

In (b) each (Oscar-status-specific) record documents the experience in each (age, period) 'rectangle'[6] traversed, i.e., the number of years spent in that rectangle , and the Vital status (0 if alive, 1 if dead) at the end of these years.[7] Because the Lexis program is written for generic *transitions ('events')* *of any type (not necessarily bad ones)*, this status variable is called `lex.Xst`, which refers to the status (in our example *vital* status, 0 alive, 1 dead) at the performer's 'exit' (pardon the pun, but the '`X`' in '`Xst`' stands for an *epidemiologic* 'exit' from the Lexis diagram, and the '`st`' stands for status). The other key variable is `lex.dur`, which refers to the `dur`ation or length of the performer's time-slice.

In (c), which is formed by summing the performer-time `lex.dur` and the `lex.Xst` over all transits through the same sex-age-period-O cell, the two sums are the *total p-t* and *total deaths* in this cell – remember that a sum of 0's and 1's is a count of the number of 1's.

i. Use dataset version (c) to compare the death rates in the performer-years lived as nominees (reference category, `w.cat=0`) with those lived as winners (index category, `w.cat=1`), by fitting the following multiplicative (i.e. 'rate ratio') model[8] to the numbers of deaths in each sex-age-period-Oscar (shortened to s-a-p-O here, in order to fit the equation into one line) 'cell'.

$$Rate_{cell} = Rate_{ref.cell} \times M_{s:ref} \times M_{a:ref} \times M_{p:ref} \times M_{O:ref},$$

where the $ref.cell$ is a suitably chosen reference 'corner' cell (Clayton and Hills' terminology), and each M (the rate 'Multiplier') is short for Mortality Rate Ratio (MRR), – the theoretical, unknown, to be estimated, ratio of the mortality rate in the category[9] of the determinant in question relative to the reference category of that determinant.

For fitting purposes, you translate the *epidemiologic* (rate) model above into the following *statistical* model

$$E[\#deaths] = e^{\{logRate_{ref} + logM_s \times s + logM_a \times a + logM_p \times p + logM_O \times O + \log(PT)\}},$$

so that

$$\log\{E[\#deaths]\} = \beta_{ref} + \beta_s \times s + \beta_a \times a + \beta_p \times p + \beta_O \times O + \log(PT).$$

Writing out both models lets you match the coefficients from the fitted statistical (`R`) model with the fitted parameter value(s) of interest in the epidemiological (rate) model. (def'n.: *epidemiologist*: a student of *rates*).

[3] For reasons jh can better explain in person, this differs slightly from that analyzed in the Redelmeier article.

[4] You are asked to the analyses with (c), which is named `aggregated-Lexis-rectangles.txt`. Nowadays, with fast computers and lots of live memory / disk storage space for large datasets, you *could* do the analysis using (b). Since it uses finer subdivisions of age and calendar period, you would get get slightly different answers, and you would probably choose to model age and calendar-time with (functions of) continuous variables, rather than with a very large number of indicator variables – 'dummy' variables, if you insist on that meaningless term – for the finer age- and calendar-period categories.

[5] The name of datafile (d), `aggregated-Lexis-rectangles-r.txt`, has the suffix '`-r`' to denote it as the 'Redelmeier' allocation of the performer-time.

[6] This terminology is from Lexis, who tended to use squares, e.g., 5-year age bands and 5-year calendar-year bands: since death rates vary faster over ages than over calendar time, you want to make the age-bands (i.e., the age-matching) quite narrow: thus jh formed rectangles that are 1 (age) year high by 10 (calendar) years wide, so in effect each slice was 1 year long: you could rerun the time-slicing program with other 'cuts.'

[7] If you want to see how these split records were created, you can look at and run the `R` code shown in the resources. It uses the `Lexis` package that is available from the `R` site, and developed by Carstensen (`R` 'Epi' package `http://staff.pubhealth.ku.dk/~bxc/Epi/`). See also the `survSplit` function in the `survival` package – we used this to split the time in the COMPARE (stents) study. One of the students in bios602 discovered two other options. One is a standalone Windows program, from `http://epi.klinikum.uni-muenster.de/pamcomp/pamcomp.html`; the other is the `pyears` function in the Survival package in R (jh doesn't remember if Survival is part of the default R installation, or needs to be added). **Stata** users: there is a time-slicing function used in conjunction with survival analyses.

[8] One could, and would if need be, refine this model further, e.g. by refining the relationship of rates with age, and allowing for the possibility of different effects of O in males and females...

[9] Or *level*, if we model the variable as an interval variable.

ii. Write out the fitted multiplicative model in the same way as Clayton and Hills did in Table 22.7 in their Introduction to Regression chapter of their Statistical Models for Epidemiology textbook. Comment on the MRR for the 'years lived as a winner' vs. 'years lived as a nominee' contrast.

iii. Comment on the fitted effects of gender[10], age and calendar time, and whether they 'fit' with what you expect, and have seen in other datasets.[11]

iv. From dataset (c) calculate the total performer-time lived as a nominee ('$PT_{nominee}$'), and the total performer-time lived as a winner ('PT_{winner}'). Compare these with the corresponding values calculated from the 'Redelmeier' version, i.e., from dataset (d). Comment.[12]

v. Fit the same multiplicative model fitted in (i) to the data in dataset (d). Compare the fitted 'O' effect in this dataset – where `w.cat` is a fixed-from-the-outset variable – with what you found in the (McGill) version – where `w.cat` is a time-dependent variable. Comment.

vi. How would Mantel have analyzed these data? The R code file in resources includes some that allows you to convert datafile (c) into a form where you can treat sex, age and calendar period as stratifying variables – it puts the 'exposed' PT and deaths in the exposed PT in the same data-record as those for the un-exposed PT in the same stratum, making it easy to obtain the stratum-specific products, and to obtain the numerator and denominator sums used to calculate the ratio in formula 8.5 – déjà vu – in Rothman2002.

Use this re-arranged dataset to calculate this Mantel-Haenszel mortality rate ratio. How does it compare with the one obtained from Poisson regression?

[10]Even though we used the term 'sex' above, one could make a good argument for preferring the term 'gender' in this context: Google 'gender vs. sex'.

[11]The effects of gender, age and calendar time are secondary here, but if you do choose to represent age and calendar-time as linear (continuous) variables, make sure you report their effects correctly – they should broadly 'line up' with the fitted effects when using indicator variables.

[12]For the principle behind the correct allocation of person-time, and early examples of incorrect P-T allocation, see section 3.1 of Volume II of Breslow and Day's text, available in the resources for the bios602 course. See also the material on 'immortal-time' bias in the 634 website

— Postscript —

The original 2001 Annals of Internal Medicine article continues to be cited as authorative ... most Google searches on the topic of longevity and fame ignore any corrections. It may be like John Haldeman (who worked for Richard Nixon during the Watergate affair) said, "Once the toothpaste is out of the tube, it's hard to get it back in!"

Harvard charges customers to subscribe to their Newsletter...

`http://www.health.harvard.edu/press_releases/oscar_winners`

WORKED EXAMPLE – perceived age and mortality rates..

```
   male third age.cat n.deaths p.years   rate log.rate
1     1     1       1       25   800.3 0.0312  -3.4661
2     1     2       1       21   792.3 0.0265  -3.6304
3     1     3       1       49   672.0 0.0729  -2.6184
4     1     1       2       39   456.6 0.0854  -2.4602
5     1     2       2       30   465.7 0.0644  -2.7423
6     1     3       2       42   440.3 0.0954  -2.3498
7     1     1       3       42   327.9 0.1281  -2.0550
8     1     2       3       46   321.1 0.1433  -1.9431
9     1     3       3       54   271.1 0.1992  -1.6135
10    0     1       1       10   781.2 0.0128  -4.3582
11    0     2       1       23   752.5 0.0306  -3.4879
12    0     3       1       33   710.5 0.0464  -3.0695
13    0     1       2       18   598.5 0.0301  -3.5041
14    0     2       2       27   576.7 0.0468  -3.0615
15    0     3       2       31   531.7 0.0583  -2.8421
16    0     1       3       42   658.7 0.0638  -2.7526
17    0     2       3       74   587.4 0.1260  -2.0716
18    0     3       3       69   517.1 0.1334  -2.0141
```

Ordinary least squares [and Gaussian variation]

```
> ols.fit = lm(log.rate~male+age.cat+third,data=ds)

> round(exp(ols.fit$coefficients),3)

(Intercept)        male     age.cat       third

      0.006       1.609       1.977       1.406

.............
```

Maximum Likelihood [and Poisson variation]

```
 > glm.fit = glm(n.deaths~male+age.cat+third,
+              family=poisson, offset=log(p.years),data=ds)

> round(exp(glm.fit$coefficients),3)

(Intercept)        male     age.cat       third

      0.007       1.544       1.942       1.365
```

WORKED EXAMPLE – events following insertion of a stent (tx=0:1g vs 1=2g)

```
   tx mid.interval Pt.days events
1   0            5    8806     33
2   1            5    8820     18
3   0           15    8692      2
4   1           15    8762      4
5   0           25    8680      0
6   1           25    8747      3
7   0           35    8663      2
8   1           35    8720      1
9   0           45    8650      0
10  1           45    8710      0
...
71  0          355    8190      0
72  1          355    8410      0
73  0          365     819      0
74  1          365     841      0

> summary(fit.glm)

glm(formula = events ~ mid.interval + tx, family = poisson,
        data = subset(dta, mid.interval >= 15),    <<<<<<<<<<<<
        offset = log(Pt.days))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8245  -1.4187  -0.1642   0.5623   1.9373

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.5354392  0.2334336 -36.565   <2e-16 ***
mid.interval -0.0009553  0.0010625  -0.899    0.369
tx           -0.2713860  0.2161583  -1.255    0.209
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

    Null deviance: 85.319  on 71  degrees of freedom
Residual deviance: 82.909  on 69  degrees of freedom
AIC: 206.52

> exp(fit.glm$coefficients)

 (Intercept) mid.interval        tx
     0.00020      0.99905      0.76
```