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:

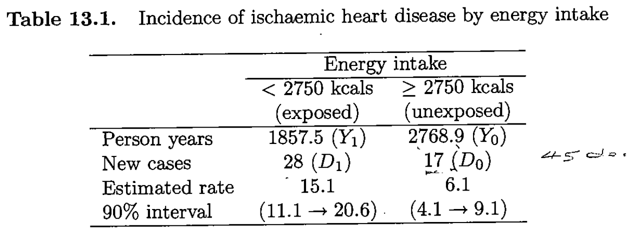

Table 13.1. Incidence of ischaemic heart disease by energy intake

	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)

45

> 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 *Unpublished data.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).

   This exercise was not part of the assignment, but here, in any case, are the results of the fitting ...

```
# Clayton Hills, example 22.6 p221

# DATA and age-category indicators

cases=c(4,2,5,12,8,14);  # order
pt=c(607.9,311.9,1272.1,878.1,888.9,667.5)
e=c(0,1,0,1,0,1) # exposure indicator, note the order
# 2 age-category indicators # ... ''roll your own''
a50=c(0,0,1,1,0,0) #  age 40 as reference category
a60=c(0,0,0,0,1,1)

# * could also use the gl function (below)
#   to set up variables that have a pattern
#   (or use the as.factor to create indicator -- 'dummy' --
#   variables for categorical variables). But you
#   have more control' if you set them up 'manually'

# no need to think of exposure as a ‘factor’ since
# it is already 0/1 and can be included as a linear
# (continuous) variable

########################################
# to fit additive (ie ID Ratio) model
#######################################

# for rate DIFFERENCE models,

# rate = rate.ref + (linear fn. of x's), so...

# cases = rate * PT

# ie.   = rate.ref*PT + PT*(linear fn. of x's), so...
```

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

```
# need to 'multipy through..' by PT
# and so terms in model become the products of
# each 'x' in the rate model,  with PT.

# use POISSON variation
# and IDENTITY link. IDENTITY just means that you
# are asking that the (untransformed)
# E[#cases] = linear predictor = [B0] + B1.X1 + B2.X2 + ..
# you are not modelling some transform of E[#cases]
# (for the multiplicative model below, you will model
# a transformation , ie the log, of E[#cases] as a
# function of the linear predictor

# remember... (? = shorthand for "coefficient to be fitted")

#  rate = ?rate.ref + ? * e + ? * a50 + ? * a60, so

#      multiplying across by pt... and expanding...

#  cases = ?rate.ref*pt + ?*e*pt + ?*a50*pt + ?*a60*pt, so

# make products with pt

e.pt   =   e*pt
a50.pt = a50*pt
a60.pt = a60*pt

# have a look

cbind(cases,pt,e,a50,a60, e.pt,a50.pt,a60.pt)

> cbind(cases,pt,e,a50,a60, e.pt,a50.pt,a60.pt)

     cases     pt e a50 a60  e.pt a50.pt a60.pt
[1,]     4  607.9 0   0   0   0.0    0.0    0.0
[2,]     2  311.9 1   0   0 311.9    0.0    0.0
[3,]     5 1272.1 0   1   0   0.0 1272.1    0.0
[4,]    12  878.1 1   1   0 878.1  878.1    0.0
[5,]     8  888.9 0   0   1   0.0    0.0  888.9
[6,]    14  667.5 1   0   1 667.5    0.0  667.5

fit.add=glm(cases ~ -1 + pt + e.pt + a50.pt + a60.pt ,
    family=poisson(link="identity"))

summary(fit.add)
```

```
Deviance Residuals:
      1        2        3        4        5        6
 0.4616  -1.2160  -0.1352   0.2763  -0.3110   0.4682

Coefficients:
        Estimate Std. Error z value Pr(>|z|)
pt      0.005176   0.002718   1.904   0.0569 .
e.pt    0.008432   0.003138   2.687   0.0072 **
a50.pt -0.001003   0.003130  -0.320   0.7487
a60.pt  0.004850   0.003911   1.240   0.2149
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:    Inf  on 6  degrees of freedom
Residual deviance: 2.1024  on 2  degrees of freedom
AIC: 32.226

> str(fit.add)
List of 30 [seveal omtted]

 $ coefficients     : Named num [1:4]  0.00518  0.00843 -0.00100  0.00485
  ..- attr(*, "names")= chr [1:4] "pt" "e.pt" "a50.pt" "a60.pt"
 $ residuals        : Named num [1:6]  0.854 -2.244 -0.309  0.932 -0.912 ...
  ..- attr(*, "names")= chr [1:6] "1" "2" "3" "4" ...
 $ fitted.values    : Named num [1:6]  3.15  4.24  5.31 11.07  8.91 ...
  ..- attr(*, "names")= chr [1:6] "1" "2" "3" "4" ...
...
 $ model            :'data.frame': 6 obs. of  5 variables:
  ..$ cases : num [1:6] 4 2 5 12 8 14

SO... match up the numbers with the ?'s in the model...

? ref.rate  =  0.005176
? e         =  0.008432
? a50       = -0.001003
? a60.pt    =  0.004850

#  rate = ?rate.ref + ? * e + ? * a50 + ? * a60, so

# (all rates are events/1py)

#  0.0051 + 0.0084 if exposed  -0.0010 if age50 + 0.0048 if age 60
```

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

```
# again [ notation: ? = beta to be fitted, X = multiplication ]

# rate = ?rate.ref X (exp[? * e]) X (exp[? * a50]) X (exp[? * a60])

# cases = rate * pt = exp[lin. comb'n of ?'s & regressors] * pt

# log(cases) = [lin. comb'n of ?'s & regressors] + log(pt)

# log(cases) = [usual ?0 + ?1 x1  + ?2 x 2 ... ] + offset

#  By default, when you specify Poisson variation, the LOG link is used,
#
# E[rate] = exp(?0) X exp(?1 x1) X exp(?2 x2) ....
#
# so just match up the coefficients with the output


fit.mult=glm(cases ~ e + a50 + a60,family=poisson,offset=log(pt))

beta.fitted = fit.mult$coefficients ;  beta.fitted

(Intercept)           e         a50         a60
 -5.4176721   0.8696587   0.1290097   0.6920480

> round(  exp(beta.fitted[1]), 4)
(Intercept)
     0.0044 <-----------------  events /1 PY  for ref. profile
     ======
>
> round(  exp(beta.fitted[2:4]), 2)
   e  a50  a60
2.39 1.14 2.00  <-------------  rate-multipliers (ie rate ratios)
==== ==== ====

> round(fit.mult$fitted.values,1)

   1    2    3    4    5    6
 2.7  3.3  6.4 10.6  7.9 14.1
```

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)`

```
> fit.mult=glm(cases ~ e + age,family=poisson,offset=log(pt))
>
> beta.fitted = fit.mult$coefficients
>
> beta.fitted
(Intercept)           e         age
-5.58975028  0.86489191  0.04112415
>
> round(  exp(beta.fitted[1]), 4)
(Intercept)
     0.0037 <---corner ( intercept) rate .. at first AGE in data !!! )
>
> round(  exp(beta.fitted[2:3]), 2)
   e  age
2.37 1.04 <-- 1.04 PER YEAR  of Age (vs. per dECADE cat. above)
====
2.37 (vs. 2.39.. expect little diff with diff. age representations)
```

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

```
(Intercept)           e         age     < ------ beta.fitted
-7.44033714  0.86489191  0.04112415

> round(  exp(beta.fitted[1]), 4)
(Intercept)
      6e-04 <--- 6/10000 PY ..  at AGE 0,  long way from data!!

> round(  exp(beta.fitted[2:3]), 2)
   e  age
2.37 1.04 <-- again, per YEAR of age.. for 10Y, use exp(10*0.0411)
====
WHEN DATA ARE MOVED LEFT/RIGHT ON X-AXIS, SLOPE W.R.T. X UNCHANGED
```

[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 ?

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.

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 ‘$\underline{M}$ultiplier’) 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.

[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.

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*).

```
# ? = beta to be fitted,

#  s a p O: indicators for (male)sex, age, period, whether time spent as Oscar winner

# X = multiplication...

# rate = ?rate.ref X (exp[? * s]) X (exp[? * a]) X (exp[? * p]) X (exp[? * O])

# cases = rate * pt = exp[lin. comb’n of ?’s & regressors] * pt

# log(cases) = [lin. comb’n of ?’s & regressors] + log(pt)

# log(cases) = [usual ?0 + ?1 x1  + ?2 x 2 ... ] + offset

> head(ds)

  a.cat p.cat m.cat w.cat lex.Xst lex.dur
1     6     2     0     0       0     5.0
2     7     2     0     0       1     1.5
3    10     2     0     0       0     2.0
4     2     3     0     0       0     2.0
5     3     3     0     0       0     3.0
6     4     3     0     0       0     7.0

> tail(ds)

    a.cat p.cat m.cat w.cat lex.Xst lex.dur
462    13    10     1     1       0       4
463    14    10     1     1       0       5
464    15    10     1     1       0      10
465    16    10     1     1       0       5
466    17    10     1     1       0       3
467    18    10     1     1       0       1

> fit.age.int.scale = glm(lex.Xst ~ m.cat + a.cat + p.cat
+             + w.cat, offset=log(lex.dur),family=poisson,data=ds)
> fit.age.int.scale$coefficients
(Intercept)       m.cat       a.cat       p.cat       w.cat
 -8.5885225   0.3375658   0.4391931  -0.1317224  -0.1854718
```

```
> round(exp(fit.age.int.scale$coefficients[1]),5)
(Intercept)
    0.00019  <<<<< "corner" rate: UNITS deaths per PY (19 per 10000 PY)

    NOTE: ‘corner’ is at a.cat=0; in the study the smallest a.cat was 6!

> round(exp(fit.age.int.scale$coefficients[2:5]),2)
m.cat a.cat p.cat w.cat
 1.40  1.55  0.88  0.83  [Rate-Multipliers]
   ^     ^    ^      ^
   ^     ^    ^      ^ rates are 17 % lower in performer-time lived as a winner
   ^     ^    ^
   ^     ^    ^  rates were 12% lower in each successive calendar decade
   ^     ^    ^
   ^     ^  rates were 55% higher in each successive 5-year age band
   ^
   ^ rates were  40% higher in males than females

 NOTE the use of causal-neutral wording..
 Many are tempted to say ‘as x increases, y increases’
 The neutral wording ‘y is higher when x is higher’  reminds the reader to ask why!

 SOME OF YOU took the fast route and fitted the age-and-period-as-categorical
 version give in the R code on the website.

 That code was put there to (a) show you how to include a categoriacl variable if you
 are too pressed to ‘roll-your-own’ indicator (’dummies’ to others) variates.

For those of you who went ahead and fitted age and period as categorical variables,
here are the data you were ‘fitting to’ as far as age is concerned:

The 1 to 19 are the 19 age bands, 5 years wide each. I think the first
age band is for ages 5-9, and includes on child actor, who, because
the PY is 4, must have been nominated at age 6 or so.

> cases=aggregate(ds$lex.Xst, by=list(ds$a.cat),FUN="sum")
> pt = aggregate(ds$lex.dur, by=list(ds$a.cat),FUN="sum")
> both=cbind(cases,pt)

> t(both[1:10,]) ;  t(both[11:19,])
         1  2  3     4   5    6      7    8    9   10
Group.1  1  2  3   4.0   5    6    7.0    8    9   10 << age bands
x        0  0  0   2.0   1    2    6.0    7   14   16 << numbers of deaths
x.1      4 43 95 263.5 735 1441 2049.5 2358 2489 2411 << numbers of PY

           11   12   13   14  15  16  17 18 19
Group.1    11   12   13   14  15  16  17 18 19  << ditto for age bands 11 to 19
x          24   30   35   47  42  50  32 15  4  << numbers of deaths
x.1      2220 1968 1658 1291 911 542 220 72 10  << numbers of PY

SO YOU SHOULDN’T BE SURPRISED THAT IF YOU USE AS A CORNER
(REF category) A BAND WITH 0 DEATHS in 4 PY, the intercept will
be very unstable!!  log(rate) = -16.0048  with SE of 2995 !!

WORST STILL, IF YOU USE THIS BAND AS THE REF. BAND, THEN ALL OF
THE OTHER beta’s for other age categories are DIFFERENCES in log(rate)
between these other bands and the ref. band.. since the ref. beta is unstable,
all of the difference will also be unstable !!

SO you should choose as a reference band one that has lots of information..
(see after this printout, below)
```

```
>   summary(glm(lex.Xst ~ m.cat + as.factor(a.cat) + as.factor(p.cat)
+                       + w.cat, offset=log(lex.dur),family=poisson,data=ds))

Coefficients:
                    Estimate Std. Error   z value Pr(>|z|)
(Intercept)         -16.0048  2995.2190    -0.005 0.995737   <<<<<<<<
m.cat                 0.3659     0.1131     3.235 0.001217 **
as.factor(a.cat)2    -0.7470  3316.8495 -0.000225 0.999820
as.factor(a.cat)3    -1.0992  3190.9705 -0.000344 0.999725
as.factor(a.cat)4    13.5283  2995.2189     0.005 0.996396
as.factor(a.cat)5    11.8910  2995.2190     0.004 0.996832
as.factor(a.cat)6    11.8216  2995.2189     0.004 0.996851
as.factor(a.cat)7    12.6073  2995.2188     0.004 0.996642
as.factor(a.cat)8    12.6146  2995.2188     0.004 0.996640
as.factor(a.cat)9    13.2639  2995.2188     0.004 0.996467
as.factor(a.cat)10   13.4384  2995.2188     0.004 0.996420
as.factor(a.cat)11   13.9427  2995.2188     0.005 0.996286
as.factor(a.cat)12   14.3057  2995.2188     0.005 0.996189
as.factor(a.cat)13   14.6546  2995.2188     0.005 0.996096
as.factor(a.cat)14   15.2288  2995.2188     0.005 0.995943
as.factor(a.cat)15   15.4980  2995.2188     0.005 0.995872
as.factor(a.cat)16   16.2175  2995.2188     0.005 0.995680
as.factor(a.cat)17   16.7190  2995.2188     0.006 0.995546
as.factor(a.cat)18   17.1382  2995.2188     0.006 0.995435
as.factor(a.cat)19   17.9220  2995.2188     0.006 0.995226
as.factor(p.cat)3    -3.0222     1.2311    -2.455 0.014096 *
as.factor(p.cat)4    -2.3698     1.0413    -2.276 0.022855 *
as.factor(p.cat)5    -2.3877     1.0293    -2.320 0.020359 *
as.factor(p.cat)6    -2.5774     1.0267    -2.510 0.012057 *
as.factor(p.cat)7    -2.5282     1.0231    -2.471 0.013466 *
as.factor(p.cat)8    -2.5671     1.0228    -2.510 0.012075 *
as.factor(p.cat)9    -3.0014     1.0249    -2.928 0.003407 **
as.factor(p.cat)10   -5.1151     1.4273    -3.584 0.000339 ***
w.cat                -0.1719     0.1204    -1.427 0.153462

#############

# redo the age-bands so the ref band is band 16, with 50 deaths

# 16->1 17->2 18->3 19->4    1->5  2->6 ... 15->19
ds$age.band16.as.ref =
 (ds$a.cat >= 16)*(ds$a.cat-15) + (ds$a.cat<16)*(ds$a.cat +4)

ds[1:10,]

   a.cat p.cat m.cat w.cat lex.Xst lex.dur age.band16.as.ref
1      6     2     0     0       0     5.0                10
2      7     2     0     0       1     1.5                11
3     10     2     0     0       0     2.0                14
4      2     3     0     0       0     2.0                 6
5      3     3     0     0       0     3.0                 7
6      4     3     0     0       0     7.0                 8
7      5     3     0     0       0    44.0                 9
8      6     3     0     0       0    53.0                10
9      7     3     0     0       0    47.0                11
10     8     3     0     0       0    23.0                12

>   summary(glm(lex.Xst ~ m.cat + as.factor(age.band16.as.ref) + as.factor(p.cat)
+                       + w.cat, offset=log(lex.dur),family=poisson,data=ds))
```

```
Call:
glm(formula = lex.Xst ~ m.cat + as.factor(age.band16.as.ref) +
    as.factor(p.cat) + w.cat, family = poisson, data = ds, offset = log(lex.dur))

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.0683290  -0.6245200  -0.2501901  -0.0001009   2.6790100

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                       0.2127     1.0313   0.206 0.836609
m.cat                             0.3659     0.1131   3.235 0.001217 **
as.factor(age.band16.as.ref)2     0.5015     0.2269   2.210 0.027090 *
as.factor(age.band16.as.ref)3     0.9207     0.2954   3.116 0.001831 **
as.factor(age.band16.as.ref)4     1.7045     0.5274   3.232 0.001231 **
as.factor(age.band16.as.ref)5   -16.2175  2995.2188  -0.005 0.995680
as.factor(age.band16.as.ref)6   -16.9645  1424.8351  -0.012 0.990500
as.factor(age.band16.as.ref)7   -17.3167  1100.4349  -0.016 0.987445
as.factor(age.band16.as.ref)8    -2.6892     0.7290  -3.689 0.000225 ***
as.factor(age.band16.as.ref)9    -4.3265     1.0138  -4.268 1.98e-05 ***
as.factor(age.band16.as.ref)10   -4.3959     0.7284  -6.035 1.59e-09 ***
as.factor(age.band16.as.ref)11   -3.6102     0.4373  -8.255  < 2e-16 ***
as.factor(age.band16.as.ref)12   -3.6029     0.4079  -8.832  < 2e-16 ***
as.factor(age.band16.as.ref)13   -2.9536     0.3066  -9.634  < 2e-16 ***
as.factor(age.band16.as.ref)14   -2.7791     0.2910  -9.550  < 2e-16 ***
as.factor(age.band16.as.ref)15   -2.2748     0.2512  -9.055  < 2e-16 ***
as.factor(age.band16.as.ref)16   -1.9118     0.2333  -8.193 2.54e-16 ***
as.factor(age.band16.as.ref)17   -1.5629     0.2220  -7.039 1.94e-12 ***
as.factor(age.band16.as.ref)18   -0.9887     0.2044  -4.838 1.31e-06 ***
as.factor(age.band16.as.ref)19   -0.7195     0.2097  -3.431 0.000602 ***
as.factor(p.cat)3                -3.0222     1.2311  -2.455 0.014096 *
as.factor(p.cat)4                -2.3698     1.0413  -2.276 0.022855 *
as.factor(p.cat)5                -2.3877     1.0293  -2.320 0.020359 *
as.factor(p.cat)6                -2.5774     1.0267  -2.510 0.012057 *
as.factor(p.cat)7                -2.5282     1.0231  -2.471 0.013466 *
as.factor(p.cat)8                -2.5671     1.0228  -2.510 0.012075 *
as.factor(p.cat)9                -3.0014     1.0249  -2.928 0.003407 **
as.factor(p.cat)10               -5.1151     1.4273  -3.584 0.000339 ***
w.cat                            -0.1719     0.1204  -1.427 0.153462
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 821.96  on 466  degrees of freedom
Residual deviance: 299.57  on 438  degrees of freedom
AIC: 741.85

NOW, relative to the (log) rate in the new ref. band, we can see the increasing
(log) death rates in the original age bands 17-19 (i.e. in new 2 3 4)

The large negative beta for the new age band 5 (youngest) is still unstable,
(indeed there were no deaths in the 3 youngest age bands)
but from new band 8 onwards (old 12 onwards) we see A CLEAR MONOTONE PATTERN.

The successive (de)increments in the log(rate) with successive categories
 suggest that interval (‘continuous’) versions of these are adequate --
 and they make for a far more parsimonious model.

THE TAKE-HOME MESSAGE
---------------------
CHOOSE REF. CATEGORY CAREFULLY, so you get to see what is in the data.
```

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.

Females	period: age band	**ref.**	next	next	etc.
	↑	↑			
	next	$\times 1.55^2$			
	next	$\times 1.55$			
	ref.	$\boxed{0.00019 \text{ deaths/py}}$	$\times 0.88$	$\times 0.88^2$	→

Males: Each of above entries multiplied by 1.4
Oscar: Each of above entries multiplied by 0.83

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]

- We saw an increment of 7-10% per year in all-cause mortality in other datsets (eg Danish) and so 1..55 for 5 years is similar ($1.09^5 = 1.54$)

There have been dramatic falls in age-specific all-cause mortality in the adult-age bands over the 20th century, and so the direction makes sense. From 1970 to 2000, age-specific mortality rates in Quebec fell by approx. 50% – but the aging of the population nullified it, in the sense that the crude (not adjusted for age changes) death rates in 1970 and 2000 are about the same!

Based on the death rates, men should pay 40-60% higher life insurance premiums! This is also déjà vu.

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]

```
              McGill                                                   Redelmeier
> aggregate(ds$lex.dur,by=list(ds$w.cat),FUN="sum")   > aggregate(ds.r$lex.dur, ...)

  Group.1     x                                            Group.1     x
1       0 14558                                             1       0 13638
2       1  6223                                             2       1  7143

About 1000 PY (920 to be exact) of what Redelmeier allocated as PY lived by Oscar
winners are in fact PY lived as nominees.. this explains why his death rates for
what he calls the Oscar "group" are lower than they should be (and his rate ratio
of 0.77 (below) so impressive). These 1000 years are ''immortal'' in the sense that
by definition (they refer to years of performers who subsequently won an Oscar)
they cannot contain deaths -- if the performer died in these years, (s)he could
never be Winner!

You might think this is an inconsequential example, with no serious impact on
epidemiology, but when this type of immortal time bias  involves some high-profile
studies involving the role of statins in the prevention of diabetes, its more serious.
See for example, Problem of immortal time bias in cohort studies: example using statins
for preventing progression of diabetes. LE Levesque, JA Hanley, A Kezouh, S Suissa
BMJ  2010;340:b5087, doi: 10.1136/bmj.b5087, (Published 12 March 2010)
```

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.

```
> fit.age.int.scale = glm(lex.Xst ~ m.cat + a.cat + p.cat
+             + w.cat, offset=log(lex.dur),family=poisson,data=ds.r)
>
> round(exp(fit.age.int.scale$coefficients[1]),5)
(Intercept)
    0.00019
>
> round(exp(fit.age.int.scale$coefficients[2:5]),2)
m.cat a.cat p.cat w.cat
 1.41  1.55  0.88  0.77 <<<<<<<< more extreme than the 0.83
```

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?

```
> # split the dataset into 2 sets of records,
>
> # one for the performer time spent as winner
```

[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

```
>   index.cat=ds[ds$w.cat==1,]; index.cat[1:3,]
    a.cat p.cat m.cat w.cat lex.Xst lex.dur
258     4     2     0     1       0       2
259     7     2     0     1       0       1
260     4     3     0     1       0       4
>
> # and one for the performer time spent as nominee
>
>   ref.cat=ds[ds$w.cat==0,];   ref.cat[1:3,]
  a.cat p.cat m.cat w.cat lex.Xst lex.dur
1     6     2     0     0       0     5.0
2     7     2     0     0       1     1.5
3    10     2     0     0       0     2.0
>
> # put these datasets side by side
>
> # 2 variables with same name, so use suffix to distinguish them
> # .w for winner time  .n for ’not-winner’ (nominee) time
>
> # lex.Xst.w will denote deaths in winner-time
> # lex.dur.w will denote ’as winner’ p-t
>
> # lex.Xst.n will denote deaths in nominee-time
> # lex.dur.n will denote ’as nominee’ p-t
>
>   ds.for.mh=merge(index.cat,ref.cat,by=c("a.cat","p.cat","m.cat"),
+                      suffixes=c(".w",".n"), sort=TRUE) ;
>
> head(ds.for.mh) ; tail(ds.for.mh)
  a.cat p.cat m.cat w.cat.w lex.Xst.w lex.dur.w w.cat.n lex.Xst.n lex.dur.n
1    10    10     0       1         0        11       0         0        21
2    10    10     1       1         0         5       0         0        14
3    10     3     1       1         0        17       0         0        17
4    10     4     0       1         0        16       0         0        44
5    10     4     1       1         0        34       0         2        85
6    10     5     0       1         0        59       0         0       111
    a.cat p.cat m.cat w.cat.w lex.Xst.w lex.dur.w w.cat.n lex.Xst.n lex.dur.n
197     9     7     0       1         0        76       0         1       199
198     9     7     1       1         0        57       0         2       153
199     9     8     0       1         0        57       0         2       170
200     9     8     1       1         0        43       0         1       198
201     9     9     0       1         0        77       0         0       181
202     9     9     1       1         0        49       0         1       126

# calculate #.exposed.cases*unexposedPT/PT
#       and #.unexposed.cases*exposedPT/PT     # ... fill in the formula

ds.for.mh$num.mh = ds.for.mh$lex.Xst.w * ds.for.mh$lex.dur.n /
                     ( ds.for.mh$lex.dur.w + ds.for.mh$lex.dur.n);

ds.for.mh$den.mh = ds.for.mh$lex.Xst.n * ds.for.mh$lex.dur.w /
                     ( ds.for.mh$lex.dur.w + ds.for.mh$lex.dur.n);
head(ds.for.mh)

> ds.for.mh[10:15,]
   a.cat p.cat m.cat w.cat.w lex.Xst.w lex.dur.w w.cat.n lex.Xst.n lex.dur.n    num.mh    den.mh
10    10     7     0       1         0        86       0         3       202 0.0000000 0.8958333
11    10     7     1       1         0        75       0         3       134 0.0000000 1.0765550
12    10     8     0       1         1        79       0         2       137 0.6342593 0.7314815
13    10     8     1       1         0        38       0         1       190 0.0000000 0.1666667
14    10     9     0       1         1        68       0         1       215 0.7597173 0.2402827
15    10     9     1       1         0        58       0         0       155 0.0000000 0.0000000 << notice that this stratum is uninformative w.r.t. the RateRatio
```

```
> sum(ds.for.mh$num.mh) ;  sum(ds.for.mh$den.mh) ;
[1] 59.21302
[1] 72.52263
>
> round(sum(ds.for.mh$num.mh)  /  sum(ds.for.mh$den.mh), 2) ;

[1] 0.82 <<<<<< hazard (or rate) ratio:

   v. close to (more model-based) h.r from Poisson regression



 LET’S NOT stop ’standing on the shoulders of giants’ [To use Phil Cole’s words from
 foreward to the Breslow and Day textbook, where he pays tribute to the
 trailblazers in the statistical design and anaysis of epidemiology studies ]
 just because the methods are 51 years ‘old’ and their inventors now dead.
```

— 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`

—

See Links to material re “Immortal-time” bias at bottom of webpage on regression for incidence rates... including a topic that has been in the news recently...

Do statins delay diabetes onset? Lévesque, Hanley, Kezouh, Suissa 2010

http://www.bmj.com/cgi/content/full/340/mar12_1/b5087