## 1. Analysis of IHD data in Table 22.6 of Clayton & Hills

In the 2008.02.22 class, we discussed how to fit an '*additive*' rate model, but because jh hadn't had time to set up the products of each '$X$' with the PT variable, we didn't actually fit this model (the code in the one extra page appended to the photocopy of the Clayton and Hills Ch 23 was only for fitting multiplicative models). The `R` code (now available under the resources for 'Regression models for (incidence) rates.') has been updated to include that for fitting the additive model. There are also additional notes interspersed with the code.

i. Fit an additive model[1], and present the results in the same format as Table 22.7 of Clayton and Hills (Ch 22 was handed out on 2008.02.08, and is also available in the resources for the bios602 course).

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 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. (Sex-specific) All-cause death rates: Québec 1971 *vs.* 2002

If male, do the 1971 vs. 2002 comparison within males; if female do the comparison within females (dataset/Rcode in Resources - subset to these 2 years; `dac` = no. deaths all causes). Limit analysis to the age-span 40-85.

Remember that the population numbers are estimates as of July 1 of that year. Assume the population size in the age-band is constant over the year, so the the number of *person-years* giving rise to the deaths is $PT = MidYearPopulation \times 1year$. Thus, what we are dealing here is an open population (NOT a cohort), with some persons crossing from one age band to another, or coming into an age band from out-of-province, or going out-of-province, during the year. You can think of a 5-year-age-band *times* 1-year-time-band as a $5 \times 1$ 'rectangle' in the Lexis 'space.'

i. Calculate and compare the crude all-cause mortality rates for 1971 and 2002.

ii. Calculate the "directly" standardized all-cause mortality rates for 1971 and for 2002, using as weights (as the standard) an average of the 1971 and 2002 age structures. Comment on your findings.

iii. Take the 'corner' (cf Clayton & Hills) as the age-band 40-45 (reference-age) in 1972 (reference-year). Fit the multiplicative rate regression model

$$Rate_{age,year} = Rate_{corner} \times MRR_{age:ref-age} \times MRR_{year:ref-year},$$

where MRR is shorthand for 'Mortality Rate Ratio.'

Interpret your findings, and compare with your results in (i).

*To decide check whether the relationship of the log rates with age can be represented using a simple linear age term, e.g., using the mid-point of each age-band as a single 'interval' / 'continuous' age term, rather than as a more complex one that relies on indicator ('dummy') variables for age categories, plot the log-rates versus age.*

iv. Fit the model

$$Rate_{year} = Rate_{ref-year} \times MRR_{year:ref-year},$$

and interpret the fitted coefficients and the antilogs of these coefficients.

## 3. (2002) All-cause death rates: Québec Males *vs.* Females

Limit your analysis to the year 2002, and the age-span 40-85.[3]

i. Calculate and compare the crude all-cause mortality for males and females.

ii. Calculate the "directly" standardized rates for all-cause mortality for males and for females, using as weights (as the standard) an average of the male and female age structures. Compare with (i), and comment.

[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) 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 ?

[3] Although we focus here on just one calendar year, you can easily imagine a 3-variable regression model involving age, calendar year, and gender.

iii. Using 40-45 year old females as the 'corner' category, and age as a linear variable, fit the model

$$Rate_{age,sex} = Rate_{corner} \times MRR_{age:ref-age} \times MRR_{sex:ref-sex},$$

and interpret the fitted coefficients and the antilogs of these coefficients.

iv. Plot the fitted log-rates against age.

v. You have just fitted a *proportional-rates* model, i.e., one with a constant $MRR_{male:female}$ over ages, so that the 2 sets of log-rates, plotted vs. age, are a constant distance apart (i.e., the log-rate curves are 'parallel').

To see if this is a good fit, you could plot the actual and the fitted rates on the same graph and judge the fit 'by eye.'

In addition, you could fit a model in which the fitted straight lines will not be parallel, by adding a *male* × *age* product term. Do so, and superimpose the 2 fitted lines under this more complex model, containing an interaction term, and compare with those from the simpler model that does not. Comment.

*It might take more complex sex-specific form to reproduce the rate vs age curves adequately. However, Gompertz* (1779-1865) found that this 'log-rates are linear in age' (model to be accurate over a wide age-span).*

**See* `http://en.wikipedia.org/wiki/Gompertz-Makeham_law_of_mortality`

## 4. 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 the 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 Resources you will find (a) the Oscar data set[4] with one data-record per performer (b) a dataset (with approx. 20,000 records) in which each the performer's 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) have been aggregated into 'sex-age-period-Oscar' cells, with 5-year age-bands and 10 year calendar-year-bands,[5] and (d) a file similar to (c), but where *all* of a performer's 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.[6]

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'[7] 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.[8] 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.

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

[5] Do the analysis 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.

[6] 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.

[7] 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.'

[8] 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/`). 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.

i. To compare the death rates in the performer-years lived as nominees (reference category, `w.cat=0`) versus those lived as winners (index category, `w.cat=1`), fit the following multiplicative (i.e. 'rate ratio') model[9] 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 $MMR$, which in turn is short for Mortality Rate Ratio, the (theoretical, unknown, to be estimated) ratio of the mortality rate in the category[10] of the variable in question relative to the reference category of that 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*).

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 MMR for the 'years lived as a winner' vs. 'years lived as a nominee' contrast.

iii. Comment on the fitted effects of gender[11], age and calendar time, and whether they 'fit' with what you expect, and have seen in other datasets.[12]

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.[13]

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?

vii. Use this same dataset to calculate separate Mantel-Haenszel mortality rate ratios for actors and actresses. Based just on the numbers of deaths involved, do you think they are statistically significantly different?

If you wanted to pursue this effect-modification numerically, you could use the formula to obtain the SE of each rate ratio (or rather the SE of the log-rate-ratio). The formula is given in section 3.6(d) of Breslow and Day Volume II. It is quite tedious to do by hand, but quite easy with `R` or Excel.

[9] 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...

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

[11] 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'.

[12] 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.

[13] 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 (sign in with campus\your-das-name).