(1-2): "Rosuvastatin to Prevent Vascular Events in Men and Women with Elevated CRP" [Ridker, nejm nov 20, 2008]; (3): Clayton & Hills chapter 22.

In the 1st assignment, you analyzed a large dataset that reflects as closely as possible the two reported cumulative incidence curves, and the censoring pattern for the primary end point. For Q1-2, focus on a reduced dataset (`TxYears.txt`, attached) with 10 observations, 1 per year for each treatment arm. Each row or 'cell' consists of the statin arm indicator, the year (1,2, ... 5), the number of events, and the number of PY of observation in that 'cell'. Change 'Year' to 0.5, 1.5, ... , 4.25 (the last window is from $t = 4$ to $t = 4.5$).

1 Pattern of incidence rates vs. time of follow-up (t) in Control arm of JUPITER study

i. Fit the smooth-in-time multiplicative ID (rate) model[1] – epidemiologic –

$$ID(t) = ID_{t=0} \times \exp(\beta t) = \exp(\log[ID_{t=0}] + \beta t) = \exp(\beta_0 + \beta_1 t),$$

to the data from the placebo arm.

ii. Interpret the coefficients of the – statistical – model you fitted by software, and convert them to the parameters of the epidemiologic (ID) model.

iii. Plot the fitted $ID(t)$ function as a smooth function of t. In `R`, e.g, ...

```
time=seq(0,5,0.1); fitted.ID = ...  ; plot(time,fitted.ID)
```

Estimate – by eye/computer – the area under this fitted $ID(t)$ curve, and from it calculate the estimated 5-year cumulative incidence (5-year risk).

iv. Based on the fitted model, how much higher is $ID_{t=2.0}$ than $ID_{t=0}$? Calculate a 95%CI for this ratio. Hint: CI for ID ratio = exp[CI for log(ID ratio)]. Do the same for $ID_{t=3.2}$ vs. $ID_{t=1.2}$.

v. Calculate a 95%CI for $ID_{t=5}$, To start, calculate $V = Var\{\log[\widehat{ID_{t=5}}]\}$; from it, obtain a CI for the log; then convert to the CI for its antilog:

$$\exp\{\log[\widehat{ID_{t=5}}] \mp 1.96 \times V^{1/2}\}.$$

The formula for V is an example of the *general* formula for the variance of a linear combination of two parameter estimates, since $\log[\widehat{ID_{t=5}}] = \widehat{\beta_0} + 5 \times \widehat{\beta_1}$. You probably learned the formula for the variance of a linear combination of two uncorrelated estimates, but maybe not of two correlated estimates, or in general, two correlated random variables Y_1 and Y_2, with variances V_1 and V_2, and covariance V_{12}. The general formula for the sum of a constant c_1 times Y_1 and a constant c_2 times Y_2 is[2]

$$Var[c_1 \times Y_1 + c_2 \times Y_2] = c_1^2 V_1 + c_2^2 V_2 + 2c_1c_2V_{12}.$$

In our example, $c_1 = 1,\ Y_1 = \widehat{\beta_0},\ \ c_2 = 5,\ Y_2 = \widehat{\beta_1}$.

Typically, the standard output from a regression model gives $SE_1 = V_1^{1/2}, SE_2 = V_2^{1/2}, \ldots$ – so one can back-calculate V_1, V_2 etc. – but not the covariance term(s). But one can get both the variances and covariances directly from the `(summary(fit))$cov.unscaled` portion of the `summary` object – see the structure (`str`) of the `summary` object.

Repeat the 95%CI calculation for t nearer the middle of the studied range, i.e., $ID_{t=2.5}$ and at the other extreme $ID_{t=0}$.

vi. From the fitted counts of numbers of events (obtainable from the same objects) calculate a X^2 goodness of fit statistic. Note that this statistic is calculated from the actual frequencies (numbers of cases), not from the observed and fitted rates (or in the case of logistic regression, not from the observed and fitted proportions or percentages). Because we fitted 2 parameters to 5 datapoints, the *df* for the X^2 is 5-2=3.

vii. Peruse the other components of the objects produced by fitting and by `summary`, and comment on any that you recognize.

2 Rate ratios and rate differences; (non)-proportionality

i. Fit the smooth-in-time multiplicative ID (rate) model – epidemiologic –

$$ID(t) = ID_{t=0,tx=0} \times \exp(\beta_t t + \beta_{tx} I_{tx}) = \exp(\log[ID_{0,0}] + \beta_t t + \beta_{tx} I_{tx}),$$

where I_{tx} is an Indicator[3] variable for the statin arm, i.e. taking on the value 1 if the cell represents the index category (PT contributed by those 'assigned to the statin'), and 0 otherwise ('assigned to placebo').

ii. Interpret the coefficients of the – statistical – model fitted by the software, and convert them to the parameters of the epidemiologic (ID) model.

[1] The link opposite the march 5 entry in JH's `c634` site, reachable from `http://www.biostat.mcgill.ca/hanley` contains `R (glm)` and `SAS (genmod)` code that does this for other datsets, and some `R` code for this dataset. There is also a longer set of notes on fitting 'rate regression' models.

[2] For a linear combination 3, it is $c_1^2V_1 + c_2^2V_2 + c_3^2V_3 + 2c_1c_2V_{12} + 2c_1c_3V_{13} + +2c_2c_3V_{23}$.

[3] Called a 'dummy' variable in less sophisticated circles.

iii. On the same graph, plot the fitted $ID(t)$ function as 2 smooth functions of t, one for the statin arm, one for placebo.

iv. Calculate the fitted ID ratio $\left(\frac{ID_{tx=1}}{ID_{tx=0}}\right)$ at $t = 2$ and again at $t = 5$. Comment on the proportionality or otherwise of the fitted ID's. Also, calculate the fitted difference $(ID_{tx=1} - ID_{tx=0})$ at these two t values.

v. On 1 graph, plot the fitted $\underline{\log}[ID(t)]$ function as 2 smooth functions of t, one for the statin arm, one for placebo. What are the implications of the choice of scale on the presence or absence of 'effect modification'?[4]

vi. From the fitted counts of numbers of events, calculate a X^2 g-o-f statistic. Because we fitted 3 parameters to 10 datapoints, $df = 10 - 3 = 7$.

vii. Fit the smooth-in-time $\log[ID(t)]$ model

$$\log[ID(t)] = \log[ID_{t=0,tx=0}] + \beta_t t + \beta_{tx} I_{tx} + \beta_{product} ProductTerm,$$

where $ProductTerm = t \times I_{tx}$, and on the same graph plot the 2 fitted $ID(t)$ functions, 1 for each arm. What is the ID Ratio $(ID_{tx=1}/ID_{tx=0})$ at $t = 2$ and at $t = 5$? Comment.

viii. Fit *separate* smooth-in-time $\log[ID(t)] = \log[ID_0] + \beta_t t$ models for the 2 arms (you have already fitted the one for the placebo arm). Compare them with the 4-parameter model fitted in the previous sub-question.

ix. (Pas déjà vu) Fit the <u>additive rates</u> (constant ID *difference*) model:

$$ID(t) = ID_{t=0,tx=0} + \gamma_t t + \gamma_{tx} I_{tx} = \gamma_0 + \gamma_t t + \gamma_{tx} I_{tx},$$

interpret the fitted coefficients, draw the 2 fitted ID functions, and get a sense of the closeness of the fit.

Hint: as is explained in the 1-page handout "Regression Models for Rates - Summary 2008.02.22", the trick for fitting an additive rate model via glm is to first expand the model for the expected number of events in each cell

$$E[\#events] = ID_{t_{mid}} \times PT = \{ID_{0,0} + \gamma_t t_{mid} + \gamma_{tx} I_{tx}\} \times PT,$$

where t_{mid} is 0.5, 1.5, etc.. Expanded, this becomes

$$E[\#events] = \gamma_0 \times \underline{PT} + \gamma_t \times \underline{t_{mid} \times PT} + \gamma_{tx} \times \underline{I_{tx} \times PT}.$$

In this case therefore, we are directly modelling $E[\#events] = \mu_{events}$, and so we use the "`identity`" link that leaves the μ untransformed.

The 3 'regressor variables (underlined) in this statistical model are PT, and its products with the 2 regressor terms (t_{mid} & I_{tx}) in the ID model.

Notice also that $\gamma_0 \times \underline{PT}$ is not constant from cell to cell and so, unlike in most regression models, there is no 'constant intercept'. Even though we have called it γ_0 to show that it is the ID at (0,0), in fact it is applied to a column of PT values – and these *vary* from observation to observation.[5] So in fact this is a 3-parameter regression model with an *intercept of zero* – this makes sense since if there is no PT, there are no cases! To *force the intercept to be zero*, we specify a '-1' in the model.

Thus the `R` syntax for fitting this 3-parameter additive ID model is

$$\texttt{glm}(events \sim -1 + z1 + z2 + z3, family = poisson(link = \text{``}identity''\text{)}),$$

where

$$z1 = PT;\ z2 = t_{mid} \times PT;\ z3 = I_{tx} \times PT.$$

Exercise: Use the equations above to identify what the 3 coefficients represent in the (epidemiologic) ID model (as opposed to in the fitted "statistical" model that has fitted numbers cases on the left side).

3 C & H example

i. *(Using an analogy with the method you used in Q2(i), but with age as a categorical variable, represented by two indicator variables that you are better off creating yourself)* Fit the (multiplicative) regression model in Table 22.5 of Clayton and Hills chapter 22 to the dataset shown in Table 22.6 of this same chapter, and check the estimates against the results shown in Table 22.7.[6] Note in Table 22.5 C&H's use of the word '*corner*' for the rate in the *reference* cell.

[4] 'Effect modification' is a more expressive term than 'interaction'. We use 'interaction' (product) terms in statistical models so that the linear predictor can involve products of the primary variates, but I urge you to avoid the term interaction when speaking to outsiders. The term effect modification immediately conveys the message that there are 'different slopes (or effects) for different folks'; the term interaction does not. The term effect modification immediately implies <u>3</u> variables: the X ('exposure', 'agent', etc), the Y ('outcome', 'mean', 'risk', 'rate', etc) and the modifier M (age, sex, genetic subtype etc). However, interaction in everyday parlance just involves 2 items/actors. Example, courtesy of Miettinen : "love makes time pass; time makes love pass."

[5] You will sometimes see an `R` regression model with a '1' as the first term on the right hand side e.g. $y \sim 1 + X1 + x2$ when what is meant is that the expected value is $\beta_0 \times 1 + \beta_1 \times X1 + \beta_2 \times X2$: the software makes a design matrix with 3 columns of predictor variables, a column of 1's, a column of the $X1$ values, and a column of the $X2$ values – even though technically calling the first column the 'variable' $X0$' is stretching it a bit, since the $X0$ values have no variation.

[6] The resources opposite the march 5 entry in JH's `c634` site, reachable from `http://www.biostat.mcgill.ca/hanley` has `R (glm)` and `SAS (genmod)` code that does this.