Summary of Poisson Models for Mesothelioma Incidence Projections: as used in HSE models

Timothy Wyant

12/21/13

Contents

1 The Poisson distribution

Counts – such as the number of mesothelioma cases in a year – are often modeled using the Poisson distribution.

$$f(k; \lambda) = \Pr(Y = k) = \lambda^k \frac{e^{-\lambda}}{k!}$$

This distribution has mean λ and variance λ, where $\lambda > 0$.

Sometimes the counts are related to an exposure variable t - amount of elapsed time, number of square miles, or number of people *P*. In such cases, the Poisson distribution is typically expressed as:

$$f(k; \lambda, P) = \Pr(Y = k) = (\lambda P)^k \frac{e^{-\lambda P}}{k!}$$

In simple terms, if two countries were similar in their adaptation and use of asbestos over time, but country A had twice the population as country B, we would expect that there would be twice as many mesothelioma cases in country A as in country B.

In fitting observations using Poisson regression using a function like R's $glm()$, exposure variables are accounted for in the model statement by including offset(log(P)) as a model term.

The sum of independent Poisson variables is Poisson. So if Y_1 is Poisson with mean $\lambda_1 P$, and Y_2 is Poisson with mean $\lambda_2 P$, then $Y = Y_1 + Y_2$ is Poisson with mean $(\lambda_1 + \lambda_2)P$.

If we have n observed counts y_i from a Poisson-distributed Y, with no exposure variable P, the log of the likelihood function is (ignoring constant terms that do not depend on λ):

$$l(\lambda, y) = \sum_{i=1}^{n} y_i \log \lambda - n\lambda$$

Sometimes the independent Poisson distributions are not identical. One simple example is where there are different exposure variables P_i for each y_i, and the expected value of each y_i is λP_i. In this case the log-likelihood is:

$$l(\lambda|y,P) = \sum_{i=1}^{n} (y_i \log(\lambda P_i) - \lambda P_i)$$

As another example, suppose the expected value of each y_i depends on an independent variable x_i. If Y is the number of mesothelioma cases,for example, X could be year of occurence T. (This would be a very simple and not very useful model, but suffices for illustrative purposees.) A standard form of the log-likelihood given n observations of x and y would be:

$$l(\alpha, \beta|y,P,x) = \sum_{i=1}^{n} (y_i \log(\lambda_i P_i) - \lambda_i P_i)$$

where

$$\lambda_i = \ \exp(\alpha + \beta x_i)$$

In this specification, I have made the commonly used assumption that the log of λ is a linear combination of independnt variables. In the parlance of Generalized Linear Models, and R's *glm*() function, the "link function" is the natural log.

2 Fitting the HSE model for mesothelioma incidence – structure

HSE (Health Safety Executive – a UK government occupational health agency) models for national mesothelioma incidence can be interpreted as Poisson models.

A certain number of asbestos cases Y that occur among people age A in year T are due to exposures l years earlier, in year $t = T - l$, when these people were age $a = A - l$. The number of cases from this earlier exposure can be modeled as Poisson:

$$f(k; \lambda_l, P_{A,T}) = \Pr(Y = k) = (\lambda_l P_{A,T})^k \frac{e^{-\lambda_l P_{A,T}}}{k!}$$

More to the point, the number of mesothelioma cases $Y_{A,T}$ due to occupational exposure among the number of living people $P_{A,T}$ is distributed as Poisson, with mean:

$$\sum_{l=1}^{A-1} \lambda_{l,A,T} P_{A,T}$$

In other words, l is the lag of years back from A. In standard fits of the HSE model for mesothelioma incidence the means $\lambda_{l,A,T}$ are assumed to have the form:

$$\lambda_{l,A,T} = \exp(\alpha + W(a = A - l) + D(t = T - l) + k \log(\ \max(\ 0, l + 1 - L)) \text{ where:}$$

$L =$ 10 years
$W =$ is a smooth function of a,
$D =$ is a smooth function of t, and
$k =$ the exponent that reflects the exponential growth in mesothelioma risk as time since exposure increases. The exponent k in HSE models should generally be between 2 and 3.

L is the minimum latency period for an exposure to cause a case of mesothelioma, and is assumed to be 10. If l is less than 10, the quantity $\log(\ \max(\ 0, l+1-L))$ is -Inf. The quantity exp(-Inf) is taken to be zero, so the Poisson mean is zero as well.

W(a) is a function that captures the notion that the intensity of occupational exposure in a population varies with a=age – usually going up sharply as people enter the workforce in their late 'teens and early 20s, staying high in the 20s and early 30s, and declining from then on. In my fits of the HSW model, I construct a spline basies for a, usually with at least three degrees of freedom to capture asymmetry. In R parlance, for three degrees of freedom, $W(a) = ns(a, df = 3)$.

Similarly, D(t) is a functon that captures the notion that the intensity of occupational exposure in a population varies with t=year – usually going up in the 1930s, peaking sometime in the period 1950-1970, and declining and eventually dropping close to zero (for developed nations) sometime in the period 1980-2000. In my fitting, I typically construct a spline basis for t, just as I do for a.

For simplicity of exposition, I will assume in what follows that I have used 3 degrees of freedom for each spline, so that the smooth function W of a in the specifications of $lambda$ is a linar combination of $[a_1, a_2, a_3]$ (the spline basis vectors), and similarly the smooth function D of t is a linear combination of $[t_1, t_2, t_3]$.

Then, for any population age A and year T, the Poisson mean for mesotheilioma cases generated in year T by exposures in year T-l looks like this:

$$\begin{aligned}\lambda_{l,A,T}P_{A,T} = \exp(&\alpha + \omega_1 * a1 + \omega_2 * a1 + \omega_3 * a1 + \\ &\delta_1 * t1 + \delta_2 * t2 + \delta_3 * t3 + k\ \log(\ \max(\ 0, l+1-L))P_{A,T}\end{aligned}$$

and the Poisson mean for the total number of mesothelioma cases at age A in year T that are due to previous occupational exposure is:

$$\lambda_{A,T}P_{A,T} = (\sum_{l=1}^{n} \lambda_{l,A,T})P_{A,T})$$

3 Fitting the HSE model for mesothelioma incidence – implementation

Assume you have i=1,...,n observations of P_i (population size) and Y_i (number of mesothelioma cases) at age A_i in year T_i. The goal is to maximize this log-likelihood:

$$l(\alpha, \omega, \delta, k|Y, P, x) = \sum_{i=1}^{n}(Y_i \log(\lambda_i P_i) - \lambda_i P_i)$$

where λ_i is λ_{A_i,T_i} from the final equation from the preceding chapter.

Step 1.

Let:

$N = \sum_{i=1}^{n}(A_i - 1)$
$K =$ the number of parameters to be fit; this is 8, if W and D each have three degrees of freedom

N is the total number of age-year occupational exposure combinations that produce the the mesothelioma counts over all the A and T combinations.

Build a working matrix WM with N rows, and columns for: $A, T, l, a, t, 1, a_1, a_2, a_3, y_1, y_2, y_3,$ $\log($ $\max(0, l + 1 - L))$ and assume it is sorted by A, T, and l.

β_{Kx1} is a vector of values for the eight parameters. These values will either be start values for the iterative fitting routines, or the final fitted values.

X_{NxK} is a subset of WM that has a column for each of the eight independent variables, including a column of 1's for the intercept term α.

$Aggregator_{nxN}$ is a matrix in which each row is a vector of 1's and 0's that identify the exposure years in WM associated with each of the A,T combinations, sorted in A,T order,

Then the λs for each of the n A,T combinations can be specified as:

$$\lambda_{nx1} = Aggregator_{nxN} * \exp(X_{NxK}\beta_{Kx1})$$

Then, given these λs, we can calculate the log-likelihood as

$$l(\lambda|y, P) = \sum_{i=1}^{n}(y_i \log(\lambda P_i) - \lambda P_i)$$

The function that yields the log-likelihood is called many times during numeric maximization using an R function like mle(). Specifying the log-likelihood in terms of matrix multiplications, as above, yields fast execution times for each invocation of the function.

4 Miscellaneous notes

4.1 Forcing the estimated total mesothelioma cases to equal the observed

4.2 Background mesothelioma incidence rates

4.3 Halflife – clearance of asbestos fibers from the lungs

4.4 Adjusting for early underdiagnosis