SORT 42 (2) July-December 2018, 237-252 DOI: 10.2436/20.8080.02.76

Poisson excess relative risk models: new implementations and software

Manuel Higueras[1,2,3] and Adam Howes[2]

Abstract

Two new implementations for fitting Poisson excess relative risk methods are proposed for assumed simple models. This allows for estimation of the excess relative risk associated with a unique exposure, where the background risk is modelled by a unique categorical variable, for example gender or attained age levels. Additionally, it is shown how to fit general Poisson linear relative risk models in R. Both simple methods and the R fitting are illustrated in three examples. The first two examples are from the radiation epidemiology literature. Data in the third example are randomly generated with the purpose of sharing it jointly with the R scripts.

MSC: 62J02

Keywords: Radiation epidemiology, Poisson non-linear regression, improper priors, R programming

1. Introduction

The excess relative risk (ERR) represents the additional risk of disease (e.g., leukaemia, brain tumour) per unit of exposure (e.g., absorbed dose of ionising radiation). In a linear ERR model with d exposures, the risk is modelled by

$$e^{\eta}\left(1+\sum_{j=1}^{d}\beta_j D^{(j)}\right),$$

where each parameter β_j is the ERR associated with the absorbed dose $D^{(j)}$. The risk is represented by the product of the background risk term, e^{η}, and the term within parenthesis, which is the relative risk. Poisson linear ERR models can be used to calculate the ERR in longitudinal cohort studies with active follow-up. It is assumed that

[1] Departamento de Matemáticas y Computación, Universidad de La Rioja, Edificio CCT - C/ Madre de Dios 53, 26006 Logroño (La Rioja), Spain.

[2] Basque Center for Applied Mathematics, Bilbao, Spain.

[3] Institute of Health and Society, Newcastle University, Newcastle upon Tyne, UK.

Received: January 2018
Accepted: November 2018

$$C_i \sim \text{Pois}\left(PY_i e^{\eta_i}\left(1+\sum_{j=1}^{d}\beta_j D_i^{(j)}\right)\right), \tag{1}$$

where C_i and PY_i are the number of disease cases and the number of person-years of follow-up, and $D_i^{(j)}$ is the mean dose (weighted by the person-years) of exposure j for stratum $i = \{1,\ldots,n\}$ respectively (BEIR VII Phase 2, 2006). The most common situation in ERR models is to have only one exposure variable. More complicated ERR models with effect modification of the dose-response are also often reported, e.g. Grant et al. (2017).

The background risk can be modelled by m covariates, *i.e.* $\eta_i = \alpha_0 + \sum_{k=1}^{m}\alpha_k x_i^{(k)}$. These covariates are usually time-dependent variables, e.g. attained age or transplant status. The model (1) is not the canonical log-linear Poisson model (McCullagh and Nelder, 1989). Since it mixes both log-linear and linear terms it is a generalised non-linear model.

In this work, Poisson ERR models with simple forms are studied to obtain estimates in closed or almost closed form. This allows calculations to be made faster and more accurate. As an alternative to other implementations in the literature, such as Epicure (Preston et al., 1993) and SAS (SAS Institute Inc., Cary, North Carolina) (Richardson, 2008), the software R (R Core Team, 2017) was used to fit general Poisson ERR models. Three applied examples are detailed, and the data and R scripts of the third example are included as supplementary material.

2. Simple ERR model

A simple ERR model may be defined by assuming one exposure, $d = 1$, and that the background risk linear predictor, η_i, is of the form $\alpha_1 + \sum_{k=2}^{K}\alpha_k \mathbf{1}_{\{k\}}(x_i)$, where x_i represents a categorical variable with K levels. This model is simple, with only one exposure, and one categorical covariate in the background risk term.

Following these assumptions

$$C_i \sim \text{Pois}\left(PY_i e^{\alpha_1+\sum_{k=2}^{K}\alpha_k \mathbf{1}_{\{k\}}(x_i)}\left(1+\beta D_i\right)\right). \tag{2}$$

Let $\vec{\alpha} = \{\alpha_1,\ldots,\alpha_K\}$ and $X = \{C, PY, D, x\}$, then the likelihood of the parameter set $\Theta = \{\vec{\alpha}\beta\}$ is given by

$$\mathrm{L}(\Theta|X) = \prod_{i=1}^{n}\frac{[PY_i e^{\alpha_1+\sum_{k=2}^{K}\alpha_k \mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i)]^{C_i}\exp(-PY_i e^{\alpha_1+\sum_{k=2}^{K}\alpha_k \mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i))}{C_i!} \tag{3}$$

and the log-likelihood is

$$
\begin{aligned}
l(\Theta|X) = \quad & \log(\mathrm{L}(\Theta|X)) = \sum_{i=1}^{n}\left[C_i(\log PY_i + \alpha_1 + \sum_{k=2}^{K}\alpha_k \mathbf{1}_{\{k\}}(x_i) + \log(1+\beta D_i))\right] \\
- \quad & \sum_{i=1}^{n} PY_i \mathrm{e}^{\alpha_1 + \sum_{k=2}^{K}\alpha_k \mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i) - \sum_{i=1}^{n}\log C_i!.
\end{aligned}
\tag{4}
$$

For this model two implementations are proposed, one is frequentist and the other is Bayesian. The frequentist implementation provides a closed form of the profile likelihood of β and the Bayesian provides the marginal posterior for β also in a closed form.

2.1. Profile likelihood and maximum likelihood estimator

A profile likelihood CI (PLCI) for the ERR parameter is preferred to the typical Wald CI because the likelihood function of ERR models is usually non-normal in shape. Let $\mathrm{L}(\Theta|X)$ be the likelihood function as in 3, then the profile likelihood of β is

$$L_1(\beta|X) = \max_{\theta} \mathrm{L}(\theta, \beta|X).$$

The $(1-a)\cdot 100\%$ PLCI are the values of β that meet the requirement

$$\log(L_1(\beta|X)) > \hat{l} - \chi^2_{1,1-a}/2,$$

where $\hat{l} = l(\hat{\Theta}|X)$ is the maximum value of the log-likelihood function and $\chi^2_{1,1-a}$ is the $1-a$ quantile of a chi-squared distribution with 1 degree of freedom. Note that $L_1(\beta|X)$ is the likelihood of a Poisson GLM: $C \sim \mathrm{Pois}\,(PY(1+\beta D)\mathrm{e}^{\eta})$ where $PY(1+\beta D)$ is the offset. In general, for only one exposure the profile likelihood of the ERR is the likelihood of a Poisson GLM with canonical logarithm link.

Assuming a simple model as (2), the profile likelihood for β can be calculated by solving

$$
\begin{cases}
\dfrac{\partial l}{\partial \alpha_1} = & S - \mathrm{e}^{\alpha_1}\left(T_1 + \displaystyle\sum_{k=2}^{K} T_k \mathrm{e}^{\alpha_k}\right) = 0 \\
\dfrac{\partial l}{\partial \alpha_k} = & S_k - T_k \mathrm{e}^{\alpha_1+\alpha_k} = 0,\ k = \{2,\dots,K\}
\end{cases},
$$

where $S = \sum_{i=1}^{n} C_i$, $S_k = \sum_{i|x_i=k} C_i$, and $T_k = \sum_{i|x_i=k} PY_i(1+\beta D_i)$. Let $\vec{\alpha}(\beta) = \{\alpha_1(\beta), \dots, \alpha_K(\beta)\}$ then the profile likelihood for β is $L_1(\beta) = L(\vec{\alpha}(\beta), \beta|X)$ where

$$
\begin{aligned}
\alpha_1(\beta) = \quad & \log\left(S - \sum_{k=2}^{K} S_k\right) - \log(T_1), \\
\alpha_k(\beta) = \quad & \log(S_k) - \log(T_k) - \alpha_1(\beta),\ k = \{2,\dots,K\}.
\end{aligned}
$$

To obtain the maximum likelihood estimators of the parameters, the partial derivative of the log-likelihood with respect β is evaluated at $\vec{\alpha} = \vec{\alpha}(\beta)$, *i.e.*

$$\left.\frac{\partial l}{\partial \beta}\right|_{\vec{\alpha}=\vec{\alpha}(\beta)} = \sum_{i=1}^{n} \frac{C_i D_i}{1+\beta D_i} - \mathrm{e}^{\alpha_1(\beta)} R_1 - \mathrm{e}^{\alpha_1(\beta)} \sum_{k=2}^{K} \mathrm{e}^{\alpha_k(\beta)} R_k = 0,$$

where $R_k = \sum_{i|x_i=k} PY_i D_i$. This equation is solved numerically to get the estimator $\hat{\beta}$ and the rest of the estimators are $\hat{\vec{\alpha}} = \vec{\alpha}(\hat{\beta})$.

The likelihood ratio test p-value for null hypothesis $\beta = 0$ is

$$\mathrm{P}\left(\chi_1^2 > l(\vec{\alpha}(\hat{\beta}), \hat{\beta}|X) - l\left(\vec{\alpha}(0), 0|X\right)\right),$$

where χ_1^2 is a chi-squared distribution with 1 degree of freedom.

It is possible that the PLCI bound does not converge. In this situation, the Wald-type CI bound is usually reported. This can be calculated by the Hessian matrix,

$$H(\vec{\alpha}, \beta) =$$
$$\begin{bmatrix} -\mathrm{e}^{\alpha_1}\left(T_1 + \sum_{k=2}^{K} \mathrm{e}^{\alpha_k} T_k\right) & -\mathrm{e}^{\alpha_1+\alpha_2} T_2 & -\mathrm{e}^{\alpha_1+\alpha_3} T_3 & \dots & -\mathrm{e}^{\alpha_1+\alpha_K} T_K & -\mathrm{e}^{\alpha_1}\left(R_1 + \sum_{k=2}^{K} \mathrm{e}^{\alpha_k} R_k\right) \\ -\mathrm{e}^{\alpha_1+\alpha_2} T_2 & 0 & 0 & \dots & 0 & -\mathrm{e}^{\alpha_1+\alpha_2} R_2 \\ -\mathrm{e}^{\alpha_1+\alpha_3} T_3 & 0 & 0 & \dots & 0 & -\mathrm{e}^{\alpha_1+\alpha_3} R_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\mathrm{e}^{\alpha_1+\alpha_K} T_K & -\mathrm{e}^{\alpha_1+\alpha_K} R_K & & & & \\ -\sum_{i=1}^{n} \frac{C_i D_i^2}{(1+\beta D_i)^2} & & & & & \end{bmatrix},$$

and evaluating it at the maximum likelihood estimator, *i.e.* $H(\vec{\alpha}(\hat{\beta}), \hat{\beta})$. The variance-covariance matrix is $-H(\vec{\alpha}(\hat{\beta}), \hat{\beta})^{-1}$.

2.2. Posterior ERR

Bayesian analysis combines prior information, in the form of probability distributions, with the likelihood function of an assumed model, providing posterior results as probability distributions too. The continuous version of Bayes' theorem establishes

$$\mathrm{P}(\Theta|X) = \frac{\mathrm{L}(\Theta|X)\mathrm{P}(\Theta)}{\int \mathrm{L}(\Theta|X)\mathrm{P}(\Theta)\mathrm{d}\Theta}, \tag{5}$$

where Θ is the continuous parameter set, X is the observed data set, $\mathrm{L}(\Theta|X)$ is the likelihood function, $\mathrm{P}(\Theta)$ is the prior probability density function of Θ and $\mathrm{P}(\Theta|X)$ is the posterior probability density of Θ given data X. See, for instance Christensen et al. (2011), for further description.

Following model (2): X, Θ and $\mathrm{L}(\Theta|X)$ are as stated in Section 2. Assuming all α_k's and β are independent, the prior probability density is

$$\mathrm{P}(\Theta) = \mathrm{P}(\beta)\prod_{k=1}^{K}\mathrm{P}(\alpha_k).$$

It is also assumed that all α_k's priors are non informative, such that the probability is the same for all the values in the support of the parameters. This leads to the following improper uniform priors:

$$\alpha_k \quad \sim \mathscr{U}(-\infty,+\infty), k = \{1\dots K\} \tag{6}$$

and a prior for β open to any distribution with support bounded below by $-1/\max(D)$, to ensure the Poisson intensity is positive. The Bayesian framework affords the definition of improper prior distributions.

Applying Bayes' theorem (5), the posterior of Θ is

$$\begin{aligned}
&\mathrm{P}(\Theta|X) \propto \mathrm{P}(\beta)\cdot\mathrm{L}(\Theta|X)\\
&\propto \mathrm{P}(\beta)\prod_{i=1}^{n}(PY_i\mathrm{e}^{\alpha_1+\sum_{k=2}^{K}\alpha_k\mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i))^{C_i}\exp(-PY_i\mathrm{e}^{\alpha_1+\sum_{k=2}^{K}\alpha_k\mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i))\\
&= \mathrm{P}(\beta)\cdot\exp\left(S\alpha_1-\mathrm{e}^{\alpha_1}\sum_{i=1}^{n}PY_i\mathrm{e}^{\sum_{k=2}^{K}\alpha_k\mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i)\right)\cdot\\
&\cdot\prod_{i=1}^{n}(PY_i\mathrm{e}^{\sum_{k=2}^{K}\alpha_k\mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i))^{C_i}.
\end{aligned} \tag{7}$$

The goal here is to get the marginal posterior of the ERR, the posterior distribution of β. Let $\vec{\alpha}_{-1} = (\alpha_2,\dots,\alpha_K)$, the first step is to calculate the joint marginal posterior of $(\vec{\alpha}_{-1},\beta)$ which it is proportional to the integral of expression (7) over α_1, *i.e.*

$$\begin{aligned}
\mathrm{P}(\vec{\alpha}_{-1},\beta|X) \propto\ & \mathrm{P}(\beta)\int_{-\infty}^{+\infty}\mathrm{L}(\Theta|X)\mathrm{d}\alpha_1\\
=\ & \mathrm{P}(\beta)\left[\prod_{i=1}^{n}(PY_i\mathrm{e}^{\sum_{k=2}^{K}\alpha_k\mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i))^{C_i}\right]\\
\cdot\ & \int_{-\infty}^{+\infty}\exp\left(S\alpha_1-\mathrm{e}^{\alpha_1}\sum_{i=1}^{n}PY_i\mathrm{e}^{\sum_{k=2}^{K}\alpha_k\mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i)\right)\mathrm{d}\alpha_1
\end{aligned}$$

$$\begin{aligned}
&= \mathrm{P}(\beta)\frac{\prod_{i=1}^{n}(PY_i\mathrm{e}^{\sum_{k=2}^{K}\alpha_k\mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i))^{C_i}}{\left[\sum_{i=1}^{n}PY_i\mathrm{e}^{\sum_{k=2}^{K}\alpha_k\mathbf{1}_{\{k\}}(x_i)}(1+\beta D_i)\right]^{S}}(S-1)! \\
&\propto \frac{\mathrm{P}(\beta)\left[\prod_{i=1}^{n}(1+\beta D_i)^{C_i}\right]\mathrm{e}^{\sum_{k=2}^{K}S_k\alpha_k}}{\left[\sum_{i|x_i=1}PY_i(1+\beta D_i)+\sum_{k=2}^{K}\left(\mathrm{e}^{\alpha_k}\sum_{i|x_i=k}PY_i(1+\beta D_i)\right)\right]^{S}}.
\end{aligned} \tag{8}$$

Then the marginal posterior of the ERR is proportional to the multiple integral of Expression (8) over $\vec{\alpha}_{-1}$,

$$\begin{aligned}
\mathrm{P}(\beta|X) = &\int_{\vec{\alpha}_{-1}}\mathrm{P}(\vec{\alpha}_{-1},\beta|X)\mathrm{d}\vec{\alpha}_{-1} \\
\propto\ &\mathrm{P}(\beta)\left[\prod_{i=1}^{n}(1+\beta D_i)^{C_i}\right]\int_{\vec{\alpha}_{-1}}\mathrm{e}^{\sum_{k=2}^{K}S_k\alpha_k} \\
&\cdot\left[\sum_{i|x_i=1}PY_i(1+\beta D_i)+\sum_{k=2}^{K}\left(\mathrm{e}^{\alpha_k}\sum_{i|x_i=k}PY_i(1+\beta D_i)\right)\right]^{-S}\mathrm{d}\vec{\alpha}_{-1} \\
= &\frac{\mathrm{P}(\beta)\left[\prod_{i=1}^{n}(1+\beta D_i)^{C_i}\right]\left[\sum_{i|x_i=1}PY_i(1+\beta D_i)\right]^{\sum_{k=2}^{K}S_k-S}}{\prod_{k=2}^{K}\left[\sum_{i|x_i=k}PY_i(1+\beta D_i)\right]^{S_k}} \\
&\cdot\frac{(S_2-1)!}{\prod_{i=1}^{S_2-1}S-i}\prod_{k=3}^{K}\frac{(S_k-1)!}{\prod_{i=1}^{S_k-1}S-\sum_{j=2}^{k-1}S_k-i} \\
\propto\ &\frac{\mathrm{P}(\beta)\left[\prod_{i=1}^{n}(1+\beta D_i)^{C_i}\right]\left[\sum_{i|x_i=1}PY_i(1+\beta D_i)\right]^{\sum_{k=2}^{K}S_k-S}}{\prod_{k=2}^{K}\left[\sum_{i|x_i=k}PY_i(1+\beta D_i)\right]^{S_k}}.
\end{aligned} \tag{9}$$

Consequently,

$$\mathrm{P}(\beta|X) = \frac{\mathrm{P}(\beta)\left[\prod_{i=1}^{n}(1+\beta D_i)^{C_i}\right]\left[\sum_{i|x_i=1} PY_i(1+\beta D_i)\right]^{\sum_{k=2}^{K} S_k - S}}{N\prod_{k=2}^{K}\left[\sum_{i|x_i=k} PY_i(1+\beta D_i)\right]^{S_k}}, \tag{10}$$

where N is the normalizing constant

$$N = \int_0^{+\infty} \mathrm{P}(\beta)\left[\prod_{i=1}^{n}(1+\beta D_i)^{C_i}\right]\left[\sum_{i|x_i=1} PY_i(1+\beta D_i)\right]^{\sum_{k=2}^{K} S_k - S} \prod_{k=2}^{K}\left[\sum_{i|x_i=k} PY_i(1+\beta D_i)\right]^{-S_k} \mathrm{d}\beta, \tag{11}$$

that is calculated by numerical integration (there is no analytical solution). The probability density (10) does not have a recognizable form, but this is not unusual when dealing with Bayesian analysis.

The integrals in expressions (8) and (9) are calculated by recursive integration by parts.

3. Poisson ERR fitting in R

Cohort studies in radiation epidemiology are usually huge, and hence maximum likelihood estimation of the model parameters is computationally intensive. This computational load increases for the calculation of the profile likelihood confidence intervals.

As mentioned in Section 1 a general ERR model has the form

$$C_i \sim \mathrm{Pois}\left(PY_i \mathrm{e}^{\alpha_0 + \sum_{k=1}^{m}\alpha_k x_i^{(k)}}\left(1 + \sum_{j=1}^{d}\beta_j D_i^{(j)}\right)\right). \tag{12}$$

Let $\vec{\alpha} = \{\alpha_0, \ldots, \alpha_m\}$ and $\vec{\beta} = \{\beta_1, \ldots, \beta_d\}$, the log-likelihood function of of the parameter set $\Theta = \left\{\vec{\alpha}, \vec{\beta}\right\}$

$$
\begin{aligned}
l(\Theta|X) = & \sum_{i=1}^{n}\left[C_i(\log PY_i + \alpha_0 + \sum_{k=1}^{m}\alpha_k x_i^{(k)} + \log\left(1+\sum_{j=1}^{d}\beta_j D_i^{(j)}\right)\right] \\
& - \sum_{i=1}^{n} PY_i \mathrm{e}^{\alpha_0+\sum_{k=1}^{m}\alpha_k x_i^{(k)}}\left(1+\sum_{j=1}^{d}\beta_j D_i^{(j)}\right) - \sum_{i=1}^{n}\log C_i!.
\end{aligned} \tag{13}
$$

The gradient of the log-likelihood function can be efficiently defined by the following expressions

$$
\begin{aligned}
\frac{\partial l}{\partial \vec{\alpha}} &= \vec{S} - [(PY \circ (1+\vec{\beta}\cdot\mathscr{D}) \circ E)\cdot A], \\
\frac{\partial l}{\partial \vec{\beta}} &= [C \oslash (1+\vec{\beta}\cdot\mathscr{D})]\cdot\mathscr{D} - (PY\circ E)\cdot\mathscr{D},
\end{aligned} \tag{14}
$$

where

$$
\mathscr{D} = \begin{bmatrix} D_1^{(1)} & D_1^{(2)} & \dots & D_1^{(d)} \\ D_2^{(1)} & D_2^{(2)} & \dots & D_2^{(d)} \\ \vdots & \vdots & \ddots & \vdots \\ D_n^{(1)} & D_n^{(2)} & \dots & D_n^{(d)} \end{bmatrix}, \; A = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(m)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(m)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(m)} \end{bmatrix},
$$

$$
E = \exp\left(\vec{\alpha}\cdot A^{\mathrm{T}}\right), \; \vec{S} = C\cdot A,
$$

and operators $\circ$ and $\oslash$ represents the Hadamard product and division respectively.

In cases where the PLCI bound does not converge, the Hessian can be calculated using the following second-order partial derivatives of the log-likelihood to calculate the Wald-type CI bound,

$$
\begin{aligned}
\frac{\partial^2 l}{\partial\alpha_t\partial\alpha_q} &= -\mathrm{e}^{\alpha_0} T_{t,q},\; t,q = \{0,\dots,m\}, \\
\frac{\partial^2 l}{\partial\alpha_t\partial\beta_q} &= -\mathrm{e}^{\alpha_0} R_{t,q},\; t = \{0,\dots,m\},\; q = \{1,\dots,d\}, \\
\frac{\partial^2 l}{\partial\beta_t\partial\beta_q} &= -\sum_{i=1}^{n}\frac{C_i D_i^{(t)} D_i^{(q)}}{\left(1+\sum_{j=1}^{d}\beta_j D_i^{(j)}\right)^2},\; t,q = \{1,\dots,d\},
\end{aligned}
$$

where

$$T_{t,q} = \sum_{i=1}^{n} PY_i \left(1 + \sum_{j=1}^{d} \beta_j D_i^{(j)}\right) x_i^{(t)} x_i^{(q)} \mathrm{e}^{\sum_{k=1}^{m} \alpha_k x_i^{(k)}},$$

$$R_{t,q} = \sum_{i=1}^{n} PY_i D_i^{(q)} x_i^{(t)} \mathrm{e}^{\sum_{k=1}^{m} \alpha_k x_i^{(k)}},$$

In R (version 3.5.1), by means of the `maxLik()` function from the `maxLik` package (Henningsen and Toomet, 2011) (version 1.3.4), model (12) can be fitted by defining the log-likelihood function (13). For faster and more accurate results, the gradient function implemented as in (14) can be included in the `maxLik()` function.

The R script for the results in Section 4.3 is provided as supplementary material, as a reference for the R implementation of model (12) fitting.

ERR models are usually fitted by Epicure, a very specialised proprietary software, which is the gold standard in radiation epidemiology practice. In recent years, some studies have been published using SAS, e.g. Journy et al. (2015), but there is not a SAS Stored Process for this aim. However, there are some SAS macros for fitting ERR models and calculating PLCI's, e.g. in Richardson (2008) for Poisson models by means of `PROC NLMIXED`. In Grant et al. (2017), an R routine was developed to analyze the Life Span Study data of a-bomb survivors in Hiroshima and Nagasaki, by means of the `gnm()` function in package `gnm` (Turner and Firth, 2018). There is also an R package called `linERR` which fits ERR models for censored survival data (Morinña, 2016).

This is proposed as a free licence and open source alternative of Epicure's `AMFIT` module, which is used to fit Poisson ERR models. Moreover, the R routines in Grant et al. (2017) also cover this purpose, in fact they also allow to fit more complex models with dose-effect modification.

The previous step to fitting the Poisson ERR model is to generate the person-years table. These tables are created by stratifying by categories of different variables, e.g. attained age, the original censored data. For each cell of the table, the accumulated person-years and events are calculated. In Epicure the module `DATAB` generates these tables. Further work in this project includes the creation of an R package with tools to fit Poisson ERR models, calculate PLCI's and generate person-years tables. Function `pyears()` in the `survival` package (Therneau, 2015) builds person-time tables, but for non-dynamic exposures.

4. Practical examples

Two applied examples for data from the literature are given. The third example is the application of the proposed implementations here to a subset from the first example data set. This third example is presented to facilitate reproducible and replicable research, because the data sets of the first two examples are not shareable.

The ERR is per mGy (milligray) in all examples shown here.

4.1. Pearce et al. 2012

Pearce et al. (2012) analysed the risk of leukaemia and brain tumours in young patients who were first underwent computed tomography (CT) scans in National Health Service hospitals in England, Wales, or Scotland in a 23 years retrospective cohort study. In the leukaemia follow-up, there were 74 leukaemia diagnosis for 178,604 patients, and a total of 1,720,984 person-years. The person-year table was built assuming 2 years exclusion and lag periods.

A Poisson ERR model is assumed with unique exposure (the accumulated ionising radiation dose), and the background risk is modelled by

$$\eta = \alpha_1 \mathbf{1}_{a_i<5} + \alpha_2 \mathbf{1}_{5\leq a_i<20} + \alpha_3 \mathbf{1}_{20\leq a_i<30} + \alpha_4 \mathbf{1}_{30\leq a_i<35} + \alpha_5 \mathbf{1}_{a_i\geq 35}$$

where a_i is the attained age. This model has the same form as the simple model in Section 2: one exposure and baseline rate modelled by a unique categorical variable.

Following the implementation in Section 2.1, the maximum likelihood estimate of the ERR is $\hat{\beta} = 0.0362$ and its 95% PLCI is $(0.0052, 0.1198)$ with p-value 0.0097. These values match with those shown in Pearce et al. (2012).

Following the implementation in Section 2.2 and considering $\beta \sim \mathscr{U}(-1/\max(D) = -0.0015, +\infty)$, Figure 1 shows the posterior density function of the ERR following Equation (10). The modal posterior ERR value is 0.0361 and its 95% highest posterior density (HPD) interval is (0.0023, 0.1460).

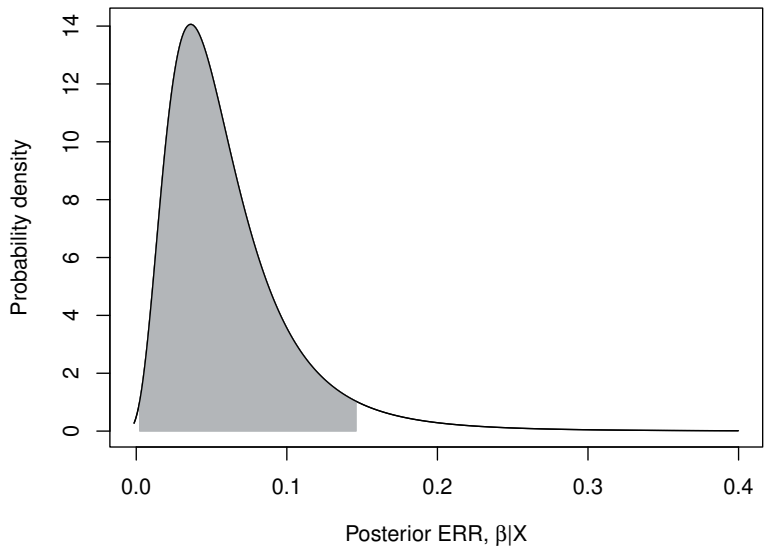

Figure 1: *Posterior probability density of the ERR (solid line) and its 95% HPD (shaded grey) in Section 4.1.*

One of the big advantages of the Bayesian framework is that it is possible to calculate the posterior probability of a parameter being contained by a given interval. For instance in this case there is a posterior probability of 0.5111 for the ERR being greater than 0.050.

In order to compare this model with improper flat priors to a model with informative priors, a model with the following priors is assumed

$$\begin{aligned} \vec{\alpha} &\sim \mathcal{N}([-10,0,0,0,0]^{\mathrm{T}}, 0.1 \cdot I_5), \\ \beta &\sim \text{Gamma}(1.1, 5). \end{aligned} \tag{15}$$

I_5 is the identity matrix of size 5. The parametrization of the multivariate normal distribution is represented by the mean vector and precision matrix, and for the gamma distribution by the shape and rate values. The posterior distribution of the ERR is drawn using JAGS (version 4.3.0) (Plummer (2003)). The modal posterior ERR value is 0.0377 and its 95% HPD interval is (0.0042, 0.114). This MCMC model has 2 chains of 50,000 iterations after 1000 burning iterations and thinning interval 10. It is computational intensive, it takes around 20 hours.

Although both the Bayesian and the frequentist methods provide estimation and uncertainty results, when comparing them it is important to note that they represent different foundational approaches. In particular, the frequentist method assumes that the parameter is a fixed value and the maximum likelihood estimator is a random variable whereas the Bayesian method assumes the opposite.

4.2. Harbron et al. 2018

Harbron et al. (2018) analysed the risk of leukaemia and lymphoma in patients who underwent cardiac catheterizations while aged 22 years or younger. There were 36 cases for 9,467 patients, and a total of 74,405.88 person-years at risk in this study. Doses were lagged by 2 years. The exclusion period was also 2 years.

To calculate the ERR, a Poisson ERR model was assumed with unique exposure with background risk as

$$\eta = \alpha_1 \mathbf{1}_{a_i<5} + \alpha_2 \mathbf{1}_{5\leq a_i<10} + \alpha_3 \mathbf{1}_{10\leq a_i<15} + \alpha_4 \mathbf{1}_{15\leq a_i<20} + \alpha_5 \mathbf{1}_{20\leq a_i<25} + \alpha_6 \mathbf{1}_{a_i\geq 25} + T_i$$

where a_i is the attained age and T_i represents the status of organ transplantation. Note that this model does not have the same structure as the simple model in Section 2.

In Harbron et al. (2018) this model was fitted in R as stated in Section 3. The maximum likelihood estimate of the ERR is $\hat{\beta} = 0.0180$, and its 95% PLCI is $(-0.0021, 0.0961)$ with p-value 0.1084.

Assuming a simple model with background risk modelled only by the transplant status, the results for the two methods are:

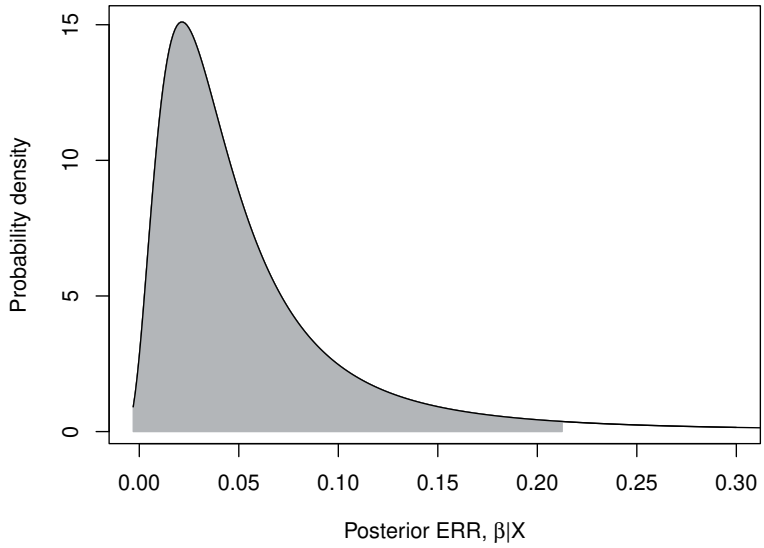

Figure 2: *Posterior probability density of the ERR (solid line) and its 95% HPD (shaded grey) in Section 4.2.*

- Following the method in Section 2.1 the maximum likelihood estimate of the ERR is $\hat{\beta} = 0.0214$ and its 95% PLCI is $(-0.0008, 0.1049)$ with p-value = 0.0661.
- Following the method in Section 2.2 and considering $\beta \sim \mathscr{U}(-1/\max(D) = -0.0030, 1)$, Figure 2 shows the posterior density function of the ERR following Equation (10). The modal posterior ERR value is 0.0215 and its 95% HPD is $(-0.0030, 0.2125)$, and $\mathrm{P}(\beta|X > 0.050) = 0.4091$.

4.3. Sub-cohort

A 14,000 random row subset of the person-years table from the leukaemia analysis in Section 4.1, with information of accumulated person-years, weighted mean accumulated dose, sex and weighted mean attained age was generated. In this sub-cohort there are 9 leukaemia cases in a total of 158,953.3 person-years.

A Poisson ERR model is assumed, with unique exposure and background risk modelled by $\eta = \alpha_0 + \alpha_1 a_i$, where a_i is the attained age. The attained age is not a categorical variable, so this model does not have the same structure as the simple model in Section 2.

Fitting this model in R as stated in Section 3, the maximum likelihood estimate of the ERR is $\hat{\beta} = 0.0247$ (it agrees with the result returned by `gnm()`), and its 95% PLCI is $(-0.0553^*, 0.3341)$ with p-value 0.4535. The symbol * denotes the bound is Wald-type.

To check the effect of gender on the ERR, an interaction between the dose and the sex is added to the previous model, *i.e.* the ERR term is $(1 + \beta_1 D_i + \beta_2 F_i D_i$ where F_i is the indicator of female patient.

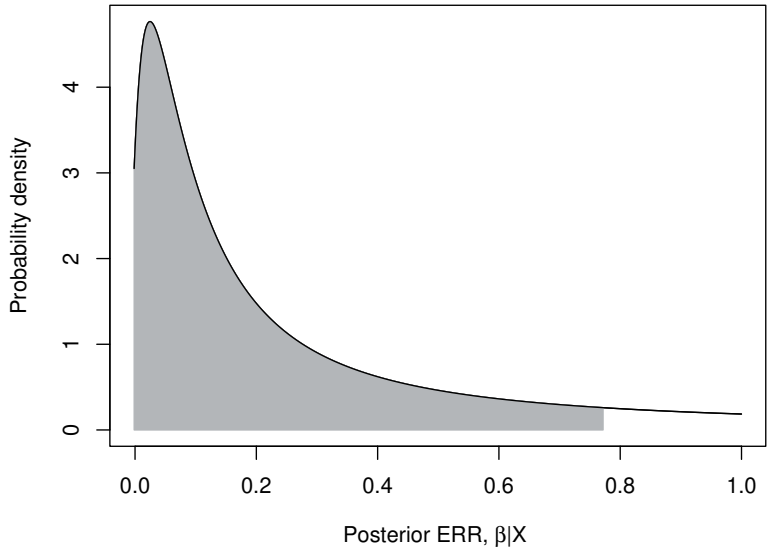

Figure 3: *Posterior probability density of the ERR (solid line) and its 95% HPD (shaded grey) in Section 4.3.*

This can be fitted as a model with two exposures, D_i and F_iD_i, and the ERR results 0.0039 for male and 0.0541 for female, this is $\hat{\beta}_1 = 0.0039$ and $\hat{\beta}_2 = 0.0502$, but the female effect is not significant because the likelihood ratio test p-value for testing $\beta_2 = 0$ is 0.3059.

Now, assuming a simple model with background risk modelled by three categories of attained age, *i.e.*

$$\eta = \alpha_1 \mathbf{1}_{a_i<10} + \alpha_2 \mathbf{1}_{10\leq a_i<15} + \alpha_3 \mathbf{1}_{a_i\geq 15},$$

the results for the two methods are:

- Following the method in Section 2.1 the maximum likelihood estimate of the ERR is $\hat{\beta} = 0.0247$ and its 95% PLCI is (-0.0584^*, 0.3659) with p-value = 0.3884.
- Following the method in Section 2.2 and considering $\beta \sim \mathscr{U}(-1/\max(D) = -0.0015, 1)$, Figure 3 shows the posterior density function of the ERR following Equation (10). The modal posterior ERR value is 0.0247 and its 95% HPD interval is (-0.0015, 0.7717), and $\mathrm{P}(\beta|X > 0.050) = 0.7724$. If $\beta \sim \mathrm{Gamma}(1.1, 5)$, the modal posterior ERR value is 0.0234 and its 95% HPD interval is (0, 0.3294). Analogously to example at Section 4.1, an MCMC model is applied to draw the posterior of the ERR, assuming the same priors (with the difference of the dimension of $\vec{\alpha}$, *i.e.* $\vec{\alpha} \sim \mathscr{N}([-10,0,0,0,0]^{\mathrm{T}}, 0.1 \cdot I_3)$), the modal posterior ERR value is 0.0346 and its 95% HPD interval is (0.0001, 0.3073).

5. Conclusion

The simple methods presented here for estimating the ERR in radiation epidemiology follow-up studies are easy to implement. Although these models have restricted forms, they cover a wide range of situations. For instance, the leukaemia analysis in Pearce et al. (2012) was performed with this type of model. Additionally, they can be used to get sensible initial values for fitting ERR models with more complex structures.

R is an open-source statistical software program with a free license and large user community. As such, it is well suited for the development of reproducible and replicable research. In this work an R script for fitting Poisson ERR models is shared and guidelines for implementing ERR models in R are given in Section 3.

Further work in this project will lead to the development of an R package with tools to fit Poisson ERR models, build person-years tables with time-dependent variables, and calculate PLCI's. This package will have application in radiation epidemiology follow-up studies.

Acknowledgments

This research was supported by the Basque Government through the BERC 360 2014-2017 and the Spanish Ministry of Economy and Competitiveness MINECO and FEDER: BCAM Severo Ochoa excellence accreditation SEV-2013-0323 and MINECO Challenges MTM2017-82379-R.

MH is grateful for the help of Dr. Richard W. Harbron and Prof. Mark S. Pearce, from Newcastle University; and to the reviewers for their valuable reviews.

References

Committee to Assess Health Risks from Exposure to Low Levels of Ionizing Radiation (2006). *Health Risks from exposure to low levels of ionizing radiation. BEIR VII Phase 2*. Washington: The National Academies Press.

Christensen, R., Johnson, W., Brasncum, A. and Hanson, T.E. (2011). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. Boca Raton: Champman & Hall/CRC Press.

Grant, E.J., Brenner, A., Sugiyama, H., Sakata, R., Sadakane, A., Utada, M., Cahoon, E. K., Milder, C. M., Soda, M., Cullings, H. M., Preston, D. L., Mabuchi, K. and Ozasa, K. (2017). Solid Cancer Incidence among the Life Span Study of Atomic Bomb Survivors: 1958–2009. *Radiation Research*, 187(5), 513–537.

Harbron, R.W., Chapple, C.-L., O'Sullivan, J.J., Lee, C, McHugh, K., Higueras, M. and Pearce, M.S. (2018). Cancer incidence among children and young adults who have undergone x-ray guided cardiac catheterization procedures. *European Journal of Epidemiology*, 33(4), 393–401.

Henningsen, A. and Toomet, O. (2011). `maxLik`: A package for maximum likelihood estimation in R. *Computational Statistics*, 26(3), 443–458.

Journy, N., Rehel, J.-L., Ducou Le Pointe, H., Lee, C., Brisse, H., Chateil, J.-F., Caer-Lorho, S., Laurier, D. and Bernier, M.-O. (2015). Are the studies on cancer risk from CT scans biased by indication? Elements of answer from a large-scale cohort study in France. *British Journal of Cancer*, 112(1), 185–193.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, 2nd edition. Boca Raton: Champman & Hall/CRC Press.

Morinña, D. (2016). *linERR: Linear Excess Relative Risk Model*, version 1.0, URL: https://CRAN.R-project.org/package=linERR.

Pearce, M.S., Salotti, J.A., Little, M.P., Mchugh, K., Lee, C., Kim, K.P., Howe, N.L., Ronckers, C.M., Rajaraman, P., Craft, A.W., Parker, L. and Berrington de González, A. (2012). Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. *Lancet*, 380(9840), 499–505.

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*.

Preston, D.L., Lubin, J.H., Pierce, D.A. and McConney, M.E. (1993). *Epicure: user's guide*. Seattle: Hirosoft International Corporation.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. URL https://www.R-project.org/.

Richardson, D.B. (2008). A simple approach for fitting linear relative rate models in SAS. *American Journal of Epidemiology*, 168(11), 1333–1338.

Therneau, T. (2015). *A Package for Survival Analysis in S*, version 2.38, URL: https://CRAN.R-project.org/package=survival.

Turner, H. and Firth, D. (2018). *Generalized nonlinear models in R: An overview of the gnm package*, version 1.1-0, URL: https://cran.r-project.org/package=gnm.