Expected cases tool methodology

1. Models

Expected cases are calculated from a generalized linear model[1] considering a Negative Binomial or a Poisson distribution using an age-drift model[2] and a model with different age slopes[3]. Four models are finally tested.

Table 1. Models fitted

Y_{ij} distribution	Model specification	Model
$Y_{it} \sim \text{NegBin}\,(\mu_{it}, \upsilon)$	$\ln(E\,(Y_{it})) = \ln(P_{it}) + \alpha_i + \beta t$	Negative Binomial age-drift
	$\ln(E\,(Y_{it})) = \ln(P_{it}) + \alpha_i + \beta_i t$	Negative Binomial age-specific slope
$Y_{it} \sim \text{Poiss}\,(\lambda_{it})$	$\ln(E\,(Y_{it})) = \ln(P_{it}) + \alpha_i + \beta t$	Poisson age-drift
	$\ln(E\,(Y_{it})) = \ln(P_{it}) + \alpha_i + \beta_i t$	Poisson age-specific slope

Where:

- Y_{it} are the number of cases in the i-th age group and the year or central year period t
- $E(Y_{it})$ are the expected cases in the i-th age group and the year or central year period t
- P_{it} are the population at risk in the i-th age group and the year or central year period t
- α_i are the age groups coefficients
- β is the period or year coefficient
- β_i are the slopes for each age group considered
- The $\ln(P_{it})$ term is named offset. When introduced, the expected value of the incidence or mortality rate is indirectly modeled: $y_{it} = E(Y_{it})/P_{it}$
- The rescaled period is introduced. For example, if we are working with central intervals and we have 1985, 1990, 1995 they are transformed by subtracting the minimum, so we work with 0, 5 and 10

2. Akaike Information Criterion (AIC)[4]:

The AIC is a tool used for model selection. Given a data set, the best model from the fitted models is the one that has the lower AIC value.

The AIC is based in the concept of entropy: It is offering a relative measure of the information lost when a given model is used to describe reality. It describes the trade off between bias and variance.

The AIC is calculated as:

$$\mathrm{AIC} = 2\mathrm{p} - 2\ln(\mathrm{L})$$

where:

- k is the number of parameters in the statistical model

- L is the maximized value of the likelihood function for the estimated model

3. Goodness of Fit (GOF)[1,5]

The GOF of a statistical model describes how well it fits a set of observations. Measures of GOF typically summarize the discrepancy between observed values and the values expected under the model.

We use the Pearson's chi-square test to assess the goodness of fit. So we need to know what the Pearson residual is.

The Pearson residual is defined as:

$$\mathrm{r_i} = \frac{\mathrm{y_i} - \mu_\mathrm{i}}{\sqrt{\mathrm{V}(\mu_\mathrm{i})}}$$

It is just the raw residual (the number of cases in the i-cell minus its fitted value) scaled by the estimated standard deviation of y_i.

Finally, the Pearson X^2 goodness-of-fit statistic is defined as:

$$\mathrm{X}^2 = \sum \mathrm{r_i^2}$$

This statistic can then be used to calculate a p-value by comparing the value of the statistic to a chi-square distribution. The number of degrees of freedom is the number of cells (n) minus the number of parameters (p). So, the Pearson's chi-square test with a significance level α is:

$$\begin{cases} \mathrm{H_0}: \mathrm{X}^2 \sim \chi^2_{\mathrm{n-p,1-\alpha}} \\ \mathrm{H_1}: \mathrm{No\ H_0} \end{cases}$$

The absence of significance indicates that the model fits.

4. Process for model selection

For the model selection the following algorithm is used:

Step 1: All models (Table 1) are fit for each cause and sex.

Step 2: The model with minimum AIC is pre-selected.

Step 3: Situations:

3.1. The model pre-selected fits (p-value of GOF > 0.05). This will be the final model. The process ends and the method for selection is set to “Minimum AIC”.

3.2. The model pre-selected does not fit, therefore other models are evaluated.

3.2.1. If one or more models fit, the one selected will be that with higher p-value. The process will end and the method for selection is set to “Maximum GOF”.

3.2.2. If any of these models do fit, select that with minimum AIC. The process ends and the method for selection sets to “The model does not fit”.

When any group-sex does not adjust any model, in the table of adjusted models we will observe "The model does not fit ". This means the estimates for this particular group-sex returned in the file must not be used.

5. Trends and prediction graphs

The trends graph of the ASRs is represented by smooth curves which are obtained with the *loess* function of the R statistic software[6]. This function fits a local polynomial regression[7] for the ASRs.

Trends and predictions graphs can be returned only when there are future data predictions.

If no model fits for a particular group-sex, Expected cases tool does not return its graph.

REFERENCES

1. McCullagh P, Nelder J. Generalized Linear models. 2nd ed. Boca Raton (US): Chapman & Hall/CRC; 1989. 511 p.

2. Dyba T. A simple non-linear model in disease incidence prediction. Stat Med. 1997;16:2297–309.

3. Dyba T, Hakulinen T. Comparison of different approaches to incidence prediction based on simple interpolation techniques. Stat Med. 2000;19(13):1741–52.

4. Akaike H. A new look at the statistical model identification. IEEE Trans Autom Control. 1974;19(6):716–23.

5. Breslow NE, Day NE. Statistical methods in cancer research. Vol. II. The design and analysis of cohort studies. IARC Sci Publ. Lyon: International Agency for Research on Cancer; 1987;82:1–406.

6. R Core Team. R: A language and environment for statistical computing. [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2015. Available from: https://www.r-project.org/

7. Cleveland WS, Grosse E, Shyu M. Chapter 8. Local regression models. In: Chambers JM, Hastie TJ, editors. Statistical Models in S. Chapman&Hall/CRC; 1992. p. 309–76.