Poisson Regression in Epidemiology

Various authors [3, 9, 11, 12] have noted that **Poisson regression** can be used to analyze cohort survival data (*see* **Cohort Study**). This formulation also leads to a unification of **risk** estimation based on internal comparison of rates among members of a cohort with various exposure levels and classical epidemiologic methods based on external rates that yield standardized mortality ratios or standardized incidence ratios [2, 5] (*see* **Standardization Methods**).

Poisson regression is an important alternative to **partial-likelihood**-based analysis of the **proportional hazards** model (*see* **Cox Regression Model**) and to parametric analyses of such models (*see* **Survival Analysis, Overview**) for two main reasons. First, it provides an efficient and intuitive method for dealing with cumulative exposures and other **time-dependent covariates** and for allowing risk to depend on multiple time scales (e.g. attained age, time since exposure, or calendar time). Secondly, it facilitates the consideration of a broad range of risk models including those that allow for the direct parametric description of baseline rates, absolute excess rates, and **relative risks**.

Breslow & Day [4] offer a general discussion of the use of Poisson regression in the analysis of cohort survival data. Some of the most extensive applications of these methods have involved studies of radiation effects on mortality and cancer incidence in the atomic bomb survivors [14].

Poisson Regression of Survival Data

The data from cohort survival studies typically consist of information on whether or not the event of interest occurred, the event or censoring time, t, and a vector of possibly time-dependent covariates, $\mathbf{z}$, for each cohort member. Since interest centers on **hazard rates** it is natural and useful for the purposes of analysis or summarization to reorganize such data into an event–time table defined by a cross-classification over a set of time intervals and covariate categories. The data for each cell in such a table include the total number of events, c_{is}, the total time (person-years) at risk, R_{is}, and representative values of the covariates, z_{is} for time period i and category s. For each cell the ratio of the number of events to the time at risk is a crude hazard rate. The analysis involves **regression** methods to smooth these rates as a function of time and other **covariates**.

When such tables are produced as simple summaries of a data set, it is common to limit the number of time periods and other factors used to define the table. However, for modeling rates it is appropriate to use detailed tables with many cells based on a relatively fine **stratification** over time and other factors. For example, a rate table to be used in an analysis of an occupational cohort study (*see* **Occupational Epidemiology**) might be defined in terms of age, year, age at first exposure, sex, and cumulative exposure with hundreds or even thousands of cells. An event–time table for a **clinical trial** might involve follow-up time, age at entry, sex, and treatment. Although not usually necessary in practice, the methods can be applied to a table based on individual subjects where the only grouping is on time. This suggests the close connection between the use of Poisson regression methods for the analysis of rates and the Andersen–Gill **counting process** method [1] for analysis of hazard functions.

If it is assumed that the hazard, λ_{is} is constant within each cell, then the expected number of events in the cell is given by

$$\mathrm{E}(c_{is}) = R_{is} \times \lambda_{is}.$$

In terms of a parametric function, $\lambda(t_i, z_{is}, \theta)$ for the rates, the log **likelihood** for the survival data under the piecewise constant hazard assumption is

$$\sum_{l,s} c_{is} \ln(\lambda(t_l, z_{is}, \theta)) - R_{is} \times \lambda(t_i, z_{is}, \theta),$$

which is equivalent to the log likelihood that would arise if the event counts in the table were independent **Poisson** random variables. Thus, Poisson regression can be used to estimate the parameters in this model.

With this approach, modeling rates in terms of time is straightforward since, in contrast to Cox regression, there is no distinction between time-dependent and time-independent covariates. This is because the time-dependent computations are carried out when the event – time table is constructed and are not repeated each time a model is fitted.

Using External Rates or Expected Cases

In some situations one has external data on the expected rates λ_q^e stratified by time and other factors (e.g. age, calendar time period, and sex but not exposure or treatment related factors). In this case, it is possible to compute the expected number of cases for each cell in the table as $C_{is}^e = R_{is} \times \lambda_q^e$, where $q = q(is)$ denotes the external rate strata corresponding to cell is. In this case, Poisson regression can be used to model the relative hazard, ρ_{is}, since

$$\mathrm{E}(c_{is}) = C_{is}^e \times \rho_{is} = R_{is} \times \lambda_q^e \times \rho_{is}.$$

When **person-years** are replaced by expected numbers of cases, this type of analysis is known as the subject-years method or standardized mortality ratio (SMR) regression [4].

Models for Rate Regression

Following the pioneering work of Cox [8], the most commonly used hazard function model is the **log-linear** proportional hazards model

$$\lambda(t, z, \theta) = \lambda_0(t, \alpha) \times \exp(\beta z). \tag{1}$$

Here λ_0 is a baseline hazard for an individual with covariate $z = 0$.

Other models are also important, however. For example, in **dose–response** studies it is often useful to consider models in which the **excess relative risk** is a linear function of dose d; that is

$$\lambda(t, z, \theta) = \lambda_0(t, \alpha) \times (1 + \beta d).$$

Preston [16] has described a flexible general class of parametric **additive hazard models** of the basic form

$$\lambda_0(t, \alpha, z_0) + \lambda_{\mathrm{EAR}}(t, \beta, z_1) \tag{2}$$

and

$$\lambda_0(t, \alpha, z_0)[1 + \lambda_{\mathrm{ERR}}(t, \beta, z_1)], \tag{3}$$

in which λ_0 represents the baseline or background rates and λ_{EAR} and λ_{ERR} describe the excess absolute or excess relative risks. In these models baseline rates are usually assumed to be loglinear functions of the covariates while the excess risks are modeled as linear or products of linear and loglinear functions of the covariates.

One reason for the popularity of the Cox regression model is that it allows one to focus (perhaps too much) on the **relative risk** while treating the baseline hazard as completely unspecified. A similar simplification is possible in the analysis of **relative risk models** for rates using Poisson regression. This is accomplished by the inclusion of a **multiplicative** parameter for each time interval leading to models such as

$$\tau_i \exp(\beta z) \quad \text{or} \quad \tau_i(1 + \beta d). \tag{4}$$

This approach can also be extended to allow stratification over additional factors, in which case the model is similar to the stratified Cox regression model. Preston et al. [17] describe an efficient **algorithm** for models with large numbers of stratum parameters.

Parameter Estimation and Inference

Parameter estimates for Poisson regression models are computed using **maximum likelihood** methods. Models in which the rates depend on the parameters through a linear function βz, are **Generalized Linear Models (GLM)**. Parameter estimates for GLMs can be computed using iteratively reweighted **least squares** with person-years (or cases for subject-years analysis or standardized mortality/incidence ratio regression) as an "offset". These methods are available in all of the major statistical packages including GLIM, SAS, and S-PLUS (*see* **Software, Biostatistical**). However, the more general rate function models such as (3) and (4) are not GLMs. In this case, it is necessary to make use of special software to define the likelihood and possibly its derivatives. The Epicure package [17] is designed to work with models in the general class described by (1)–(4) above.

Inference about parameters of interest can be carried out using the standard asymptotic methods, including Wald, score, and **likelihood ratio tests**. However, because of the nonlinear nature of the models and, in many applications, the limited information on **excess risks**, asymptotic **standard errors** and hence **hypothesis tests** and **confidence intervals** based on Wald tests can be misleading. Score or likelihood ratio tests and **profile-likelihood**-based confidence intervals should be emphasized when working with additive hazard models.

An important issue concerns the assessment of **goodness of fit** for Poisson regression models derived from detailed event–time tables. Because rate modeling often involves relatively rare events and event–time tables with many cells, the rates or the number of events in each cell of the table can be quite small. In this case, neither the global deviance nor the Pearson **chi-square statistic** provides reasonable guidance as to goodness of fit. The total deviance is often much smaller than the putative **degrees of freedom** (the number of cells in the table minus the number of free parameters in the model). Pregibon [15] developed generalized regression **diagnostics** that can be used for regression models in exponential families. While such diagnostics may be useful in looking for lack of fit and other problems with fitted models [10], they should be interpreted with caution since the underlying data are not independent Poisson counts. In view of these issues, the most effective general method for the assessment of goodness of fit when using Poisson regression to analyze rates is to make use of likelihood ratio tests designed to detect specific departures from models of interest, such as time dependence or nonlinearity, or to make use of **Akaike's criterion** or related statistics to compare alternative (possibly nonnested) models.

Creating Event–Time Tables

The creation of an adequate event–time table is often the most difficult aspect of carrying out analyses of rates using Poisson regression. Among other features, an ideal program for the construction of event–time tables would:

1. allow for categorization on multiple time scales (age, year, length of follow-up, etc.), as well as multiple time-independent and time-dependent factors with variable length intervals in each of these scales;
2. allow for late entry, disjoint follow-up intervals, and multiple events;
3. include procedures for the computation of and categorization on time-dependent quantities;
4. allow computation and storage of counts for multiple event types along with representative values (often time-at-risk weighted means) for covariates of interest for each cell in the table;
5. have efficient procedures for handling the large, sparse tables that can arise when one stratifies on multiple time scales;
6. be able to deal with the data structures that can arise in describing complex exposure histories; and
7. facilitate the incorporation of external rates.

Several computer programs are currently available for the creation of event–time tables. However, many of these programs, e.g. OCMAP [6] and O/E [13], are designed for specific applications and are of limited use in more general problems. Procedures for the creation of such tables in the major statistical programs are extremely limited or nonexistent. The DATAB module in Epicure [17] and "Person-years" [7] are probably the most flexible general-purpose programs for event–time tabulation available at this time. Hopefully, there will be major improvements in this area over the next few years.

Summary

Poisson regression is a powerful tool for the analysis of rates from cohort survival studies that facilitates simple, straightforward analyses of temporal patterns, baseline risks, excess relative or **absolute risks**, and other aspects of hazard functions that may be difficult to assess with other methods. The application of Poisson regression requires that data on individual subjects be organized into event–time tables stratified on time and other factors of interest and, for the most interesting models, specialized software capable of dealing with nonlinear Poisson regression models is also required. The tools needed to conduct these analyses are available today but it is likely that they will be more fully developed in the years to come.

References

[1] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[2] Berry, G. (1983). The analysis of mortality by the subjects years method, *Biometrics* **39**, 173–184.

[3] Breslow, N.E. (1974). Covariance analysis of censored survival data, *Biometrics* **30**, 89–99.

[4] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. II: *The Design and Analysis of Cohort Studies*. IARC Scientific Publications 82, International Agency for Research on Cancer, Lyon.

[5] Breslow, N.E., Lubin, J.H., Marek, P. & Langholz, B. (1983). Multiplicative models and cohort analysis, *Journal of the American Statistical Association* **78**, 1–12.

[6] Caplan, R.J., Marsh, G.M. & Enterline, P.E. (1983). A generalized effective exposure modeling program for assessing dose-response in epidemiologic investigations, *Computers in Biomedical Research* **16**, 587–596.

[7] Coleman, M. (1986). Cohort study analysis with a FORTRAN computer program, *International Journal of Epidemiology* **15**, 134–137.

[8] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[9] Frome, E. (1983). The analysis of rates using Poisson regression models, *Biometrics* **39**, 665–675.

[10] Frome, E. & Morris, M.D. (1989). Evaluating goodness of fit of Poisson regression models in cohort studies, *American Statistician* **43**, 144–147.

[11] Holford, T.R. (1980). The analysis of rates and survivorship using log-linear models, *Biometrics* **36**, 299–305.

[12] Laird, N. & Oliver, D. (1981). Covariance analysis of censored survival data using log-linear models, *Journal of the American Statistical Association* **76**, 231–240.

[13] Monson, R.R. (1974). Analysis of relative survival and proportional mortality, *Computers in Biomedical Research* **7**, 325–332.

[14] Pierce, D.A., Shimizu, Y., Preston, D.L., Vaeth, M. & Mabuchi, K. (1996). Studies of the mortality of atomic bomb survivors. Report 12, Part I. Cancer Mortality 1950–1990, *Radiation Research* **146**, 1–27.

[15] Pregibon, D. (1981). Logistic regression diagnostics, *Annals of Statistics* **9**, 705–724.

[16] Preston, D.L. (1990). Modeling radiation effects on disease incidence, *Radiation Research* **124**, 343–344.

[17] Preston, D.L., Lubin, J., Pierce, D.A. & McConney, M.E. (1993). *Epicure, Users' Guide*. Hirosoft International, Seattle.

DALE L. PRESTON