

Statistical Methods in Studies of Radiation Risk

Dale Preston
Department of Statistics
Radiation Effects Research Foundation
Hiroshima, Japan

Introduction

Ionizing radiation is one of the most thoroughly studied environmental hazards and modern statistical methods are essential to our developing understanding of the health effects of radiation exposure. The long term follow-up of the atomic bomb survivors in Hiroshima and Nagasaki, which has been underway for 50 years and will continue into the next century, is one of the largest long-term epidemiological follow-up studies ever undertaken. In addition to mortality follow-up of more than 90,000 exposed victims of the bombings, the survivor studies involve a longitudinal clinical follow-up study of a cohort of 15,000 survivors, which will soon enter its 40th year. The atomic bomb survivor studies are not the only large scale investigations of the long term health effects on radiation on humans. Other large studies involve nuclear industry workers from several countries (Cards et al. 1995, Gilbert et al. 1993), cohorts of people exposed to radiation for diagnostic or therapeutic purposes (Darby, Boice et al. 1985, Howe, 1995), studies of the effects of the Chernobyl accident and of other populations in the former Soviet Union with significant radiation exposures (Kossenko and Degteva, 1994; Koshurnikova et al 1994).

The United Nations and other national and international organizations routinely produce major reports synthesizing the results of studies of radiation effects UNSCEAR, ICRP 60, BEIR V). Because of the public concerns about the health effects of radiation the results of these studies have a direct and significant impact on policies regarding radiation protection.

Analyses of data from studies of radiation effects involve a broad range of interesting and challenging statistical issues. Statistical problems that arise in radiation studies include: errors in dose estimates often with highly asymmetric or non-continuous distributions (Pierce et al. 1990); over-dispersed count data in the analysis of chromosome aberrations or other biomarkers of radiation exposure (Ochi and Prentice 1984, Moore 1986); calibration problems that arise in efforts to use biological outcomes to estimate unknown exposures; longitudinal methods for the analysis of growth curves or clinical measurement data (Wong et al); image analysis and pattern recognition in the examination of parent-child trios to search for DNA-level evidence of radiation-induced mutations.

One area in which studies of radiation effects have lead to the development of useful statistical methods concerns regression analyses of cohort survival data. Partial likelihood methods for proportional hazards models in which the relative risk is (usually) modeled as a log-linear function of covariates with an arbitrary model for the baseline hazard have played a central role in the analysis of survival data since the seminal paper by Cox (1972) and the book by Kalbfleisch and Prentice (1980). In studies of radiation effects partial likelihood methods and proportional hazard models were used by Gilbert and Marks (1979) in analyses of data on US nuclear workers and by Prentice et al. (1983) for analyses of radiation effects in the atomic bomb survivors. However, since the early 1980's alternative models and estimation procedures based on Poisson regression methods have been adopted for use in studies of radiation effects on mortality and morbidity and are now being used in an increasing number of non-radiation applications.

In this paper I introduce a class of generalized hazard function models that has proven to be extremely useful in the analysis of radiation effects, and describe how Poisson regression methods applied to detailed tabulations of cases and person-years can be used to fit these models. Issues related to the construction of suitable person-year tables are also discussed. Although this presentation focuses on the use of these models and methods in studies of radiation effects, it should be clear that the models and methods are applicable to a wide range of problems.

Poisson regression and hazard models for radiation effects

The classical proportional hazards model has the form

$$\lambda(t)e^{\beta z}$$

where $\lambda(t)$ represents the baseline or background hazard as a non-parametric function of time t , z is a vector of covariates, and $e^{\beta z}$ is the relative risk function. The most common (though rarely used) extensions to this model involve: allowing the background hazard to depend on some stratification variables; the inclusion of time dependent covariates in the relative risk model; and replacing the log-linear relative risk function by a linear model of the form $(1 + \beta z)$, where βz is a linear model for the excess relative risk.

While classical proportional hazards modeling methods are useful in modeling radiation effects in the atomic bomb survivors and other large cohorts there are several compelling reasons to consider alternative approaches. First, because of the large size of many of these studies and the nature of the models of interest, which often involve time-dependent covariates, there is a need to develop more efficient computational procedures. In addition, experimental data and theoretical considerations have suggested the use of alternatives to the simple log-linear models for the relative risk.

The following equation is an example of a relative risk model of interest in radiation studies:

$$\lambda_{sp}(a) (1 + \beta_s d e^{\gamma \ln(t)}) \quad (1)$$

In this model the baseline hazard ($\lambda_{sp}(a)$) is a function of attained age (a) and depends on strata defined by sex (s) and calendar time (p). The dose effect is described in terms of the excess relative risk, which depends linearly on dose (d) with a sex-dependent slope and log-log-linearly on time since exposure (t). The model (1) often must be extended to allow for non-linear dose response functions or for the consideration of other factors that may modify the radiation effect, such as age at exposure. In many radiation studies the situation is further complicated by the fact that dose is a relatively complex time-dependent covariate.

Model (1) has several features that distinguish it from commonly used proportional hazards models. First, the model involves multiple time-scales (age, calendar time, and time since exposure) and time-dependent covariates. In addition, the form of the model is more complex than the log-linear or simple additive $(1 + \beta z)$ models available in the standard computer packages for proportional hazards modeling.

Even with suitably flexible modeling software (such as Epicure [Preston et al. 1993]) and today's powerful computers, fitting a model like (1) to a large cohort without some grouping of the data is time consuming and discourages the exploratory modeling that can lead to potentially important insights into the data being considered.

As noted above, Poisson regression methods for survival data can be used to overcome almost all of the problems that arise in working with models such as (1). The basic ideas underlying the use of Poisson regression for survival data were outlined by Holford (1980), Frome (1983), and Pierce and Preston (1984). In particular if one assumes that the hazard function is constant on strata defined by time and other factors then the likelihood for the censored data is equivalent to the likelihood that arises when the number of events in each stratum (C_i) are treated as independent Poisson random variates with

$$E(C_i) = \lambda_i P_i$$

where λ_i is the stratum (or cell)-specific hazard and P_i is the total time at risk (usually person-years) in the stratum. If the data for each cell also include a vector of covariates ($\mathbf{z}_i$) that describe representative values for the various time scales (for example attained age, time since first exposure, and calendar period) and other covariates (dose, sex, age at first exposure) for the cell, then one can use Poisson regression methods to fit models that describe the variation in the hazard function over strata with time and other factors.

One of the strengths of the Poisson regression approach to the analysis of survival data is that once a suitable cross tabulation has been created, parameter estimation and inference for models involving multiple time scales or other time-dependent covariates requires no more computational effort than are required for models that do not have such factors. Creation of a person-year tabulation requires roughly the same amount of computational effort as one iteration in the fitting of a model with time-dependent covariates. Because the Poisson regression method can be viewed as fitting models to crude rates it can be easier to test hypotheses and to explain results than is the case for other methods commonly used in the analysis of cohort survival data.

An appealing feature of partial likelihood methods is that they allow one to focus on the relative risk without any explicit modeling of the background hazard, which can be quite difficult for ungrouped survival data. There are efficient Poisson regression algorithms (Preston et al. 1993) that allow one to model relative risks for grouped survival data using highly stratified models for the background hazards. When using these algorithms the background stratification factors include categories for the primary time scale (e.g. age) and other factors of interest (e.g. sex, place of residence, birth cohort, calendar time period).

In analyses of radiation data attention is usually focused on the radiation effect but it can also be important to describe and understand the background hazard function. In contrast to the situation for ungrouped survival data, when using Poisson regression methods to model grouped survival data, it is possible, and indeed quite straightforward to specify and fit fully parametric models, such as $e^{\alpha_s + \tau_s p + \theta_s \ln(a) + \eta_s \ln^2(a)}$, for the baseline hazard.

Because Poisson regression methods allow one to fit parametric models for the full underlying hazard, it is simple and straightforward to consider non-proportional hazards models. One class of models of interest in radiation effects are absolute risk (risk difference) models. The following equation is an example of "simple" background plus excess rate model:

$$e^{\alpha_s + \tau_s p + \theta_s \ln(a) + \eta_s \ln^2(a)} + (\beta_1 d + \beta_2 d^2) \quad (2)$$

In this model the excess absolute rate is assumed to be a quadratic function of dose but not vary with time (after some minimal latent period). Until about 15 years ago, constant absolute risks like (2) were of considerable interest in radiation studies. With the introduction of relative risk modeling methods it has been shown that the excess absolute risk is not constant in time and attention has focused on relative risk models like (1). However, current analyses of the atomic bomb survivor data (Pierce et al. 1996) suggest that time-dependent models for the excess absolute risk such as

$$e^{\alpha_s + \tau_s p + \theta_s \ln(a) + \eta_s \ln^2(a)} + \beta d e^{\gamma \ln(a)} \quad (3)$$

can describe the excess risks for cancers other than leukemia as well as more complex relative risk models. Fitting a model like (3) using Poisson regression methods is no more difficult than fitting relative risk models if one has access to suitable Poisson regression software.

In some studies of radiation effects rates in the study population are compared to those for an external unexposed. As described by Breslow et al. (1983), standardized mortality ratio (SMR) regression methods can be used for this purpose. Given population rates, e.g. r_{sap} where the rates depend on sex, age at death, and calendar time period, one might be interested in an SMR-regression model like

$$\theta_s r_{sap} (1 + \beta_s d e^{\gamma \ln(a)})$$

Fitting such a model using standard proportional hazards modeling software would be a challenge but fitting this models (or its risk difference equivalent) is quite simple using good Poisson regression software (e.g. the AMFIT program in Epicure).

Person year computation for Poisson regression

As suggested above Poisson regression methods can be extremely useful for the analysis of censored survival data. They are especially useful and efficient when working with multiple time-scales or time-dependent covariates or with SMR-regression. The key to using Poisson regression methods in survival analysis is the creation of the person year tabulation. As noted above, in order to use Poisson regression methods the data for individual cohort members must be summarized in a person-year (or more generally event-time) table.

The cells (records) in this table are defined by a cross classification over categories of the factors of interest which would typically include at least one and often several time scales (e.g. attained age, time since first exposure, length of employment), factors which do not vary with time (e.g. sex, race, age at first exposure, place of first exposure, dose in the case of a single acute exposure), and possibly time-dependent covariates (dose in the case of occupational exposures). In most cases each cell in a person year table will include data from several subjects. However, there is, in principal, no problem in working with data in which there is a separate record for subject for each period at risk. This latter data structure is quite similar to data structure used in analyzing survival data using counting process methods (Andersen et al. 1992).

The data for each cell in the table include the total time at risk in the cell (typically in person years), the number of events in the cell (in some studies there can be multiple types of events in a cell, e.g. deaths from various causes); and representative values for the covariates in then cell. These representative values (or summary variables) are typically person-year (or person) weighted means for all people ever at risk in the cell. It can also be useful to count the number of people ever at risk in each cell and the number of people who had there first period at risk in this cell. For SMR regression it is also necessary to include the appropriate population rate(s) as a covariate for each cell.

In creating a person year table for survival analysis, it is important that the resolution be fine enough not to hide important variability in the risk. For example if one has reason to believe that the hazard function is changing quite rapidly early in the follow-up period and much less rapidly later in the follow-up period, it would be important to allow rather narrow intervals for shortly after the exposure with, perhaps, broader intervals at later times. When working with time-dependent covariates, it is also important to make sure that the resolution used in updating a person year table is fine enough to insure accurate assessment of the time-dependent covariates. For example, in a study in which one factor of interest is the number of children a women has the use of five year intervals for time or age can lead to an inaccurate

categorization of the number of children. One can deal with such problems by either explicitly using a finer resolution for the time scale(s) or, if the program allows, setting the resolution to be used in updating the table to be finer than the resolution specified by the time categories used to define the table.

Person-year tables for survival analysis can, and generally will, have a large number of cells. In recent work with the atomic bomb survivor data the person-year tables generally have 5,000 to 25,000 cells. In smaller studies it is not uncommon for the number of cells in a reasonable person-year table to exceed the number of cohort members.

There are few general purpose programs for the creation of the complex person year tables. The DATAB program in Epicure is the most powerful widely available general purpose person-year tabulation program. DATAB can be used to create tables with a broad range of summary variables that involve categories defined in terms of multiple time scales, fixed variables, and complex time-dependent variables. DATAB also includes a facility for the addition of external rates for use in SMR regression.

Inference and goodness of fit in Poisson regression analyses of survival data

As noted above, by assuming piecewise constant hazards it is possible to use ordinary Poisson regression methods to model hazard functions. An appropriate level of grouping for many analyses can involve a large number of cells with the expected number of events being quite small in many if not all cells in the table. In addition, in many problems, including radiation effects studies based on large cohorts, the number of informative events may be quite small so there is a need to be concerned with the adequacy of the asymptotics that underlie commonly used procedures for hypothesis testing and inference.

When working with highly stratified data simple comparison of the total deviance to its degrees of freedom is not useful as a summary measure of goodness of fit. In particular, the total deviance is often much smaller than its degrees of freedom (number of cells minus number of free parameters). This problem arises because the deviance corresponds to the difference between the log-likelihood for the model being fit and the log-likelihood for a saturated model (i.e. a model that includes a parameter for each cell in the table) and for finely stratified data the log-likelihood for the saturated model depends on arbitrary aspects of the grouping over factors not included in the model of interest. Because of the problems with the global deviance (or Pearson χ^2 statistic), assessment of the fit of grouped data survival models should rely on likelihood ratio tests (deviance differences) for hypotheses related to specific aspects of the lack of fit. Such tests may involve test for trend or the use of categorical factors.

For non-nested models, i.e. pairs of models in which the parameters of one model are not simply a subset of parameters in the other, comparison of Akaike information criterion (Akaike, 1973; McCullagh and Nelder 1990 pp 90-91) values can be useful. The Akaike information criterion is defined as the deviance plus two times the number of fitted parameters in a model.

The most commonly used methods for hypothesis testing and the construction of confidence intervals for specific parameters involve the use of Wald statistics, which for a single parameter are based on the ratio of the parameter estimate to its asymptotic standard error. The validity of these tests is based on the assumption that the profile log-likelihood is well-approximated by a quadratic function of the parameter(s). When working with limited amounts of data or non-linear regression models it is wise to use asymptotic methods which do not rely on assumptions about the shape of the profile likelihood. In such cases one should use likelihood ratio tests and profile likelihood-based confidence bounds (Cox and Hinkley, 1974 p. 218). Plots of the of the profile likelihood can provide useful insights into the nature of uncertainties in parameter estimates. Computation of likelihood-based bounds and the evaluation of profile likelihood functions can require a lot of computation, but if one is using software that provides direct support for profile likelihood evaluation the computations are quite practicable on current workstations and high-end personal computers.

Concluding remarks

Poisson regression methods for grouped survival data are an important tool for the analysis of censored data. By making it possible to model the hazard function directly this approach allows the analyst to work with a broad range of useful models that are difficult or even impossible to fit using standard parametric or semi-parametric methods. In this paper I have presented some examples of general hazard function models from a class of relative and absolute risk models that are central to modern analyses of data on the long-term health effects of radiation exposure. Models of this type should also be useful in other fields.

The obstacle to the widespread use of these methods is the creation of detailed person year tabulations. Powerful, general purpose programs for the creation of such tables are available (e.g. DATAB), but additional work is needed to make it easier to create person year tables for situations that involve complex time-dependent exposures or other time-dependent factors.

References

- Akaike H. (1973) Information theory and an extension of the maximum likelihood principle. in *Second International Symposium on Information Theory*. (eds. Petrov BN, Czaki F) Akademiai Kiado, Budapest pp 267-281.
- Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1992). *Statistical Models based on Counting Processes*. New York: Springer-Verlag.
- Boice, J.D.Jr., Day, N.E., Anderson, A., Brinton, L.A., Brown, R., Choi, N.W., Clarke, C.M.P., Curtis, R.E., Flannery, J.T., Hakama, M., Hskulinen, T., Howe, G.R., Jensen, O.M., Kleinerman, R.A., Magnin, D., Magnus, K., Makela, K., Malker, B., Nelson, N., Patterson, F., Pompe-Kirn, V., Primic-Zakelj, M., Prior, P., Ravnihar, B., Skeet, R.G., Skjerven, J. E., Smith, P.G., Sok, M., Spengler, R.F., Storm, H.H., Stovall, M., Tomkins, G.W.O. and Wall, C. (1985). Second cancers following radiation treatment for cervical cancer. An international collaboration among cancer registries. *J Natl Cancer Inst.* 74, 955-975.
- Breslow, N.E., Lubin, J.H., Marek, P. and Langholz, B. (1983). Multiplicative models and cohort analysis. *J. Amer Stat. Assoc.* 78, 1-12.
- Cardis, E., Gilbert, E.S., Carpenter, L., Howe, G., Kato, I., Armstrong, B.K., Beral, V., Cowper, G., Douglas, A., Fix, J., Fry, S.A., Kaldor, J., Lave, C., Salmon, L., Smith, P.G., Voelz, G.L. and Wiggs, L.D. (1995). Effects of low doses and low dose rates of external ionizing radiation: cancer mortality among nuclear industry workers in three countries. *Radiat. Res.* 142(2), 117-132.
- Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman and Hall.
- Cox, D.R. (1972). Regression models and lifetables. *J.R. Stat Soc. B* 34, 187-220.
- Weiss, H.A., Darby, S.C., Fearn, T. and Doll, R. (1995). Leukemia mortality after X-ray treatment for ankylosing spondylitis. *Radiat. Res.* 142 (1), 1-11.
- Frome, E.L. (1985). The analysis of rates using Poisson regression models. *Biometrics* 39, 665-674.
- Frome, E.L. and Checkoway, H. (1985). Use of Poisson regression models in estimating incidence ratios and rates. *Am. J. Epidemiol.* 121, 309-323.
- Gilbert, E.S. and Marks, S. (1979). An analysis of the mortality of workers in a nuclear facility. *Radiat. Res.* 79, 122-148.

- Gilbert, E.S., Cragle, D.L. and Wiggs, L.D. (1993). Updated analyses of combined mortality data for workers at the Hanford Site, Oak Ridge National Laboratory, and Rocky Flats Weapons Plant. *Radiat. Res.* 136 (3), 408-421.
- Holford, T.R. (1980). The analysis of rates and of survivorship using log-linear models. *Biometrics* 36, 229-306, 1980.
- Howe, G.R. (1995). Lung cancer mortality between 1950 and 1987 after exposure to fractionated moderate-dose-rate ionizing radiation in the Canadian fluorescopy cohort study and a comparison with lung cancer mortality in the atomic bomb survivors study. *Radiat. Res.* 142, 295-304.
- Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.
- Koshurnikova, N.A., Buldakov, L.A., Bysogolov, G.D., Bolotnikova, M.G., Komleva, N.S. and Peternikova, V.S. (1994). Mortality from malignancies of the hematopoietic and lymphatic tissues among personnel of the first nuclear plant in the USSR. *The Science of the Total Environment* 142, 19-23.
- Kossenko, M.M. and Degteva, M.O. (1994). Cancer mortality and radiation risk evaluation for the Techa river population. *The Science of the Total Environment* 142, 73-89.
- McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models 2nd Edition*. New York: Chapman and Hall.
- Moore, D. F. (1986) Asymptotic properties of moment estimators for overdispersed counts and proportions. *Biometrika* 73, 583-588.
- Ochi, Y. and Prentice, R. L. Likelihood inference in a correlated probit regression model, *Biometrika* 71, 531-543.
- Pierce, D.A., Preston, D.L. and Ishimaru, T. (1983). A method for analysis of cancer incidence in Japanese A-bomb survivors with application to acute leukemia. *RERF Technical Report 15-83*, Hiroshima (Japan): RERF.
- Pierce, D.A. and Preston, D.L. (1984) Hazard function modeling for dose response analyses of cancer incidence data in A-bomb survivor data. In *Atomic bomb survivor data: utilization and analysis*, Prentice, R.L. and Thompson, D.J. (eds), 51-66. Philadelphia, SIAM.
- Pierce, D.A., Shimizu, Y., Preston, D.L., Mabuchi, K. and Vaeth, M. (1996). Studies of the mortality of A-bomb survivors. Report 12, Part I. Cancer: 1950-1990. *Radiat. Res.* (in press).
- Pierce, D.A., Stram, D.O. and Vaeth, M. (1990). Allowing for random errors in radiation exposure estimates for the atomic bomb survivors data. *Radiat. Res.* 123, 275-284.
- Prentice, R.L., Yoshimoto, Y. and Mason, M.W. (1983). The relationship of cigarette smoking and radiation exposure to cancer mortality in Hiroshima and Nagasaki. *J Natl Cancer Inst.* 70, 611-622.
- Preston, D.L., Lubin, J.H., Pierce, D.A. and McConney, M.E. (1993) *Epicure Users Guide*. Seattle: Hirosoft International Seattle.
- Wong, F.L., Yamada, M., Sasaki, H., Kodama, K. and Akiba, S. (1993). Adult Health Study Report 7. Noncancer disease incidence in the atomic-bomb survivors, 1958-86 (examination cycles 1-14). *Radiat. Res.* 135, 418-430.