On statistical modeling of incidence of competing events, with application to labor mobility analysis

Petr Volf[1]

Abstract. The contribution deals with the problem of competing risks (of competing events) in the statistical events occurrence analysis. In such a setting, one event excludes the occurrence of the other. Moreover, their latency may be dependent. Therefore, instead the analysis of marginal distributions (or intensities) of events, it is more convenient to model their real incidence, via so called incidence function. We present methods of such an incidence analysis and illustrate it on an example with unemployment data.

Keywords: mathematical statistics, event-history analysis, incidence, unemployment study.

JEL classification: C41, J64
AMS classification: 62N02, 62P25

1 The event-history analysis

The beginning of statistical event-history analysis can be traced back to the life-tables construction by actuaries and demographers several centuries ago. Later on, remarkable progress of theory and methodology is due so called survival analysis developed mainly in areas of biostatistics, medical statistics and also technical reliability analysis. The main characteristic of the model of an event occurrence is the rate or intensity with which the events occur. In biostatistics, the goal is usually in observing the time until a single non-repeatable event (e. g. death). In contrast, in the field of social or demographic studies, several kinds of events may be followed (e. g. transition among several labor states, several important events in the life), some of them repeatable (e. g. change of job or marriage). Event-history data gives the type of event together with the time at which it happened, so that each observed event has also its "mark" indicating what (and to whom) occurred.

In the area of social sciences, the event-history models are considered for instance in monographs of Tuma and Hannan [8], Gourieroux [2] or Winkelmann [9]. A fundamental paper of Heckman and Singer [4] is also devoted to the continuous time models (i. e. models of intensities) for econometrics duration analysis. The latest ideas and techniques in the area of statistical event-history analysis are connected with the theory and application of the stochastic counting processes (cf. Andersen et al. [1]).

1.1 Counting process – definition and examples

A multivariate counting process $N(t) = \{N_1(t), N_2(t), \ldots, N_n(t)\}$ is a stochastic process having n components $N_i(t)$, each of them counting the number of a (registered, observed) specified event. So that it can describe n types of events, or follow the occurrence of only one event, for n objects. The time t runs as a rule from 0 to some finite T at which the observation is terminated and the data are collected. It could be the calendar time, in other cases it could be a kind of relative time (as the age). It is assumed that $N_i(0) = 0$ at the beginning and that N_i has jump of size $+1$ at the moment when the event is observed. As sometimes we are not able to observe complete history, some events remain non-recorded, they are *censored*.

Further, let $I_i(t)$ be the indicator process, $I_i(t) = 1$ if events of i-th object can be observed at moment t, $I_i(t) = 0$ otherwise. In other words, $N_i(t)$ is exposed to the risk of the count only if $I_i(t) = 1$. Let there

[1]Institute of Information Theory and Automation, Prague, e-mail: volf@utia.cas.cz

exist a nonnegative function

$$h_i(t) = \lim_{d\to 0} \frac{P\{N_i(t+d^-) - N_i(t) = 1 | I_i(t) = 1\}}{d}.$$

It is called the hazard function (or hazard rate). Then $\lambda_i(t) = I_i(t)\, h_i(t)$ is the *intensity*. Finally, define the *cumulative hazard rate* $H_i(t) = \int_0^t h_i(s)\, ds$ and *cumulative intensity* $L_i(t) = \int_0^t \lambda_i(s)\, ds$. Then the process $M_i(t) = N_i(t) - L_i(t)$ is a martingale with zero mean, non-correlated increments, $M_i(t)$ are non-correlated mutually.

2 Problem of competing risks

Let us recall first an example from Han and Hausman [3]. The data contain $n = 1\,055$ records of people who have lost their job and are either searching for a new one or waiting for reemployment by the previous employer. The duration of unemployment is the variable of our interest, so that the time from the onset of unemployment is the reference time t. We assume that the histories of different individuals are independent each of other, therefore the individual times may be shifted in such a way that $t = 0$ is the moment of the loss of job. Each record contains values T_i, δ_i, $\boldsymbol{X}_i$, where T_i is the moment of a new employment (then $\delta_i = 1$) or the moment of recall ($\delta_i = 2$) or the moment of censoring ($\delta_i = 0$). The time scale is given in weeks, from 0 to 70 weeks. $\boldsymbol{X}_i$ denotes a set of covariates (constant in the time), e. g. age at $t = 0$, years of schooling (continuous variables), sex, race, marital status, occupation, industry and some others discrete categorical variables. The task is to estimate hazard rates of obtaining a new job and of the recall – so that two competing risks are examined (and the censoring is the third way of the end of the record of each person). In [3] a parametric model of unemployment duration is considered. In the present paper a more general semi- and non-parametric setting will be employed.

The analysis of cases with multiple competing events is complicated by the fact that the occurrence of one event excludes the occurrence of others. In this sense, times to events are dependent, and it is hard to model such a dependence – just because estimated marginal distributions lack reasonable interpretation. Of course, we can consider cause – specific hazard functions for events $j = 1, 2, \ldots, K$ (in the example we had $K = 2$),

$$h_j(t) = \lim_{d\to 0} \frac{P(t \le T < t + d,\, \delta = j \,|\, T \ge t)}{d},$$

or overall hazard rate for all events together,

$$h(t) = \lim_{d\to 0} \frac{P(t \le T < t + d \,|\, T \ge t)}{d}.$$

Here we assume that the hazard rates do not depend on i, they are the same for each person. If we integrate them, we obtain cumulated hazard rates $H_j(t)$, $H(t)$, and also overall survival function $S(t) = P(T > t) = \exp(-H(t))$. Notice that $h = \sum_{j=1}^K h_j$. Instead of marginal survival functions $S_j(t) = \exp(-H_j(t))$ representing an ideal nonrealistic case when only j-th risk is present, we are more interested in modeling actual incidence of event j, via so called cumulative incidence functions

$$F_j^*(t) = P(T \le t, \delta = j) = \int_0^t S(s) \cdot h_j(s)\, \mathrm{d}s.$$

Notice that it has certain properties of distribution function, but its $\lim F_j^*(t) = P(\delta = j) < 1$ if t tends to infinity. Further, it holds that $S(t) = 1 - \sum_{j=1}^K F_j^*(t)$. It is seen that marginal $F_j(t) = 1 - S_j(t)$ overestimates actual incidence of event j.

2.1 Estimation method

All cumulative hazard rates can be estimated standardly by the Nelson–Aalen estimator, namely

$$\hat{H}_j(t) = \int_0^t \sum_{i=1}^n \frac{\mathrm{d}N_{ij}(s)}{I(s)}, \qquad \hat{H}(t) = \sum_{j=1}^K \hat{H}_j(t), \tag{1}$$

where again i is the index of object (person) and j of type of event (in the case of presence of censoring from right side, index $j = 0$, for value $\delta_i = 0$, denotes an end of observation caused by censoring). Overall survival function can then be estimated by the Kaplan Meier estimator, or directly as

$$\hat{S}(t) = \exp(-\hat{H}(t)).$$

Asymptotic properties of estimates of incidence functions

$$\hat{F}_j^*(t) = \int_0^t \hat{S}(s)\,\mathrm{d}\hat{H}_j(s) \tag{2}$$

follows from the properties of good asymptotic properties of $\hat{S}$ and $\hat{H}_j$ and are derived for instance in Lin [6]. In general, limit distribution of $\sqrt{n}(\hat{F}_j^*(t) - F_j^*(t))$ is that of Gauss random process, with estimable covariance structure. As it is not a martingale, further inference (e.g. statistical tests) is not easy. Notice, however, that in the simplest case without censoring $F_j^*(t)$ and $S(t)$ correspond, at each fixed t, to probabilities in a multinomial distribution, the estimates correspond to relative occurrence, so that their properties simplify. In general, confidence regions for statistical testing are obtained by a Monte Carlo random generation.

3 Regression models

Generally, the intensity can depend on some explanatory variables, covariates. This dependence is modeled via regression models. The values of covariates can again be given by an observed random process depending on time. Let us denote it $X_i(t)$, for i-th object (person). Then, the regression model for intensity assumes that the random point process behavior is governed by a (bounded and smooth, say) hazard function $h(t, x)$ from $[0, T] \times \mathcal{X}$ to $[0, \infty)$, where $\mathcal{X}$ is the domain of values of $X(t)$. The intensity is then

$$\lambda_i(t) = h(t, X_i(t)) \cdot I_i(t),$$

so that it is actually a random process, too. Corresponding theory as well as the methodology of statistical analysis is collected in many papers and monographs devoted to statistical survival analysis, see for instance [1] and [5].

3.1 Examples of regression models

The idea to separate common hazard rate from the influence of covariates led to the multiplicative model, called also the proportional hazard model,

$$h(t, x) = h_0(t) \cdot \exp(b(x)).$$

Function $h_0(t)$ is the baseline hazard function, $b(x)$ is the regression (response) function. If the response function is parametrized, we obtain semiparametric Cox's model. Its most popular form assumes that $h(t, x) = h_0(t) \exp(\beta x)$.

Alternatives are for instance the Aalen's additive regression model or the accelerated time model used frequently in reliability analysis. In the present paper the Cox's model is utilized, hence we shall employ 'maximum partial likelihood' estimators of parameters β and the Breslow-Crowley estimator of the increments of cumulated baseline hazard function $H_0(t) = \int_0^t h_0(s)ds$. Function $h_0(t)$ is then obtained by kernel smoothing of these increments.

3.2 Estimation in Cox's model

In the setting of counting processes, the conditional likelihood (given intensities $\lambda_i(t)$) is

$$\mathcal{L}^c = \prod_{i=1}^{n} \left[\prod_{t<T} \lambda_i(t)^{dN_i(t)} \exp\left\{ - \int_0^T \lambda_i(t)\,dt \right\} \right].$$

Here $\lambda_i(t) = I_i(t) \cdot h(t, X_i(t))$, $dN_i(t) = 1$ just at the moment of count of $N_i(t)$, $dN_i(t) = 0$ otherwise.

In the case of multiplicative model, the Cox's partial likelihood function is available for the estimation of $b(x)$ independently on $h_0(t)$. The logarithm of partial likelihood is

$$\ell^P = \sum_{i=1}^{n} \int_0^T \log \left[\frac{\exp b(X_i(t))}{\sum_{j=1}^{n} \exp b(X_j(t))\, I_j(t))} \right] dN_i(t),$$

and an "ideal" estimate of function $b(x)$ should maximize it. Thus, in the framework of Cox's model where $b(x) = \beta x$, we search for optimal parameters β. Estimates are asymptotically normal, that is why we are able to compute the test statistics having approximately Gaussian or chi-squared distribution.

3.3 Incidence in regression model

This problem is analyzed for instance in Scheike and Zhang [7]. When the value of covariate is x and is constant in time, the cumulative incidence functions have the form

$$F_j^*(t;x) = P(T \le t,\, \delta = j \,|\, x) = \int_0^t \exp \left[- \int_0^{s^-} \sum_{j=1}^{k} h_j(u;x)\, \mathrm{d}u \right] \cdot h_j(s;x)\, \mathrm{d}s,$$

where, as before, the first term represents overall survival function (when the value of covariate equals x) and $h_j(s;x)$ is a cause j specific hazard function. So that each component of F_j^* can be estimated by an appropriate procedure. For instance if $h_j(t;x)$ follow Cox's model, the methods described in preceding part can be employed. Naturally, it is even more difficult to establish (at least asymptotic) properties of incidence functions estimates. An overview of contemporary state of the art and ideas of further methodology development are collected also in [7].

In the case of time-dependent covariates it is seen that $F_j^*(t;x)$ depend on the whole path of $x(s), s < t$. So that we can predict the incidence only if we are able to predict the development of covariates.

There are, naturally, many other interesting questions concerning the influence of covariates to incidence, and mutual relationship of incidence of different events. Thus, in medical studies it is expected that the decrease of risk of one cause (of a disease) can lead to increased incidence of another cause (though the overall incidence decreases). In the situation of our example with unemployment data, the dependence of competing risks could be quite different. We shall illustrate it in the next part.

4 Example

The data of the following example are artificial, however they reflect a situation typical e.g. for academic institutions in the beginning of 90-ties (and also at present), namely a forced temporal reduction of staff. The data are available at http://simu0292.utia.cas.cz/volf/MME 2010/.

Let us follow, during a period $[0, T]$, $T = 120$ months, a "fate" of $n = 185$ employees of a company. Let the unvoluntary departure ("firing") and voluntary change of job be two events of our interest denoted by $j = 1, 2$, no other censoring is involved, the rest of people remains with the company up to T, where the observation period ends. Assume that the i-th employee joined the company at a time $S_i \in [0, T]$. At that moment the corresponding counting processes $N_{ji}(t)$ started to be at risk of jump. Here $i = 1, ..., n$, $j = 1, 2$, we denote T_{ji} times of occurrence of j-th event for person i. We actually observe $T_i = min_j T_{ji}$, so that the indicator $I_i(t) = 1[S_i < t \le T_i]$ is common for both risks. Most of records have $T_i = T$, i. e. the majority of employees were still with the company at the end of the study. There were together 59 events of type 1 and 25 of type 2.

Figure 1. shows estimated cause specific hazard rates obtained by a smoothing technique from Nelson-Aalen estimates (1), and also their sum. It is seen that during the first 8 years the overall hazard rate is almost constant. However, in last years there were 2 periods of forced reduction of the company staff, the intensity of "voluntary" departures reacted with a slight delay (uncertainty of present job increased the search for a new one).

Then, Figure 2. displays estimated marginal distribution functions F_j (dashed) and incidence functions F_j^* - see (2). As expected, the latter are lower than marginal ones, with difference increasing with time.

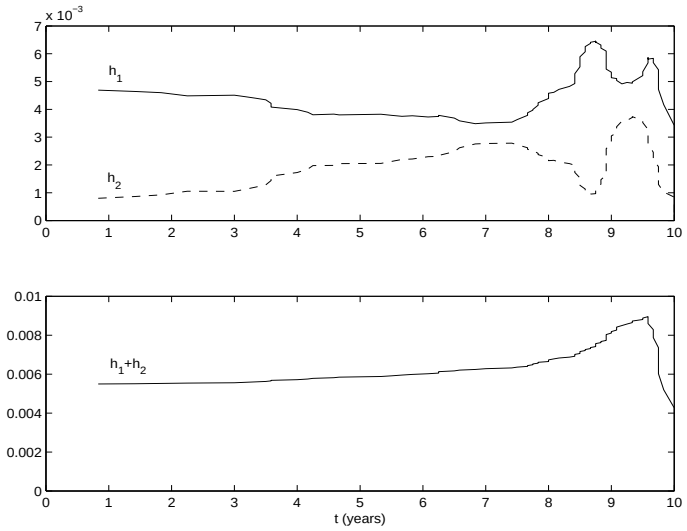

Figure 1: Estimated hazard rates for events 1 and 2 (above), estimate of overall hazard rate (below)

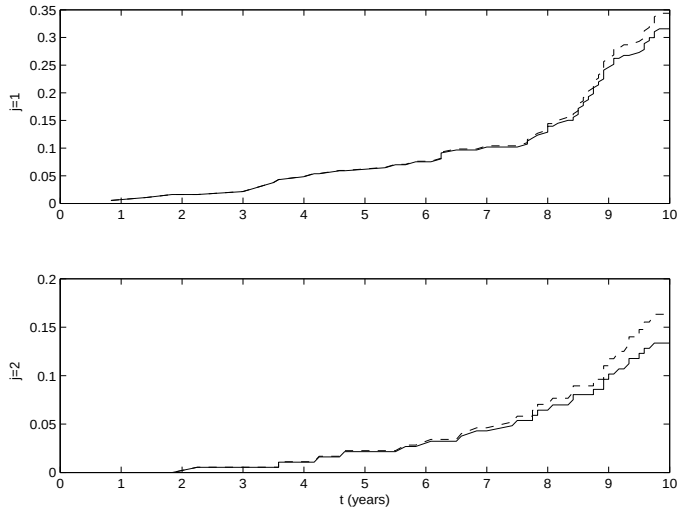

Figure 2: Estimates of marginal distribution functions F_j (dashed) and incidence functions F_j^* (full), for event 1 (above) and event 2 (below)

4.1 Analysis in Cox's model

Both risks to leave the job depend on a number of other attributes, on properties of the employee as well as on the situation of the company. Let us consider, in our illustrative example, just four "subjective" covariates entering the model. Let $\boldsymbol{X}_i(t)$ be $(X_{1i}(t), X_{2i}(t), X_{3i}, X_{4i})$, where X_{1i} is the age of the i-th employee at time t, X_{2i} is the length of previous employment in the company, up to t. Further, X_{3i} is the job category. There were 5 categories, from the highest (1) to the lowest (5) qualification (see the data), $X_{4i} = 1$ for male, $= 2$ for female employee. X_3 and X_4 are discrete and constant in the time, while both X_1 and X_2 are time dependent. Let for $k = 1, 2$ $X_{ki} = X_{ki}(T_i)$ be the value of k-th covariate for i-th person at the moment T_i. Then the value of this covariate at time $t < T_i$ is $X_{ki}(t) = \max\{0; X_{ki} - (T_i - t)\}$. Further, let us define $I_i(t) = 1$ for $t \in (\max\{0; T_i - X_{2i}\}, T_i]$ – the period during which the person i has been with the company, $I_i(t) = 0$ otherwise.

Let us analyze the data in the framework of Cox's model. Table 1 displays the partial likelihood-based estimates of parameters β_k, $k = 1, \ldots, 4$, together with the P-values P_k of the test of hypothesis $\beta_k = 0$. The results from Table 1 suggest the following conclusions: For the first event ($j = 1$) the hypothesis that the risk does not depend significantly on a covariate is rejected for components X_2 (just on 10% level of the test) and X_4. People with longer record in the company were more likely to leave, women had higher

risk of forced leave. When the second event ($j = 2$) is considered, the hypothesis of negligible dependence is rejected for X_2, too, and for X_3 (voluntary departures of people with higher qualification were more frequent).

	$j = 1$		$j = 2$	
k	β_k	P_k	β_k	P_k
1	0.0075	0.567	-0.0259	0.220
2	0.0712	0.098	0.1428	0.044
3	0.1124	0.388	-0.4773	0.053
4	0.7474	0.005	-0.2057	0.622

Table 1: Estimated Cox's model parameters and P-values of test $\beta_k = 0$

5 Conclusion

The problem of competing risks has been studied and the difference between expected marginal occurrence and real incidence of events has been analyzed. Further, the evaluation of incidence in the framework of a regression model for events intensities, and also the response of one event incidence to the change of intensity of competing event were discussed. The method has been illustrated on an example with unemployment data, where two competing risks of voluntary and forced change of job were considered. naturally, the approach can find application in other areas dealing with count data and occurrence of competing events.

Acknowledgements

The research is supported by the grant of GA CR No 402/10/0956.

References

[1] Andersen, P.K., Borgan, O., Gill, R.D., and Keiding N.: *Statistical Models Based on Counting Processes.* Springer, New York, 1993.

[2] Gourieroux, Ch.: Un modele de recherche optimale d'employ. In: *Analyse Statistique des Durees de Vie, Les Journs d'Etude en Statistique,* Univ. d'Aix–Marseille II, 1988.

[3] Han, A. and Hausman J.A.: Flexible parametric estimation of duration and competing risk models, *J. of Applied Econometrics* **5** (1990), 1–28.

[4] Heckman, J.J. and Singer, B.: Economic duration analysis, *Journal of Econometrics* **24** (1984), 63–132.

[5] Kalbeisch, J.D. and Prentice, R.L.: *The Statistical Analysis of Failure Time Data.* Wiley, New York, 2002.

[6] Lin, D.Y.: Non-parametric inference for cumulative incidence fumctions in competing risks studies, *Statistics in Medicine* **16** (1997), 901–910.

[7] Scheike, T.H. and Zhang, M.: *Flexible competing risks regression modelling and goodness-of-fit.* Research Report 08/03, Dept. of Biostatistics, Univ. of Copenhagen, 2008.

[8] Tuma, N.B. and Hannan, M.T.: *Social Dynamics: Models and Methods.* Academic, Orlando, Fla, 1984.

[9] Winkelmann, R.: *Econometric Analysis of Count Data.* Springer, New York, 2003.