Incompatibility Between Hazard and Logistic Regression in Modeling Competing Risks[*]

Gebrenegus Ghilagaber[†]

Abstract

This paper examines the relationship between outcomes from separate hazard modeling of two competing risks and those based on hazard-modeling of the overall risk combined with logistic-regression on the conditional probability of the risk of interest. The formulation presented in this paper generalizes for continuous time models and allows for non-categorical covariates. Our analytical results show that the two approaches are incompatible because they address different issues. Such a straightforward explanation of the incompatibility between the two approaches is expected to prompt investigators to focus on identifying situations of when the issue addressed by each approach may be of substantive interest.

1 Introduction

Many investigations involve analysis of data representing time to occurrence of a certain event like birth, death, marriage, divorce, or migration. In most practical situations such events arise due to one of several causes as is the case in death due to cancer or heart-failure. Although multiple causes can be considered as single to examine the overall rate of occurrence, public policies

[*] An earlier version of this paper has been presented in Session **C-14 (Survival Analysis III - Competing Risks)** at the XXI^{st} *International Biometric Conference*, Freiburg - Germany, July 21 - 26, 2002. The author is gratefull to B. Carstensen and P. Royston for valuable comments on that occassion.

[†] Address for Correspondence: Department of Statistics, Stockholm University, SE-106 91 Stockholm - Sweden. E-mail. Gebre@stat.su.se

can benefit from knowledge of the relative rate associated with each unique cause and the effects of explanatory variables on occurrence rates due to specific causes.

The most common approach to analyze time-to-event data with multiple causes of failure has been to begin by defining what are known as cause-specific hazard functions. In modeling the rate associated with each cause, study units are treated as censored at the occurrence of events due to any of the other causes. Han & Hausman (1990), Liefbroer (1991), Narendranathan & Stewart (1991; 1993) are some examples in the applied literature that follow this procedure.

Some investigators, on the other hand, argue that the suitability of this approach depends on the circumstances. Hachen (1988), for instance, provides a substantive argument and holds that modeling cause-specific hazard rates is appropriate only when the determinants of an event are likely to vary across the causes.

Further, it may be noted that the influence that covariates have on the overall hazard work via their impacts on the individual hazard rates and the fact that these rates add up to the overall hazard. Thus, separate analysis of the competing risks may give considerable insights into the determinants of the individual hazard rates, without providing any insight into the impacts of these determinants on the total hazard rate. We may get to know a lot about what influences death-risks due to various causes without necessarily learning very much about the effects on total death-risks. Even if the cause-specific hazard functions satisfy certain model specifications like the proportional hazards the overall risk may not have a similar representation with its relative risks.

For reasons of this nature, a hazards model for the overall risk is often specified in its own right and is analyzed along with a logistic specification of the conditional probability of the risk of interest (Sörensen, 1983 and Ghilagaber, 1998 are few examples). Such approach has also been recommended as an alternative when modeling cause-specific hazard rates is inappropriate (Hachen, 1988). A further support to this approach has been the argument that an understanding of the timing, calendaring and patterning of the phenomenon under investigation may be studied by analyzing the overall hazard rate; and how this rate divides itself among its constituent parts may be reflected by analyzing the conditional probabilities. Based on this, some, mainly empirical, studies have even maintained that the two approaches in effect lead to the same conclusions with regard to covariate-effects on cause-specific hazard rates (Tuma & Hannan, 1984; Hachen, 1988; Ghilagaber, 1998).

In the present work we attempt to have a closer look at the relationship between the two approaches. Our analytical investigation reveals that the two approaches may lead to numerically equivalent outcomes but are inconsistent because they address different issues. In the case of two competing risks, for instance, the cause-specific hazard framework seeks to answer if a covariate effect on each cause-specific hazards is significant while the logistic regression case asks if the difference between this covariate's effect on the two risks is significant. We therefore suggest that efforts be diverted to identifying real-life situations of when the questions addressed by each approach may be of substantive interest.

The hazard-rate framework is introduced in the next section. In Section 3, we outline the problem at hand and display simple formulations to demonstrate the logical inconsistency between hazard- and logistic-regression in modeling competing risks. The last section summarizes the contents of the paper.

2 Hazard Models

2.1 The hazard-rate framework

A central concept in the analysis of data representing times to occurrence of some specified event is the hazard function. Such a function, commonly denoted by $\lambda(t)$, is defined as the instantaneous rate at which the event occurs at a specific point t of a (non-negative) time variable T :

$$\lambda(t) = \lim_{\Delta t \longrightarrow 0} \frac{P\left[t < T \leqslant t + \Delta t | T \geq t\right]}{\Delta t} \tag{1}$$

Hazard rates can vary not only over time, but also among individuals within a population. Thus, one objective in the analysis of time-to-event data is to draw inferences about the influence of covariates on the hazard function.

In his influential paper, Cox (1972) proposed a model where a covariate z affects the hazard function in a multiplicative manner according to

$$\lambda(t, z) = q(t) \exp(z\beta) \tag{2}$$

where $q(t)$ is an unspecified base-line function of time and β is an unknown parameter representing the effect of the covariate z. The factor $exp(z\beta)$ describes the hazard of failure for an individual with z **relative** to that of a standard (with $z = 0$). Details on estimation and tests on β may be found in Cox (1975). While there are other parametric specifications of the base-line hazard function these are not of immediate relevance for the present work.

2.2 Cause-specific hazard rates and the conditional probability of a specific cause

In the presence of R ($R \geq 2$) causes of failure indexed by r ($r = 1, ..., R$), let the (random) variable C represent the cause of failure. A cause-specific hazard function may then be written as

$$\lambda_r(t) = \lim_{\Delta t \longrightarrow 0} \frac{P\left[t < T \leqslant t + \Delta t, C = r | T \geq t\right]}{\Delta t}, \qquad r = 1, 2, ..., R \tag{3}$$

so that $\lambda_r(t)$ is the instantaneous rate of occurrence of the event due to cause r at time t and in the presence of $R - 1$ other causes. Assuming each event type to be unique we get the overall hazard rate as

$$\lambda(t) = \sum_{r=1}^{R} \lambda_r(t) \tag{4}$$

Dividing (4) by (3) leads to

$$\pi_r(t) = \frac{\lambda_r(t)}{\lambda(t)}, \qquad r = 1, 2, ..., R \tag{5}$$

which is an expression for the conditional probability that the event is due to cause r, given that one of the events has occurred.

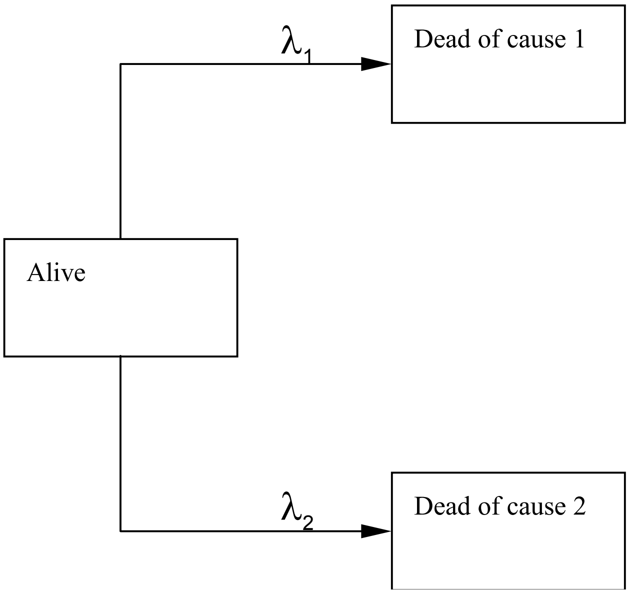

Figure 1:

3 Incompatibility between hazard- and logistic-regression

3.1 The Problem

Assume that one wants to analyze a two-cause competing risk model, illustrated in Figure 1. Denote the two cause-specific hazard rates by $\lambda_1(t)$ and $\lambda_2(t)$, respectively; define the overall hazard rate

$$\lambda(t) = \lambda_1(t) + \lambda_2(t), \tag{6}$$

and introduce the conditional probability due to cause 1,

$$\pi_1(t) = \frac{\lambda_1(t)}{\lambda(t)}. \tag{7}$$

It is worth noting that all four functions ($\lambda_1, \lambda_2, \lambda$, and π) will be known as soon as two of them are specified. In principle, therefore, the analysis of the behavior indicated in Fig. 1 can be based on a study of any two of these four functions. This is attractive and is perhaps the reason why many investigators feel at ease to substitute one pair of functions by another pair and proceed believing that the resulting insights are invariant across the pairs. However, one choice of two functions for analysis may lead to outcome that is inconsistent with that from another similar-looking pair of functions.

Below, we present two such choices of function pairs and demonstrate inconsistencies between them. The choices to be presented are (1) hazards regression analysis of $\lambda(t)$ followed by logistic regression of $\pi_1(t)$, and (2) separate hazards regression analysis of $\lambda_1(t)$ and $\lambda_2(t)$. Our discussion will be limited to the case where the cause-specific hazard rates have the usual Cox-type proportional-hazards format as in equation (2) with a single fixed covariate, namely,

$$\lambda_k(t,z) = q_k(t) \exp\left\{\beta_k z\right\}, \qquad\qquad k = 1,2 \tag{8}$$

Extensions to more general formats and/or to models with more than one (fixed or time-varying) covariates are straightforward.

3.2 Hazards regression of $\lambda(t)$ combined with logistic regression of $\pi_1(t)$

Suppose we specify a hazards model for $\lambda(t)$:

$$\lambda(t,z) = q(t) \exp\left\{\beta z\right\} \tag{9}$$

and a logistic regression model for $\pi_1(t)$:

$$\log it\left[\pi_1(t,z)\right] = \ln\left\{\frac{\pi_1(t)}{1-\pi_1(t)}\right\} = \alpha_t + \beta^* z \tag{10}$$

which, upon solving for $\pi_1(t)$ yields

$$\pi_1(t,z) = \frac{1}{1+\exp\left[-\left(\alpha_t + \beta^* z\right)\right]} \tag{11}$$

Let us now see how the specification of the cause-specific hazards would look like. Given $\lambda(t,z)$ and $\pi_1(t,z)$ one can obtain $\lambda_1(t,z)$ as

$$\begin{aligned}
\lambda_1(t,z) &= \lambda(t)\pi_1(t) \\
&= \frac{q(t)\exp\{\beta z\}}{1+\exp\{-(\alpha_t+\beta^* z)\}} \\
&= \frac{q(t)\exp\{\beta z\}}{1+\exp\{-(\alpha_t+\beta^* z)\}}\frac{\exp\{\alpha_t+\beta^* z\}}{\exp\{\alpha_t+\beta^* z\}} \\
&= \frac{q(t)\exp\{\alpha_t+(\beta^*+\beta)z\}}{1+\exp\{\alpha_t+\beta^* z\}} \\
&= \frac{q(t)\exp(\alpha_t)\exp[(\beta^*+\beta)z]}{1+\exp\{\alpha_t+\beta^* z\}}
\end{aligned}$$

or

$$\lambda_1(t,z) = \frac{q_1(t)\exp(\beta_1)z}{1+\exp\{\alpha_t+\beta^* z\}} \tag{12}$$

with $q_1(t) = q(t)\exp(\alpha_t)$ and $\beta_1 = \beta^* + \beta$, equalities which may hold in the case of, for instance, piecewise constant base-line specification.

While the numerator in (12) looks like the usual proportional hazards specification in (8), the formula in (12) is marred by the second term in the denominator, which makes the coefficients much harder to interpret than the relative risks we are used to. It is thus clear that (8) and (12) are inconsistent and that analysis of $\lambda(t)$ and $\pi_1(t)$ does not lead to the same insights of the phenomena being studied as those gained from a hazards analysis of $\lambda_1(t)$ and, by extension, $\lambda_2(t)$.

3.3 Separate hazards regression analysis of $\lambda_1(t)$ & $\lambda_2(t)$

We now present a more straightforward explanation of the incompatibility between the two approaches.

Let the hazard rate due to cause 1 be given by the proportional hazards model

$$\lambda_1(t,z) = q_1(t)\exp(\beta_1 z) \tag{13}$$

and the hazard rate due to cause 2 be given by the proportional hazards model

$$\lambda_2(t,z) = q_2(t)\exp(\beta_2 z) \tag{14}$$

where, as defined earlier, $q_1(t)$ and $q_2(t)$ denote the base-line functions for the two cause-specific hazard rates.

Invoking equations (4) and (5), the conditional probability due to cause 1, $\pi_1(t)$, may be obtained as

$$\begin{aligned}
\pi_1(t,z) &= \frac{\lambda_1(t,z)}{\lambda(t,z)} \\
&= \frac{\lambda_1(t,z)}{\lambda_1(t,z)+\lambda_2(t,z)} \\
&= \frac{q_1(t)\exp(\beta_1 z)}{q_1(t)\exp(\beta_1 z)+q_2(t)\exp(\beta_2 z)}
\end{aligned}$$

$$= \frac{1}{1+\left\{\frac{q_2(t)}{q_1(t)}\exp\left[\left(\beta_2-\beta_1\right)z\right]\right\}} \tag{15}$$

Comparing the expression for the conditional probability given in (15) with its previous expression given in (11) we note that these two expressions coincide if

$$\exp(-\alpha_t) = \frac{q_2(t)}{q_1(t)} \tag{16}$$

and

$$\beta^* = \beta_2 - \beta_1. \tag{17}$$

It may be shown that the equality in (16) is approximately satisfied if the base-line hazard functions are piecewise constant. The equality in (17), on the other hand, says that the two approaches ask totally different questions. The cause-specific hazard approach asks about the size (β_k) of covariate effects for each risk, while the logistic-regression approach asks about the *difference* (β^*) in the size of a covariate's effects across the two risks.

Thus, empirically subtracting estimates related to one risk from those related to the other risk should give the estimates obtained from logistic regression. This is true up to a rounding error, and has been documented in, among others, Ghilagaber (1998).

However, concluding, on the basis of such empirical results, that inferences drawn from the two approaches are identical may be misleading because, as we noted above, the questions addressed by each approach are different. Tests of significance on the estimates β_1 and β_2 (in the cause-specific hazard framework) ask if each of these quantities differs significantly from zero, while tests of significance for $\beta^* = \beta_1 - \beta_2$ (the logistic regression case) ask if their *difference* is significantly different from zero.

The difference of a covariate's effects on the two hazard rates may be small or large depending on the direction of the effects. For instance, if a covariate's effect is very large but does not vary across the type-specific hazards, this would imply that the difference is close to zero which, in turn, implies that the covariate has no effect on the conditional probability. On the other hand, even insignificantly small effects on the hazard rates that are in opposite directions may lead to a larger difference which, in turn, leads to a large effect on the conditional probability.

In other words, covariates which significantly affect one or both of the hazard rates may not have any effect on the conditional probability. Conversely, covariates with insignificant effects on one or both of the hazard rates may turn out to have significant effect on the conditional probability.

These simple and straightforward reflections show that a hazards analysis of the overall risk $\lambda(t)$ combined with a logistic regression of the odds $\frac{\pi_1}{1-\pi_1}$ can give insights whose nature is totally different from those gained via separate hazards analysis of the individual risks $\lambda_1(t)$ and $\lambda_2(t)$.

4 Summary

We have examined two approaches to modeling multiple causes of failure. One of them employs a standard competing risk hazard regression model, while the other uses hazards regression of the overall risk combined with logistic regression to model the type of failure *conditional* on the fact that an event due to one of the causes in question occurred.

It is shown that the outcomes obtained from the two approaches are numerically equivalent, up to some mild reparametrizations. However, and more importantly, it is also shown that the two approaches ask different questions - one about the size of covariate effects for each cause of failure, the other about the difference in the size of a covariate effect across the two causes.

As differences of a covariate's effects on the two hazard rates may be small or large depending on the direction of the effects it is possible to get contradictory results from the two approaches. In other words, the type of insight one gains through analysis of the overall risk and the odds of a specific cause on the one hand, and those gained from analyzing the individual hazard rates on the other, can be totally different and often contradictory.

As a final remark we hope the present results contribute towards realizing the distinctions between these two approaches and prompt investigators to focus on the issue of when the question addressed by each approach might be of substantive interest.

References

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

[2] Cox, D. R. (1975), Partial Likelihood. *Biometrika* **62**, 269-276

[3] Ghilagaber, G. (1998), Analysis of survival data with multiple causes of failure: a comparison of hazard- and logistic-regression models with application in demography. *Quality & Quantity* **32**, 297-324.

[4] Hachen, D. S. (1988), The competing risks model: a method for analyzing processes with multiple types of events. *Sociological Methods & Research* **17**, 21-54.

[5] Han, A., and J. A. Hausman (1990), Flexible parametric estimation of duration and competing risks models. *Journal of Applied Econometrics* **5**, 1-28.

[6] Liefbroer, A. C. (1991), The choice between married and unmarried first union by young adults: a competing risk analysis. *European Journal of Population* **7**, 273-298.

[7] Narendranathan, W., and M. B. Stewart (1991), Simple methods for testing for the proportionality of cause-specific hazards in competing risks models. *Oxford Bulletin of Economics and Statistics* **53**, 331-340.

[8] Narendranathan, W., and M. B. Stewart (1993), Modeling the probability of leaving unemployment: competing risks models with flexible base-line hazards. *Applied Statistics* **42**, 63-83.

[9] Sörensen, A. (1983), Women's employment patterns after marriage. *Journal of Marriage and the Family* **45**, 311-321.

[10] Tuma, N. B., & M. T. Hannan (1984), *Social Dynamics: methods and models.* Orlando: Academic Press