Developments in Statistics
Andrej Mrvar and Anuška Ferligoj (Editors)
Metodološki zvezki, 17, Ljubljana: FDV, 2002

Covariate Effects in Periodic Hazard Rate Models

Ulrich Pötter[1] and Kai Kopperschmidt[2]

Abstract

Labour market participation, consumer behaviour, and many other phenomena exhibit strong periodic patterns that result from cyclic behaviour, constraints on the timing of events, or seasonal variation. While these periodicities can generally be neglected when dealing with small data sets or coarsely grouped event times, they pose challenges to the analysis of large data sets with precise recordings. It seems natural to require that statistical models used in the analysis of such data sets reproduce any underlying periodicities. In particular, the conditional hazard rate given covariates should be periodic for all possible values of the covariates. We show that this requirement severely restricts the class of covariate effects models.

We define periodicities by points of zero crossings of the derivative of the hazard rate. We then develop the concepts of hazard envelope and essential extrema. These allow the construction of classes of covariate effect models with time varying coefficients that respect the underlying periodic structure.

1 Introduction

Labour market participation, consumer behaviour, and many other phenomena exhibit strong periodic patterns that result from cyclic behaviour, constraints on the timing of events, or seasonal variation. These phenomena become apparent when large data sets with precise recordings of the timing of events become available. Figure 1 exhibits the hazard rate of the inter-purchase time of an 1-litre ice-cream package. The estimate is based on data provided by the German Homescan Panel of A.C. Nielsen. The data contain information on the day of purchases for some 8.400 households over a period of three years.

The (discrete) daily hazard rate oscillates with maxima at 7, 14, 21 days and so on. It is plausible to assume that the reason for this behaviour is the weekly

[1] University of Bochum, Bochum, Germany; Ulrich.Poetter@ruhr-uni-bochum.de.

[2] A.C. Nielsen, Frankfurt, Germany; Kai.Kopperschmidt@Germany.ACNielsen.com.

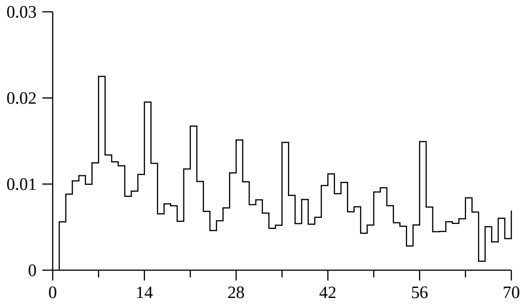

Figure 1: Hazard rate of inter-purchase times (days) of ice-cream packages.

purchase schedule of most households. This argument is supported by the fact that these patterns occur across sociodemographic subgroups, e.g., regardless of whether there are children present in the household or not, cf. Figure 2.

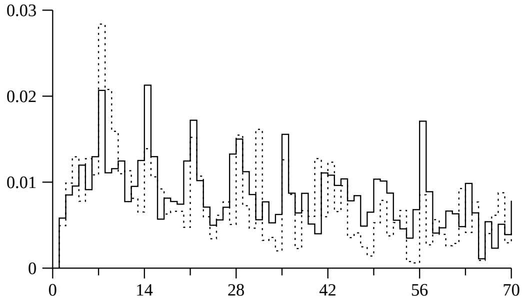

Figure 2: Conditional hazard rate of inter-purchase times (days) of ice-cream packages. Households without children: solid line, Households with children: dotted line.

As a second example, Figure 3 presents the estimated hazard rate of job durations in Germany for the years 1975 to 1990. The estimate is based on a subsample of records of the social security administration (see Bender et al., 1996) covering some 400,000 job spells. The hazard rate shows large annual peaks, somewhat smaller quarterly peaks and also monthly peaks. However, the number of job durations away from these peaks is still considerable. Again, the findings are similar across subgroups (e.g., men and women). An obvious reason for this pattern is that institutional and juridical regulations in general restrict ending a job to the end of a quarter or to the end of a (calendar) year. Of course, such regulations ought to be the same for all socio-demographic subgroups. In fact, the impact of the regulations is extremely strong: Figure 4 depicts the total number of job exits per calendar day for the period 1990–1999. The numbers are based on the complete data of the social security administration. The number of job exits not coincident with a month's end is generally below 100.

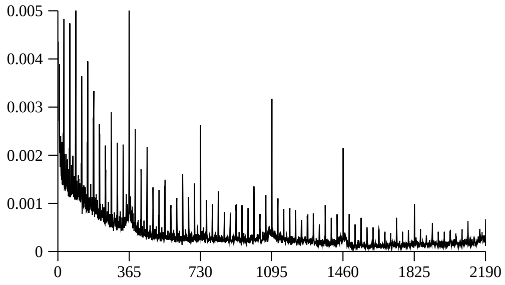

Figure 3: Hazard rate of job durations (days) in Germany 1975–1990.

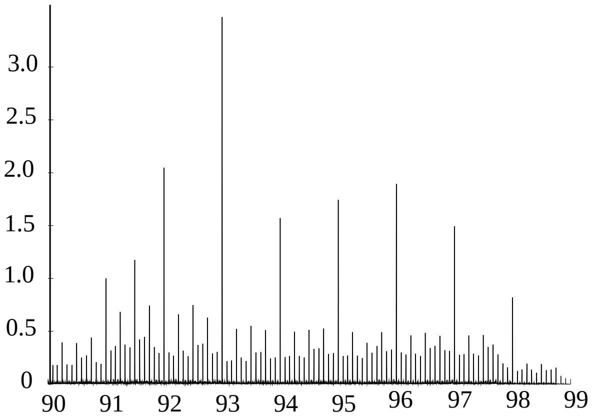

Figure 4: Total number (in million) of job exits by day in Germany 1990–1999. Number for 31.12.1992 truncated.

2 Marginal and conditional hazard rates

In both examples, *strong* external influences cause both the conditional and the marginal hazard rates to oscillate. These influences are *common* in the sense that local maxima and minima appear at the same times for the marginal as well as for the conditional hazard rates. Regression models should account for this periodic behaviour: The conditional hazard rates implied by the models should exhibit the same periodicities for all values of the covariates. Moreover, the periodicities of the conditional hazard rates should be the same as those of the implied marginal hazard rates. Surprisingly, however, *no* regression model strictly satisfies these requirements.

Let $T > 0$ and X be random variables on a common probability space representing duration and covariate information. Denote by $\lambda(t) = f(t)/(1 - F(t-))$ and $\lambda(t|x) = f(t|x)/(1 - F(t-|x))$ the marginal and conditional hazard rates. Here, the conditioning is on the events $\{X = x\}$, $f(t)$ and $f(t|x)$ are the marginal and conditional densities, and $F(t)$ and $F(t|x)$ are the marginal and conditional cumulative distribution functions.

For simplicity, we assume that $\lambda(t|x)$ is twice continuously differentiable with respect to t. If $t_1, t_2, \ldots$ are the locations of minima and maxima of $\lambda(t|x)$, then $\dot{\lambda}(t_i|x) = 0$, $i = 1, 2, \ldots$, where $\dot{\lambda}(t|x)$ is the derivative of the conditional hazard rate with respect to t. A possible though rather strict formulation of the above requirements becomes: There is a sequence $0 < t_1 < t_2 \ldots$ such that

$$\dot{\lambda}(t_i) = 0 = \dot{\lambda}(t_i|x), \text{ for } i = 1, 2, \ldots \text{ and for all } x. \tag{2.1}$$

To see that in fact no non-trivial model can satisfy this condition, we need to consider the relation between marginal and conditional hazard rates and their derivatives. The marginal hazard rate is given by a time-dependent "convex combination" of the conditional hazard rates:

$$\lambda(t) = E\big(\lambda(t|X)\big|T \geq t\big) \tag{2.2}$$

where the expectation is with respect to the distribution of X conditional on the event $\{T \geq t\}$.

Differentiating this relation leads to

$$\dot{\lambda}(t) = E\big(\dot{\lambda}(t|X)\big|T \geq t\big) + \big[\lambda(t)^2 - E\big(\lambda(t|X)^2\big|T \geq t\big)\big] \tag{2.3}$$

When the derivatives $\dot{\lambda}(t_i|x)$ vanish for all x, the first term becomes 0. By Jensen's inequality, the second term is negative unless $\lambda(t|x)$ is constant in x. Thus, if all derivatives of conditional hazard rates vanish at a point t_i, then the derivative of the marginal hazard rate has to be negative.

To illustrate, consider two subgroups distinguished by the covariate $X \in \{0, 1\}$ and assume

$$\begin{aligned} P_0 &:= P(X = 0) = \frac{1}{2} = P(X = 1) =: P_1, \\ \lambda_0(t) &:= \lambda(t|X = 0) = \frac{5}{4} + \sin t, \\ \lambda_1(t) &:= \lambda(t|X = 1) = 2 \cdot \lambda_0(t), \end{aligned}$$

Then

$$\lambda(t) = \frac{\lambda_0(t) \cdot \exp\big(-\Lambda_0(t)\big) + \lambda_1(t) \cdot \exp\big(-\Lambda_1(t)\big)}{\exp\big(-\Lambda_0(t)\big) + \exp\big(-\Lambda_1(t)\big)},$$

where

$$\Lambda_i(t) := \int_0^t \lambda_i(s)\,ds, \ i = 0, 1,$$

are the cumulative hazard rates of the subgroups.

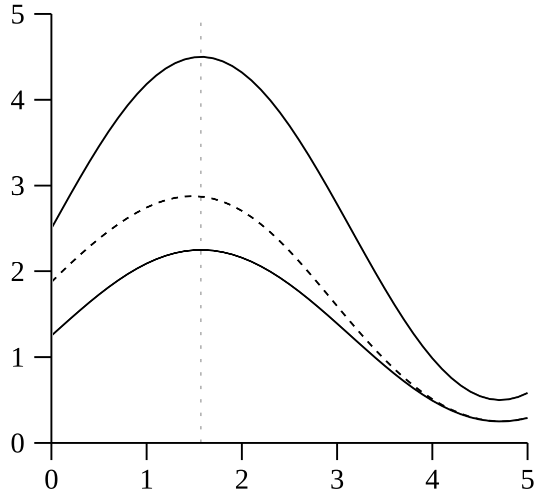

Figure 5: Due to the time-dependence of (2.2), the marginal hazard rate (dashed line) does not have the same extrema as the conditional hazard rates (solid lines).

Figure 5 displays the different extrema of the marginal hazard rate $\lambda(t)$ and the conditional hazard rates $\lambda_0(t)$ and $\lambda_1(t)$. Although $\lambda(t) \in [\lambda_0(t), \lambda_1(t)]$ holds due to (2.2), the time-dependence of the convex combination causes the derivative $\dot{\lambda}$ to vanish at different times than $\dot{\lambda}_0$ and $\dot{\lambda}_1$.

3 Hazard envelopes

On the other hand, if the conditional hazard rates $\lambda_i(t)$, $i = 0, 1$, oscillate *strongly* and *commonly* in the sense that (w.l.o.g.)

$$\begin{aligned} \ddot{\lambda}_0(t_{2i}) &< 0, & \ddot{\lambda}_1(t_{2i}) &< 0, \\ \ddot{\lambda}_0(t_{2i-1}) &> 0, & \ddot{\lambda}_1(t_{2i-1}) &> 0, \\ \lambda_0(t_{2i}) &> \lambda_1(t_{2i-1}), & \lambda_0(t_{2i}) &> \lambda_1(t_{2i+1}), \ \forall i \geq 1, \end{aligned}$$

hold, then the marginal hazard rate $\lambda(t)$, being a pointwise convex combination of $\lambda_0(t)$ and $\lambda_1(t)$ as well as a differentiable function of t, must have a local maximum in each interval (t_{2i-1}, t_{2i+1}), $i \geq 1$, and a local minimum in each interval interval (t_{2i}, t_{2i+2}), $i \geq 1$. Qualitatively speaking, $\lambda(t)$ oscillates as well, cf. Figure 6.

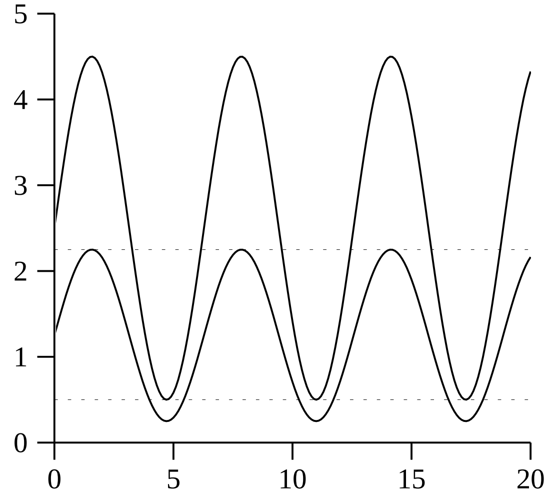

Figure 6: The marginal hazard rate necessarily oscillates between the two dashed lines.

Thus the marginal hazard rate will oscillate in a similar way as the conditional hazard rates if the latter have common minima and maxima and if they oscillate strongly enough. To bound the behaviour of the marginal hazard rate, we introduce the concept of the *hazard envelope* $(\underline{\lambda}(t), \overline{\lambda}(t))$:

$$\underline{\lambda}(t) := \inf_x\{\lambda(t|x)\}\ ,\ \overline{\lambda}(t) := \sup_x\{\lambda(t|x)\} \tag{3.1}$$

Note that in general neither $\underline{\lambda}(t)$ nor $\overline{\lambda}(t)$ need to correspond to any member of the family of conditional hazard rates. Nevertheless, it follows from (2.2) that the marginal hazard rate at a given t is bounded by the extreme points of the conditional hazard rates. Thus

$$\underline{\lambda}(t) \le \lambda(t) \le \overline{\lambda}(t)\ \forall t \tag{3.2}$$

Suppose next that maxima of the hazard envelope occur at even numbered times t_{2i} while minima occur at odd numbered times t_{2i-1}. Suppose further that the conditional hazard rates have common minima and maxima at t_{2i} and $t_{2i=1}$, respectively. We say that the conditional hazard rates have an *essential* maximum at t_{2i} if

$$\overline{\lambda}(t_{2i-1}) < \underline{\lambda}(t_{2i}) > \overline{\lambda}(t_{2i+1}) \tag{3.3}$$

We say that the conditional hazard rates have an *essential* minimum at t_{2i+1} if

$$\underline{\lambda}(t_{2i}) > \overline{\lambda}(t_{2i+1}) < \underline{\lambda}(t_{2i+2}) \tag{3.4}$$

An essential maximum implies at least one maximum of the marginal hazard rate in the interval (t_{2i-1}, t_{2i+1}), while an essential minimum implies at least one minimum of the marginal hazard rate in the interval (t_{2i}, t_{2i+2}).

4 Consequences for model choice

We are now in the position to formulate more reasonable requirements for regression models in situations with strong periodicities: Suppose there is a sequence of time points $t_{2i}, i \geq 1$ at which maxima of the hazards are to occur. Think of the weekly maxima in the hazard rate of inter-purchase times or the quarterly maxima in the hazard rate of job durations. Then one might want to restrict attention to models of covariate effects that, firstly, admit maxima of the conditional hazard rates at the t_{2i} for all values of the covariates, that, secondly, admit the existence of common minima of the conditional hazard rates at some sequence of times t_{2i-1}, and that thirdly, admit essential maxima at all the t_{2i} even in the presence of non-trivial covariate effects.

Consider the class of proportional hazards models with

$$\lambda(t|x) = \lambda_0(t)\psi(x\beta),\ \psi(x\beta) > 0 \tag{4.1}$$

For this class the envelope hazard coincides with certain conditional hazard rates if the support of the distribution of the covariates is compact. The situation is then very similar to that depicted in Figures 5 and 6. The first and second requirements are easily met. In fact, they simply depend on the choice of an appropriate baseline hazard rate $\lambda_0(t)$. Whether or not a local maximum is essential will depend both on the extend of covariate effects and the amplitude of the baseline hazard rate. Thus proportional hazard rate models are certainly feasible candidate models.

But what happens if one wants, for some good reason, use non-proportional hazards models? Consider the accelerated failure time model. This model posits a scaling effect of covariates: Suppose that T_x is a random variable representing duration conditional on the covariate value x. Suppose further that there is a random variable T_0 on the same probability space as T_x such that

$$T_x = T_0/\psi(x\beta),\ \psi(x\beta) > 0 \tag{4.2}$$

and such that the T_0 have identical distributions for all values of the covariates. The conditional hazard rates are then of the form:

$$\lambda(t|x) = \psi(x\beta) \cdot \lambda(\psi(x\beta) \cdot t),\ \psi(x\beta) > 0 \tag{4.3}$$

But in such a model, maxima and minima of conditional hazard rates for different values of the covariates cannot coincide, except in the trivial case of no covariate effect, $\psi(x\beta) \equiv 1$.

Does this rule out the use of accelerated failure time models and many other non-proportional hazards models? Not if one is prepared to allow the effects of covariates to change with time. But how does one allow for time-varying covariate effects without destroying the defining features of the accelerated failure time model? After all, if one allows for general time dependent effects $\beta(t)$ and plugs this into

(4.3), then the "scaling the time axis" property is destroyed, while there is no obvious way of plugging a time indexed $\beta(t)$ into (4.2).

There is, however, a quite natural way to define changing covariate effects that respects the scaling interpretation of accelerated failure time models. One has to change the global " change of scale" interpretation of covariate effects into a local property at a point in time. That can be done by using derivatives. Starting with the "scale change" interpretations in terms of random variables as in (4.2), one can consider the derivative of the baseline duration with respect to duration with covariate value x. That derivative should be influenced, at a point in time, by the covariate effect at that same point in time. One might thus write

$$\left.\frac{\partial t_0}{\partial t_x}\right|_{t_x=u} = \psi(x\beta(u)) \tag{4.4}$$

But then

$$t_0 = \int_0^{t_x} \psi(x\beta(u))\, du =: \Psi(t_x; \bar{\beta})$$

where $\bar{\beta}$ contains the covariate information and the changes in covariate effects. Therefore

$$T_x = \Psi^{-1}(T_0; \bar{\beta})$$

The hazard rate corresponding to this model of covariate effects is

$$\lambda(t|x) = \psi(x\beta(t)) \cdot \lambda_0(\Psi(t; \bar{\beta})) \tag{4.5}$$

Note that this differs from the naive idea to plug in some $\beta(t)$ into (4.3) while it preserves the interpretation of the effects of covariates as a (local) scaling of the time axis.[3]

With this definition of varying covariate effects it is now easy to exhibit versions of the accelerated failure time model that do respect the requirements formulated at the beginning of this section. If we choose

$$\begin{array}{rcl} \Psi(t_{2i}; \bar{\beta}) & = & t_{2i} \\ \dot{\Psi}(t; \bar{\beta}) & > & 0 \\ \ddot{\Psi}(t_{2i}; \bar{\beta}) & = & 0 \text{ and } \dot{\lambda}_0(t_{2i}) = 0 \end{array}$$

then

$$\dot{\lambda}(t_{2i}|x) = \dot{\psi}(x\beta(t_{2i})) \cdot \lambda_0(\Psi(t_{2i}; \bar{\beta})) + \psi(x\beta(t_{2i}))^2 \cdot \dot{\lambda}_0(\Psi(t_{2i}; \bar{\beta})) = 0$$

[3] A version of this model for time-dependent covariates has been proposed by Cox and Oakes (1984, p. 67). Robins and Tsiatis (1992) developed an estimator for this model.

With this choice of $\Psi(t), \beta(t)$, and $\lambda_0(t)$, all conditional hazard rates will have the same points of maxima as the baseline $\lambda_0(t)$. As in the case of proportional hazards models, whether the maxima are essential depends on the amplitude of $\lambda_0(t)$ and on the covariate effects process $\psi(x\beta(t))$. But since the envelope hazard will in general not coincide with any of the conditional hazard rates, one needs to compute the envelope hazard explicitly.

References

[1] Bender, S., Hilzendegen, J., Rohwer, G., and Rudolph, H. (1996): *Die IAB–Beschäftigtenstichprobe 1975–1990. Beiträge zur Arbeitsmarkt und Berufsforschung 197.* Institut für Arbeitsmarkt und Berufsforschung, Nürnberg.

[2] Cox, D.R. and Oakes, D. (1984): *Analysis of Survival Data.* Chapman and Hall, London.

[3] Robins, J. and Tsiatis, A.A. (1992): Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika*, **79**, 311–319.