THE COX REGRESSION MODEL FOR CLAIMS DATA IN NON-LIFE INSURANCE

BY

NIELS KEIDING

Department of Biostatistics, University of Copenhagen

AND

CHRISTIAN ANDERSEN and PETER FLEDELIUS

ATP PensionService A/S, Hilleroed, Denmark

ABSTRACT

The Cox regression model is a standard tool in survival analysis for studying the dependence of a hazard rate on covariates (parametrically) and time (nonparametrically). This paper is a case study intended to indicate possible applications to non-life insurance, particularly occurrence of claims and rating

We studied individuals from one Danish county holding policies in auto, property and household insurance simultaneously at some point during the four year period 1988-1991 in one company The hazard of occurrence of claims of each type was studied as function of calendar time, time since the last claim of each type, age of policy holder, urbanization and detailed type of insurance Particular emphasis was given to the technical advantages and disadvantages (particularly the complicated censoring patterns) of considering the nonparametrically underlying time as either calendar time or time since last claim. In the former case the theory is settled, but the results are somewhat complicated The latter choice leads to several issues still under active methodological development. We develop a goodness-of-fit criterion which shows the lack of fit of some models, for which the practical conclusions might otherwise have been useful.

1. INTRODUCTION

Individual rating in non-life insurance may be based on exogenous variables (age of policy holder, urbanization) but in auto insurance various schemes for dynamical individual rating based on endogenous information (previous claim career) are well established. A possible further development of such procedures would be to base rating on endogenous variables for more than

ASTIN BULLETIN, Vol 28, No 1, 1998, pp 95-118

one type of non-life insurance. This would – as all such schemes – require an extensive knowledge base, and to focus ideas we studied the example of household, property and auto insurance. The joint development in time of the occurrences of claims of these three types is conveniently phrased in terms of the theory of event history analysis which has developed rapidly during the last decade, cf. Blossfeld et al (1989) and Blossfeld and Rohwer (1995) for good surveys with social science applications and Andersen et al (1993) for a general treatise with many practical examples, primarily from biostatistics.

In this report we indicate some initial possibilities as well as difficulties in carrying out such a programme Restricting attention to claim occurrence (i.e disregarding claim size) we want to capture the occurrence in time of claims as function of fixed exogenous covariates (age of policy holder, urbanization) and several time variables: calendar time and times since recent claims of each type. There is an active current literature on choice of time scales in statistical models for repeated events, cf. Lawless and Thiagarajah (1996), Lawless (1998) and Oakes (1998).

Our main tool will be versions of the Cox (1972a) regression model for event history data, see Andersen et al. (1993, Chapter VII). In this "semiparametric" model, one time variable is selected as "underlying" and modelled "nonparametrically" while other time variables as well as fixed exogenous covariates are modelled parametrically See Prentice et al. (1981) for an early exposition of alternative time scales in Cox models for repeated events and Oakes (1998) for an excellent concise survey. The Cox model is introduced in Section 3 and two alternative choices of underlying time variable are considered in Section 4 (calendar time) and 5 (time since last claim). Whereas calendar time as underlying time variable leads to a relatively standard application of Cox regression methodology, it will turn out to be rather less standard to consider time since last claim. A brief discussion is contained in Section 6.

The methodology is illustrated on data from a Danish insurance company, introduced in Section 2.

2. DATA

The present case study is based on data from a Danish insurance company. Between 1 January 1988 and 31 December 1991, 15,718 persons across Denmark at least once simultaneously held household, property and auto policies in this company. We study the 1,904 persons from the county of Fyn, in which Odense is by far the largest city. For each person and each type of policy is known

- the start and the end of the policy if within 1988-1991. If there were several policies of the same type within 1988-1991, only the latest was kept in the routine records on which we work.
- age (but not sex) of policy holder
- urbanization

- for household: coverage (amount)
- for auto: coverage
- date and size of claims.

In this study we focused attention on claims that led to payments This means that we removed claims of size 0. We made no other use of claim size.

3. The Cox regression model for event history analysis

For each type $h = 1, 2, 3$ (household, property, auto) and policy holder i the intensity of having a claim at time t is denoted $\lambda_{hi}(t)$. Here t can be calendar time (cf. Section 4) or time since the last claim of a similar type (cf. Section 5), with a special definition necessary if there has not (yet) been such a claim. A third possibility would be that t was time since taking out the policy. We explain later why we do not consider the latter possibility relevant here.

The Cox regression model now postulates that

$$\lambda_{hi}(t) = \alpha_{0h}(t) \exp\left[\beta_h' Z_{hi}(t)\right] Y_{hi}(t)$$

where $\alpha_{0h}(t)$ is a freely varying so-called underlying intensity function common to all policy holders i but specific to insurance type h The indicator $Y_{hi}(t)$ is 1 if policy holder i is at risk to make a claim of type h at time t, 0 otherwise The covariate process $Z_{hi}(t)$ indicates fixed exogenous as well as time-dependent endogenous covariates. The fixed covariates considered are year of birth of policy holder and urbanization of residence, which in practice equals 1 for city (Odense) and 0 for rural (rest of Fyn). The time-dependent covariates indicate duration since last claim of each type (which can and will be parameterized in various ways). Finally the vector β_h contains the regression coefficients on the covariates $Z_{hi}(t)$

Statistical inference in the Cox regression model is primarily based on maximum partial likelihood, which in the generality necessary for this application was surveyed by Andersen et al. (1993, Chapter VII) in the framework of *counting processes*. The regression coefficients β_h are estimated by maximizing the partial likelihood

$$L(\beta_h) = \prod_j \frac{\exp(\beta_h' Z_{hi(j)}(T_{hj}))}{\sum_i Y_{hi}(T_{hj})=1 \exp(\beta_h' Z_{hi}(T_{hj}))}$$

where $T_{h1} < T_{h2} < \ldots$ are the times of claims of type h, policyholder $i(j)$ claiming at time T_{hj}. Large sample results are available to justify the application of the inverse Hessian of the log partial likelihood as approximate covariance matrix for $\hat{\beta}_h$. Because of the time-varying covariates the necessary algorithms are rather elaborate, although we were able to perform all computations on a medium-sized PC using StatUnit (Tjur, 1993). The computations may also be performed in standard packages such as BMDP, SAS or S-plus, or via the Poisson regression approach of Lindsey (1995).

For the *underlying intensity* $\alpha_{0h}(t)$ it is well-established that a natural estimator of the integrated intensity

$$A_{0h}(t) = \int_0^t \alpha_{0h}(u)du$$

is given by the step function (the "Breslow" estimator)

$$\hat{A}_{0h}(t) = \sum_{T_{hj} \le t} \frac{1}{\sum_i Y_{hi}(T_{hj})=1 \exp(\hat{\beta}_h' Z_{hi}(T_{hj}))}$$

where $T_{h1} < T_{h2} < \ldots$ are the times of claims of type h and $\hat{\beta}_h$ the maximum partial likelihood estimator of β_h.

Unfortunately $\hat{A}_{0h}(t)$ is less than optimal in communicating important features of the structure of $\alpha_{0h}(t)$; it is often desirable to be able to plot an estimate of α_{0h} itself We shall here use *kernel smoothing* (which in the context of estimating the intensity in the multiplicative intensity model for counting processes was incidentally pioneered by the actuary Ramlau-Hansen (1983)). This estimates $\alpha_{0h}(t)$ by

$$\hat{\alpha}_{0h}(t) = \sum_{j\ t-b<T_{hj}<t+b} K\left(\frac{t - T_{hj}}{b}\right) \Delta \hat{A}_{0h}(T_{hj})$$

where b is the *bandwidth*, K a *kernel function*, here restricted to $[-1, 1]$ and $\Delta \hat{A}_{0h}(T_{hj}) = \hat{A}_{0h}(T_{hj}) - \hat{A}_{0h}(T_{hj-1})$, $T_{h0} = 0$. We choose here the Epanechnikov kernel $K(x) = 0.75(1 - x^2)$. For more documentation, see again Andersen et al. (1993, pp. 483 and 507-509).

Despite its considerable flexibility, the Cox regression model is not assumption-free, the most important assumptions being that of *proportional hazards* and that of *log-linearity* of effect of regressors. There is a well-developed battery of goodness-of-fit procedures available, cf. Andersen et al. (1993, Section VII.3), and several of these methods have been used in the present case-study (never indicating deviation from model assumptions). However, space prevents us from documenting these here.

4 Cox regression of claim intensity
Calendar time as underlying time variable

Our first choice of underlying time scale is *calendar time*, which is always observable and whose association with variations in claim intensity may form an interesting object of study. Technically, the counting process approach elaborated by Andersen et al. (1993, Section III.4) easily allows for entry and exit of policies from observation (the "Aalen filter") in this situation.

However, an important purpose of this study was to ascertain the observability and possible extent of the association of claim intensity to the duration(s) since earlier claim(s), and it is less obvious how to account for these. Because of the relatively limited period of observation (4 years) it was necessary to make several pragmatic choices. First, the dependence on earlier claims was operationalized as dependence on duration since latest claim, and this was achieved by defining the indicator covariates

[1-90]: There has been a claim less than 90 days ago.
[91-180]: The latest claim was between 91 and 180 days ago.
[181-270]: The latest claim was between 181 and 270 days ago.
[271-360] The latest claim was between 271 and 360 days ago.
[> 360] There has been no claim during the past 360 days.

Since the database contains no information on claims before 1988, these covariates would not all be observable early in the period. We therefore decided to use 1988 as run-in year, only for collecting information on earlier claims.

A further problem was the many instances where a new policy was taken out within 1988-1991 In case no claims happened, the above covariates would remain unobservable for 360 days, which forced us to add the covariate

[no inf.]. policy (of this type) was taken out less than 360 days ago and during that time there were no claims.

4.1. Household claims in calendar time

For household claims the relevant covariates were: year of birth of policyholder (categorized in three groups separated by 1 January 1938 and 1 January 1948), urbanization (Odense vs rest of Fyn) and duration since last claim of each type as described above All groups of covariates were of statistical significance and the estimated model had regression coefficients as given in Table 4.1.

It is seen that compared to the "no information" situation when no claim has happened after a recently taken out policy, knowledge of a recent *household* claim during the recent 0-9 months increases the risk of a new household claim by a factor ranging from $e^{0.562} = 1.8$ to $e^{0.808} = 2,2$, i.e , a factor of about 2. On the other hand knowledge of claim-free career of one year decreases the risk by the (statistically insignificant) factor of 0.9.

Past *property* claims have effects according to a similar pattern, although the effects are smaller, except for very recent property claims ($e^{0.629} = 1.9$), some of which may be caused by the same events that caused the household claim. Unfortunately the database cannot identify such cases, which would in principle violate the proportional hazards assumption of the Cox regression model

TABLE 4.1

REGRESSION COEFFICIENTS IN REDUCED COX MODEL FOR HOUSEHOLD CLAIMS

Covariate	*Estimate*	*Standard error*	*P*
Household[no inf.]	0	–	–
Household[1-90]	0.562	0.277	0.043
Household[91-180]	0.725	0.275	0.008
Household[181-270]	0.808	0.275	0.003
Household[271-360]	0.206	0.303	0.496
Household[> 360]	−0.105	0.243	0.665
Property[no inf.]	0	–	–
Property[1-90]	0.629	0.197	0.001
Property[91-180]	0.178	0.219	0.416
Property[181-270]	0.107	0.225	0.663
Property[271-360]	0.287	0.225	0.202
Property[> 360]	−0.132	0.161	0.413
Auto[no inf.]	0	–	–
Auto[1-90]	0.224	0.209	0.284
Auto[91-180]	0.301	0.208	0.148
Auto[181-270]	0.258	0.217	0.234
Auto[271-360]	−0.187	0.260	0.473
Auto[> 360]	−0.144	0.148	0.330
Born[> 1947]	0	–	–
Born[1938-1947]	0.015	0.086	0.860
Born[< 1938]	−0.406	0.100	0.000
Rural	0	–	–
City	0.381	0.076	0.000

Past *auto* claims show overall significance, although the effect of each period is small, generally in a similar pattern as for the other types of insurance.

The *age* pattern has a decreased intensity for older policy-holders (intensity factor $e^{-0.406} = 0.7$) while the two younger groups are very similar; finally urbanization generates the expected gradient with an increased risk in the city ($e^{0.381} = 1.5$).

The underlying intensity is estimated as described in Section 3, using 3 different bandwidths for illustration, see Fig. 4.1. It is not easy to conclude much from the somewhat irregular pattern except perhaps a slight general decrease. The boundary effects at the start and the end of the studied period are statistical artefacts deriving from the kernel estimation approach.

It may be noticed from Table 4.1 and the following tables that several of the patterns of dependence on time since last claim might be simplified. As an example in Table 4.1, the regression coefficients Auto[1-90], Auto[91-180]

and Auto[181-270] look rather similar, as do Auto[271-360] and Auto[> 360]. However, there is no obviously consistent pattern across types of claims and types of risk indicators, so we have refrained from conducting what would in any case be post-hoc attempts at statistical identification of such patterns.

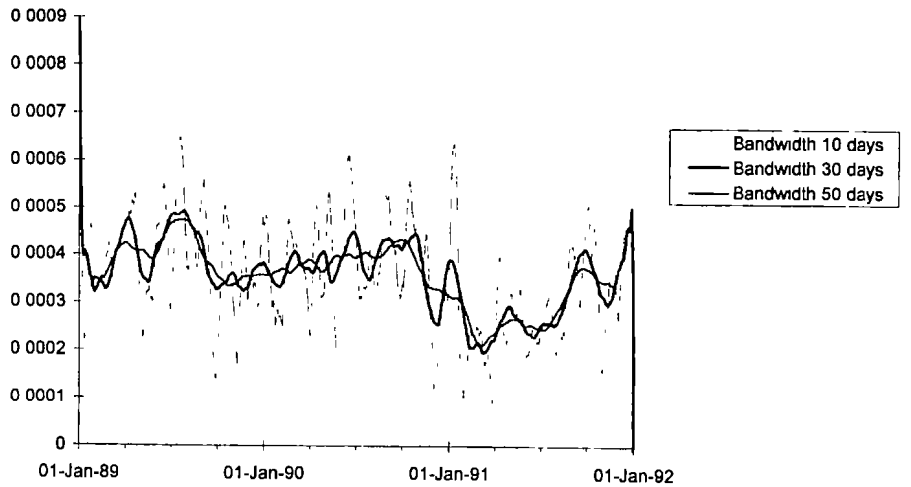

FIGURE 4 1 Kernel smoothed underlying intensities for household claims

4.2. Property claims in calendar time

For property insurance there is a series of optional additional coverage possibilities, which are all included as specific indicator covariates fire, glass, insects, wash basins, pipe, rot

The estimates of the reduced model are given in Table 4.2. Note that urbanization is statistically insignificant and that there is an unusual age pattern, the middle-aged having a somewhat lower risk than the young and the old. In the interpretation of the age effect it is however particularly important to keep in mind the specially selected population each person must have had all three types of policies simultaneously at some point during 1988-1991, this restricts consideration to better situated people.

Of the optional additional coverage, only glass and pipe coverage are retained as risk variables, both clearly increasing the risk. That fire does not appear is related to the fact that almost all policies chose that option. For duration since last claim the general pattern is similar to the earlier one, although one must notice that there is never a significantly lower risk than that of [no inf.], which (as we shall discuss more fully below) will limit the practical applicability of the results.

TABLE 4 2

REGRESSION COEFFICIENTS IN REDUCED COX MODEL FOR PROPERTY CLAIMS

Covariate	*Estimate*	*Standard error*	*P*
Household[no inf]	0	-	-
Household[1-90]	0 485	0 229	0 034
Household[91-180]	0 302	0 240	0 208
Household[181-270]	0 345	0 240	0 151
Household[271-360]	0 032	0 260	0 902
Household[> 360]	−0 080	0 192	0 676
Property[no inf]	0	-	
Property[1-90]	0 524	0 206	0 011
Property[91-180]	0 334	0 217	0 124
Property[181-270]	0 206	0 224	0 357
Property[271-360]	0 281	0 224	0 210
Property[> 360]	−0 180	0 181	0 320
Auto[no inf]	0	-	-
Auto[1-90]	0 501	0 184	0 006
Auto[91-180]	0 262	0 202	0 194
Auto[181-270]	0 182	0 210	0 387
Auto[271-360]	0 267	0 211	0 205
Auto[> 360]	0 026	0 141	0 851
Born[> 1947]	0		-
Born[1938-1947]	−0 196	0 079	0 013
Born[< 1938]	−0 061	0 079	0 438
Glass	0 411	0 140	0 003
Pipe	0 185	0 072	0 010

The underlying intensity is estimated in Fig 4 2 and shows a dramatic peak in early 1990, apparently traceable to extreme weather conditions

4.3. Auto claims in calendar time

In addition to the standard covariates, auto claims are expected to depend on whether or not there is auto comprehensive coverage and whether or not a certain "free claim" allowance is included in the policy.

The estimates of the reduced model are given in Table 4 3, where it is immediately noticed that, perhaps contrary to expectation, auto comprehensive coverage does not increase risk of claim for this population of insures. Note also the age pattern, generally unusual for auto insurance with

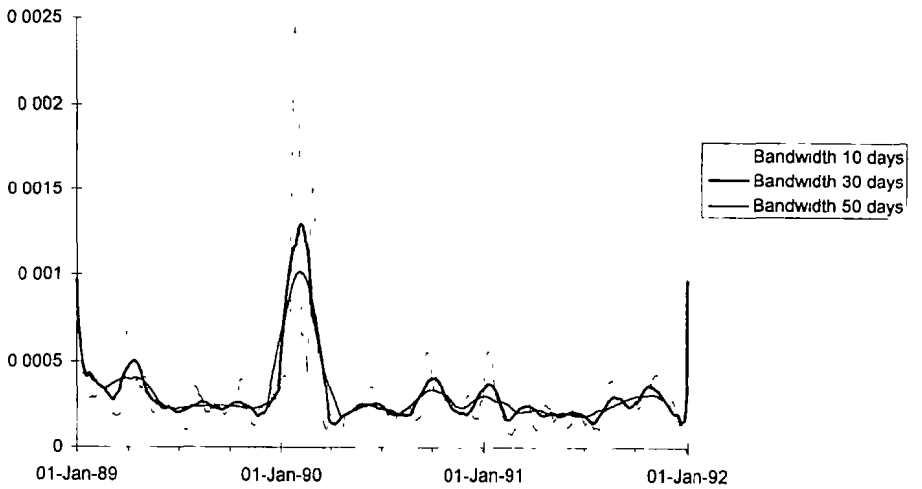

FIGURE 4 2 Kernel smoothed underlying intensity for property claims

maximal risk among the middle-aged policy-holders. (Note that there are no data to account for size of household, and note once again the specially selected population.)

TABLE 4 3

REGRESSION COEFFICIENTS IN REDUCED COX MODEL FOR AUTO CLAIMS

Covariate	*Estimate*	*Standard error*	*P*
Household[no inf]	0	-	-
Household[1-90]	0 388	0 245	0 114
Household[91-180]	0 303	0 251	0 226
Household[181-270]	0 304	0 252	0 227
Household[271-360]	0 493	0 244	0 043
Household[> 360]	0 001	0 193	0 995
Auto[no inf]	0	-	-
Auto[1-90]	0 730	0 259	0 005
Auto[91-180]	0 862	0 257	0 001
Auto[181-270]	0 738	0 264	0 005
Auto[271-360]	0 618	0 273	0 024
Auto[> 360]	0 294	0 231	0 203
Born[> 1947]	0	-	-
Born[1938-1947]	0 100	0 079	0 209
Born[< 1938]	−0 140	0 087	0 106
Free claim	1 048	0 083	0 000

The patterns regarding duration since last claim show no overall effect of recent property claims and some effect (increase) on risk of recent household claim. As expected, recent auto claims considerably increase the risk of a further auto claim, as does the "free claim" option (no penalty in premium scale after a claim)

The underlying intensity (Fig. 4.3) indicates some seasonality with peaks in the winter and the summer, however this pattern is rather irregular.

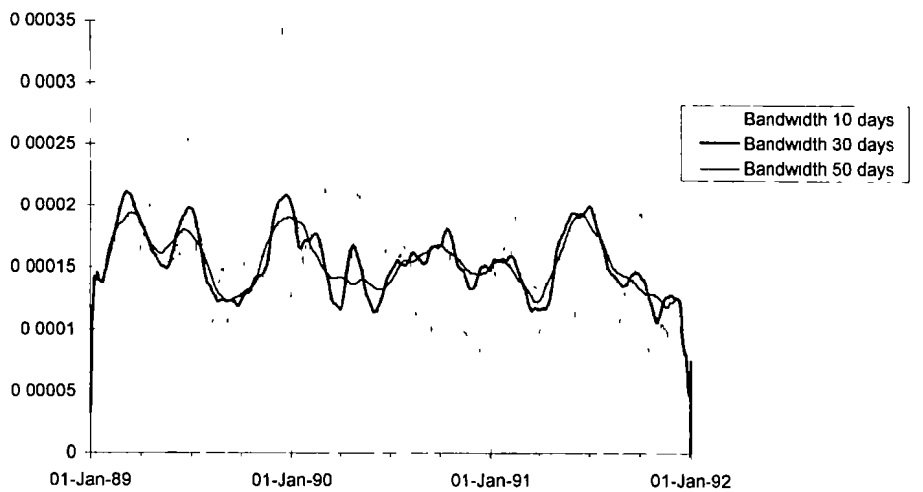

FIGURE 4 3 Kernel smoothed underlying intensity for auto claims

4.4. Preliminary conclusions: calendar time as underlying variable

Two problems are common to all analyses so far. First, the unstable nature of the population of policies during the relatively short observation window of four years make the desired allowance for time since earlier claims difficult to achieve in practice. The general reference category of [no inf.], meaning that a policy of the relevant type was taken out less than a year ago and there have not yet been claims to that policy, in all cases carries a very low risk for new claims of the type under study This relative low-risk behaviour of new policyholders is obviously difficult to integrate into a reward system for faithful customers In this connection it must be emphasized that the routine nature of our database (which may well be typical of such databases) did not allow the distinction between genuinely new policies and "bureaucratical" renewals initiated by the company or the policyholder in order to update conditions.

Secondly, some of our concrete results point to the rather special selection procedure underlying the present database: all policyholders were required to have held all three types simultaneously at least once in 1988-1991 As an example, think of the rather biased selection of young policyholders!

5. Cox regression of claim intensity: Using duration rather than calendar time as basic time variable

In the discussion so far it has become obvious that we need to reason in several time variables: calendar time as well as duration(s) since recent claim(s). At least because of the possibility that there have not yet been any claims, we may also need the time since the policy was taken out. When using the Cox regression model such as introduced in Section 3

$$\lambda_{hi}(t) = \alpha_{0h}(t) \exp\left[\beta_h' Z_{hi}(t)\right] Y_{hi}(t)$$

one may choose one of these time scales as "basic" (= t) and handle the other(s) as (time-dependent) covariates $Z_{hi}(t)$. An important criterion for choosing between these possibilities is the additional flexibility in the description offered by the "nonparametric" underlying intensity $\alpha_{0h}(t)$. We actually saw in Section 4 that various indications regarding seasonal patterns appeared in the graphs of Figs. 4.1-3

Another criterion is ease of handling special observation plans. When calendar time is used, the exact time is always known for each policy-holder, in contrast to what is the case for duration since last claim. We discussed the latter problem at the beginning of Section 4, where we constructed time-dependent covariates to account for durations since earlier claims.

However, both prior expectation and our experiences so far point to the importance of time since last claim as decisive time variable, for which the maximal modelling flexibility offered by the nonparametric part of the Cox model would be useful. To discuss an adequate statistical analysis in this time-scale, consider first the simple situation without covariates, which is a renewal process.

5.1. Estimation of renewal processes observed in a fixed time window

Let $X_1, X_2,$ be independent random variables (durations) with distribution functions $F_1, F_2 = F_3 = . = F$, assumed to have finite expectations μ_1 and μ and density functions $f_1 = F_1'$ and $f = F'$. Let $S_n = X_1 + ... + X_n, n = 1, 2, ..$ and the stochastic process (a *renewal process*)

$$N_t = \sum_{n=1}^{\infty} I\{S_n \leq t\},$$

the number of durations since time 0. If $f_1 = (1-F)/\mu$ the process is *stationary*. Observing a renewal process in an interval $[t_1, t_2]$ amounts to observing the renewal times (claims) $T_j \in [t_1, t_2]$ or equivalently $(N_t - N_{t_1},\ t \in [t_1, t_2])$. Let T_j be the first renewal after t_1, i.e. $N_{T_j} = N_{t_1} + 1$. Then $T_j - t_1$ is called the *forward recurrence time*, and if the process is stationary, this has *density* function $(1-F)/\mu$.

Observing a renewal process in an observation window $[t_1, t_2]$ involves four different elementary observations:

1. Times x_i from one renewal to the next, contributing the density $f(x_i)$ to the likelihood.
2. Times from one renewal to t_2, right-censored observations of F, contributing factors of the form $1 - F(t_2 - T_j)$ to the likelihood.
3. Times from t_1 to the first renewal (forward recurrence times), contributing, in the stationary case, factors of the form $(1 - F(T_j - t_1))/\mu$ to the likelihood.
4. Knowledge that no renewal happened in $[t_1, t_2]$, being right-censored observations of the forward recurrence time, contributing in the stationary case a factor

$$\int_{t_2 - t_1}^{\infty} (1 - F(u))du/\mu.$$

In the stationary case the resulting maximum likelihood estimation problem is well understood. Vardi (1982) derived an algorithm (a special case of the EM-algorithm) in a discrete-time version of the problem, and Soon and Woodroofe (1996) gave an elaborate and very well-written discussion in continuous time. McClean and Devine (1995) conditioned on seeing at least one renewal in $[t_1, t_2]$, excluding observations of type 4 and restricting attention to observations of type 3 right-truncated at $t_2 - t_1$, i.e. with density

$$(1 - F(u - t_1))/\left(1 - \int_0^{t_2 - t_1} F(v)\, dv\right)$$

Again an EM-type algorithm is feasible.

In our situation we need to be able to generalize the estimation method from iid variables to the Cox regression model, and we would also prefer to avoid the stationarity condition required for inclusion of the (uncensored and censored) forward recurrence times of type 3 and 4.

This is possible by restricting attention to (uncensored and censored) times since a renewal, that is, observations of type 1 and 2. As discussed repeatedly by Gill (1980, 1983), see also Aalen and Husebye (1991) and Andersen et al. (1993, Example X.1.8), the likelihood based on observations of type 1 and 2 is identical to one based on independent uncensored and censored life times from the renewal distribution F. Therefore the standard

estimators (Kaplan-Meier, Nelson-Aalen) from survival analysis are applicable, and their usual large sample properties may be shown (albeit with new proofs) to hold.

The above analysis is sensitive to departures from the assumption of homogeneity between the iid replications of the renewal process. Restricting attention to time since first renewal will be biased (in the direction of short renewal times) if there is unaccounted heterogeneity, as will the re-use of second, third, ... renewals within the time window. As always, incorporation of observed covariates may reduce the unaccounted heterogeneity, but the question is whether this will suffice

5.2. Cox regression of duration since last claim

The Cox (1972a) proportional hazards regression model for survival analysis was implemented by Cox (1972b) in the so-called *modulated renewal processes*, for which the hazard of the renewal distribution is assumed to have a similar semiparametric decomposition. This model has received much less attention than the survival analysis model and its event history analysis generalization (Prentice et al., 1981, Andersen and Gill, 1982, Andersen et al , 1993, Chapter VII), although Kalbfleisch and Prentice (1980) and Oakes and Cui (1994) discussed estimation. Careful mathematical-statistical analysis was provided by Dabrowska et al. (1994) and Dabrowska (1995), who showed that if the covariates depend on no other time variables than the backward recurrence times, then the 'usual' asymptotic results of the Cox partial (or profile) likelihood may be proved.

In the present case we have the additional complication of observing through a fixed (calendar) time window. Inclusion of likelihood factors of types 3 and 4 is then intractable, but if the model were true (in particular, if the observed covariates sufficiently account for individual heterogeneity), valid inference may be drawn from the reduced likelihood based on time since first claim (factors of types 1 and 2)

Finally, we want to incorporate time-dependent covariates not depending on the backward recurrence time only (for example, in the analysis of household claims we want to incorporate times since the last property or auto claim) and the analysis is then no longer covered by Dabrowska's asymptotic results.

As pointed out at the end of the last section, if there is unaccounted heterogeneity the expected bias by restricting attention to time since first renewal will be in the direction of short renewal times, and this will be even worse if times since second, third etc renewal times are also included. We build a goodness-of-fit criterion on this intuition, as follows.

5.3. A goodness-of-fit criterion for the Cox modulated renewal process observed through a fixed time window

We assume that the occurrence of claims of type h for policy holder i at duration t since last claim of that type is governed by a Cox regression model with intensity

$$\lambda_{hi}(t) = \alpha_{0h}(t) \exp\left[\beta_h' Z_{hi}(t)\right] Y_{hi}(t)$$

with interpretation as before. For this model Dabrowska (1995) proved asymptotic results for the 'usual' profile likelihood based inference, under the crucial assumption that the covariates $Z_{hi}(t)$ depend on time only through (the backwards recurrence time) t. (Obviously a full model will require an additional specification of occurrence of the first claim of type h after the policy is taken out.)

The claim occurrences are viewed through a fixed time window, but under the model valid inference may be based on the likelihood composed of the product of contributions from the distribution of time from first to second claim, second to third claim, and so on, the last being right-censored The expected deviation from the model is that time from claim $j = 1$ is longer than times from claims $j = 2, 3, \ldots$. We therefore extend the model to the Cox regression model

$$\lambda_{hij}(t) = \alpha_{0hj}(t) \exp\left[\beta_{hj}' Z_{hi}(t)\right] Y_{hi}(t).$$

In practice the regression coefficients β_{hj} and the underlying intensities $\alpha_{0hj}(t)$ after claim j are assumed identical for $j = 2, 3, \ldots$. A good evaluation of the fit of the Cox model can be based on first assessing identity of regression coefficients ($\beta_{h1} = \beta_{h2}$) and then, refitting in a so-called stratified Cox regression model with identical β_{hj} but freely varying $\alpha_{h0j}(t)$ over j, comparing the underlying intensities ($\alpha_{0h1}(t) = \alpha_{0h2}(t)$) after first and after later claims. For the first hypothesis a standard log partial likelihood ratio test may be performed, for the second we use graphical checks as surveyed by Andersen et al. (1993, Section VII. 3) Further development of this goodness-of-fit approach might follow the lines of Andersen et al (1983)

5.4. Household claims by duration since last such claim

The relevant covariates are the same as listed in Section 4 1 except of course that duration since last household claim is now described in the non-parametric part of the Cox model rather than by time-dependent covariates. Table 5.1 shows the final model after elimination of non-significant covariates It is noted that the result is rather simpler than that represented by Table 4 1 since in addition to time since last household claim, also time since last auto claim and age have disappeared.

TABLE 5 1

REGRESSION COEFFICIENTS IN REDUCED COX MODEL FOR HOUSEHOLD CLAIMS

Covariate	*Estimate*	*Standard error*	*P*
Property[no inf]	0	–	–
Property[1-90]	0 659	0 199	0 001
Property[91-180]	0 118	0 243	0 623
Property[181-270]	0 281	0 238	0 237
Property[271-360]	0 211	0 266	0 428
Property[> 360]	−0 140	0 165	0 394
Rural	0	–	–
City	0 251	0 103	0 015

The remaining covariates, time since last property claim and urbanization, have similar effects (particularly for the former) as before, and similar remarks apply.

The underlying intensity is estimated in Fig. 5.1 for the first three years (thereafter the random variation dominates) A clear decrease is seen: the longer the duration since the last household claim, the lower the intensity of a new one.

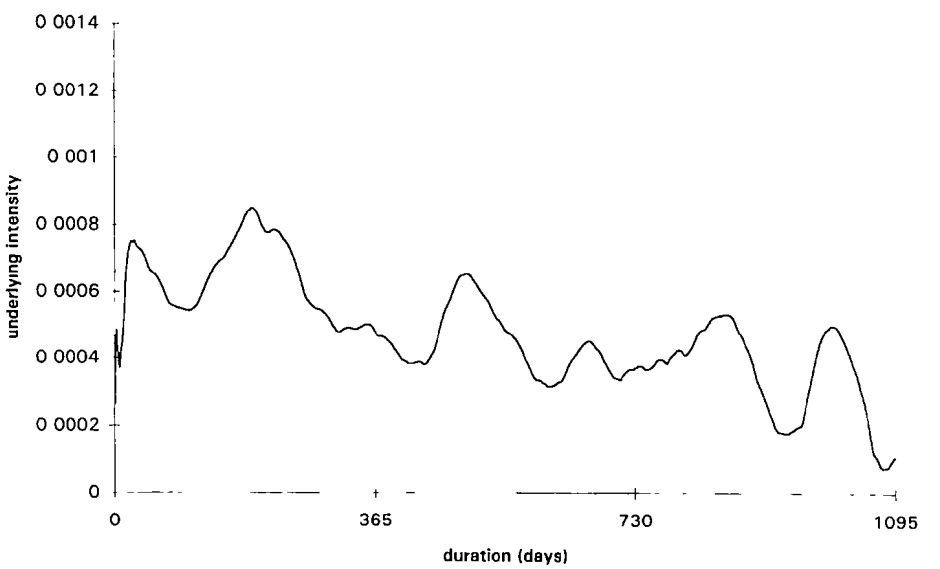

FIGURE 5 1 Kerned smoothed underlying intensity for household claims (bandwidth 50 days)

Fitting the stratified model specified in the previous section to the covariates of Table 5.1 leads to insignificantly different regression parameter estimates after first and after later claims ($\chi^2 = 8.87, f = 6$). To compare the estimates of underlying intensities $\alpha_{h0j}(t)$ between times since first claim and times since later claims, Fig. 5.2 shows integrated intensity estimates against time, whereas Fig 5.3 shows integrated intensity estimates against one another Both plots indicate good agreements so that the model, and hence the above interpretation, would seem acceptable

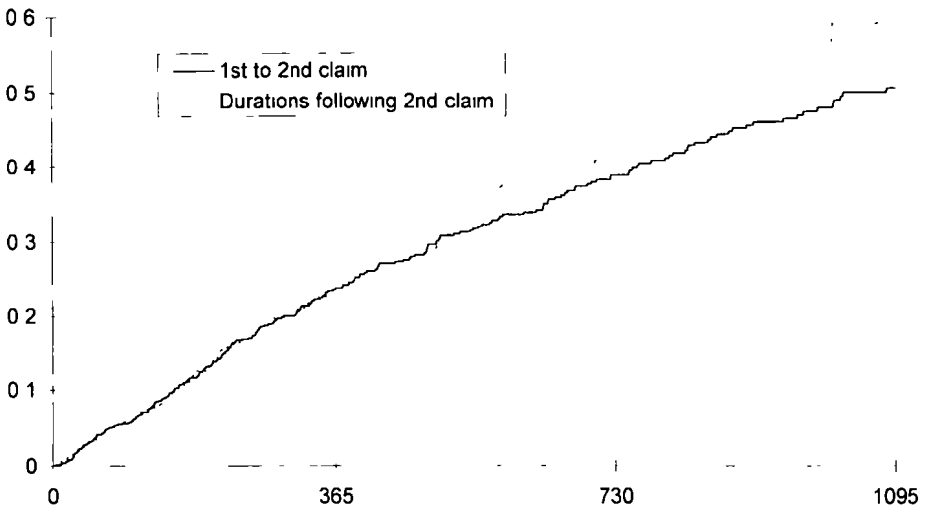

FIGURE 5 2 Estimated integrated underlying intensities for household claims

5.5. Property claims by duration since last such claim

In a similar fashion Table 5.2 shows the final model after elimination of non-significant covariates. (A likelihood ratio test for no effect of time since last household claim gave $P = .01$.)

TABLE 5 2

REGRESSION COEFFICIENTS IN REDUCED COX MODEL FOR PROPERTY CLAIMS

Covariate	*Estimate*	*Standard error*	*P*
Household[no inf]	0	-	-
Household[1-90]	0 198	0 208	0 340
Household[91-180]	0 321	0 213	0 131
Household[181-270]	0 110	0 236	0 634
Household[271-360]	−0 140	0 269	0 602
Household[> 360]	−0 253	0 157	0 106

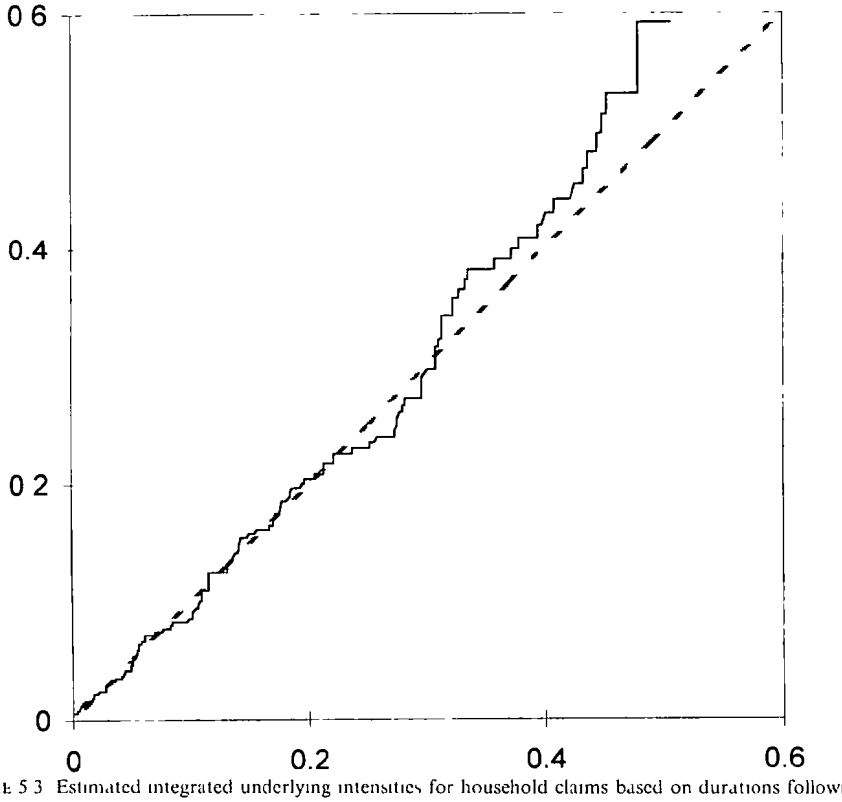

FIGURE 5 3 Estimated integrated underlying intensities for household claims based on durations following second claim plotted against those based on the (possibly right censored) duration from first to second claim

As for household claims, we get a much simpler description in the present time-scale, the only remaining covariate being time since last household claim The effect of this covariate in qualitatively similar to what it was in Table 4.2. The underlying intensity (Fig. 5 4) is decreasing. The gradient between best and worst customers (expressed by range of variation of regression coefficients) is smaller than for household claims, corresponding to common expectation.

For the goodness-of-fit test the identity of regression coefficients was again easily accepted ($\chi^2 = 0.73$, $f = 5$), but here the unfortunate bias in the direction of shorter durations after second and further claims is clearly visible from Figs. 5.5 and 5 6. The model must be judged as not fitting and the above conclusions cannot be sustained

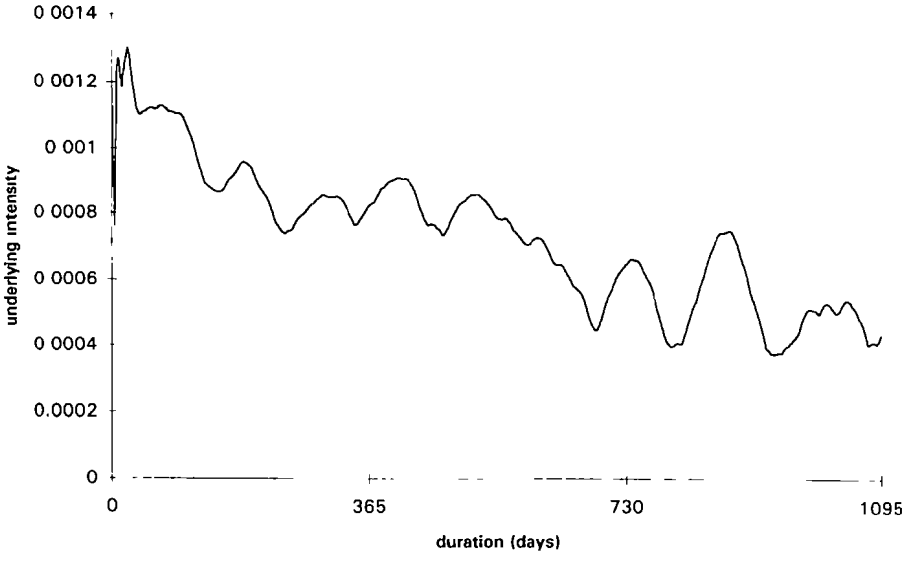

FIGURE 5 4 Kernel smoothed underlying intensity for property claims (bandwidth 50 days)

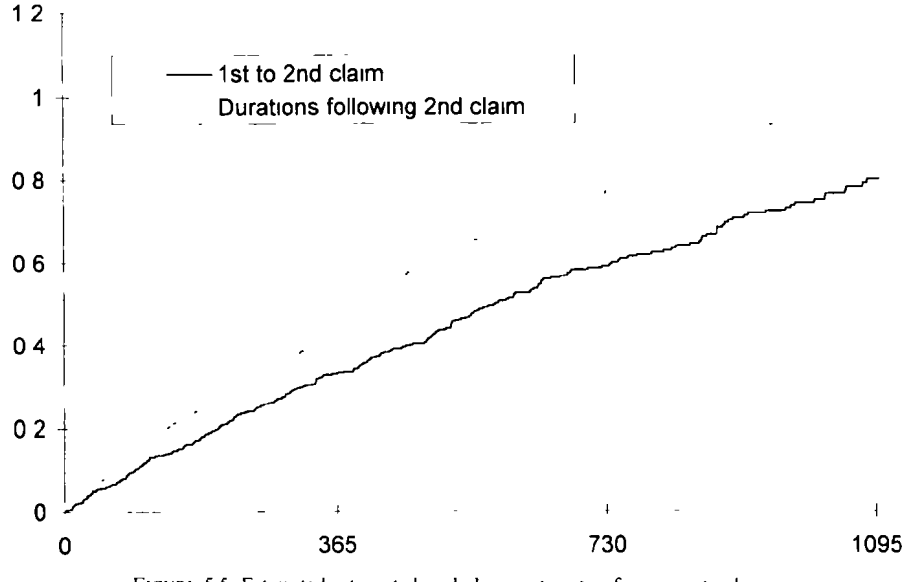

FIGURE 5 5 Estimated integrated underlying intensities for property claims

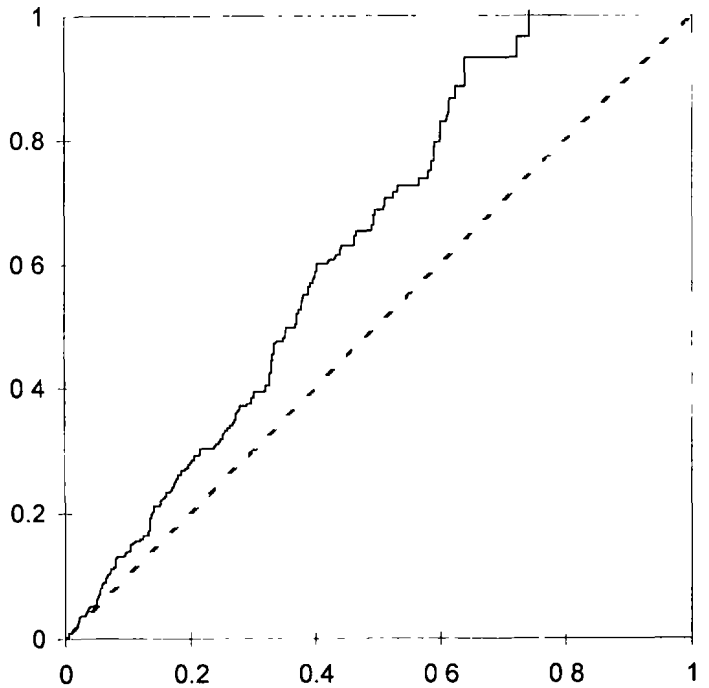

FIGURE 5 6 Estimated integrated underlying intensity for property claims based on durations following second claim plotted against those based on the (possibly right censored) duration from first to second claim

5.6. Auto claims by duration since last such claim

Finally, Table 5.3 documents the result of fitting the Cox regression model to time since last auto claim, using the covariates listed in Section 4, particularly Section 4.3, and eliminating statistically insignificant covariates

TABLE 5 3

Regression coefficients in reduced Cox model for auto claims

Covariate	*Estimate*	*Standard error*	*P*
Household[no inf]	0	–	–
Household[1-90]	0 304	0 205	0 139
Household[91-180]	0 295	0 218	0 175
Household[181-270]	0 053	0 240	0 826
Household[271-360]	0 032	0 251	0 897
Household[> 360]	−0 334	0 155	0 031
Auto comprehensive	−0 405	0 148	0 005
Free claim	0 320	0 121	0 008

Compared to Table 4.3, we necessarily have lost time since last auto claim, but furthermore, age is no longer significant while, most surprisingly, auto comprehensive coverage seems to *decrease* the risk of the next auto claim by a factor of $e^{-405} = 0.67$ We can only interpret the latter phenomenon with reference to a peculiar selection of policyholders who choose comprehensive coverage. The underlying intensity (Fig. 5 7) shows a clear decrease.

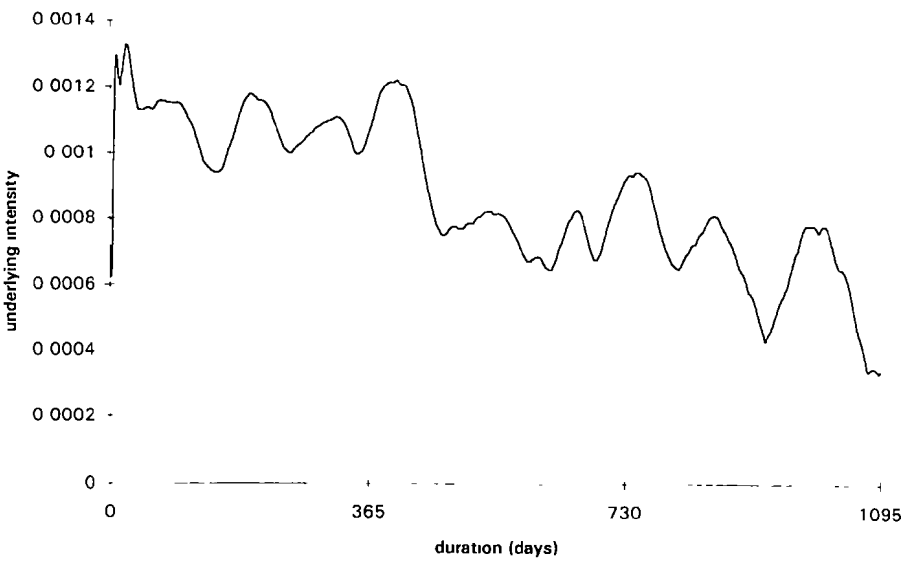

FIGURE 5 7 Kernel smoothed underlying intensity for auto claims (bandwidth 50 days)

The result of the goodness-of-fit test is very similar to that for household insurance above: regression coefficients are easily identical ($\chi^2 = 2\,26$, $f = 7$), but the expected bias is immediately obvious from Figs. 5.8 and 5.9 The model must thus be considered poorly fitting, and the results cannot be sustained.

5.7. Preliminary conclusions: duration as underlying time variable

The two basic difficulties mentioned in Section 4 4 were not removed by changing to duration as basic time variable. Furthermore, technical problems of estimation (as well as reluctance to postulate stationarity) forced us to omit all durations already running at the start of observation

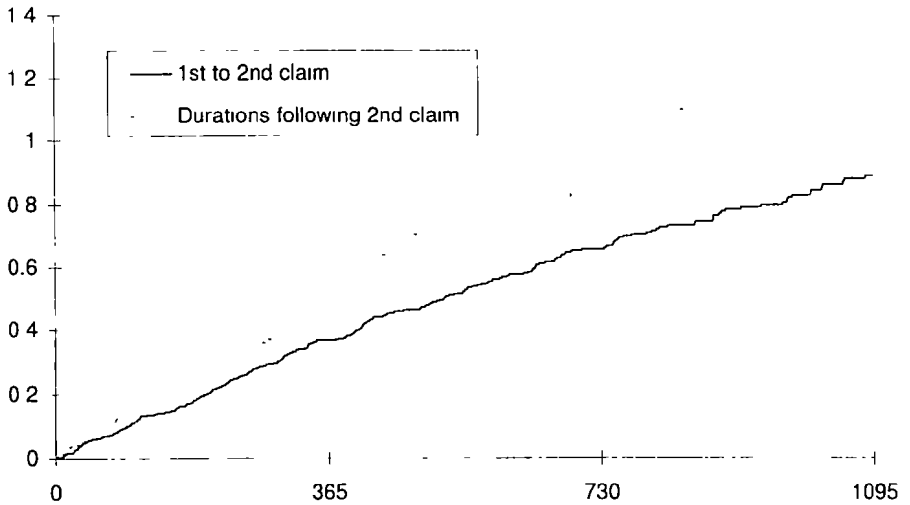

FIGURE 5 8 Estimated integrated underlying intensities for auto claims

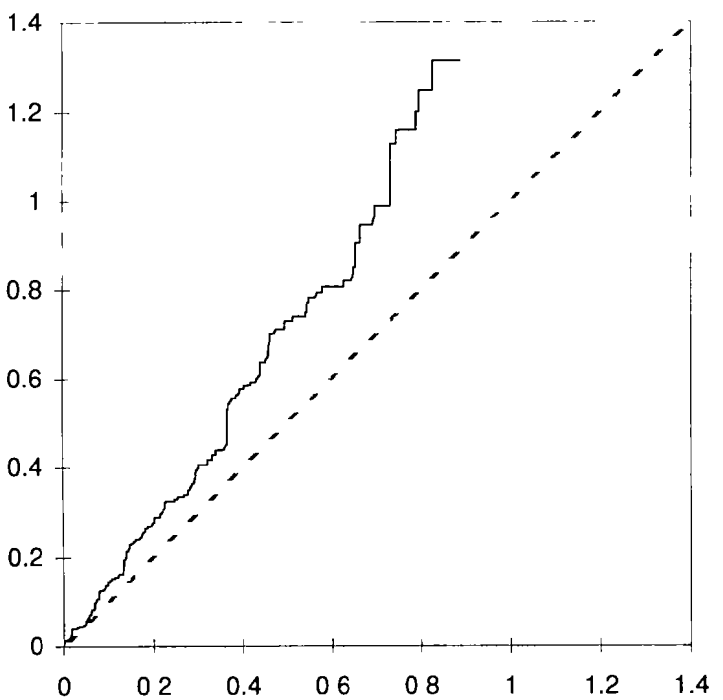

FIGURE 5 9 Estimated integrated underlying intensities for auto claims based on durations following second claim plotted against those based on the (possibly right censored) duration from first to second claim

1 January 1988 or when a new policy was taken out. Even based on these reduced data, we were able to construct a goodness-of-fit criterion that rejected the Cox regression model for property and auto claims, while household claims seemed to be amenable to analysis by this approach.

In any case the analysis performed in this section is in practice restricted to what happens during the first three years after a claim, and it is impossible to extrapolate from here to the situation before the first claim or long after a claim, both of which carry an important weight in practice.

6. Discussion and conclusion

The purpose of this report was to demonstrate some possibilities of recently developed tools in event history analysis in describing routinely collected data on non-life insurance claim histories, with the long-term aim of individualizing rating To simplify matters we ignored claim size but attempted to handle such presumably realistic difficulties as relatively short collection period (4 years), many bureaucratic renewals and the special selection pattern arising from the desire to simultaneously study household, property and auto insurance in the same company.

Our basic tool was an event history generalization of the proportional hazards model due to Cox (1972a) for survival data, see Andersen et al (1993, Chapter VII) for a detailed exposition

A central feature has been the choice of time origin. The primary choice was to use calendar time as underlying time in the Cox regression model, which necessitated a run-in period for assessing time since last claim but otherwise allowed detailed identification of effects of fixed (exogenous) and time-varying (endogenous) covariates, in most but not all cases yielding results in good accordance with expectation.

A more experimental choice was to use time since last claim as underlying time in the Cox regression model, tying to Cox's (1972b) modulated renewal process The mathematical-statistical theory of this model is rather less settled (Dabrowska, 1995). We develop in Section 5 a necessary (but by no means sufficient) goodness-of-fit criterion which, for property and auto claims, is violated even for our restricted data after first claim. Although the use of time since last claim as underlying time variable does have advantages, particularly in leading to much simpler regression models, it will so far have to be considered to be under development. The goodness-of-fit investigation indicated residual unaccounted heterogeneity, for which some kind of frailty modelling (Oakes 1992, 1998, Hougaard 1995, Scheike et al. 1997) might be fruitful.

Several of the difficulties and shortcomings listed in Sections 4.4 and 5.7 refer to the routine nature of the database that we used (and which we believe to be typical). Further attempts at employing such techniques in this

context should perhaps make an effort to obtain better tuned databases, to further calibrate and explain the tools before they are released with practical ambitions.

REFERENCES

AALEN, O O and HUSEBYE, E (1991) Statistical analysis of repeated events forming renewal processes *Statistics in Medicine*, **10**, 1227-1240

ANDERSEN, P K , BORGAN, Ø , GILL, R D and KEIDING, N (1993) *Statistical Models Based on Counting Processes*, Springer Verlag, New York

ANDERSEN, P K , CHRISTENSEN, E , FAUERHOLDT, L and SCHLICHTING, P (1983) Evaluating prognoses based on the proportional hazards model *Scand J Statist* , **10**, 141-144

ANDERSEN, P K and GILL, R D (1982) Cox's regressing model for counting processes A large sample study *Ann Statist* , **10**, 1100-1120

BLOSSFELD, H -P , HAMERLE, A and MAYER, K U (1989) *Event History Analysis Statistical Theory and Application in the Social Sciences* Lawrence Erlbaum, Hillsdale, NJ

BLOSSFELD, H -P and ROHWER, G (1995) *Techniques of Event History Modeling* Lawrence Erlbaum, Mahwah, NJ

COX, D R (1972a) Regression models and life tables (with discussion) *J R Statist Soc* , B **34**, 187-220

COX, D R (1972b) The statistical analysis of dependencies in point processes In *Stochastic Point Processes*, Ed P A W Lewis, pp 55-66 John Wiley, New York

DABROWSKA, D M (1995) Estimation of transition probabilities and bootstrap in a semiparametric Markov renewal model *Nonparametric Statistics*, **5**, 237-259

DABROWSKA, D M , SUN, G and HOROWITZ, M M (1994) Cox regression in a Markov renewal model An application to the analysis of bone marrow transplant data *J Amer Statist Association*, **89**, 867-877

GILL, R D (1980) Nonparametric estimation based on censored observations of a Markov renewal process *Z Wahrsch verw Geb* , **53**, 97-116

GILL, R D (1983) Discussion of the papers by Helland and Kurtz *Bull Internat Statist Inst* , **50**, 239-243

HOUGAARD, P (1995) Frailty models for survival data *Lifetime Data Analysis*, **1**, 19-38

KALBFLEISCH, J D and PRENTICE, R L (1980) *The statistical analysis of failure time data* Wiley, New York

LAWLESS, J F (1998) Repeated events *Encyclopedia of Biostatistics*, **5**, 3783-3787

LAWLESS, J F and THIAGARAJAH, K (1996) A point-process model incorporating renewals and time trends, with application to repairable systems *Technometrics*, **38**, 131-138

LINDSEY, J K (1995) Fitting parametric counting processes by using log-linear models *Appl Statist* , **44**, 201-212

MCCLEAN, S and DEVINE, C (1995) A nonparametric maximum likelihood estimator for incomplete renewal data *Biometrika*, **82**, 791-803

OAKES, D (1992) Frailty models for multivariate event times In Klein, J P and Goel, P K (eds) *Survival analysis State of the art*, pp 371-379 Netherlands, Kluwer

OAKES, D (1998) Duration dependence *Encyclopedia of Biostatistics*, **2**, 1248-1252

OAKES, D and CUI, L (1994) On semiparametric inference for modulated renewal processes *Biometrika*, **81**, 83-90

PRENTICE, R L , WILLIAMS, B J and PETERSON, A V (1981) On the regression analysis of multivariate failure time data *Biometrika*, **68**, 373-379

RAMLAU-HANSEN, H (1983) Smoothing counting process intensities by means of kernel functions *Ann Statist* , **11**, 453-466

SCHEIKE, T H , MARTINUSSEN, T and PETERSEN, J H (1997) Retrospective ascertainment of recurrent events An application to time to pregnancy Res Rep 97/14, Dep Biostat , Univ Copenhagen

SOON, G and WOODROOFE, M (1996) Nonparametric estimation and consistency for renewal processes *Journal of Statistical Planning and Inference*, **53**, 171-195

TJUR, T (1993) The StatUnit manual University of Copenhagen, Institute of Mathematical Statistics

VARDI, Y (1982) Nonparametric estimation in renewal processes *Ann Statist*, **10**, 772-785

Dr. NIELS KEIDING
Institute of Public Health
Department of Biostatistics
University of Copenhagen
3 Blegdamsvej
DK-2200 Copenhagen N
Denmark
Tel. +45 35 32 79 01
Fax +45 35 32 79 07

CHRISTIAN ANDERSEN
PETER FLEDELIUS
ATP PensionService A/S
Kongens Vaenge 8
DK-3400 Hilleroed
Denmark