Frailty modelling of time-to-lapse of single policies for customers holding multiple car contracts

Marion Haugen[a*] and Tron Anders Moger[b]

[a]Norwegian Computing Center, P.O. Box 114 Blindern, NO-0314 Oslo, Norway

[b]Department of Health Management and Health Economics, University of Oslo, P.O. Box 1089 Blindern, NO-0318 Oslo, Norway

[*]Corresponding author. E-mail: marion.haugen@nr.no. Phone: (+47) 22 85 26 02. Fax: (+47) 22 69 76 60.

Abstract

Haugen M, Moger TA. Frailty modelling of time-to-lapse of single policies for customers holding multiple car contracts. Scandinavian Actuarial Journal. Corporate customers often hold multiple contracts and this might give dependence between the lapsing times of the single policies. We present a shared gamma frailty model in order to study the time-to-lapse of single car policies for customers holding multiple car contracts with the same insurance company, accounting for measured and time dependent covariates. Customers with the highest frailty value tend to leave the company earlier than the others and finding these is a central aspect within a company's customer relationship management strategy. We estimate conditional survival curves which illustrate the decreased survival probability of a customer after a lapse in a single car insurance policy. The individual survival curves are overestimated if the underlying association for cars with the same customer is ignored. Fitting misspecified Cox's proportional hazards model also results in an underestimation of the standard error of the parameter estimates. Keywords: frailty, intracluster dependence, non-life insurance, survival analysis.

1 Introduction

A company must generate sales to survive but the firm's sales are not only a function of new customer numbers, but also of how many existing customers are retained (Brockett et al., 2008). With a real life example from the financial services industry, Van den Poel and Lariviére (2004) illustrated that an increase in retention rate of just one percentage point may result in substantial profit increases. For customers holding multiple types of contracts with the same insurance company, e.g. house, content and car, Brockett et al. (2008) showed that a lapse in one policy may be a signal of the be-

ginning of the customer's defection to a competitor. Finding the customers most likely to leave the company is hence a crucial task in customer relationship management (CRM). Once they are identified, individual customer retention procedures can be carried out.

Survival and event history analysis are important tools in statistics. These methods have applications for instance in insurance (Brockett et al., 2008, Keiding et al., 1998, Van den Poel and Lariviére, 2004), medicine (Drzewiecki and Andersen, 1982, Jepsen et al., 2008, Maruza et al., 2012) and reliability (Wong and Tsai, 2012). Cox's proportional hazards model (Cox, 1972) and its extensions have become standard methods for survival analysis (Therneau and Grambsch, 2000). Frailty models (Hougaard, 1995, 2000, Therneau and Grambsch, 2000) add a random effect to Cox's proportional hazards model, an effect thought to act multiplicatively on the hazard function such that a large value increases the hazard. The idea is that individuals have different frailties and those who are most frail tend to experience the event of interest earlier than the others. Aalen (1988) discussed the impact of individual heterogeneity in survival analysis and illustrated how random effects can deal with it.

When observations are clustered into groups, such as families, hospitals or cities, the shared frailty model is the most often adapted model (Rondeau et al., 2012). It was introduced by Clayton (1978) and extensively studied by Hougaard (2000). In this paper we have applied the shared frailty model in order to study the time-to-lapse of single car policies for customers holding multiple car contracts. Some customers had only a single car insurance policy but many of them held multiple car contracts and the frailty term may be dependent for these. If a single car insurance policy is lapsed, there is a risk that all cars belonging to this customer will leave the portfolio.

Each customer forms a cluster or a group with a specific frailty value, which is shared among all cars in the cluster. Sharing a frailty value generates dependence within a cluster, whereas conditional on the frailty those cars are independent.

Intracluster dependence is often modelled with the gamma frailty distribution because of its simple interpretation and mathematical tractability. The closed form expressions of the unconditional survival, cumulative density and hazard function are easily derived due to the simplicity of the Laplace transform of the probability function for the gamma distribution (Hirsch and Wienke, 2012). Closed form expressions of both the univariate and the bivariate unconditional survival curves are used to graphically illustrate the effect a lapse in a single car insurance policy has on the survival probability of a customer. One then plots the survival function given that another car in the cluster has lapsed and the survival function given that another car has maintained the policy to see how the dependence within groups of cars affects the survival for different combinations of the covariates. If the underlying dependencies in the data are ignored, the individual survival curves are overestimated. These illustrations may only be carried out if we estimate the distribution of the frailty. A different approach is to use the robust standard errors of the parameter estimates. These estimates are adjusted for dependencies but we are not able to make good illustrations in the same manner as with a fitted frailty distribution. All of the estimation may be carried out by using standard software for survival analysis.

This paper is organised as follows. The data are described in Section 2. Section 3 describes the shared gamma frailty model and the estimation of model parameters, and it gives some theoretical results. The main results from our statistical analysis are presented in Section 4 and finally we

summarise our findings and discuss some remaining challenges in Section 5.

2 Material

The dataset consisted of all corporate customers in a car insurance portfolio for small and medium sized companies in the largest non-life insurance company in Norway, Gjensidige, from 1998 to 2007. There were 61.189 unique customers and 201.897 unique cars in the full dataset. The study period was fixed and by the end of the study not all cars had left Gjensidige. In survival analysis, such right censored data are easily accounted for.

The data included fixed and time dependent covariates on each car that may help to find the customers most likely to leave the company. There were no data on driver characteristics since several drivers may use one car. The covariates are measured from the beginning of the policy until the moment of lapsing or censoring (December 31, 2007). A summary of the covariates is given in Table 1, most of them are self-explanatory. The covariates DiscountC and DiscountO indicate whether a customer receives a chain store discount on the yearly premium or a discount due to membership in a specific organisation, respectively. There were four possible Covers: liability, fully or partially comprehensive insurance and accident. All cars had liability insurance, 76% of the cars with two covers had accident cover in addition and 86% of the cars with three covers also had fully comprehensive insurance. Approximately 80% of the cars had fully comprehensive insurance and 13% had partially comprehensive insurance. Categorised versions of the covariates Covers, Area, Disloyal and Usage are used in our frailty model. The subcategories for Area are specified in Table 2. Dimakos et al. (2009) showed that the loyalty to Gjensidige varied for car manufacturers and the disloyal car manufacturers applied in this analysis are determined from their results.

The use of categorisation is consistent with insurance company practice and with the literature of non-life insurance (Antonio et al., 2010, Desjardins et al., 2001).

Table 1, in about here.

The reason for the lapse in a single car insurance policy was unknown. A customer could switch one car with another. Approximately 28.000 cars were replaced within three months of their lapse and we considered these policies to be maintained by the new car. There were no data on total insurance time for the cars, neither duration nor first date. This was problematic since the lapsing probability from Gjensidige was not constant over time periods, so a misplacement of the observed time period could lead to bias in the estimation of survival. Some of the cars, maybe all, insured in 1998 have probably entered the company prior to the study start and therefore should have been left truncated to avoid this problem. The cars which are insured in 1998 are deleted (22% of the cars) and we have considered policies from 1999 to 2007. Cars with questionable information on some of the covariates were removed from the sample and therefore ignored in the analysis.

With these constraints we are left with a dataset of 48.040 unique customers and 108.274 unique cars. The customers were small and medium sized companies, 59% of them had only one car contract while 98% had at most ten cars insured. 70.751 cars left Gjensidige during the study period, 46% of the customers only had one lapse and 96% of the customers had at most five lapses. Single car policies entered and left the sample, lapses occurred without the customer leaving the company. Among the customers holding at least two car contracts (41% of the customers), 17% lapsed all the contracts within a maximum of two months, 49% lapsed some of the con-

tracts within a maximum of two months and 34% only had one lapse during the study period. The yearly turnover in the data was approximately 21%.

3 A frailty model for customers holding multiple contracts

Cox's proportional hazards model is semiparametric since the effect of covariates is estimated parametrically without specification of the distribution of the baseline hazard. The baseline hazard describes how the risk changes over time at baseline levels of covariates. In this analysis the individual hazard rate gives the risk for a lapse in a single car insurance policy per unit of time, given that the car is still insured in Gjensidige. This hazard rate varies between customers, some have a higher risk for leaving the company, and observable covariates can probably not explain all of this variation. Frailty models take dependence caused by unmeasured covariates into consideration. A frailty model may be applied when measurements that vary within the group are missing or a shared frailty model when a latent common group effect is present.

As mentioned in Section 1, one argument for our choice of frailty distribution is the simplicity of the Laplace transform of the probability function for the gamma distribution. It is also reasonable to presume that the risk for an event is skewly distributed over the customers. Therefore, we apply Cox's proportional hazards model with shared gamma frailty on the customer. The lapsing times of single car policies are assumed to be conditional independent with respect to the shared frailty, that is had we known the frailty, the events would have been independent. Let κ be a gamma distributed frailty having variance $\theta = 1/\nu$ and mean 1 for identifiability. Those clusters that

possess $\kappa > 1$ are said to be more frail for reasons left unexplained by the covariates and will have an increased risk of failure. Conversely, those clusters with $\kappa < 1$ are less frail and will tend to survive longer. Large values of θ signify closer positive relationship for the cars in a cluster and greater heterogeneity between the clusters. The Laplace transform of the probability function for the gamma distribution is $L(s) = (1 + \theta s)^{-1/\theta}$. The log of the density of κ can be written as:

$$\log[f(\kappa;\nu)] = (\nu - 1)\log(\kappa) - \nu\kappa + \nu\log(\nu) - \log\Gamma(\nu)$$

The conditional hazard rate for car j ($j = 1, \ldots, N_i$) in cluster i ($i = 1, \ldots, 48.040$) is given as

$$\lambda_{ij}(t|\kappa_i) = \kappa_i\lambda_{ij}(t) = \kappa_i\lambda_0(t)\exp\{\boldsymbol{\beta}^T\boldsymbol{X}_{ij}(t)\}, \tag{1}$$

where t throughout denotes the time in days, $\lambda_0(t)$ is the baseline hazard rate at time t, $\boldsymbol{\beta}$ is the vector of regression coefficients to be estimated, and $\boldsymbol{X}_{ij}(t)$ is the covariate vector at time t. The baseline hazard $\lambda_0(t)$ is assumed to be common to all cars in the portfolio and is estimated non-parametrically from the data. The estimation procedure is described in the last paragraph of this section. The covariate effect $\exp(\beta)$, also called the hazard ratio (HR), should be interpreted as a change in the within cluster relative risk.

Therneau et al. (2003) showed that the shared gamma frailty model can be written exactly as a penalised likelihood. Penalised models treat the frailty terms as additional regression coefficients which are constrained by a penalty function added to the log-likelihood. The hazard in (1) may be written as $\lambda_{ij}(t|\gamma_i) = \lambda_0(t)\exp\{\boldsymbol{\beta}^T\boldsymbol{X}_{ij}(t) + \gamma_i\boldsymbol{Z}_{ij}(t)\}$, where $\gamma_i = \log(\kappa_i)$

and $\boldsymbol{Z}$ is a matrix of indicator variables such that $\boldsymbol{Z}_{ij} = 1$ when car j belongs to cluster i and 0 otherwise. This hazard is estimated by a maximisation, over both $\boldsymbol{\beta}$ and γ, of the penalised partial log-likelihood

$$\begin{aligned} \text{PPL}(\boldsymbol{\beta}, \gamma; \theta) &= l(\boldsymbol{\beta}, \gamma) - g(\gamma; \theta) \\ &= l(\boldsymbol{\beta}, \gamma) + 1/\theta \sum\nolimits_i \Big[\gamma_i - \exp(\gamma_i)\Big], \end{aligned} \tag{2}$$

where l is the log of the usual Cox partial likelihood, g is a penalty function restricting the values of γ and θ is a tuning parameter, equal to the variance of κ. Typically, the purpose of g is to "shrink" γ towards zero and the amount of shrinkage is controlled with θ. For censored cases, we define δ_{ij} to be 0 if the case is right censored and 1 if the case is uncensored. Further, $R(t_{ij})$ is the risk set which consists of units (i, j) at risk of leaving Gjensidige at time t. Let $h_{ij} \equiv \boldsymbol{\beta}^T \boldsymbol{X}_{ij}(t_{ij}) + \gamma_i \boldsymbol{Z}_{ij}(t_{ij})$. The Cox partial log-likelihood is then given by:

$$l(\boldsymbol{\beta}, \gamma) = \sum_{i=1}^{48.040} \sum_{j=1}^{N_i} \delta_{ij} \left[h_{ij} - \log \left\{ \sum_{(k,l) \in R(t_{kl})} \exp(h_{kl}) \right\} \right]$$

The Newton-Raphson algorithm can be used to fit the penalised partial log-likelihood. The model fitting consists of an inner and outer loop. For given θ, the penalised model is solved with a few (usually three to five) Newton-Raphson iterations, and the corresponding value of the penalised partial log-likelihood is returned. The outer loop maximises the profile log-likelihood for θ.

Let T_{i1} and T_{i2} be the survival times of two cars of customer i. The individual survival function for these cars is given as

$$S(t_{ij}|\kappa_i) = \exp(-\kappa_i \Lambda_{ij}(t_{ij})), \tag{3}$$

where

$$\Lambda_{ij}(t) = \int_0^t \lambda_{ij}(u)du = \int_0^t \lambda_0(u)\exp\{\boldsymbol{\beta}^T \boldsymbol{X}_{ij}(u)\}du \tag{4}$$

is the cumulative baseline hazard. Define $F_1(t_{i1}, t_{i2}) = P(T_{i2} \leq t_{i2} | T_{i1} \leq t_{i1})$, the probability that car number two lapses within time t_{i2} given that car number one has lapsed within time t_{i1}, and $F_0(t_{i1}, t_{i2}) = P(T_{i2} \leq t_{i2} | T_{i1} > t_{i1})$, the probability that car number two lapses within time t_{i2} given that car number one has maintained the policy up to time t_{i1}. From these probabilities, one may derive the conditional survival for a car in cluster i at time t_{i2} (e.g. maintaining the policy the first year), given that another car in the cluster has lapsed within time t_{i1}, $1 - F_1(t_{i1}, t_{i2})$, and the conditional survival for a car at time t_{i2}, given that another car has maintained the policy up to time t_{i1}, $1 - F_0(t_{i1}, t_{i2})$. As in Moger and Aalen (2008), by using Bayes' theorem and that all cars in a cluster are independent given the frailty, one gets:

$$\begin{aligned} 1 - F_1(t_{i1}, t_{i2}) &= \frac{S_{i2}(t_{i2}) - S(t_{i1}, t_{i2})}{1 - S_{i1}(t_{i1})} \\ 1 - F_0(t_{i1}, t_{i2}) &= \frac{S(t_{i1}, t_{i2})}{S_{i1}(t_{i1})} \end{aligned} \tag{5}$$

The marginal survival function is given by $S_{ij}(t_{ij}) = L(\Lambda_{ij}(t_{ij}))$, where L is the Laplace transform of the probability function for the gamma distribution. The bivariate survival function for two cars within cluster i is easily expressed by means of L, evaluated at the total cumulative baseline hazard:

$$S(t_{i1}, t_{i2}) = L(\Lambda_{i1}(t_{i1}) + \Lambda_{i2}(t_{i2}))$$

The conditional survival curves in (5) can then be plotted and compared to the population survival function.

Information in the frailty concerning the dependence between cars with the same customer is summarised with Kendall's τ, a dependence measure widely used in Hougaard (2000). For the shared gamma frailty model Kendall's τ is given by $\theta/(2+\theta)$. If the dependence completely disappears after covariate adjustment, then the measured covariates give all information about the lapsing times of single car policies for a customer and loyal or disloyal customers can be detected from the covariates. Given the covariates, there is no association between the events for a customer.

A shared frailty model may be fitted by using standard software for survival analysis, like the coxph function in the survival library in R (Hirsch and Wienke, 2012, R Development Core Team, 2012). It fits the proportional hazards model and its frailty extension by maximising the penalised partial log-likelihood (2) (Therneau and Grambsch, 2000). A description of how time dependent covariates are handled in the coxph function is given in Fox (2002). The aforementioned functions in R do not provide standard errors for the estimated frailty variance. This is provided in the R package frailtypack (Rondeau et al., 2012), which is not applied here. Standard software may also be applied to estimate the baseline hazard $\lambda_0(t)$. We have used the basehaz function in R and then plugged the estimate into the cumulative baseline hazard (4), and further estimated the individual (3) and conditional (5) survival curves.

4 Results

First, we fit the model without covariates. The penalised partial log-likelihood with and without frailty is -713.094 and -758.948, respectively, a highly sig-

nificant frailty by the likelihood ratio test (p-value < 0.001). The frailty variance is 1.18, corresponding to a value of Kendall's τ of 0.37 which indicates large differences between customers.

The first model with covariates includes Premium, Area, Disloyal and Usage. The covariates are described in Table 1. The penalised partial log-likelihood increases to -712.707, the frailty variance is 1.16 and Kendall's τ is 0.37. These covariates do not explain much of the unobserved heterogeneity in the data. When we also include the covariates Bonus, DiscountC, DiscountO, Covers and DueDate, the penalised partial log-likelihood increases to -709.203, the frailty variance and Kendall's τ is 1.01 and 0.34, respectively. Compared to the first model, there is less degree of heterogeneity among the clusters. The covariate selection was partly motivated by what was available of covariates in our dataset and this is the final model.

The upper part of Table 2 gives the results for the final model. All the covariates are significant at 5% significance level. The effect of a change in any of the covariates may be calculated. For example, an increase in the Premium from 6.000 to 8.000 Norwegian Kroner (NOK) gives a 2% increased risk for lapsing a single car insurance policy within the cluster ($\exp\{9.9\text{e-}4 \cdot (80\text{-}60)\} \approx 1.02$). If the Bonus is increased from 60% to 80%, the risk for a lapse in a single car insurance policy is reduced with 34% ($\exp\{\text{-}2.097 \cdot (0.8\text{-}0.6)\} \approx 0.66$). The Disloyal car manufacturers (i.e. Audi, BMW and Chrysler) have an increased risk of 18% compared to the loyal car manufacturers. The solid lines in Figure 1 show the fitted survival curves of a car with a Premium of approximately 6.000 NOK and 60% Bonus, the remaining covariate values are given under the heading "Car 2" of Table 3. The dashed line to the left in Figure 1 shows that an increase in the Premium with 2.000 NOK gives an almost identical survival curve for this car,

whereas the dashed line to the right shows that an increase in the Bonus to 80% results in a considerably increased survival probability.

Table 2, in about here.

Figure 1, in about here.

The lower part of Table 2 gives the results from a corresponding Cox model without a frailty term. The penalised partial log-likelihood for this model is -751.056. When the underlying dependence between the lapsing times of single car policies for a customer is ignored, the standard error of the parameter estimates is underestimated. As expected, most of the regression parameters for the Cox model are smaller than for the shared frailty model. The estimates are population average effects unlike the individual effects from the shared frailty model.

The frailty distribution is shown in the histogram to the left in Figure 2. To the right in the figure, the turnover rate for remaining and removed policies are given in solid and dashed lines, respectively. Remaining (removed) policies are cars with a frailty value equal to or lower (higher) than the one given on the x-axis. The plot shows that the frailty values may be applied to find cars with a high or low risk for leaving Gjensidige. For example, from the histogram we know that a cut-off value of two on the frailty will remove 5% of the customers, corresponding to 9% of the cars. Cars with a frailty value above two have a turnover rate on 98% and removing them from the portfolio results in a decrease in the overall turnover rate from 65% (70.751 of 108.274 cars) to 62% (61.375 of 98.732 cars). The decrease in premium income by removing these cars is 8%, which must be weighed against the cost of having disloyal customers. For new customers, the frailty may be treated as a missing variable and be imputed by, for example, the mean of

the frailty values for customers with similar covariate values. The estimated frailty value may then be used to determine whether the customer should be included in the portfolio or not.

Figure 2, in about here.

Figure 3 shows the fitted survival curves (3) of four cars belonging to customers holding at least two car contracts and having different estimated frailty values, specified in the headings of the figure. The cars have been selected for illustration and the covariate values are given in Table 3. The Premium varies from approximately 2.600 NOK to 30.000 NOK, only car number two has a discount due to a membership in a specific organisation and all cars have three Covers. The solid lines are the survival curves from the shared frailty model and the dashed lines are the survival curves from the Cox model without a frailty term. The figure illustrates that the survival curves are overestimated if the frailty is neglected, more for some cars than others.

Figure 3, in about here.

Table 3, in about here.

The solid lines in Figure 4 show the Kaplan-Meier curve based on all data. The dashed and dotted lines show conditional survival curves for the shared frailty model, found by setting $t_{i1} = 365$ in (5). In the illustration, we have used the same cars as in Figure 3 and the second car for the customer has been selected such that the covariate values correspond with the ones given in Table 3. The figure clearly shows the effect a lapse within the cluster has on the probability for maintaining the policy in Gjensidige. For all cars, the survival curve of the car given that another car in the cluster

has lapsed within one year (dashed line) is well below the survival curve of the car given that another car in the cluster has maintained the policy the first year (dotted line). If the conditional survival curves are below the Kaplan-Meier curve, e.g. upper left plot, the covariate values of the car have a negative effect on survival. With more beneficial covariate values, a lapse within the cluster results in a much smaller effect on survival, e.g. upper right plot.

Figure 4, in about here.

5 Discussion

For customers holding multiple car contracts, the time-to-lapse of single car policies may be modelled with a shared frailty model where the random effect is a continuous variable describing excess risk or frailty for the customers. The unexplained heterogeneity is caused by the natural assumption that customers respond differently to price changes, resulting in a necessity for adjustment of both known and unknown covariates. The frailty term represents the neglected common covariates for the cars of a customer and this is an advantage of the frailty model. The effect of, for example, premium is conditional on both known and unknown covariates, in contrast to the Cox model which gives premium effects conditional on the known covariates only. We had customers holding the same type of contracts in our dataset. More generally, the shared frailty model may be applied to customers holding multiple types of contracts, such as house, content and car.

Random effects are very often present in survival data. Henderson and Oman (1999) investigated the consequences of ignoring frailty when present,

fitting misspecified Cox's proportional hazards model to the marginal distributions. They concluded that a misspecification can lead to an underestimation of covariate effects and inaccuracy in fitted survival curves by quite important and significant amounts. The amount of bias in regression coefficients depends in magnitude on the variability of the frailty term and the form of the frailty distribution. For the data analysed in this paper, we have seen that a Cox model results in underestimated standard error of the parameter estimates and overestimated fitted survival curves.

The frailty models are far from perfect. The choice of the frailty distribution is mainly due to mathematical convenience and different frailty distributions give quite different dependence structures. The gamma frailty model describes high late dependence (Hougaard, 1995) so the dependence is most important for late events. Possible frailty distributions in the R function frailty are the gamma, lognormal and log-t distributions. In contrast to the gamma distributions, the Laplace transforms of the probability functions for the lognormal and log-t distributions are theoretically intractable so approximation or numerical integration must be applied for probability results. The mathematical tractability of the gamma distribution is an argument for our choice of frailty distribution.

Some covariates that might have a great influence on a customer's decision to leave Gjensidige were not available for this study. The last offered premium to the customer before the lapse probably had a great influence on the decision to leave Gjensidige. The reason for the lapses was unknown. Some cars could be lapsed because the firm had to be downsized or costs had to be reduced, regardless of the premium. Lastly, we had no data on confounding variables such as competitors' campaigns, newspaper and TV headlines, and economic cycles. Without these internal and external covari-

ates, it is difficult to know if the observed lapses are caused by the yearly premium and consequently that the premium effects are correctly estimated.

The main aim of this paper has been to illustrate the inaccuracies in the model fit when a frailty term is present but ignored in a car insurance portfolio with customers holding multiple contracts. In a more detailed analysis it would have been natural to focus more on the covariate selection by using a more thorough grouping of the categorical covariates and adding interaction effects between, for example, yearly premium and car manufacturer. Model assessment could have been performed by out-of-sample and out-of-time prediction (Günther et al., 2011).

6 Acknowledgements

Marion Haugen was supported by Statistics for Innovation, $(\text{sfi})^2$. We are grateful to Gjensidige and Lars Kvifte for providing the data and giving us a better overview of the data. We thank Ingunn Fride Tvete at Norwegian Computing Center and Xeni Kristine Dimakos at the Norwegian Agency for Lifelong Learning for helping to plan the project and for keeping touch with Gjensidige. We greatly appreciate the insightful comments from Professor Ørnulf Borgan at Department of Mathematics at the University of Oslo, Clara-Cecilie Günther at Norwegian Computing Center and an anonymous referee. Their valuable suggestions have led to great improvements to the quality of the paper.

References

Aalen, O. O. (1988). Heterogeneity in survival analysis. *Stat. Med.*, 7(11):1121–1137.

Antonio, K., Frees, E. W., and Valdez, E. A. (2010). A multilevel analysis of intercompany claim counts. *Astin Bull.*, 40(1):151–177.

Brockett, P. L., Golden, L. L., Guillen, M., Nielsen, J. P., Parner, J., and Perez-Marin, A. M. (2008). Survival analysis of a household portfolio of insurance policies: How much time do you have to stop total customer defection? *The Journal of Risk and Insurance*, 75(3):713–737.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 34(2):187–220.

Desjardins, D., Dionne, G., and Pinquet, J. (2001). Experience rating schemes for fleets of vehicles. *Astin Bull.*, 31(1):81–105.

Dimakos, X. K., Storvik, B., and Vårdal, J. F. (2009). Kundelojalitet i PVK/BIL NL. NR-Note, SAMBA/51/08, Norwegian Computing Center.

Drzewiecki, K. T. and Andersen, P. K. (1982). Survival with malignant melanoma: a regression analysis of prognostic factors. *Cancer*, 49(11):2414–2419.

Fox, J. (2002). Cox Proportional-Hazards Regression for Survival Data: Appendix to *An R and S-PLUS Companion to Applied Regression*. http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-cox-regression.pdf.

Günther, C.-C., Tvete, I. F., Aas, K., Sandnes, G. I., and Borgan, Ø. (2011). Modelling and predicting customer churn from an insurance company. *Scand. Actuar. J.* DOI:10.1080/03461238.2011.636502.

Henderson, R. and Oman, P. (1999). Effect of frailty on marginal regression estimates in survival analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 61(2):367–379.

Hirsch, K. and Wienke, A. (2012). Software for semiparametric shared gamma and log-normal frailty models: An overview. *Computer Methods and Programs in Biomedicine*, 107(3):582–597.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime Data Anal.*, 1(3):255–273.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer, New York.

Jepsen, P., Vilstrup, H., Andersen, P. K., Lash, T. L., and Sørensen, H. T. (2008). Comorbidity and survival of Danish cirrhosis patients: A nationwide population-based cohort study. *Hepatology*, 48(1):214–220.

Keiding, N., Andersen, C., and Fledelius, P. (1998). The Cox regression model for claims data in non-life insurance. *Astin Bull.*, 28(1):95–118.

Maruza, M., Albuquerque, M. F. P. M., Braga, M. C., Barbosa, M. T. S., Byington, R., Coimbra, I., Moura, L. V., Batista, J. D. L., Diniz, G. T. N., Miranda-Filho, D. B., Lacerda, H. R., Rodrigues, L. C., and Ximenes, R. A. A. (2012). Survival of HIV-infected patients after starting tuberculosis treatment: a prospective cohort study. *The International Journal of Tuberculosis and Lung Disease*, 16(5):618–624.

Moger, T. A. and Aalen, O. O. (2008). Regression models for infant mortality data in Norwegian siblings, using a compound Poisson frailty distribution with random scale. *Biostatistics*, 9(3):577–591.

R Development Core Team (2012). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.r-project.org/.

Rondeau, V., Mazroui, Y., and Gonzalez, J. R. (2012). frailtypack: An R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software*, 47(4):1–28.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York.

Therneau, T. M., Grambsch, P. M., and Pankratz, V. S. (2003). Penalized survival models and frailty. *J. Comput. Graph. Statist.*, 12(1):156–175.

Van den Poel, D. and Lariviére, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European J. Oper. Res.*, 157(1):196–217.

Wong, J.-T. and Tsai, S.-C. (2012). A survival model for flight delay propagation. *Journal of Air Transport Management*, 23:5–11.

List of Tables

List of Figures

4 Conditional survival curves of four cars showing the effect a lapse has on survival. The Kaplan-Meier curve (solid line) is based on all data. The dashed line shows the survival of the specified car given that another car in the cluster has lapsed within one year. The dotted line shows the survival of the specified car given that another car in the cluster has maintained the policy the first year. Covariate values for the cars are specified in Table 3.

Corresponding author:

Marion Haugen, Norwegian Computing Center, P.O. Box 114 Blindern,

NO-0314 Oslo, Norway. Phone: (+47) 22 85 26 02.

Table 1: Description of measured covariates.

Covariate	Description
Premium	Yearly premium per 100 NOK (range [0.55, 2086.00]).
Bonus	Percentage no-claims bonus (range [-0.8, 0.8]).
DiscountC	Chain store discount (0 = No, 1 = Yes).
DiscountO	Organisational discount (0 = No, 1 = Yes).
Covers	Number of covers (1, 2, 3, 4).
Area	Counties of Norway, grouped into five subcategories.
Disloyal	Disloyal car manufacturer (0 = No[a], 1 = Yes[b]).
Usage	Area of usage (1 = Private, 2 = Taxi, 3 = Various use[c]).
DueDate	Less than 30 days until due date (0 = No, 1 = Yes).

[a] Volkswagen, Toyota, Ford, Opel, etc.
[b] Audi, BMW, Chrysler.
[c] For example driving school, rental or craftsmen.

Table 2: Estimated covariate effects with standard errors (SE) from the shared frailty model and the Cox model. Hazard ratios (HR) with 95% confidence intervals.

	Estimate	SE	HR (95% CI)
Shared frailty model			
Premium	9.9e-4	1.1e-4	1.001 (1.001-1.001)
Bonus	-2.097	0.031	0.123 (0.116-0.130)
DiscountC	-0.357	0.047	0.699 (0.638-0.767)
DiscountO	-0.385	0.036	0.680 (0.634-0.729)
Covers = 2	-0.253	0.040	0.777 (0.719-0.839)
Covers = 3	-0.518	0.038	0.596 (0.553-0.641)
Covers = 4	-0.975	0.069	0.377 (0.330-0.432)
Reference level: Akershus, Oslo			
Area = 2[a]	-0.170	0.023	0.844 (0.807-0.882)
Area = 3[b]	0.106	0.023	1.112 (1.064-1.163)
Area = 4[c]	-0.339	0.021	0.712 (0.684-0.742)
Area = 5[d]	-0.102	0.024	0.903 (0.862-0.947)
Disloyal	0.168	0.018	1.183 (1.141-1.226)
Usage = 2	-0.072	0.027	0.930 (0.883-0.981)
Usage = 3	-0.010	0.012	0.990 (0.966-1.014)
DueDate	-0.731	0.009	0.482 (0.473-0.490)
Cox model			
Premium	5.4e-4	8.4e-5	1.001 (1.000-1.001)
Bonus	-1.744	0.022	0.175 (0.167-0.183)
DiscountC	-0.163	0.027	0.849 (0.805-0.896)
DiscountO	-0.263	0.020	0.769 (0.739-0.799)
Covers = 2	-0.209	0.030	0.811 (0.764-0.861)
Covers = 3	-0.430	0.028	0.651 (0.615-0.688)
Covers = 4	-0.900	0.061	0.406 (0.361-0.458)
Reference level: Akershus, Oslo			
Area = 2[a]	-0.109	0.012	0.897 (0.876-0.918)
Area = 3[b]	0.044	0.012	1.045 (1.021-1.069)
Area = 4[c]	-0.242	0.011	0.785 (0.768-0.802)
Area = 5[d]	-0.099	0.012	0.906 (0.884-0.928)
Disloyal	0.203	0.014	1.225 (1.191-1.259)
Usage = 2	-0.003	0.017	0.997 (0.965-1.030)
Usage = 3	-0.122	0.009	0.885 (0.870-0.900)
DueDate	-0.668	0.008	0.513 (0.505-0.521)

[a]Østfold, Hedmark, Oppland
[b]Buskerud, Vestfold, Telemark, Aust-Agder, Vest-Agder
[c]Rogaland, Hordaland, Sogn and Fjordane, Møre and Romsdal
[d]Sør-Trøndelag, Nord-Trøndelag, Nordland, Troms, Finnmark

Table 3: Covariate values of four cars used for illustration.

Covariate	Car 1	Car 2	Car 3	Car 4
Premium	300.90	59.75	67.93	26.11
Bonus	0.2	0.6	0.7	0.6
DiscountC	0	0	0	0
DiscountO	0	1	0	0
Covers	3	3	3	3
Area	1	2	3	4
Disloyal	1	0	0	0
Usage	3	3	1	1
DueDate	1	1	0	0

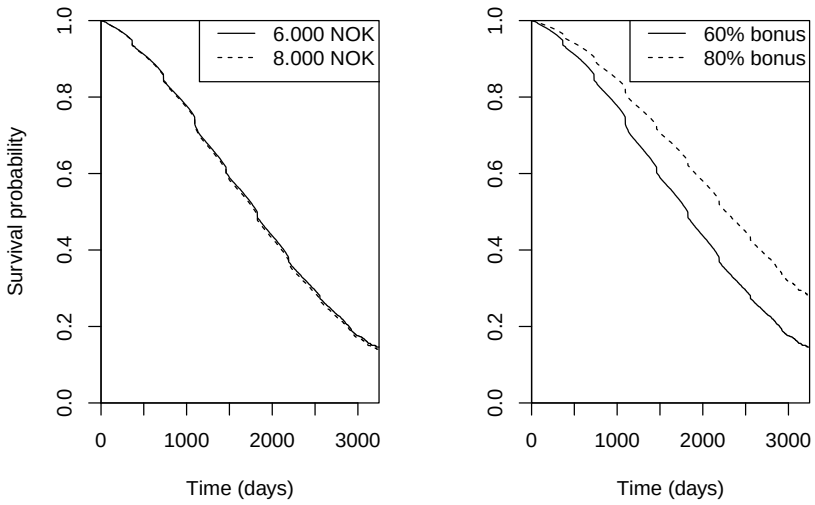

Figure 1: Fitted survival curves showing, on the left, the effect of an increased yearly premium from 6.000 NOK to 8.000 NOK and, on the right, an increased bonus from 60% to 80%. Covariate values for the car are specified under the heading "Car 2" of Table 3.

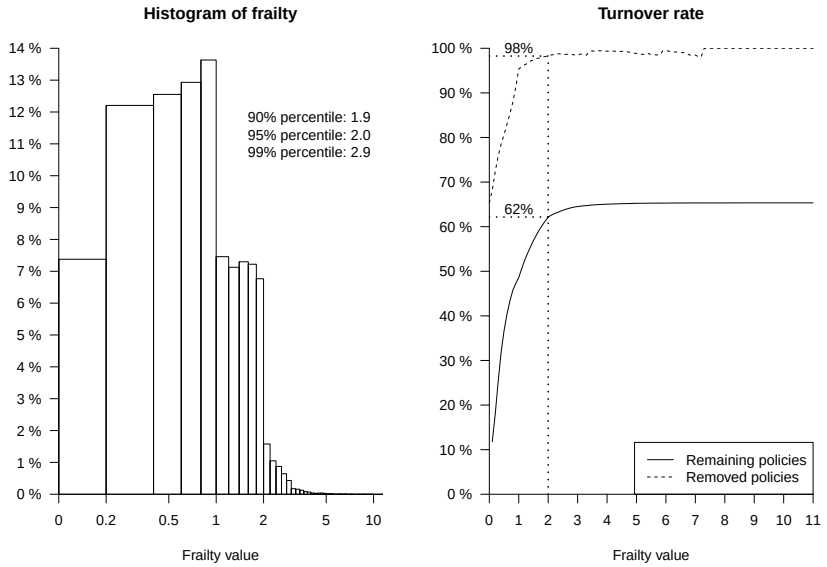

Figure 2: On the left, a histogram of the frailty showing the proportion of customers with frailty values specified by the bars. The solid line on the right gives the turnover rate for cars with a frailty value equal to or lower than the specified value on the x-axis ("remaining policies"). The dashed line gives the turnover rate for cars with a frailty value above the specified value on the x-axis ("removed policies").

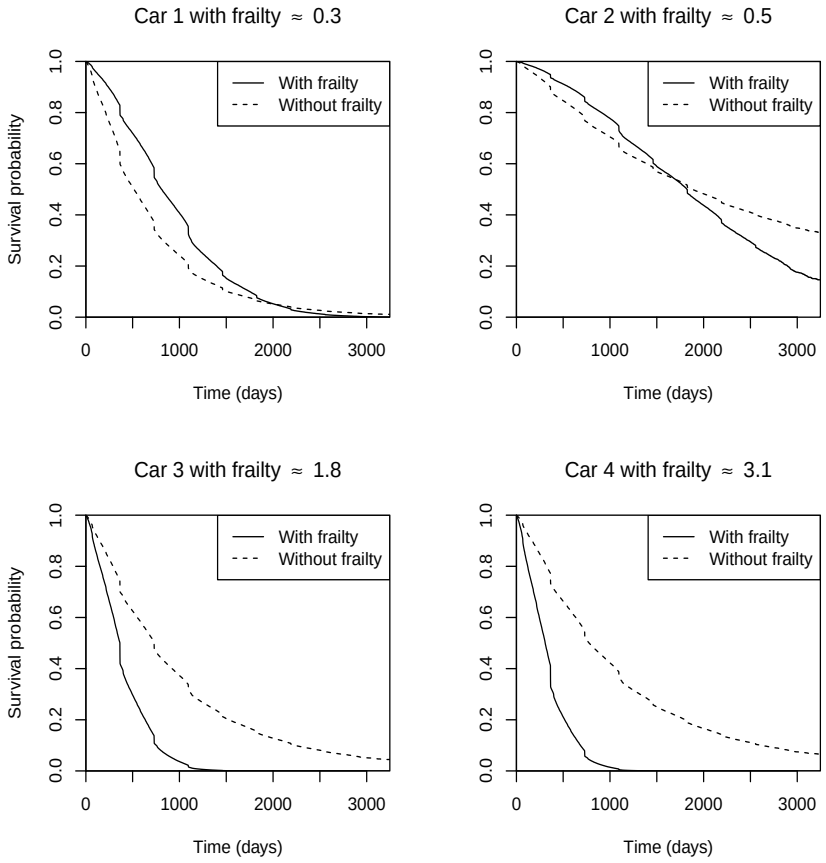

Figure 3: Individual survival curves of four cars from a Cox model with (solid line) and without (dashed line) frailty. Covariate values for the cars are specified in Table 3.

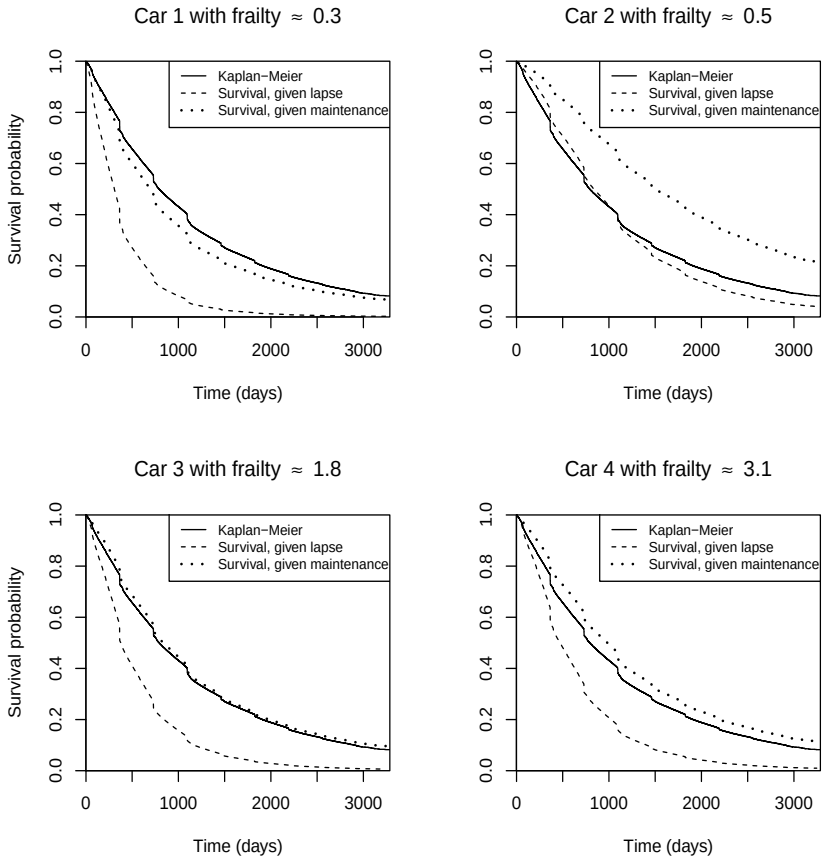

Figure 4: Conditional survival curves of four cars showing the effect a lapse has on survival. The Kaplan-Meier curve (solid line) is based on all data. The dashed line shows the survival of the specified car given that another car in the cluster has lapsed within one year. The dotted line shows the survival of the specified car given that another car in the cluster has maintained the policy the first year. Covariate values for the cars are specified in Table 3.