Stockholms
universitet

Estimating transfer probabilities for pension policies

Oskar Vedin

Masteruppsats i försäkringsmatematik
Master Thesis in Actuarial Mathematics

Masteruppsats 2019:9
Försäkringsmatematik
Juni 2019

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Master Thesis **2019:9**
http://www.math.su.se

Estimating transfer probabilities for pension policies

Oskar Vedin*

June 2019

Abstract

Many pension insurance policies in the Swedish life insurance market have the possibility to transfer the policy between companies. Since the insurance company typically have high acquisition costs for these policies a transferred policy out of the company could lead to a net loss on that policy. In this thesis we use logistic regression to estimate the probability of transfer for such policies on both simulated and real data. We find that we require a very large number of observations to get reliable estimates and that classifications measures like the ROC curve can give misleading results. We also see find that modeling continuous covariates from grouped observations, such as whole policy years, is preferable to using continuous observations.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: oskar.vedin@gmail.com. Supervisor: Filip Lindskog.

Acknowledgements

I would like give my deepest gratitudes to my supervisor Prof. Filip Lindskog at the Institute of Mathematics at Stockholm University for his guidance and patience. Without him this thesis would not have been possible.

I would also like to thank Chief Actuary Hanna Scherp as well as other colleagues at Movestic Livförsäkring AB for valuable insights and lessons, in this and other topics, during my time at the company.

In loving memory of my late father.

Contents

1 Introduction

In the Swedish life insurance market there exists mainly two different types of life insurance savings contracts, pension insurance policies and endowment policies. For both types of contracts the paid premiums gets invested in various assets or funds and the accumulated value of the policy, from here on referred to as the policy value, is later paid out to the policyholder. The premium can either be a one time premium paid at the beginning of the policy or a regular premium where premiums are paid repeatedly into the policy until the time of pay out. The premium payer can either be the same person as the policyholder or someone else. One common type of policy where the premium payer is different from the policyholder are occupational pension policies. Here the premium payer is the employer and the policyholder is the employee.

The main difference between the two types of contracts are related to how the policy value is paid out. For pension polices the start of pay out is dependent on the age of the policyholder, typically at the age 65 but earliest at the age of 55.[1] For endowment policies the time of pay out is not dependent on the policyholders age and the pay out starts according to the contract terms. How the policy value is paid out also differs between the two types of contracts. The value of a pension policy is paid out during at least five years and at most during the lifetime of the policyholder. For endowment policies the pay out can be a lump sum at the start of pay out or it can be continuously paid out during the lifetime of the policyholder.

For both types of policies there exists three different products, unit-linked insurance, equity-linked insurance and traditional life insurance. Depending on the type of product the paid premiums are invested differently. For unit-linked and equity-linked policies the policyholder chooses which funds or assets the paid premiums shall be invested in and the policy value at the time of pay out is the sum of paid premiums, investment returns and charged fees. For traditional life insurance policies the insurance company is responsible for the investments of the paid premiums. The policy value at the time of pay out for traditional policies is usually guaranteed to a percentage of the paid premiums, hence the insurance company bears all of the investment risk. Unit- and Equity-Linked policies do not typically have a guaranteed amount, so the policyholder bears all the investment risk which makes these policies very similar to a pure savings account.

The different contracts and products are illustrated in figure 1.

Both types of contracts usually include two different types of contractual options. We make the following definitions:

Free policy option. *The premium payer can at any time during the premium paying period of the contract choose to stop the premium payments, possibly with adjusted benefits.*

Applicable for both types of contracts.

Surrender option. *The policyholder can choose to withdraw all or part of the total policy value, possibly for a fee charged from the surrendered amount.*

Applicable only for endowment polices.

[1] As of the time of writing

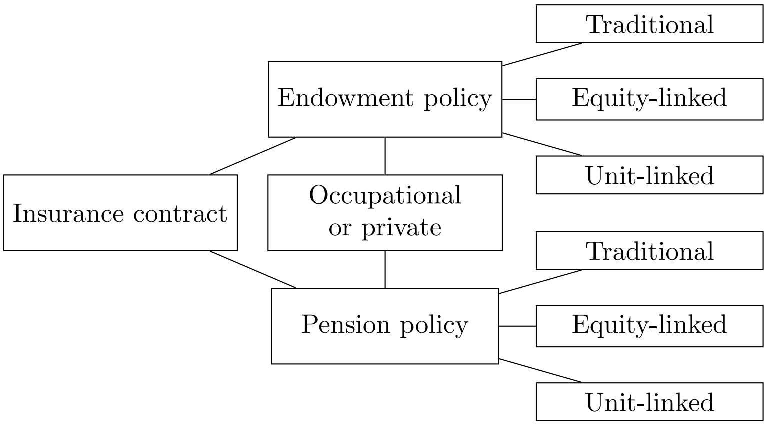


Figure 1: Illustration of the two types of life insurance contracts.

Transfer option. *The policyholder or premium payer can choose to transfer all of the policy value to a new pension policy within a different insurance company. Applicable only for pension polices.*

When the policyholder exercise the free policy option we say that the policy is paid up. Even though the free policy option is irrelevant for policies with a one time premium we shall treat such policies as paid up. A paid up policy is thus a policy where no future premiums will be paid. A paid up policy will still be paid out and the policyholder usually retains one of the other options mentioned above. A typical scenario where the free policy option is exercised is when an employee changes employer. The previous employer will then stop paying premiums into the old employees occupational pension policy. Since one employer often pay premiums into multiple occupational policies within the same insurance company, multiple occupational policies can therefore transfer simultaneously.

The time at which the policyholder can choose to exercise the surrender option is normally restricted to one year after the start date of the policy.

The fees associated with the two different lapse option above usually varies during the lifetime of the policy, with higher fees for younger polices and lower fees for older polices. Table 1 summarises the characteristics of the two different types of life insurance contracts.

Policy	**Premium**	**Pay out date**	**Pay out time**	**Free policy**	**Transfer**	**Surrender**
Pension	Regular or one time	Earliest at the age of 55	Minimum 5 years	Yes	Yes	No
Endowment	Regular or one time	Any	Any	Yes	No	Yes

Table 1: Contract characteristics

The characteristics described above are general and details can vary between companies and over time. Some of them, such as the start of pay out for pension policies, are regulated by law and have changed over time and will probably continue to change over the coming years. The purpose is not to exactly describe all possible aspects and terms of the contracts, but rather give the reader a context and examples of typical contracts and their, for this thesis, relevant features.

1.1 Risks associated with transfers and surrenders

The costs of acquiring new business for these types of contracts is usually very high for the insurance company and arises from both up-front provisions to brokers and internal costs. The insurance company expects to make a profit over the whole life time of these policies, from charges on premiums, policy value, etc. But it can take many years, or even decades before the policy turns to profit and has covered all of its initial costs. If the policy transfer, or surrender, before it has turned to profit then the insurance company will make a loss on that policy. The earlier the policy transfer the bigger the loss for the company. There exists two common methods to help reduce such losses. The transfer fee that is charged when a policy transfer will help cover some of the initial costs, but not all. These fees are often higher for young policies and lower for older policies. Another way to cover some of the initial costs is to have a clawback period for the initial provisions paid to the broker. This means that the broker has to pay back some of the initial provision if the policy transfer, or surrender within a specified time period from the start of the policy, like 5 years.

1.2 Solvency 2 and BEL

Insurance companies operation in the European Union have to value their liabilities according to the Solvency 2 directive EIOPA (2009) when calculating the solvency capital requirement (SCR). The liability towards the policyholders, the technical provisions, consists of two parts. The best estimate liability (BEL) and a risk margin. The BEL should be calculated as the discounted expected value of future cash flows. EIOPA (2015) states that

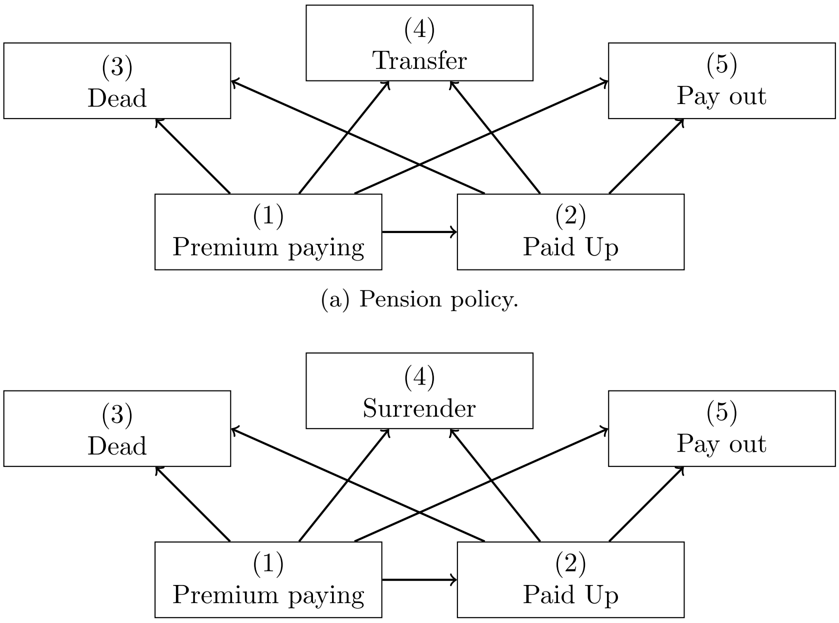
> *1.73. Insurance and reinsurance undertakings should explicitly take into account amounts charged to the policy holders relating to embedded options.*

Thus, the company has to make assumptions on the policyholders behaviour in regards to the transfer and surrender options.

In this thesis we will review the key principles of logistic regression to try and asses whether such models can be used to estimate transfer (or surrender) probabilities. The resulting models could later be used when deciding on the assumptions to be used in the cash flow model. In section 2 we will briefly review the mathematical theory of the methods we will use. We then explore these methods and fit models on both simulated and real data. Section 3 describes the used dataset and possible issues with it. In section 4 we present and compare the results from the different methods. In the last section we summarize our conclusion from section 4 and comment on the appropriateness of the methods.

2 Theory

An insurance policy of the types described in section 1 can be modeled as a multi-state Markov process where the policy jumps between different states, Alm, Andersson, Bahr, and Martin-Löf (2006, chapter 4). Different types of cash flows arises during the lifetime of the policy depending on the states and jumps of the policy. In state 1 premiums are paid into the policy, a jump from state 1 to 4 the policy value is either paid out or transferred from the insurance company etc. Figure 2 shows the different states of two simple models for typical policies where the arrows corresponds to possible jumps between states. The states 3, 4 and 5 are called absorbing states since the policy cannot jump out of these.



(a) Pension policy.

(b) Endowment policy.

Figure 2: Different states of simple models for typical pension and endowment policies.

The policy process $X(t)$ at time t take on values in the finite set J corresponding to the different states of the policy. The value of the policy process in the premium paying state at time $t = 3$ is thus $S(3) = 1$. To each of the possible jumps there exists, possible time dependent, jump intensities $\lambda_{ij}(t), i \neq j$ which describes the how the policy process jumps between different states in time. The intensity of a multi-state Markov process can be defined as

$$\lambda_{ij}(t) = \lim_{dt \to 0+} \frac{P(X(t+dt) = j | X(t) = i)}{dt}$$

This can be interpreted as the probability of jump from state i to state j in the time interval $[t, t + dt)$ when dt is small

$$P_{ij}(t) = P\left(X(t+dt) = j | X(t) = i\right) \approx \lambda_{ij}(t) dt$$

The probabilities $P_{ij}(s,t) = P(X(t) = j|X(s) = i)$ of jumping from state i to state j in the time interval $[s,t]$ are the solutions to the differential equation

$$-\frac{\partial P_{ij}(s,t)}{\partial s} = \sum_{k=1}^{5} \lambda_{ik}(t)\left(P_{kj}(s,t) - P_{ij}(s,t)\right)$$

which can be solved backwards in time with the boundary condition $P_{ij}(t,t) = \delta_{ij}$ for $s = t$, where $\delta_{ij} = 1$ if $i = j$ and zero otherwise, see Alm, Andersson, Bahr, and Martin-Löf (2006) for more details. In our case we are mostly interested in the jump probability into state 4 for some time interval (s,t)

$$P_{14}(s,t) \tag{1}$$

$$P_{24}(s,t) \tag{2}$$

Next we introduce an event time for state 4, $T_4 = \min\{t|X(t) = 4\}$. The distribution of the random variable T_4 is fully described by the jump intensities of the Markov-model.

$$P(T_4 > t) = S_4(t) = 1 - F_4(t) = 1 - P_4(0,t) = 1 - (P_{14}(0,t) + P_{24}(0,t))$$

Closely connected to the event time T_4 is the counting process $N_4(t)$ which take values in $\{0,1\}$ depending on the value of $X(t)$. When $X(t)$ jumps to state 4 at time $t = T_4$ the counting process $N_4(t)$ jumps from 0 to 1. The relationship between $N_4(t)$ and $X(t)$ is illustrated in figure 3. More details on modeling life insurance contracts as Markov-processes can be found in Alm, Andersson, Bahr, and Martin-Löf (2006, chapter 4).

Next we look at how the jump probabilities of the Markov-model relates to the expected net present value of future cash flows. Following Alm, Andersson, Bahr, and Martin-Löf (2006, p. 114-116, chapter 4.4) with some slight differences in notation, we introduce a time dependent payment flow for state i, $dB_i(t) = b_i(t)dt$ per unit of time. As an example of our simple model in figure 2a we have $b_i(t) = 0$ for states $i = 2,3,4$ corresponding to paid up, dead and transfer respectively. For the premium paying state $i = 1$ premiums are being paid and $b_1(t) > 0$. For the pay out state $i = 5$ we have $b_5(t) < 0$.

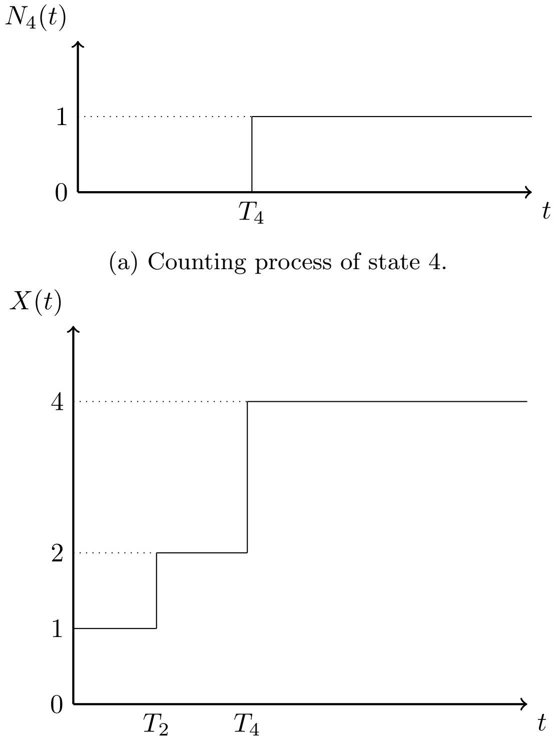
We also introduce payments $c_{ij}(t)$ when the policy process jumps from state i to j. Thus, when jumping from the premium paying state $i = 1$ or the paid out state $i = 2$ to the transferred state $i = 4$ would correspond to a pay out and $c_{ij}(t) < 0$. If we neglect the situation of a beneficiary in the case of death we only have two payments of type $c_{ij}(t)$, namely $c_{14}(t)$ and $c_{24}(t)$.

We also have the indicator variables $I_i(t) = I(X(t) = i)$, i.e. $I_i(t) = 1$ when $X(t) = i$ and zero otherwise. Let $N_{ij}(t)$ be the number of jumps from state i to j in the time interval $(0,t]$, $i \neq j$. The present value at time t of the future cash flows up to time T for one contract is thus

$$C(t,T) = \int_t^T d(t,u)\left(\sum_{j=1,5} I_j(u)b_j(u)du + \sum_{j=1,2} c_{j4}(u)dN_{j4}(u)\right) \tag{3}$$

where $d(t,u)$ is a discounting factor. The expected values of $I_j(t)$ and $dN_{jk}(t)$ conditioned on the state of the policy process is given by

$$E(I_j(t)|X(s) = i) = P_{ij}(s,t)$$



(a) Counting process of state 4.

(b) Policy process $X(t)$.

Figure 3: Relationship between the policy process and the counting process of state 4

$$E(dN_{jk}(t)|X(s)=i) = P_{ij}(s,t)\lambda_{jk}(t)dt \approx P_{ij}(s,t)P_{jk}(t,t+dt)$$

The expected value of the expression (3) conditioned on the state of the policy process at time t is thus

$$E(C(t,T)|X(t)=i) =$$
$$\int_t^T d(t,u)\left(\sum_{j=1,5} P_{ij}(t,u)b_j du + \sum_{j=1,2} P_{ij}(t,u)\lambda_{j4}(u)c_{j4}(u)du\right)$$

In practice one could approximate this expression by turning this integral to a sum for a suitable discretization and the intensities in the second sum would then be approximated by the probabilities in (1). We will now turn to models that can be used to make inference about the probabilities $P_{j4}(s,t)$ based on observations from the policy process $X(t)$.

2.1 Generalized Linear Models (GLM)

Generalized linear models are a generalization of the ordinary linear model

$$E[Y|X] = \mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

where the expected value of the response Y, conditioned on the value of the covariates $(x_1, \ldots, x_p)$, is linear. Y is assumed to be normally distributed with constant variance σ^2 in the linear model. By observing n outcomes of Y one can estimate the constants $(\beta_0, \ldots, \beta_p)$ by maximum likelihood estimation. This is done by setting the score function $U(\beta|Y,X)$ to 0 and solving for β.

$$U(\beta|Y,X) = \frac{\partial \log L(\beta|Y,X)}{\partial \beta} = \frac{\partial l(\beta|X,Y)}{\partial \beta}$$

GLM extends this ordinary linear model to responses of any distribution in the exponential dispersion family by using a link function $g()$ and the relationship

$$g\left(E[Y|X]\right) = g(\mu) = \eta = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

The ordinary linear model is thus a special case of the generalized linear models with identity link function $g(\mu) = \mu$. One commonly used link function is the natural logarithm

$$g(\mu) = \log(\mu) = \eta = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

The effects of the covariates $(x_1, \ldots, x_p)$ is thus transformed from additive on the linear predictor η to multiplicatively on $E[Y|X]$

$$\begin{aligned} E[Y|X] &= g^{-1}\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right) \\ &= \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right) \\ &= \exp\left(\beta_0\right)\exp\left(\beta_1 x_1\right)\cdots\exp\left(\beta_p x_p\right) \\ &= \gamma_0\gamma_1\cdots\gamma_p \end{aligned}$$

Another nice feature of the natural logarithm is that it maps values in the interval $(0, +\infty)$ to $(-\infty, \infty)$ which is crucial when dealing with responses in the interval $(0, \infty)$ such as counts. More details on generalized linear models can be found in McCullagh and Nelder (1989).

Generalized linear models are extensively used to model claim frequency and claim severity in non-life insurance. This can be done by using a Poisson distributed response Y_f for the number of claims, a gamma distributed response Y_s for the claim amount and the natural logarithm as link function. See Ohlsson and Johansson (2010) for more details on generalized linear models in non-life insurance.

For our purpose we would like to model the probabilities $P_{j4}(s,t)$ of a contract transferring by the usage of generalized linear models. Some examples of GLM used in life insurance are Renshaw and Haberman (1986), Briere-Giroux et al. (2010), Cerchiara, Edwards, and Gambini (2008) and Michorius (2011).

2.1.1 Logistic regression

Logistic regression can be used to model the probability $p_i = E[Y|X_i] = \mu_i$ of a Bernoulli distributed response Y with covariate values $X_i = (x_{i1}, \dots, x_{ip})^T$ by using a logit link function

$$g(p_i) = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{4}$$

$$\implies p_i = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}$$

In the standard logistic regression we thus assume a linear relationship between continuous covariates and the log-odds.

The score function and Fisher information is given by

$$U(\beta|Y, X_i) = \frac{\partial l}{\partial \beta_k} = \sum_{i=1}^{n} x_{ik}\left(y_i - \frac{\exp\left(\beta_0 + \sum_{m=1}^{p} \beta_m x_{im}\right)}{1 + \exp\left(\beta_0 + \sum_{m=1}^{p} \beta_m x_{im}\right)}\right)$$

the second order derivatives of the log-likelihood by

$$\frac{\partial^2 l}{\partial \beta_k \beta_l} = -\sum_{i=1}^{n} x_{ik} x_{il}\left(\frac{\exp\left(\beta_0 + \sum_{m=1}^{p} \beta_m x_{im}\right)}{\left(1 + \exp\left(\beta_0 + \sum_{m=1}^{p} \beta_m x_{im}\right)\right)^2}\right)$$

In matrix notation we write

$$\frac{\partial^2 l}{\partial \beta_k \beta_l} = -X_k W X_l^T$$

where

$$W(\beta) \quad = \quad \text{diag}\left[\frac{e^{\beta_0 + \sum_{m=1}^{p} \beta_m x_{1m}}}{\left(1 + e^{\beta_0 + \sum_{m=1}^{p} \beta_m x_{1m}}\right)^2}, \dots, \frac{e^{\beta_0 + \sum_{m=1}^{p} \beta_m x_{nm}}}{\left(1 + e^{\beta_0 + \sum_{m=1}^{p} \beta_m x_{nm}}\right)^2}\right]$$

and the Fisher information is thus

$$I(\beta) = -E\left[\frac{\partial^2 l}{\partial \beta_k \partial \beta_l}\right] = X_k W X_l^T$$

The diagonal matrix W is exactly the covariance matrix of the random vector $Y = (Y_1, \dots, Y_n)^T$ from which we observe outcomes $y = (y_1, \dots, y_n)^T$.

When n is large the estimates are approximately normally distributed $\hat{\beta} \overset{d}{\approx} N(\hat{\beta}, I^{-1}(\hat{\beta}))$, a fact which can be used to construct confidence intervals for the estimated parameters.

2.2 Generalized Additive Models (GAM)

The assumption of linearity for a continuous covariate in the generalized linear model might not always be satisfied in reality. One way to handle a continuous covariate is to divided it into k different subsets and thus transform it to a categorical covariate with k different levels and $k-1$ parameters to be estimated.

With this method one has to decide how to make the subdivision of intervals. Narrow intervals would get a better approximation of the underlying curve but the number of observations in these intervals can be small and the estimated parameters imprecise. Wide intervals would get accurate parameter estimates but a worse approximation of the underlying non-linear relationship.

Another way to deal with a continuous covariate is to transform the covariate x by some function $f(x)$ that more accurately approximates the relationship between the log-odds and the continuous covariate. In this section we will briefly introduce generalized additive models and cubic splines. This section is based on Ohlsson and Johansson (2010, chapter 5) where more details can be found.

In generalized additive models we no longer assume a linear relationship between the log-odds and the covariates, just an additive relationship

$$g(\mu_i) = \beta_0 + f_1(x_{i1}) + \cdots + f_p(x_{ip}) = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij})$$

between the mean and some functions f_j of the covariates x_{ij}, which makes GAM sort of an extension of the GLM. For a categorical covariate x_j with k levels we have (with level 1 as the reference level)

$$f_j(x_{ij}) = \beta_j^{(2)}\phi_j^{(2)}(x_{ij}) + \cdots + \beta_j^{(k)}\phi_j^{(k)}(x_{ij})$$

where $\phi_j^{(k)}(x_{ij}) = 1$ if $x_{ij} = k$, i.e. covariate x_{ij} is in category k and the function $f_j(x_{ij})$ is the same as in the GLM. The idea of GAM is to find a functions f_j that describes the relationship between continuous covariates and the log-odds. For this purpose we need a measurement of how well the model fits the data for given f_j, the measurement used is the deviance $D(y, \mu)$ (see section 2.3.1 for details). We also make the assumptions that the function f_j should be twice continuously differentiable and with low variation, in the sense that $\int (f_j''(x))^2 dx$ should be small. These requirements can be summarized as

$$\Delta(f_j) = D(y, \mu) + \lambda \int_a^b (f_j''(x))^2 dx, \qquad a \leq x \leq b \tag{5}$$

where the fit of the model is penalized by a high variation of f_j and the sought function minimizes this expression. The set of possible functions f_j that we will consider is called cubic splines.

A linear spline is constructed as follows. Given a set of knots $z_1, \ldots, z_k$ we define linear functions $p_m(x) = a_m + b_m x, \quad m = 1, \ldots, k-1$ with the condition that $p_{m-1}(z_m) = p_m(z_m)$, i.e. the linear functions are tied together at the endpoints. The linear spline $s(x)$ is then the resulting continuous function over the interval $[z_1, z_k]$.

$$s(x) = p_m(x), \quad z_m \leq x \leq z_{m+1};\, m = 1, \ldots k-1$$

For a quadratic spline we extend the functions $p_m(x)$ to be second degree polynomials and require that the derivatives of two connected polynomials should be the same at the internal knots, i.e. $p'_m(z_m) = p'_{m-1}(z_m), m = 2, \ldots, k-1$.

For cubic splines we have $p_m(x)$ as a third degree polynomial with the requirement that the derivative and the second derivative should be the same for connected polynomials at the internal knots, similar to the quadratic spline. We now have a twice continuous differentiable cubic spline $s(x)$ defined on the interval $z_1, \ldots, z_k$. Natural splines are splines that are extended to the interval $[a, b]$ and is linear in the interval $[a, z_1]$ and $[z_k, b]$. The reason for using natural cubic splines for the function f_j is that it minimizes the penalized deviance 5. Cubic splines are commonly parametrized as a linear combination of B-splines due to numerical properties, see Ohlsson and Johansson (2010, chapter 5). This makes interpretation of the parameters hard and instead we plot the fitted function for fixed values of the other covariates.

2.3 Model Selection

2.3.1 Deviance

One way of assessing how well a regression model fits the data is by using summary measures that are based on the residuals, $(y_i - \hat{y}_i)$. In logistic regression, the fitted values are calculated for each covariate pattern $(x_1, \ldots, x_p)$ and are compared to the observations for the same covariate pattern. In our case we would havey_j as the total number of observed transfers and $\hat{y}_j$ as the estimated number of transfers. The estimated number of transfers is calculated by $\hat{y}_j = m_j \hat{p}_j$ where m_j is the total number of observed policies for covariate pattern j and $\hat{p}_j$ is the estimated transfer probabilities for covariate pattern j. With J distinct covariate patterns we have $\sum_{j=1}^{J} m_j = n$, where n is the total number of observed policies.

One type of statistic that is used to asses the goodness-of-fit is the deviance

$$D(y, \hat{p}) = \sum_{j=1}^{J} d(y_j, \hat{p}_j)^2$$

where

$$d(y_j, \hat{p}_j) = \operatorname{sign}(y_j - \hat{y}_j) \left\{ 2 \left[y_j \log \left(\frac{y_j}{\hat{y}_j} \right) + (m_j - y_j) \log \left(\frac{m_j - y_j}{m_j - \hat{y}_j} \right) \right] \right\}^{1/2} \quad (6)$$

which is approximately χ^2-distributed with $J - p$ degrees of freedom when the suggested model have p non-redundant parameters. This statistic can thus be compared to the $\chi^2(J - p)$-distribution to test whether the model is significantly different from the saturated model. To test the significance of different covariates we can use the likelihood ratio test (LRT). Let H_f be a model with some covariates included and H_r a reduced model without one of the covariates in H_f. The likelihood ratio statistic is then $D(y, \hat{p}^{(r)}) - D(y, \hat{p}^{(f)})$ and is approximately χ^2-distributed with $f_f - f_r$ degrees of freedom where f_f is the number of non-redundant parameters of the full model and f_r is the number of non-redundant parameters of the reduced model.

The asymptotic distribution of the LRT statistic is valid by letting n go to infinity but keeping the number of parameters p in the larger model fixed. However, if the number of parameters increases with n then the χ^2-distribution is not accurate and tests based on this distribution is not valid, see Sur, Chen, and Candès (2017). In the presence of a continuous covariate we would have $J \approx n$ and since the number of parameters in the saturated model is equal to J the LRT statistic for testing one model against the saturated model is not χ^2-distributed. The LRT statistic for testing H_r against H_f is, however, χ^2-distributed since the likelihood of the saturated model cancels.

2.3.2 Hosmer-Lemeshow

Another type of statistical test that can be used for goodness-of-fit in the presence of continuous covariates or $J \approx n$is the Hosmer-Lemeshow tests. The Hosmer-Lemeshow test group the observations by their estimated probabilities, either by percentiles or fixed cut-points. Since the observations are no longer grouped by the covariate patterns we get c_k different covariate patterns in group k. The expected number of ones, e_{1k}, and zeros, e_{0k}, in group k are thus calculated as

$$e_{1k} = \sum_{j=1}^{c_k} \hat{y}_j$$

$$e_{0k} = \sum_{j=1}^{c_k} m_j(1 - \hat{p}_j)$$

and the observed ones o_{1k} and zeros o_{0k} in group k as

$$o_{1k} = \sum_{j=1}^{c_k} y_j$$

$$o_{0k} = \sum_{j=1}^{c_k} m_j - y_j$$

The statistic $\hat{C}$ can then be calculated as

$$\hat{C} = \sum_{k=1}^{g} \left[\frac{(o_{1k} - e_{1k})^2}{e_{1k}} + \frac{(o_{0k} - e_{0k})^2}{e_{0k}} \right]$$

where g is the number of groups. Hosmer and Lemeshow (1980) showed by simulations that the statistic $\hat{C}$ is approximately χ^2-distributed with $g - 2$ degrees of freedom when the model is correct and $J = n$. Hosmer, Lemeshow, and Sturdivant (2013) mention that it is likely that the approximation is valid also when $J \approx n$.

2.3.3 Classification

A third measurement that can serve as a compliment to the ones above is the ROC (Receiver Operating Characteristic) curve or more specifically the area under the ROC curve. We will briefly go trough the idea behind the ROC curve and refer the reader to Hosmer, Lemeshow, and Sturdivant (2013) for more details. We start by introducing the classification table.

A classification table is constructed as follows. Given a threshold t, we assign to observation i a one if the estimated probability p_i is greater than t or one zero if $p_i \leq t$. The new set of ones and zeros is then be compared to the actual observations and summarized in a table. Table 2 show a general classification table where c_{11} corresponds the total number of correctly classified ones, c_{01} the total number of misclassified ones, c_{10} the total number of misclassified zeros and c_{00} the total number of correctly classified zeros. e_i and o_i are the total number of estimated and observed ones and zeros, n is the total number of observations. Different ratios can be used to measure how good the model is at classifying the observations, one such ratio is the rate of correct classifications $(c_{11} + c_{00})/n$. Two other ratios that are commonly used are the sensitivity (true positive rate) c_{11}/o_1 and the specificity (true negative rate) c_{00}/o_0. The values of these ratios are of course dependent on the choice of t. In the extreme cases when $t = 0$ we would have sensitivity ≈ 1 and specificity ≈ 0 and in the other extreme case when $t = 1$ we would have sensitivity ≈ 0 and specificity ≈ 0. The ROC curve is constructed by plotting the sensitivity versus 1-specificity for cutpoints t in the interval $(0, 1)$. When sensitivity is high compared to 1-specificity the model is good at correctly classifying the observations. A model which is a great classifier would have a concave ROC curve and a model which is a bad classifier would have a ROC curve that is close to a straight line with a slope of 45 degrees. The area under the ROC curve can then be a measurement of how well the model is at classifying the outcome. See figure 4 for an illustration of a ROC curve.

	Observed		
Classified	Transfer = 1	Transfer = 0	Total
Transfer = 1	c_{11}	c_{10}	e_1
Transfer = 0	c_{01}	c_{00}	e_0
Total	o_1	o_0	n

Table 2: Classification table

How good the model is at classifying the observations depends on the distribution of the estimated probabilities for the two outcome groups, i.e. transfer and no transfer in our case. This is well illustrated in Hosmer, Lemeshow, and Sturdivant (2013, chapter 5.2.4) and we are content to mention that the model is bad at classifying the observations when these distributions are close and overlap.

The area under the ROC curve is a popular and intuitive measure to use when assessing the accuracy of a logistic model. However, depending on the purpose of the model the classification performance can be totally irrelevant, as it is in our case. To goal of our model is to correctly estimating the expected value of the binary variable and not the outcome of it.

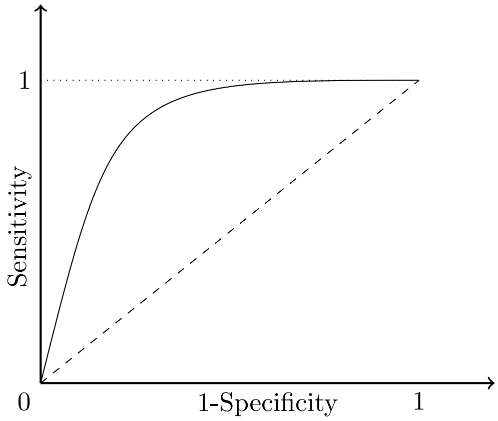


Figure 4: Illustration of ROC curve

3 Data

We will now use the above methods on both simulated and real data. We start by looking at simulated datasets and explore how different properties of the data affect the results of the logistic model.

From 2.1.1 we see that the effect of the covariates act multiplicatively on the odds of p_i

$$odds(p_i) = \frac{p_i}{1 - p_i} = \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})$$

The effect of categorical covariates will be presented on the odds scale while continuous modeled covariates will be plotted on the probability scale.

3.1 Simulations

In this section we will describe how the simulated datasets have been generated. The simulations have been made in two sections, one in which we generate Bernoulli distributed numbers as observations and one in which we generate exponentially distributed event times, a situation that more closely resembles the Markov model. These two sections will be referred to as the Bernoulli section and the Markov section respectively. The simulated data is inspired by the real data set from section 3.2 in terms of the transfer probabilities range and the effect of the policy year. The observed transfer frequencies from the real data lies somewhere in the range of 0% to 10%. We believe that the transfer frequencies of other Swedish companies lies roughly in the same interval. Transfer probabilities much higher then 10% seems unsustainable for a portfolio.

The Bernoulli section will be used to examine the following things

1. The fit of GLM vs. GAM.
2. The distribution of the LRT statistic.
3. The distribution of the Hosmer Lemeshow statistic.
4. Modeling the policy year as group observations vs. continuous observations.

And in the Markov section we will be looking at

1. The effect of competing risks

By competing risks we mean events that would prevent a transfer from happening. In other words, a jump to a state in the Markov-model from which there is no path to the transferred state. One such event could be death. This is also the situation we will be looking at.

The datasets will be generated from four different covariates, where one of them is continuous and represents the policy year. True values from which the data have been generated is presented below. The yearly probability of transfer for policy years 0-20 is determined from the graph in figure 5. The spike between year 5 and 5.5 represent corresponds to a decreased transfer fee and the end of the clawback period for initial provisions. The categorical covariates A, B and C are presented in table 3.

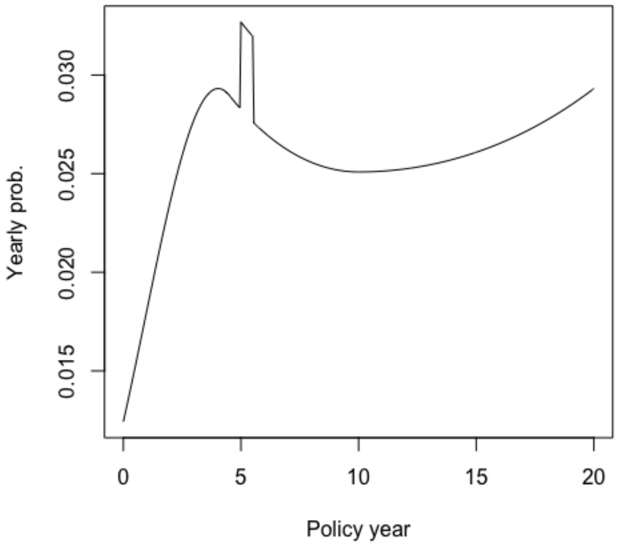


Figure 5: Yearly transfer probabilities by policy year

In the case of grouped policy years over the integers 0-20 we get $J = 21*3*3*2 = 378$ different covariate patterns. For each covariate pattern we generate m number of independent observations, equating to $n = 378m$ total number of observations. The total number of observations is thus evenly distributed over the different covariate patterns. This situation is a simplification and in reality the observations will probably be unevenly distributed over the different covariate patterns. We will be generating datasets for $m = 50, 100, 500, 1000, 2000$ which gives us datasets where $J < n$.

For the situation with policy year as continuous observations the observed policy years have been constructed such that the number of covariate patterns is the same as the total number of observations in the case with grouped policy years. For example, in the case of $m = 100$ above we get $21m$ number of observed policy years. The observed continuous policy years will be evenly distributed over the interval 0-20. We thus get $J = 21 * m * 3 * 3 * 2 = 378m$ different covariate patterns and generate 1 observations for each simulating the situation where $J = n$.

The distributions of the statistics have been obtained by simulating multiple datasets with same m and then fitting a GAM on each dataset.

In the Markov section we will only be looking at grouped policy years and $m = 1000$. For each observation we generate unique transfer times from the exponential distribution

$$f_i(x) = p_i e^{-p_i x}$$

with parameter p_i corresponding to the yearly transfer probability for covariate pattern i. We observe a transfer if the transfer time is less than or equal to 1.

In the case of competing risks we generate n unique times to death similar

Covariate	Level	Effect on odds
A	1	1.00
	2	1.10
	3	1.30
B	1	1.00
	2	1.30
	3	1.00
C	1	1.00
	2	1.00

Table 3: True value of categorical covariates for simulated data

to above but now with the parameter corresponding to the predicted one year mortality rate from DUS14 (Alm, Andersson, Bahr, Bergström, et al. (2014)) for men aged 30,40 and 50 in year 2020, see table 6. These survival times are then randomly assigned to the observations and a transfer is observed if the transfer time is less than the time to death and less than or equal to 1. This situation is illustrated in table 5 row 6.

Extracts of simulated data for each situation is presented in table 4 and 5.

Policy year	A	B	C	transfer
3	2	2	1	0
8	2	3	1	0
9	1	2	2	1
7	1	2	1	0
5	2	3	1	0
16	3	1	1	0
⋮	⋮	⋮	⋮	⋮

(a) Grouped policy years

Policy year	A	B	C	transfer
6.98	2	2	2	0
5.16	2	2	2	1
9.57	3	2	1	0
10.10	1	3	1	0
4.01	3	2	2	1
17.76	3	2	1	0
⋮	⋮	⋮	⋮	⋮

(b) Continuous policy years

Table 4: Extract from simulated data used in the Bernoulli section.

Policy year	A	B	C	Time to transfer	Time to death	Transfer w/o death	Transfer w/ death
16	1	2	2	19.01	382.32	0.00	0.00
10	2	2	1	13.36	1059.89	0.00	0.00
19	1	3	1	0.18	1105.54	1.00	1.00
0	1	3	2	17.49	1389.46	0.00	0.00
5	2	1	1	15.52	571.91	0.00	0.00
14	1	2	1	0.84	0.13	1.00	0.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 5: Extract from simulated data used in the Markov section

3.2 Real data

Next we fit a logistic regression model to a real dataset. The data consists of roughly 3 million monthly observations over 3 years with the 5 following possible covariates. Three of the covariates have been due to confidentiality.

Age	Probability (%)
30	0.035
40	0.075
50	0.205

Table 6: Expected one year mortality probability for men year 2020

1. Policy year: grouped observations of whole policy years in the range 0-13
2. Transfer fee: 2 levels
3. A: 2 levels
4. B: 3 levels
5. C: 2 levels

The reason for choosing monthly observations instead of yearly observations is to minimize the effect time varying covariates could have on the result.

The policy year will be modeled as a continuous covariate while the other covariates will be modeled as categorical. During 2018 the transfer fees was lowered and the categorical covariate Transfer fee is an indicator equal to 1 if the observations had the new lower transfer fee and zero otherwise.

4 Results

4.1 Simulations

4.1.1 Bernoulli

We start by simulating data with grouped policy years. To account for the spike at policy year 5 we have added an extra covariate in our model called "Pol. year 5" which is 1 if the policy year is in the interval $[5, 5.5]$. The estimated effects of the categorical covariates for $m = \{50, 500, 2000\}$ are presented in table 7, both for the GLM and the GAM. For the policy year we plot the fitted probabilities when covariates A=1, B=1 and C=1. The results are presented in figures **??**.

From the fitted models we make the following observations. Firstly that the estimated effects of the covariates A, B and C are extremely close between the GLM and GAM. The numbers in 7 are rounded and in reality there are small differences between the GLM and the GAM. However, for the covariate Pol. year 5 there are significant differences for $m > 50$ between GLM and GAM, with GAM being superior. Looking at the fitted probabilities for the policy year we see that the GLM fit roughly a straight line to a non-linear relationship while the GAM more accurately captures the effect of the policy year for large m It is clear that the GAM is superior when we are dealing with a non-linear relationship of one continuous covariate. We require between $m = 1000$ and $m = 2000$ for an acceptable fit of the continuous covariate. So, to get accurate estimates of the true probabilities we require many observations for each covariate. If we have the desired amount of observations the GAM clearly outperforms the GLM.

Next we look at testing the main effects of the model. This is done by the likelihood ratio test (LRT) of two hierarchical models, one model with all main effect and one model without one of the main effects. Our null is

$$H_0 : \text{No difference between the models}$$

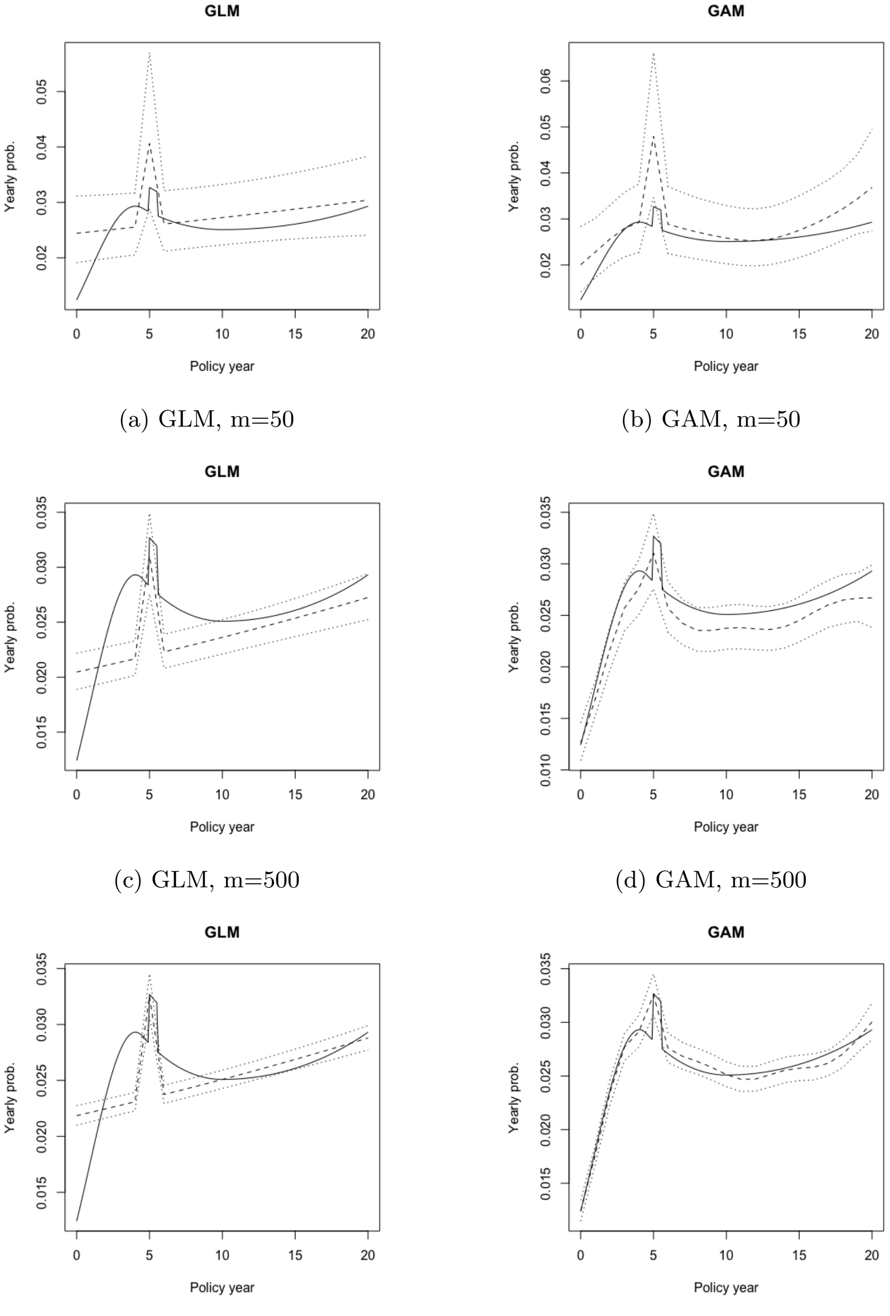
and the test statistic is χ^2-distributed under H_0. If the value of the test statistic is greater than the 0.95 quantile of the χ^2-distribution we reject the null and keep the main effect in our model. The resulting p-values are presented in tables 8. We see that that we require $m = 1000$ to keep all the significant effects for the GAM and $m = 500$ for the GLM at level $\alpha = 0.05$.

However, since the test statistic is a random variable it possible that we get another result if running the simulation again. But worth noting is that for $m = 50$ both models fail to reject the false null hypothesis. Looking at the distributions of the LRT statistic in figure 7 we see that for covariate A it does not follow the χ^2 while for covariate C it does, as expected. However, for small m the LRT statistic take values less then the 0.95 quantile of the χ^2 when testing covariate A, thus leading to the type II error.

Next we look at the situation where we observed the policy year as a continuous covariate and $J = n$. The fitted probabilities for covariates A=1, B=1 and C=1 are presented in figure 10. Again, we require many observations for a decent fit and the GAM is still superior to the GLM.

If we now look at the distribution of the LRT statistic when testing covariate C in figure 9 we see that it is still approximately χ^2-distributed, which is consistent with the statement in the end of section 2.3.1.

Instead of the LRT, we can use the Hosmer Lemeshow statistic to test if the full model is significantly different from the saturated model. This statistic



(a) GLM, m=50

(b) GAM, m=50

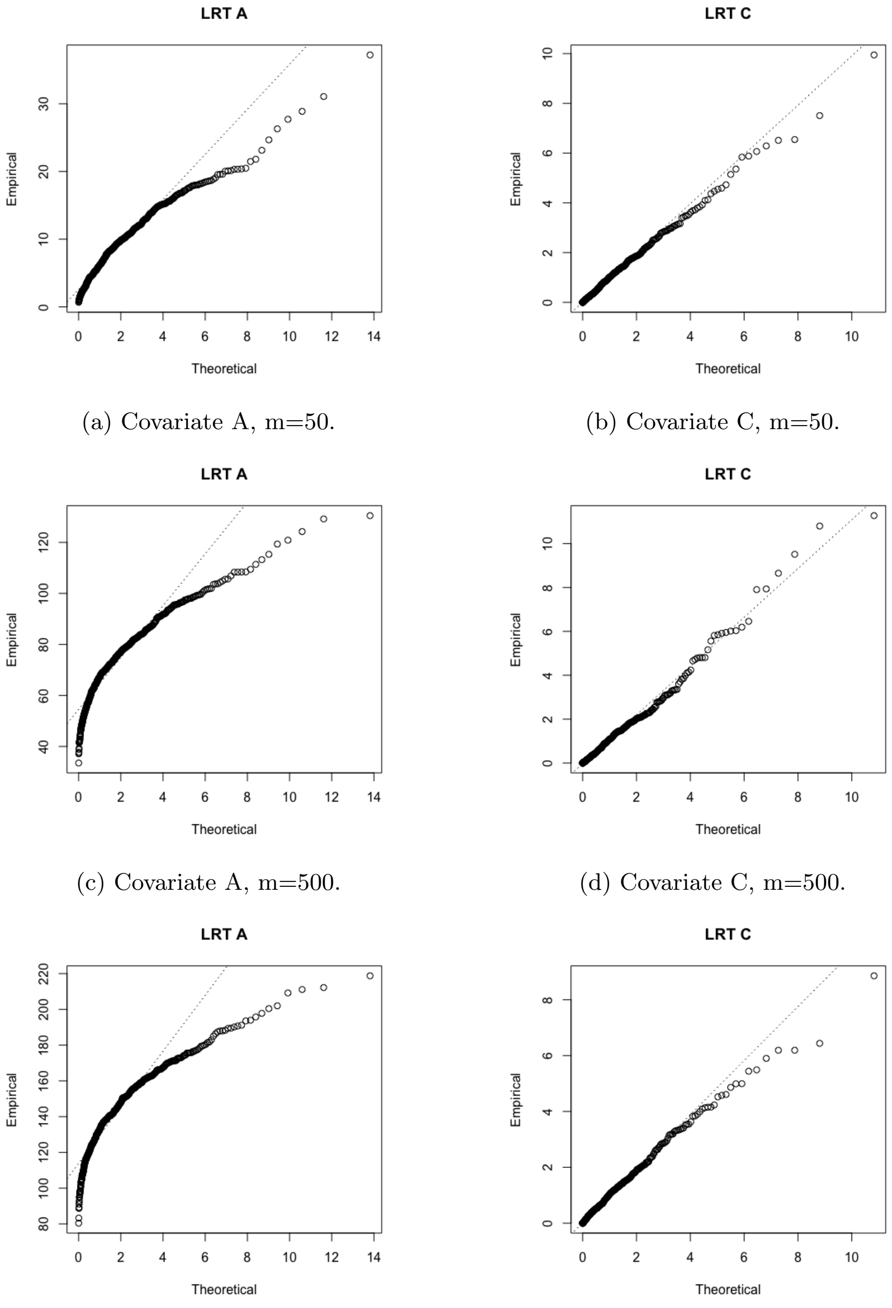
(c) GLM, m=500

(d) GAM, m=500

(e) GLM, m=2000

(f) GAM, m=2000

Figure 6: Estimated transfer probabilities for A=1, B=1 and C=1 using grouped policy years. GLM (*left*) vs. GAM (*right*). *Solid: true transfer probability, dashed: estimated transfer probability, dotted: confidence intervals of the estimates*



(a) Covariate A, m=50.

(b) Covariate C, m=50.

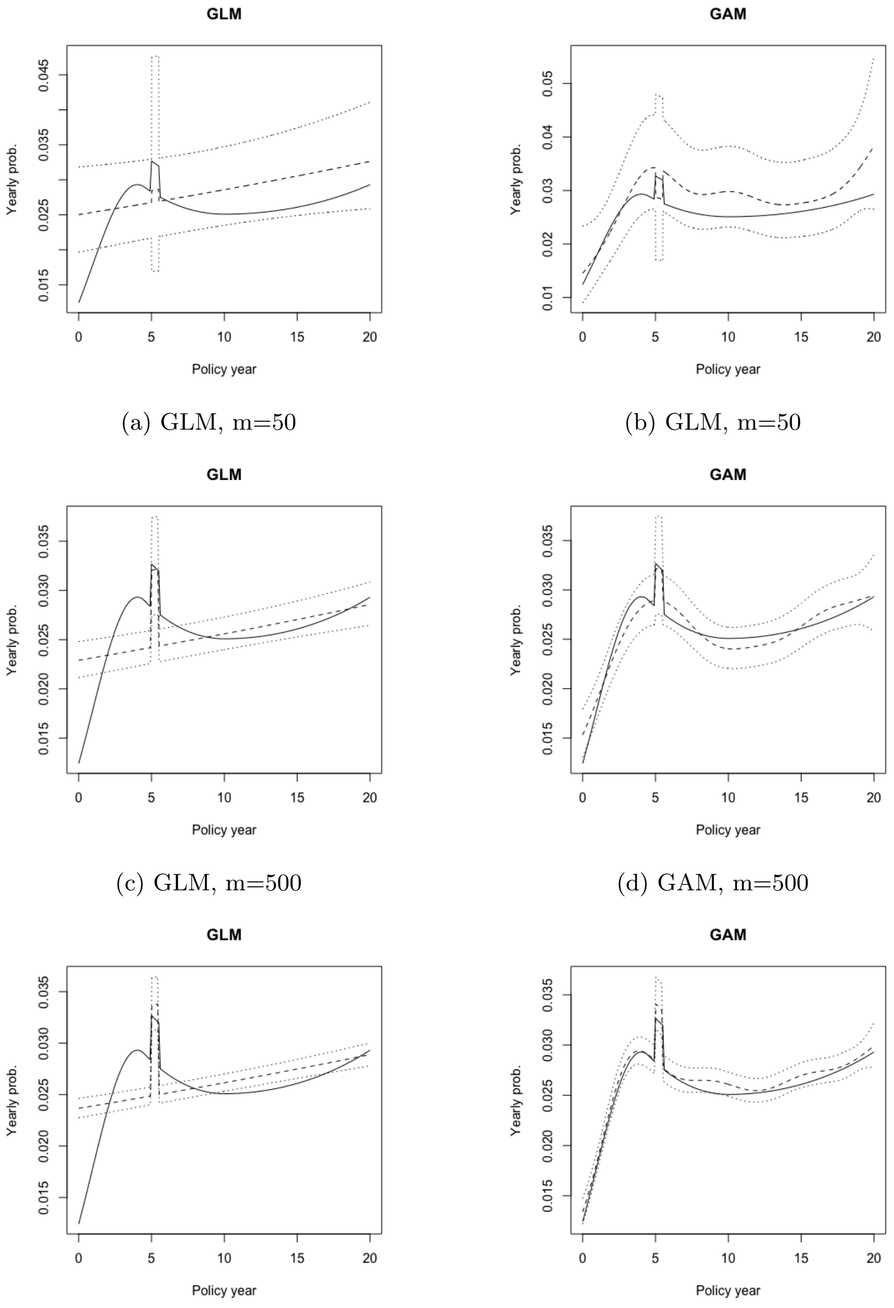
(c) Covariate A, m=500.

(d) Covariate C, m=500.

(e) Covariate A, m=1000.

(f) Covariate C, m=1000.

Figure 7: QQ-plots of the LRT statistic vs. χ^2 when testing the covariates A and C in GAM with grouped policy years.



(a) GLM, m=50

(b) GLM, m=50

(c) GLM, m=500

(d) GAM, m=500

(e) GAM, m=2000

(f) GAM, m=2000

Figure 8: Estimated transfer probabilities for A=1, B=1 and C=1 using continuous policy years. GLM (*left*) vs. GAM (*right*) $m = 50$. *Solid: true transfer probability, dashed: fitted transfer probability, dotted: confidence intervals*

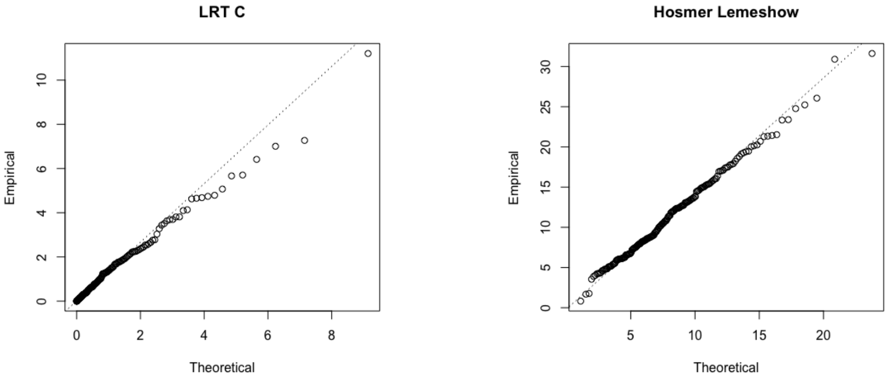


Figure 9: *Left:* QQ-plot of LRT statistic vs. χ^2 when testing covariate C for continuous policy year. *Right:* QQ-plot of Hosmer Lemeshow statistic for continuous policy year.

is supposed to be χ^2-distributed and is confirmed by figure 9. Some criticism has been raised towards this test since the choice of bins is arbitrarily and different choices can lead to different results. Testing two different GAM, one with covariate B and one without covariate B, against the saturated model using the Hosmer Lemeshow statistic for $m = 1000$ resulted in not rejecting

$$H_0 : \text{No difference between the saturated model and the proposed model}$$

in both cases. Thus, both models was determined to explain the data as well as the saturated model even though dropping B should have had a significant effect.

The ROC curves for the case with grouped policy years are displayed in figure 10. The are under the curve is roughly 0.56 for all m indicating that the model is bad at classification. However, from the fitted probabilities and effects we know that our model very accurately estimates the true probabilities of transfer when m is large. This measure is thus totally irrelevant in this setting .

4.1.2 Markov

We start by consider the situation where only transfer is possible. The estimated effects are presented in table 9 and fitted probabilities for policy year with other covariates equal to 1 in figure 11. The results are similar to the Bernoulli section with grouped policy years and $m = 1000$. Next we add the possibility of death using DUS14 as mortality assumptions. The bias in the estimated effect of the categorical covariates are presented in table 10 and the bias in the estimated probabilities per policy year for A=1, B=1 and C=1 in figure 12a. The bias is calculated as the differences between the estimates with and without the time to death. As we can see the bias is extremely small and is due to the low probability of death for these ages. Increasing the probability of death, or rather the probability of a competing risk, introduces significant bias in the estimates. This is seen in figure 12b.

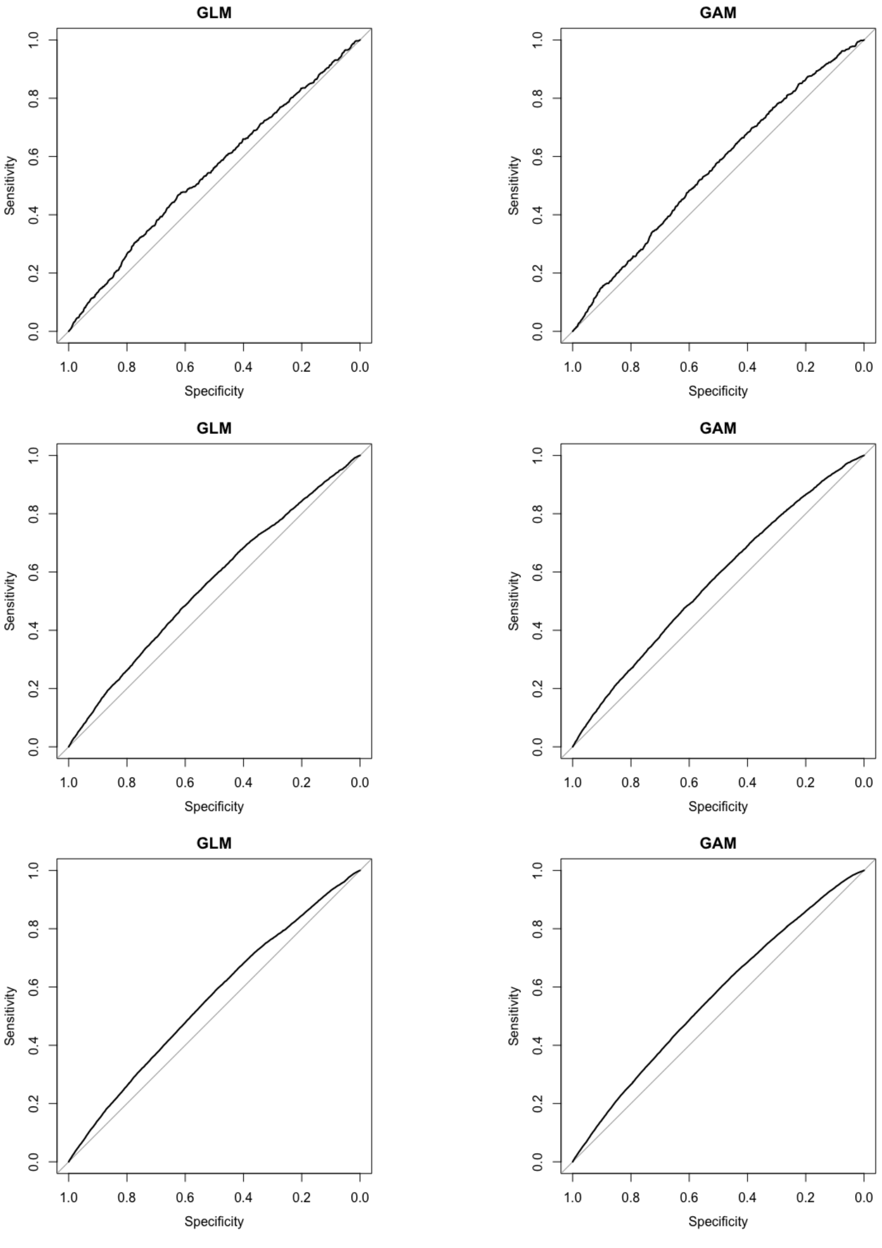


Figure 10: ROC curve using grouped policy years. GLM (*left*), GAM (*right*). The area under each curve is approximately 0.56. *Top: m=50, middle: m=500, bottom: m=2000*

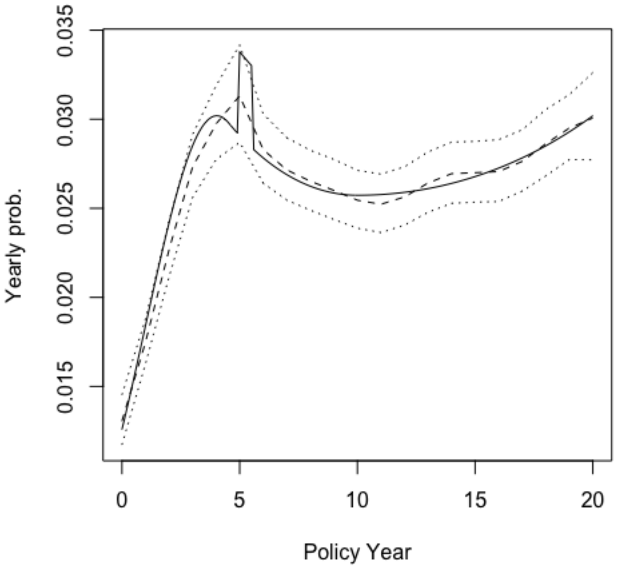
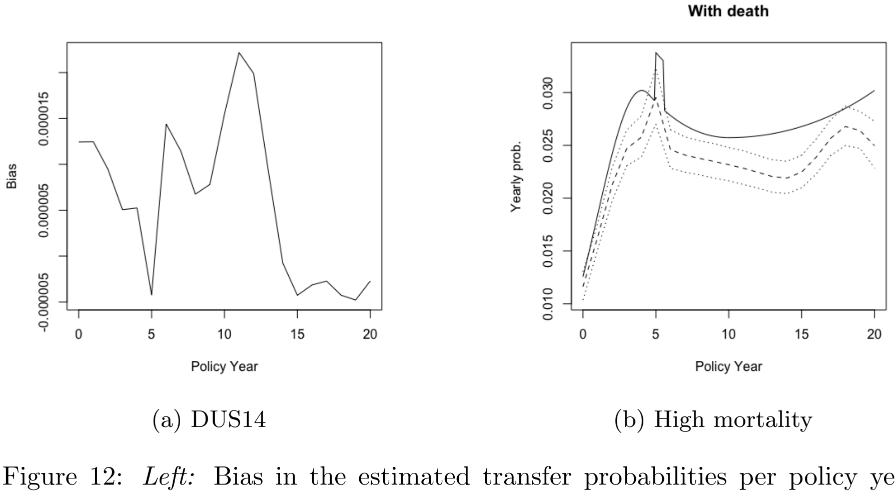


Figure 11: Estimated probabilities using GAM in Markov model. *Solid: true transfer probability, dashed: estimated transfer probability, dotted: confidence intervals*



(a) DUS14

(b) High mortality

Figure 12: *Left:* Bias in the estimated transfer probabilities per policy year when A=1, B=1 and C=1 using DUS14 as mortality assumption. *Right:* Estimated transfer probabilities per policy when A=1, B=1 and C=1 using high mortality probability (*solid: true transfer probabilities, dashed: estimated transfer probabilities, dotted: confidence intervals*).

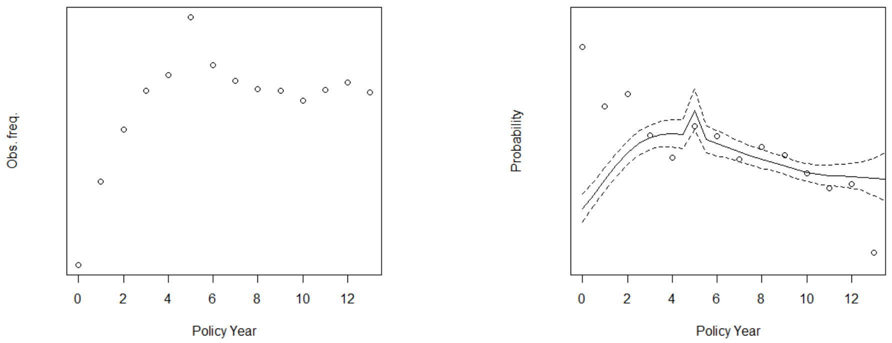


Figure 13: *Left:* Observed monthly transfer frequency for all observations per policy year. *Right:* Estimated monthly transfer probabilities for the reference group.

We conclude that the effect of mortality is negligible because of the relatively low probability of death for people in ages 20-60. In the presence of other competing risks with high probability the bias in the estimates will be significant. However, we can not think on any other competing risk that could prevent us from observing a transfer.

4.2 Real data

Given the results from the previous section we decide to use grouped policy years when fitting our model. The policy years represents whole policy years and the observed transfer frequencies over all observations are presented in figure 13. We fit a GAM with all covariates in the model and test the main effects using LRT. All main effects are significant and we keep all covariates in the model. The estimated effects of the categorical covariate are presented in table 11. Estimated probabilities by policy year for the reference group are presented in figure 13. From figure 13 it is clear that the estimated probabilities do not fit the observations well. This is an indication that the model would benefit from adding an interaction term. The relationship between the transfer probability and the policy year appears to be affected by the value of at least one other covariate. We also note that there is no spike at policy year 5 in the right plot of figure 13 and that the effect of the second level of covariate B is not significantly different from 1 since the confidence intervals contains 1. However, it could be reasonable to assume that the spike at year 5 should be included in the model even if it would turn out non-significant in a model with interactions.

The next step would probably be to add all two way interactions and test their significance.

			GLM			GAM		
Covariate	Level	True	Estimate	Lower	Upper	Estimate	Lower	Upper
A	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.10	0.97	0.78	1.19	0.97	0.78	1.19
	3	1.30	1.30	1.07	1.58	1.30	1.07	1.58
B	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.30	1.42	1.17	1.74	1.42	1.17	1.74
	3	1.00	1.16	0.94	1.43	1.16	0.94	1.43
C	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.00	0.96	0.82	1.13	0.96	0.82	1.13
Pol. year 5	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1	1.16	1.45	1.02	2.06	1.45	1.02	2.06

Covariate	Level	True	Estimate	Lower	Upper	Estimate	Lower	Upper
A	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.10	1.10	1.03	1.18	1.10	1.03	1.18
	3	1.30	1.25	1.18	1.34	1.25	1.18	1.34
B	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.30	1.32	1.24	1.40	1.32	1.24	1.40
	3	1.00	1.06	0.99	1.13	1.06	0.99	1.13
C	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.00	1.01	0.96	1.06	1.01	0.96	1.06
Pol. year 5	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1	1.16	1.41	1.26	1.57	1.13	0.99	1.29

Covariate	Level	True	Estimate	Lower	Upper	Estimate	Lower	Upper
A	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.10	1.11	1.07	1.15	1.11	1.07	1.15
	3	1.30	1.30	1.26	1.34	1.30	1.26	1.34
B	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.30	1.30	1.26	1.34	1.30	1.26	1.34
	3	1.00	1.01	0.98	1.04	1.01	0.98	1.04
C	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.00	1.00	0.97	1.02	1.00	0.97	1.02
Pol. year 5	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1	1.16	1.33	1.26	1.41	1.09	1.01	1.17

Table 7: Estimated effects of the categorical covariates with grouped policy years. *Top: m=50, middle: m=500, bottom: m=2000.*

Main effect	$m = 50$	$m = 100$	$m = 500$	$m = 1000$	$m = 2000$
A	0.279	0.000	0.000	0.000	0.000
B	0.029	0.000	0.000	0.000	0.000
C	0.533	0.091	0.532	0.318	0.618
Pol. year 5	0.486	0.553	0.000	0.000	0.000
Policy year	0.003	0.074	0.000	0.000	0.000

Main effect	$m = 50$	$m = 100$	$m = 500$	$m = 1000$	$m = 2000$
A	0.278	0.000	0.000	0.000	0.000
B	0.028	0.000	0.000	0.000	0.000
C	0.532	0.091	0.532	0.318	0.618
Pol. year 5	0.964	0.809	0.542	0.004	0.000
Policy year	0.011	0.000	0.000	0.000	0.000

Table 8: p-values LRT of main effects in GLM (*upper*) and GAM (*lower*) with grouped policy years.

Covariate	Level	True	Estimate	Lower	Upper
A	1	1.00	1.00	1.00	1.00
	2	1.10	1.08	1.03	1.13
	3	1.30	1.29	1.23	1.35
B	1	1.00	1.00	1.00	1.00
	2	1.30	1.33	1.27	1.39
	3	1.00	1.02	0.97	1.07
C	1	1.00	1.00	1.00	1.00
	2	1.00	0.99	0.95	1.02
Policy Year 5	0	1.00	1.00	1.00	1.00
	1	1.16	1.08	0.98	1.19

Table 9: Estimated effects using GAM in Markov model with $m = 1000$.

Covariate	Level	True	Estimate w/o death	Estimate w/ death	Bias
A	1	1.00000	1.00000	1.00000	0.00000
	2	1.10000	1.12557	1.12527	0.00030
	3	1.30000	1.32305	1.32183	0.00122
B	1	1.00000	1.00000	1.00000	0.00000
	2	1.30000	1.22412	1.22487	-0.00075
	3	1.00000	1.00028	1.00084	-0.00056
C	1	1.00000	1.00000	1.00000	0.00000
	2	1.00000	1.00754	1.00737	0.00017
Policy Year 5	0	1.00000	1.00000	1.00000	0.00000
	1	1.16183	1.17738	1.17537	0.00201

Table 10: Bias in estimates when using DUS14

Covariate	Estimate	Lower	Upper
Pol. year 5	1.00	1.00	1.00
	1.14	1.02	1.29
Transfer fee	1.00	1.00	1.00
	1.21	1.15	1.27
A	1.00	1.00	1.00
	0.71	0.67	0.75
B	1.00	1.00	1.00
	1.01	0.94	1.09
	1.23	1.13	1.35
C	1.00	1.00	1.00
	0.57	0.53	0.61

Table 11: Estimated effects of categorical covariates

5 Conclusions

From the simulated data we have seen that when using logistic regression to estimate transfer, or other rare events, probabilities we require a very large number of observations for reliable estimates. And in the presence of a continuous covariate with non-linear effect the GAM is preferable to the GLM because of its superior fit.

We prefer to use grouped continuous covariates when building the model because we prefer the LRT instead of the Hosmer Lemeshow when testing one model against the saturated model. The estimated probabilities when using continuous policy years are not superior to the estimates using grouped policy years. Thus, we see no advantage of using continuous policy years in the model.

The mortality in the Markov model does not have an effect on the estimates in the logistic model because of the small probability of dying in ages 20-60. However, in the presence of other competing risks with higher probabilities the estimates will have a negative bias because we observe less outcomes of the event under study, i.e. transfers.

Since the purpose of our model is to find accurate estimates of the transfer probabilities and not to predict the outcome of the Bernoulli variable Y, we conclude that classification measures of the model is irrelevant for our purpose.

On the real data we fitted a simple GAM that displayed a bad fit for the continuous covariate compared to observed outcomes. This is likely due to the fact that the relationship between the transfer probability and the policy year is affected by the value of other covariates. An interaction term would probably benefit the model.

Overall we feel that logistic regression can be used to model transfer probabilities. However, we think that the biggest limitation of the model is that we require such large number of observations for reliable estimates. Having large n (million) and many covariate could in practice probably lead to heavy computations, especially using GAM and fitting multiple splines.

One situation we mentioned briefly in section 1 is that the employer can choose to transfer multiple occupational pension policies simultaneously. This would of course mean that some of the observed are dependent, a clear violation of the underlying assumption of the GLM. However, if the number of observations that are dependent is relatively low compared to the total number of observations we believe that the effect is negligible.

References

Alm, Erik, Gunnar Andersson, Bengt von Bahr, Rikard Bergström, and Åsa Larson (2014). *Försäkrade i Sverige, livslängder och dödlighet, prognoser 2014-2070, baserade på data 2001-2012.* Svensk Försäkring. URL: `https://www.svenskforsakring.se/globalassets/rapporter/livslangder-och-dodlighet/112679_forsakrade_i_sverige_web.pdf`.

Alm, Erik, Gunnar Andersson, Bengt von Bahr, and Anders Martin-Löf (2006). *Livförsäkringsmatematik II.* Svenska försäkringsföreningen.

Briere-Giroux, Guillaume, Jean-Felix Huet, Robert Spaul, Andy Staudt, and David Weinsier (2010). *Predictive Modeling for Life Insurers; Application of Predictive Modeling Techniques in Measuring Policyholder Behavior in Variable Annuity Contracts.* Tower Watson.

Cerchiara, Rocco Roberto, Matthew Edwards, and Alessandra Gambini (2008). *GENERALIZED LINEAR MODELS IN LIFE INSURANCE: DECREMENTS AND RISK FACTOR ANALYSIS UNDER SOLVENCY II.* URL: `http://www.actuaries.org/AFIR/Colloquia/Rome2/Cerchiara_Edwards_Gambini.pdf`.

EIOPA (2009). "DIRECTIVE 2009/138/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 25 November 2009 on the taking-up and pursuit of the business of Insurance and Reinsurance (Solvency II) (recast)". In: *Official Journal of the European Union.*

— (2015). *Guidelines on the valuation of technical provisions.* URL: `https://eiopa.europa.eu/Publications/Guidelines/TP_Final_document_EN.pdf`.

Hosmer, David W. and Stanley. Lemeshow (1980). "Goodness of fit tests for the multiple logistic regression model". In: *Communications in Statistics - Theory and Methods* 9.10, pp. 1043–1069.

Hosmer, David W., Stanley. Lemeshow, and Rodney X. Sturdivant (2013). *Applied logistic regression.* Wiley.

McCullagh, P. and J.A. Nelder (1989). *Generalized Linear Models, Second Edition.* Chapman & Hall.

Michorius, Cas Z. (2011). "MODELING LAPSE RATES Investigating the Variables that Drive Lapse Rates". MA thesis. University of Twente.

Ohlsson, Esbjörn and Björn Johansson (2010). *Non-Life Insurance Pricing with Generalized Linear Models.* Springer Berlin Heidelberg.

Renshaw, A. E. and S. Haberman (1986). "Statistical analysis of life assurance lapses." In: *Journal of the Institute of Actuaries* 113.3, pp. 459–497.

Sur, Pragya, Yuxin Chen, and Emmanuel J. Candès (2017). "The Likelihood Ratio Test in High-Dimensional Logistic Regression Is Asymptotically a Rescaled Chi-Square". In: URL: `https://statweb.stanford.edu/~candes/papers/LRT.pdf`.