Université Claude Bernard – Lyon 1
INSTITUT DE SCIENCE FINANCIERE ET D'ASSURANCES

Mémoire présenté devant l'Institut de Science Financière et d'Assurances pour l'obtention du diplôme d'Actuaire de l'Université de Lyon

le 11 juillet 2011

Par : Christophe Dutang
Titre : Regression models of price elasticity in non-life insurance

Confidentialité : Oui (2 ans)

Membre du Jury I.A. :
M. Pierre THEROND

Membre du Jury I.S.F.A. :
M. Jean Claude AUGROS
M. Alexis BIENVENÜE
M. Areski COUSIN
Mme Diana DOROBANTU
Mme Anne EYRAUD-LOISEL
M. Stéphane LOISEL
Melle Esterina MASIELLO
Mme Véronique MAUME-DESCHAMPS
M. Frédéric PLANCHET
M. François QUITTARD-PINON
Mme Béatrice REY-FOURNIER
M. Christian-Yann ROBERT
M. Didier RULLIERE

Secrétariat :
Mme Marie-Claude MOUCHON

Bibliothèque :
Mme Michèle SONNIER

Entreprise :
AXA Group Risk Management

Directeur de Mémoire en entreprise :
M. Gilles HORNECKER

Invité :

Autorisation de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise

Signature du candidat

50 avenue Tony Garnier
69366 LYON Cedex 07

Abstract

Price elasticity studies analyze the effect of premium changes on customer behavior. In this memoir, we focus on its effect on the renewal of non-life insurance contracts. Methodologies developped can also be applied on new business. Every year, insurers face the recurring question of adjusting new prices. Where is the trade off between increasing premium to favour higher projected profit margins and decreasing premiums to obtain a greater market share?

This memoir aims to determine the price sensitiveness of a non life insurance portfolio taken into account individual policy features. Three markets, namely Portugal, Québec and Germany, are studied and compared. They reveal to be strongly different both in terms of insurance covers and distribution channels: two main factors of price elasticity. Three regression models have been used and compared: Generalized Linear Models, Generalized Additive Models and Survival Regression Models.

Keywords : *price elasticity*; *non-life insurance*; *regression modelling*; *generalized linear models.*

Résumé

L'élasticité prix consiste à étudier l'effet d'un changement prix sur le comportement du client. Dans ce mémoire, nous étudions l'élasticité dans le cadre de renouvellement de contrat d'assurance non-vie. Cependant, les méthodologies peuvent être aussi utilisées pour l'élasticité prix des affaires nouvelles. Chaque année, les assureurs font face à un dilemme pour établir les prix : soit augmenter soit diminuer les primes, qui impacte logiquement le profit espéré et la taille du portefeuille. Par conséquent, un compromis doit être trouvé.

Ce mémoire a pour but de déterminer la sensibilité au prix d'un portefeuille d'assurance non-vie, en tenant compte des caractéristiques individuelles des polices le constituant. Trois marché d'assurance vont être étudiés et comparés, à savoir le Portugal, le Québec et l'Allemagne. Ils vont se réveler complètement différents, aussi bien en terme de couvertures que de canaux de distributions. Trois modèles de régression vont être comparés : les modèles linéaires généralisés, les modèles additifs généralisés et les modèles de survie de régression.

Mots-clés : *élasticité prix*, *assurance non-vie*, *modèles de régression*, *modèles linéaires généralisés.*

Acknowledgements

Firstly, I would like to gratefully acknowledge Guillaume Gorge and Emmanuel Pierron, who propose this topic, as a part of a Ph. D. thesis. They are the initiators of this actuarial project.

Furthermore, I would like to thank my two succesive supervisors, Valérie Gilles and Gilles Hornecker. My memoir would not have been possible without their constant feedback.

I would like also to thank my academic supervisors, Véronique Maume-Deschamps and Stéphane Loisel, with whom I could have interesting and constructing discussions on this topic.

This memoir has been possible thanks to two wonderful open-source tools, namely the R statistical software and environment, and the typesetting and information processing software LaTeX. They deserve a place on this page!

Contents

Introduction

Price elasticity studies analyze how customers react to price changes. In this memoir, we focus on its effect on the renewal of non-life insurance contracts. The methodologies developped can also be applied to new business. Every year, insurers face the recurring question of adjusting premiums. Where is the trade off between increasing premium to favour higher projected profit margins and decreasing premiums to obtain a greater market share? We must determine a compromise to meet these contradictory objectives. Price elasticity is therefore a factor to contend with in actuarial and marketing departments in every insurance company.

Whether to target new market shares or to retain customers in the portfolio, it is essential to assess the impact of pricing on the whole portfolio. To avoid a portfolio-based approach, we must take into account the individual policy features. Moreover, the methodology to estimate the price elasticity of an insurance portofolio must be refined enough to identify customer segments. It is consequently the aim of this memoir to determine the price sensitiveness of non life insurance portfolios with respect to individual policy characteristics constituting the portfolio.

We define price elasticity as the customer's sensitivity to price changes relative to their current price. In mathematical terms, the price elasticity is defined as $e_r(p) = \frac{dr(p)}{dp} \times \frac{p}{r(p)}$, where $r(p)$ denotes lapse rate as a function of the price p. However, in this memoir, we focus on the additional lapse rate $\Delta(dp) = r(p + dp) - r(p)$ rather $e_r(p)$ since the results are more robust and a lot easier to interpret. In the following, we abusively refer $\Delta(dp)$ as the price elasticity of demand.

Chapter 1 presents the three datasets of the insurance markets, namely Portugal, Québec and Germany *. Chapter 2 studies the use of generalized linear models, while chapter 3 uses generalized additive models. Finally, chapter 4 tests the use of survival regression models. Unless otherwise specified, all numerical applications are done with the R software, R Core Team (2011).

This subject is not new in actuarial literature. Two ASTIN workshops, Bland et al. (1997), Kelsey et al. (1998), were held in the 90's to analyze customer retention and price/demand elasticity. The Shapiro & Jain (2003) book series also devoted two chapters to price elasticity: Guillen et al. (2003) used logisitic regressions, whereas Yeo & Smith (2003) took a look at neural networks. Brockett et al. (2008) should also be mentionned for their use of survival regression models.

What differentiates this memoir from previous research is the fact that we tackle the issue of price elasticity from various points of view. Not only do we focus on different markets, but we also investigate the impact of distribution channels. We have furthermore given ourselves the dual objective of comparing regression models as well as identifying the key variables needed.

∗. In this memoir, we only exploits motor datasets, but methodologies can be applied to other non-life lines.

Chapter 1

Data presentation

This chapter briefly presents the three insurance markets of the three datasets studied, namely Portugal, Québec and Germany. Each section has an identical structure: (i) insurance market presentation, (ii) data and (iii) short descriptive analysis.

1.1 Portugal market

1.1.1 Insurance market presentation

Compared to other European countries, Portugal has a quite high level of insurance penetration (premium written as percentage of gross domestic product): 3.17%. For instance, in France, the penetration rate is 3.15% as reported in Cummins & Venard (2007). As for many Europeans countries, the insurance market has been consolidating in the 90s. The market was growing at a high rate of 20% per year in term of premium. However, most of the growth comes from the life market, whereas the P&C market relatively stagnes. Nowadays, the country still suffers from the effects of the financial meltdown experienced by western countries in 2008-2009. Insurers struggle to retain business and to attract new business.

Cummins & Venard (2007) also provides useful information on the distribution channels. The policies are mainly sold through tied agents in Portugal (57%). In the Portugal market, AXA is fourth in terms of premium written for non-life insurance. The top five insurers represents 56% of the overall market (in P&C).

In motor insurance market, the third-part liability cover is mandatory. The minimum limits for personal injury and material damage are standardized: 2.5 mEUR and 0.75 mEUR in 2009 respectively.

There is no standard system for no claims discount as in other countries. But generally, they are similar between insurers: bonus can reach 50% and malus may in theory go up to 200%. Rules to be upgraded or to be downgraded are similar as in France (e.g. it takes 14 claims-free years to reach the best level). In practice, due to business reasons, the maluses are not entirely applied.

1.1.2 Portugal data

The database used for the analysis contains a very small set of variable: the policy number, the proposed premium, the last year premium, the customer choice, the gender, the driver age, the vehicle age, the policy age and the lapse reason.

Each line of the data represent a policy for a given vehicle. On this dataset, we are not able to identify if a customer has multiple vehicle insured in AXA.

The Portugal data consists of 1 year of lapse history in 2003. The table 1.1 shows the age structure of the Portugal portfolio. We can observe a cohort effect with policies aged of two and seven years old being two groups of high business.

Policy age	1	2	3	4	5	6	7	8	9	10
Frequency	6653	8731	6159	5325	3509	3774	4391	4678	3464	2970

Table 1.1: Policy age in the portfolio

1.1.3 Short descriptive analysis

To better understand the lapse and its relation to the premium, we follow with an analysis of the link between each explanatory variable and the lapse variable. All the outputs can be found in the appendix B.1.1, we just show the most useful material here. As a general comments all variables listed here are not independent from the lapse variable according to the Chi-square test.

Premium variables

Most of the portfolio seems to experience a premium decrease (cf. figure 1.1), probably due to the ageing and the market conditions. So we expect to slightly underestimate the true price elasticity of clients.

The proposed premium and price ratio (ratio of proposed premium and last paid premium) seems positevely correlated with the lapse, see table 1.2 and appendix B.1.1. The same conclusion can be drawn for last paid premium.

Undoubtly, those variables should be part of the GLM explanatory variables.

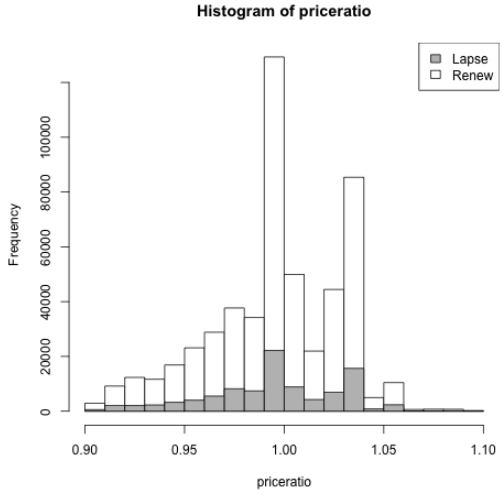


Figure 1.1: Histogram of price ratios

	(0.925,0.955]	(0.955,0.985]	(0.985,1.02]	(1.02,1.04]	(1.04,1.08]
Lapse rate (%)	18.8	20	18	18	21.9
Prop. of total (%)	8.5	20	40	29	2.5

Table 1.2: Price ratio

Customer variables

We now focus on the gender and the (driver) age variable. As the age of the customer increases, the lapse rate decreases. So the most sensitive clients seem to be the youngest clients. The gender * does not have any particular impact of the lapse, however the GLM analysis may reveal some links between the gender and lapses.

	(30,47.5]	(47.5,62.5]	(62.5,77.5]	(77.5,92.5]	FEMALE	MALE
Lapse rate (%)	20	17	14	14.6	18	19
Prop. of total (%)	38	42	17	3	20	80

Table 1.3: Driver age - Gender

Risk-specific variables

The latest variables to explore are the policy age and the vehicle age. The histograms do not reveal any pattern, see appendix B.1.1. However from the below tables, some conclusions can be derived. As the policy age decreases, the remaining clients are more and more loyal (i.e. lapse rates decrease). Unlike the policy age, the vehicle age has the opposite effect. As it increases, the lapse rate increases. One reason is that the customer may shop around for a new vehicle as well as for a new insurer.

	(2.5,5.5]	(5.5,8.5]	(8.5,11.5]	(11.5,14.5]	(14.5,17.5]
Lapse rate (%)	21	17	18	16.6	17.5
Prop. of total (%)	38	33	22	3.6	2.3

Table 1.4: Policy age

	(2.5,5.5]	(5.5,8.5]	(8.5,11.5]	(11.5,14.5]	(14.5,17.5]	(17.5,20.5]	(20.5,26.5]
Lapse rate (%)	17	18	19	20	21	21.1	39.3
Prop. of total (%)	15	21	21	16	14	8.4	4.4

Table 1.5: Vehicle age

Finally, we categorize the continuous variable for GLM regressions that follow in the next chapter.

∗. In a short future, insurers will no longer have the right to discreminate premium against the gender of the policyholder according to the directive 2004/113/CE from the European comission.

1.2 Québec market

1.2.1 Insurance market presentation

The insurance market penetration in Canada is about 5% of the Gross Domestic Product, similar to European countries but half the US market. The market is mostly deregulated but there are exceptions in motor third part liability and workers' compensation.

In Québec, private motor lines have a very special feature compared to other countries. Policies sold by insurers can only cover material damage and not bodily injuries. As well explained in Dionne et al. (2009), bodily injury covers is a public monopoly by the SAAQ *. So private insurers or mutual funds provide material damages for Québec citizen. However bodily injuries against non Québec citizen can be covered by insurers and are generally additional covers.

In terms of distribution channel, policies are mostly sold through brokers (70%), which are independent agency writers. For property and casualties (P&C), the market is not very concentrated where the top 5 players only represents 43% of the market. In 2004, AXA Canada was the ninth insurers in terms of premium written.

Citing Cummins & Venard (2007), we conclude by saying that Canada and in particular Québec is a very mature market (more than 200 years old) where insurers and mutuals play a major role not only in providing risk management for their clients but also in facilitating efficient allocation of resources, wealth management and asset protection.

1.2.2 Québec data

The Québec data consists of 4 years of lapse history between 2004 and 2007. The policy set is open in the sense that each year new policies renewing for the first time enters the data.

Each line of the data represent a policy for a given vehicle. Note that policyholder who has multiple vehicle insured in AXA a single policy number but a different vehicle number. In this case, the customer is free to cancel one, many or all vehicles.

The database contains the following variables
- Premium variables: the proposed premium, the last paid premium,
- Customer variables: the policy number, the vehicle number, the customer choice (to renew or to lapse), the gender, the driver age, the presence of a multi-vehicle discount, the cross-selling of household policy, the cover type (all risk, basic third-part liability, options) before and after renewal, a change of premium and/or wording by the broker,
- Risk-specific variables: the policy age, the vehicle age, the pricing group, the number of claims (responsible, non responsible an accident) for the last two years

In the table 1.6, we put the frequencies of policy age for the four years. It is quite easy to see the ageing of the portfolio on each diagonal.

∗. The Société de l'Assurance Automobile de Québec establishes the Québec Insurance Act in 1989.

Year	1	2	3	4	5	6	7	8	9	10
2004	44247	38877	33723	59748	54584	25085	22721	13462	9811	9707
2005	53029	35053	31843	27982	52296	47996	19935	19946	11810	8628
2006	34492	37888	23522	21876	19704	39820	36828	15546	15686	9034
2007	33591	39304	31704	19680	19118	17188	36030	33922	10455	13202

Table 1.6: Policy ages in the dataset

In the following subsection, we show a short descriptive analysis for the 2004 data. In appendix B.1.2, we provide the full descriptive analysis.

1.2.3 Short descriptive analysis

We put here only the main one-way tables, i.e. variables with the biggest impact on the lapse.

Cross-selling variables

In table 1.7, we see the strong impact of the cross-selling effect, i.e. having 2 or more vehicles insured in AXA and/or a household policy.

	Multi-vehicle discount		Having house policy	
	N	Y †	N	Y
prop. size (%)	59.16	40.84	56.46	43.54
lapse rate (%)	7.12	4.45	8.24	3.16

Table 1.7: Multi-vehicle discount / Have house policy at AXA

Claim variables

With the Québec data, we are able to observe the impact of claim history on the customer choice to lapse or not. The table 1.8 is one example of it, with the number of responsible claims during the last period of coverage.

	0	1	2+
prop. size (%)	97.44	2.50	0.06476
lapse rate (%)	5.60	7.43	11.34

Table 1.8: Last year responsible claim group

1.3 Germany market

1.3.1 Insurance market presentation

In Germany as in many industrialized countries, it is mandatory for any driver to be insured (at least) in third-part liability. Insurers are enforced by law to warn authority if a driver is uninsured. The limits for bodily injury, property damage and pecuniary loss have been standardized to 7.5 mEUR, 1 mEUR and 50 kEUR respectively.

Generally, insurance contracts are sold with add-on cover, specific deductibles, ... We consider three classes of coverage: third-part liability (TPL), full comprehensive (FC) and partial comprehensive (PC) coverages. There is a bonus-malus system in Germany, called SchadenFreiheitsrabatt *, that will be discussed in a later subsection.

In the German insurance market, private motor is mainly sold in three different ways: tied-agents (60%), brokers (15%) and direct online websites (8% †). This is a strong difference compared to Québec or Portugal markets.

Unlike other European countries, the German market had been deregulated recently, during the two years 1993/1994. Therefore, the competition rapidly increases from that year for the main private line of business. For private motor, the total gross premium income jumped from 15 billlions in 1992 of euros to 19 billions in 1993 and 21 billions in 1994.

The reduction in loss frequency and severity is an ongoing trend (as for other Western countries). This increase in safety is especially true for death numbers in traffic accident: 4477 in 2008 compared to 12000 in 1990 and 20000 ‡ in 1970. But this is counter-balanced by an increase in bodily injury claim, and still the loss experience can be potentially high.

Nowadays, the competition is still fierce among insurers and risk selection is at stack. The top 5 non life insurers only represent 41% in 2003 according to Cummins & Venard (2007). Between 2004 and 2009 the premium rates decrease while the market loss ratio increase from 86% to 96%. AXA is a leading insurer with the fourth position (in terms of premiums) behind Allianz, HUK and R+V insurers.

1.3.2 Germany data

The Germany data consists of 2 years of lapse history between 2004 and 2005 . The policy set is open in the sense that each year new policies renewing for the first time enters the data. As for other datasets, a record is a policy purchased by an individual, so an individual may have different records for the different covers he bought.

∗. meaning no claims discount.
†. The remaining 18% are sold through pyramid sales, typically with motor trade.
‡. Only for Western Germany.

This dataset is unique and very rich because it contains policies sold through different distribution channels, namely tied-agents, brokers and direct websites, see table 1.9.

	Agent	Broker	Direct
prop. size (%)	65.1	20.1	6.1
lapse rate (%)	7.4	10.3	12.1

Table 1.9: Lapse rate by channel distribution

In terms of policy age, the portfolio has a wide range of policy ages. In the table 1.10 below, we can observe that tied-agents can have a strong retention on their portfolio. More than 26% of policies sold by tied-agents are 8 years old or more.

Channel	0	1	2	3	4	5	6	7	8+
Agent	19 539	18 362	15 455	† 7 715	† 7 330	† 6 380	† 3 084	† 3 242	29 351
Broker	7 828	7 155	6 115	2 630	2 644	1 287	1 223	1 232	1643
Direct	2 524	1 501	1 734	1 539	1 719	1 984	1 506	1 197	1 573

Table 1.10: Population by policy age

These two tables confirms that selling an insurance contract is totally different wether you sell it through tied-agents, brokers or direct websites. So in the following we will separate the policies between these 3 distribution channels.

1.3.3 Short descriptive analysis

Variable list

The German data is quite rich, therefore we have the detailed features of each policy. We write below a subset of the available variables:

- Policy:
  - a dummy variable indicating the lapse,
  - the policy age,
  - the cover type (TPL, PC or FC) and the product,
  - the SF-class for PC and FC covers and the bonus evolution,
- Policyholder:
  - the policyholder age and the gender,
  - the marital status and the job group,
- Premium:
  - the last year premium, the technical premium and the proposed premium,
  - the payment frequency,
  - the market premium, i.e. the tenth lowest NB premium for a particular category,
- Car:
  - the mileage, the vehicle age,
  - the car usage, the car class,

Below, please find explanatory variables which are not in the two other datasets:
- Cross-selling:
  - the number of AXA contract in household,
  - a dummy variable on household policy,
- Claims:
  - the claim amount,
  - the claim number per year,
- Agent:
  - the cumulative rebate, the technical rebate,
  - the age difference between the agent and the policyholder.

Descriptive analysis

The full descriptive analysis can be found in appendix B.1.3. We put in table 1.11 the most impacting explanatory variables.

Claim number *	0	1	2	3	[4 - 13]
prop. size	70.59	25.29	3.60	0.44	0.092
lapse rate	13.75	13.37	16.03	12.82	35.16
Policy age †	(0,1]	(1,2]	(2,7]	(7,34]	
prop. size	24.97	16.79	34.38	23.86	
lapse rate	17.43	15.27	11.26	8.78	
Cover ‡	FC	PC	TPL		
prop. size	36.16	37.61	26.23		
lapse rate	14.26	12.64	12.79		
Bonus evolution §	down	stable	up		
prop. size	33.32	62.92	3.76		
lapse rate	16.69	11.53	12.02		
Vehicle age ¶	(0,6]	(6,10]	(10,13]	(13,18]	
prop. size	26.06	31.01	21.85	21.08	
lapse rate	15.50	13.56	12.72	10.67	

Table 1.11: Impact on lapse rates (%)

∗. coded `nbclaim08percust` in the database.
†. coded `polage` in the database.
‡. coded `cover` in the database.
§. coded `bonusevol` in the database.
¶. coded `vehiclage` in the database.

The bonus malus system

The bonus-malus system in Germany, called SchadenFreiheitsrabatt (SF in the following), aims to reveals the true risk type of a driver. It takes into account the claim (or no claim) history of the driver. The bonus-malus is characterized by a coefficient between 30% (best driver class) and 275% (worst driver class), which will be multiplied to a base premium for a given cover (TPL or damage cover).

The bonus-malus system is common for all insurers operating in Germany. And when an insured switch from one insurer to another, the full information on the SF class is transmitted between the two insurers. So the SF class represents the full claim history of the driver. Upgrading and downgrading on the SF class is governed by rules:

- a responsible claim will downgrade the SF class of the corresponding cover,
- in the case of no reported claims, the SF class is upgraded to the next level,
- a protected no-claims bonus can be applied if the class is higher than 5 such that an allowance of two claims over three years is made.

At the beginning, policyholder will start in class SF-1, SF-1/2, S, O or M. The non responsible claims include theft, natural catastrophes and glass claims.

A feature of the German bonus-malus is the distinction between TPL and damage (Full Comprehensive) cover, so that every driver has two SF classes, one for TPL and one for FC. The two classes evolve independently. If a policyholder did not start with the FC cover, then it will take the corresponding level of the SF class for TPL cover when the FC cover is purchased. In the case, a customer drops the FC cover, then the SF class will be kept in the system.

In table 1.12, we put the bonus range which can be applied for the TPL and FC covers.

SF-class	Bonus range	SF-class	Bonus range
SF 26	30 %	SF 11	45 - 50 %
SF 25	30 %	SF 10	45 - 50 %
SF 24	30 %	SF 9	45 - 50 %
SF 23	30 %	SF 8	55 - 55 %
SF 22	30 %	SF 7	55 - 60 %
SF 21	30 - 35 %	SF 6	55 - 60 %
SF 20	30 - 35 %	SF 5	55 - 65 %
SF 19	30 - 35 %	SF 4	60 - 70 %
SF 18	35 - 40 %	SF 3	70 - 80 %
SF 17	35 - 40 %	SF 2	75 - 85 %
SF 16	35 - 40 %	SF 1	90 - 100 %
SF 15	40 %	SF 1/2	115 - 140 %
SF 14	40 %	S	145 - 190 %
SF 13	40 - 45 %	O	125 - 240 %
SF 12	40 - 45 %	M	245 - 275 %

Table 1.12: Bonus range on TPL-FC covers

Chapter 2

Generalized Linear Models

The Generalized Linear Models (GLM) were introduced in the 70's to deal with non continuous and/or bounded response variables, to get rid off the limitation of linear models that must have a continuous unbounded response. GLMs are well known and well understood tools in statistics and especially in actuarial science.

The pricing and the customer segmentation could not have been as efficient in non-life insurance as it is today, without an extensive use of GLMs by actuaries. There are even books dedicated to this topic, e.g. Ohlsson & Johansson (2010). Hence the GLMs seem to be the very first choice, we can use to model price elasticity.

Furthermore, in AXA, previous memoirs also use GLMs for this topic: for instance Dreyer (2000), Sergent (2004), Hamel (2007) study motor insurance lines, while Rabehi (2007) works on household products. This chapter is divided into three sections: (i) model presentation, (ii) case studies and (iii) conclusions.

2.1 Model presentation

In this section, we present the Generalized Linear Models, tagged as GLM in the following *. GLMs were introduced in a 1972 paper of Nelder and Wedderburn, Nelder & Wedderburn (1972), but become extremely popular with the book of McCullagh and Nelder. We use McCullagh & Nelder (1989) as a guide book, since it is the reference on the topic. This section is divided into three parts: (i) theoretical description of GLMs, (ii) a clear focus on binary models and (iii) explanations on estimation and variable selection within the GLM framework.

2.1.1 Theoretical presentation

We cannot present GLMs without starting with linear models. So, the first sub-section is a short description of linear models.

∗. Note that in this document, the term GLM will never be used for general linear model.

Starting from the linear model

Let $X \in M_{np}(\mathbb{R})$ be the covariate matrix, i.e. a matrix where row contains the value of the explanatory variables for a given individual and $Y \in \mathbb{R}^k$ the vector of responses. The linear model assumes the following relationship between X and Y:

$$Y = X\Theta + \mathcal{E},$$

where Θ denotes the (unknown) parameter vector and $\mathcal{E}$ the (random) noise vector. If we made the following assumptions,

(i) white noise: $E(\mathcal{E}_i) = 0$,
(ii) homoskedasticity: $Var(\mathcal{E}_i) = \sigma^2$,
(iii) normality: $\mathcal{E}_i \sim \mathcal{N}(0, \sigma^2)$,
(iv) independence: $\mathcal{E}_i$ is independent of $\mathcal{E}_j$ for $i \neq j$,
(v) identification: $rg(X) = p < n$,

Then the Gauss-Markov theorem gives us the following results

- the least square estimator $\hat{\Theta}$ of Θ is $\hat{\Theta} = (X^T X)^{-1} X^T Y$ and $\widehat{\sigma^2} = \frac{||Y - X\Theta||^2}{n-p}$,
- $\hat{\Theta}$ is a Gaussian vector independent of $\hat{\sigma^2} \sim \chi^2_{n-p}$,
- $\hat{\Theta}$ is the unbiased estimator with minimum variance of Θ, such that $Var(\hat{\Theta}) = \sigma^2 (X^T X)^{-1}$ and $\widehat{\sigma^2}$ is the unbiased estimator of σ^2.

Let us note that the first four assumptions can be summarized in a single assumption $\mathcal{E} \sim \mathcal{N}(0, \sigma^2 I_n)$. But splitting the normality assumptions will help to identify the strong assumptions of linear models and to present the differences with GLMs.

Examples:

1. The simple linear regression $y_i = a + bx_i + \epsilon_i$ is a linear model:

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \text{and} \quad \Theta = \begin{pmatrix} a \\ b \end{pmatrix}.$$

2. The parabolic linear regression $y_i = a + bx_i + cx_i^2 + \epsilon_i$ is a linear model:

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \quad \text{and} \quad \Theta = \begin{pmatrix} a \\ b \\ c \end{pmatrix}.$$

Many properties can be derived for the linear model, notably hypothesis test, confidence interval as well as estimator convergence. See chapter 6 of Venables & Ripley (2002).

We now focus on the limitations of linear model resulting from strong assumptions. The following problems have been identified. Numerically, the computation of $\hat{\Theta}$ can be an issue if X contains colinear variables. This leads to an increase in the variance estimate and even the $\widehat{\sigma^2}$. In pratice, a solution is to test models with omitting one explanatory variable after another.

A stronger limitation is the fact the variance of the response is assumed to be the same (σ^2) for all individuals. One way to deal with this problem is to transform the response by a Box-Cox transformation. However it can be unsatifactory.

Finally the strongest limitation is the support of the response variable. By the normal assumption, Y must lies in $\mathbb{R}$, which excludes count variable (e.g. Poisson distribution) or positive variable (e.g. exponential distribution). There is no answer to that problem unless to extend the model.

In our case of a binary variable, a linear model is inadequate. Mainly for the following reasons:
- since the value of $E(Y)$ is contained within the interval $[0, 1]$, a linear predictor β . X would fall out of this range for values of X that are high enough.
- the normality hypothesis of the residuals is clearly not verified: $Y - E(Y)$ will only take two different values, $-E(Y)$ and $1 - E(Y)$. Therefore, the modelling of $E(Y)$ as a function of X needs to be changed as well as the type of error.

Toward generalized linear models

A Generalized Linear Model is characterized by three components:

1. a random component: Y_i follows a distribution of the exponential family $\mathcal{F}_{exp}(\theta_i, \phi_i, a, b, c)$ *,
2. a systematic component: the covariate vector X_i produces a linear predictor $\eta_i = X_i^T\beta$,
3. a link function $g : \mathbb{R} \mapsto S$ which is monotone, differentiable and inversible, such that $E(Y_i) = g^{-1}(\eta_i)$,

for $i \in \{1, \dots, n\}$, where θ_i is the shape parameter, ϕ_i the dispersion parameter and a, b, c three functions.

Let us note that we get back to linear models with a Gaussian distribution and an identity link function. However, there are many other distributions and link functions. Furthermore, we say a link function to be canonical if $\theta_i = \eta_i$.

There are many applications of GLM in actuarial science. The below table 2.1 lists the most common distribution with their canonical link.

Law	Canonical link	Mean	Used for
Normal $\mathcal{N}(\mu, \sigma^2)$	identity $\eta_i = \mu_i$	$\mu = X\beta$	standard linear regression
Bernoulli $\mathcal{B}(\mu)$	logit $\eta_i = \log(\frac{\mu}{1-\mu})$	$\mu = \frac{1}{1+e^{-X\beta}}$	rate modelling
Poisson $\mathcal{P}(\mu)$	log $\eta_i = \log(\mu_i)$	$\mu = e^{X\beta}$	claim frequency
Gamma $\mathcal{G}(\alpha, \beta)$	inverse $\eta_i = \frac{1}{\mu_i}$	$\mu = (X\beta)^{-1}$	claim severity
Inverse Normal $\mathcal{I}(\mu, \lambda)$	squared inverse $\eta_i = -\frac{1}{\mu_i^2}$	$\mu = (X\beta)^{-2}$	claim severity

Table 2.1: Family and link functions

Apart from the identity link function, the log link function is the most classically used link function: with this link function, the explanatory variables have multiplicative effects on the observed variable and the observed variable stays positive. Indeed, $E(Y) = \prod_i e^{\beta_i x_i}$.

∗. See appendix A.1

For example, the effect of being a young driver and owning an expensive car on average loss will be the product of the two separate effects, of the effect of being a young driver and of the effect of owning an expensive car. The log link function is central to actuarial pricing models, as it is used for modelling the frequency and the severity of claims.

Fitting procedure

To determine the vector β, we use the method of maximum likelihood. For n observations, the log-likelihood of a distribution from the exponential family is written as follows:

$$\ln(\mathcal{L}(\theta_1,\dots,\theta_n,\phi,y_1,\dots,y_n)) = \sum_{i=1}^{n}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right]. \tag{2.1}$$

Let us define $\mu_i = E(Y_i)$ and $\eta_i = g(\mu_i) = X_i\beta$, the linear prediction with i is the number of the observation, n the total number of observations.

For all i and j,

$$\frac{\partial \ln(\mathcal{L}_i)}{\partial \beta_j} = \frac{\partial \ln(\mathcal{L}_i)}{\partial \mu_i} \times \frac{\partial \mu_i}{\partial \beta_j} = (g^{-1})'(g(\mu_i)) \times \frac{y_i - \mu_i}{Var(Y_i)} X_{ij}.$$

Maximum likelihood equations are then: $\sum_i \frac{\partial \ln(\mathcal{L}_i)}{\partial \beta_j} = \sum_i (g^{-1})'(g(\mu_i)) \times \frac{y_i-\mu_i}{Var(Y_i)} X_{ij} = 0$, for all j. Therefore, we get the equations, as a function of the β_i's:

$$\sum_i \frac{\partial \ln(\mathcal{L}_i)}{\partial \beta_j} = \sum_i (g^{-1})'(X_i\beta) \times \frac{y_i - g^{-1}(X_i\beta)}{(b')^{-1}(g^{-1}(X_i\beta))} X_{ij} = 0. \tag{2.2}$$

These equations are not linear with respect to the β_is, and cannot be solved easily. As always for complex equation, we use an iterative algorithm to find the solution. In our case, most softwares use an iterative weighted least-squares method, see section 2.5 of McCullagh & Nelder (1989).

2.1.2 Binary regression

Base model assumption

We now focus on binary regression, regression where the response variable is either 1 or 0, respectively for success and failure. We cannot parametrize two outcomes with more than one parameter. So we assume

$$P(Y_i = 1) = \pi_i = 1 - P(Y_i = 0),$$

with π_i the parameter. The mass probability function can be expressed as

$$f_{Y_i}(y) = \pi_i^y (1-\pi_i)^{1-y},$$

which emphasizes the exponential family characteristic. Let us recall the first two moments are $E(Y_i) = \pi_i$ and $Var(Y_i) = \pi_i(1-\pi_i) = V(\pi_i)$.

Assuming Y_i is a Bernoulli distribution $\mathcal{B}(\pi_i)$ implies that π_i is both the parameter and the mean value of Y_i. So the link function for a binary model is expressed as follows

$$\pi_i = g^{-1}(x_i^T\beta).$$

Let us note that if individuals have identical covariates, then we can group the data and consider Y_i follows a binomial distribution $\mathcal{B}(n_i, \pi_i)$. However grouping is possible if covariates are only categorical.

As indicating in Fox (2010), the link function and the response variable can be reformulated as an unobserved variable. $\pi_i = P(Y_i = 1) = P(x_i^T\beta - \epsilon_i > 0)$. If ϵ_i follows a normal distribution (resp. a logistic distribution), we have $\pi_i = \Phi(x_i^T\beta)$ ($\pi_i = F_{\text{logistic}}(x_i^T\beta)$).

Now we can derive the log-likelihood from 2.1

$$\ln(\mathcal{L}(\pi_1, \ldots, \pi_n, y_1, \ldots, y_n)) = \sum_{i=1}^{n} \left[y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i) \right],$$

plus an omitted constant not involving π.

Link functions

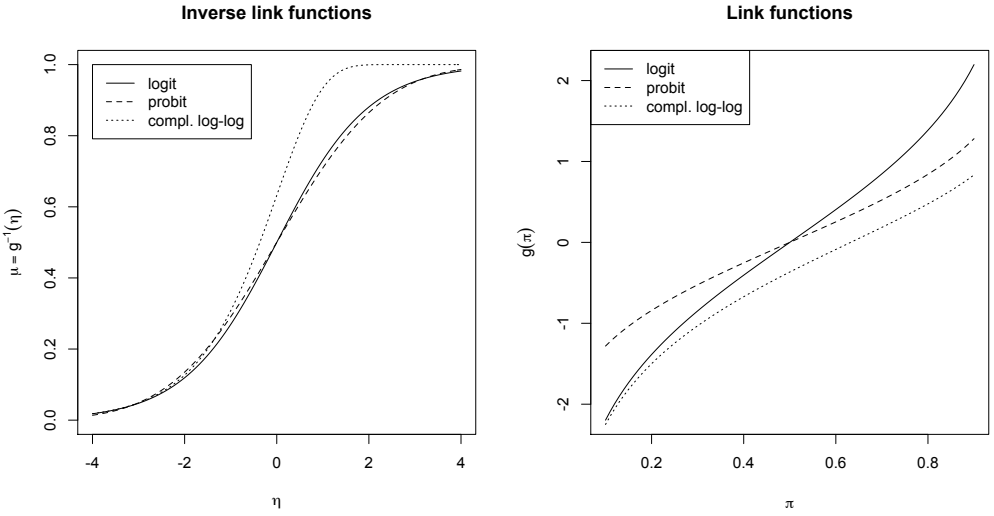


Figure 2.1: Link functions for binary regression

Generally, the following three functions are considered as link function for the binary variable

1. logit link: $g(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$ with g^{-1} being the standard logistic distribution function,
2. probit link: $g(\pi) = \Phi^{-1}(\pi)$ with g^{-1} being the standard normal distribution function,
3. complementary log-log link: $g(\pi) = \ln(-\ln(1-\pi))$ with g^{-1} being the standard Gumbel II distribution function *.

*. A Gumbel of second kind is the distribution of $-X$ when X follows a Gumbel distribution of first kind.

On the figure 2.1, we plot these three link functions and their inverse. Let us note that the first two links are symmetrical, while the last one is not. All these three functions are the inverse of a distribution function, so other link functions can be obtained using inverse of other distribution function (e.g. with the Gumbel I distribution). In addition being the canonical link function, the logit link is generally preferred because of its simple interpretation as the logarithm of the odds ratio.

Log-likelihood for canonical link

Using the expression of the variance function and the logit function ($g^{-1}(x) = \frac{1}{1+e^{-x}}$ and $(b')^{-1}(x) = x(1-x)$), 2.2 becomes

$$0 = \sum_i \frac{e^{-\eta_i}}{1+e^{-\eta_i}} \times \frac{y_i - \frac{1}{1+e^{-\eta_i}}}{\frac{1}{1+e^{-\eta_i}}\frac{e^{-\eta_i}}{1+e^{-\eta_i}}} X_{ij} = \sum_i (y_i(1+e^{-\eta_i}) - 1)X_{ij},$$

for $j = 1, \ldots, p$. These equations are called the likelihood equations. If we put it in a matrix version, we get the so-called score equation

$$X^T(Y - \mu(\beta)) = 0.$$

The Fisher information matrix for β in the case of logit link to

$$\mathcal{I}(\pi) \stackrel{\triangle}{=} -E\left(\frac{\partial^2 \ln \mathcal{L}}{\partial\beta_j \partial\beta_k}\right) = \text{diag}(\pi_i(1-\pi_i)).$$

Since we work the maximum likelihood estimator, the estimator $\hat{\beta}$ is unbiased and asymptotically Gaussian with variance matrix approximated by Fisher information $\mathcal{I}(\pi(\hat{\beta}))$ *.

2.1.3 Variable selection and model adequacy

Model adequacy

The deviance, which is one way to measure the model adequacy with the data and generalizes the R^2 of linear models, is defined by

$$D(y, \hat{\pi}) = 2(\ln(\mathcal{L}(y_1, \ldots, y_n, y_1, \ldots, y_n)) - \ln(\mathcal{L}(\hat{\pi}_1, \ldots, \hat{\pi}_n, y_1, \ldots, y_n))),$$

where $\hat{\pi}$ is the estimate of the beta vector. However for binary data, the first term is infinite. So in practice, we consider the deviance as

$$D(y, \hat{\pi}) = -2\ln(\mathcal{L}(\hat{\pi}_1, \ldots, \hat{\pi}_n, y_1, \ldots, y_n)).$$

Furthermore, the deviance is used as a relative measure to compare two models. In most softwares, in particular in R, the GLM fitting function provides two deviances: the null deviance and the deviance.

∗. see subsection 4.4.4 of McCullagh & Nelder (1989).

The null deviance is the deviance for the model with only an intercept or if not offset only, i.e. when $p = 1$ and X is a vector full of 1 *. The (second) deviance is the deviance for the model $D(y, \hat{\pi})$ with the p explanatory variables.

Another criterion introduced by Akaike in the 70's is the Akaike Information Criterion (AIC), which is also an adequacy measure of statistical models. Unlike the deviance, it aims to penalized over fitted models, i.e. models with too much parameters (compared to the length of the dataset). It is defined by

$$\mathrm{AIC}(y, \hat{\pi}) = 2k - \ln(\mathcal{L}(\hat{\pi}_1, \ldots, \hat{\pi}_n, y_1, \ldots, y_n)),$$

where k the number of parameters, i.e. the length of β. It is a useful criterion to compare two models where the number of parameters is different.

In a linear model, the analysis of residuals (which are assumed to i.i.d. Gaussian variable) may reveal that the model is unappropriate. Typically we can plot the fitted values against the fitted residuals. In GLM, the analysis of residuals is more complex, because we loose the normality assumption. Furthermore, for binary data (i.e. not binomial data), the plot of residuals are hard to interpret.

Let us study the example of Bronchitis data of Turner (2008). The data consists of 212 patients, on which we measure the presence/absence of bronchitis B for `bron`, the air pollution level in the locality of residence P for `poll` and the number of cigarettes smoked per day C for `cigs`.

```
> head(Data)
  bron cigs poll
1    0 5.15 67.1
2    1 0.00 66.9
3    0 2.50 66.7
4    0 1.75 65.8
5    0 6.75 64.4
6    0 0.00 64.4
```

Let us first regress the bronchitis indicator on all variables

$$Y = \begin{pmatrix} B_1 \\ \vdots \\ B_n \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} 1 & P_1 & C_1 \\ \vdots & \vdots & \vdots \\ 1 & P_n & C_n \end{pmatrix},$$

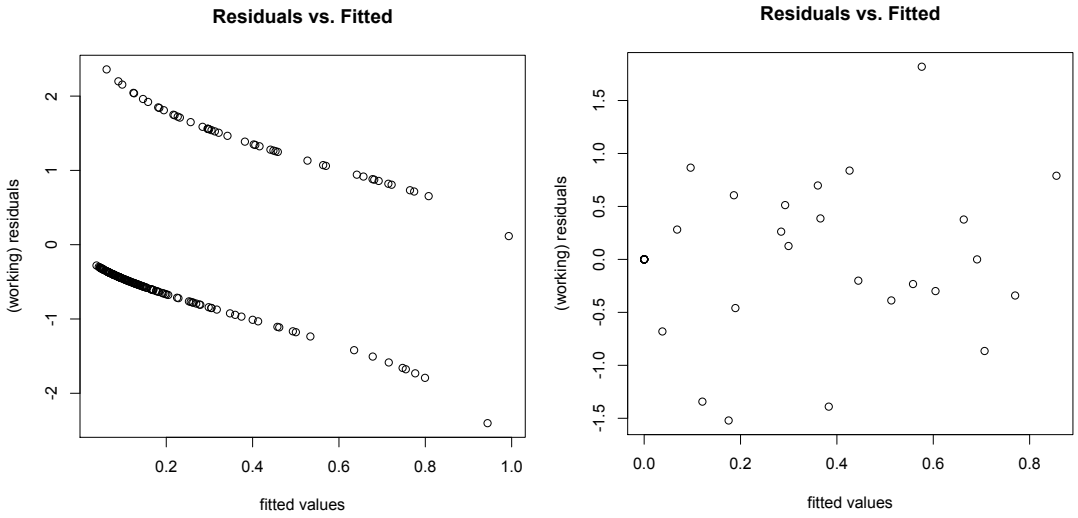
with a logit link function. We get

```
> model1 <- glm(bron ~ 1 + cigs + poll, family = binomial)
> summary(model1)

Call:
glm(formula = bron ~ 1 + cigs + poll, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4023  -0.5606  -0.4260  -0.3155   2.3594
```

∗. It means all variation comes from the random component.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.08491    2.95100  -3.417 0.000632 ***
cigs          0.21169    0.03813   5.552 2.83e-08 ***
poll          0.13176    0.04895   2.692 0.007113 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 221.78  on 211  degrees of freedom
Residual deviance: 174.21  on 209  degrees of freedom
AIC: 180.21

Number of Fisher Scoring iterations: 5
```

So the GLM fit seems good because all variables (including intercept) are significant with a very low p-value. However the plot of residuals * (see figure 2.2a) against fitted values † is quite puzzling. Two distinct curves are shown: one for ill patients and the other for healthy ones.



(a) Binary data (b) Binomial data

Figure 2.2: Analysis of residuals for binary regression

When categorizing the P variable, we lose information but we transform binary data into binomial data. This makes the fit better on this aspect, see 2.2b. So for the same data, with the same (significant) variable the two analyses of residuals lead to different conclusions. Hence conclusions of the analysis of residuals must be taken with care.

∗. $\hat{\epsilon}_i = Y_i - \hat{\pi}_i$.
†. $\hat{\pi}_i$.

Variable selection

From the normal asymptotic distribution of the estimator, we can derive confidence interval as well as hypothesis test. Therefore a p-value is available for each coefficient of the regression, which help us to keep only the most significant variable. However as removing one variable impacts the significancy of other variables, it can be quite hard to find the optimal set of explanatory variables. There are two ways to procede: either forward selection (i.e. from the null model add the most significant variable at each step) or backward elimination (i.e. from the full model remove the least significant variable at each step).

Another way to select significant explanatory variables is to use the analysis of deviance. It consists in looking at the difference of deviance between two models, i.e. ratios of likelihood. Using asymptotics distribution either chi-square of Fisher-Snedecor, a p-value can be used to remove or to keep an explanatory variable.

2.2 Case studies

Before enjoying the study of the three datasets, let us remind the purpose of the study. We want to study the individual behaviors relative premium change while taking into consideration their fundamental features in order to derive an aggregate lapse function of the whole portfolio. As we are concerned with lapse rate predictions, we are forced to use exogeneous variables.

Therefore, we need to exclude endogeneous variables from our analysis, typically the rebate granted by the broker or a dummy variable indicating if the customer drops an optional cover. Furthermore to avoid the underestimation of lapses, we need to remove records for which the broker grants a rebate or for which the customer drops an optional cover: two actions implying a price decrease.

Finally, as we focus on price, we also need to exclude the lapse by company to unbias our predictions. Those lapse do not reveal a price sensitive behavior, since the customer did not have a chance to renew its policy.

2.2.1 Portugal

Let us start with Portgual data, see 1.1.2 for details. For the GLM analysis, we have to remove the lines with missing variables or the lines with aberrant variable value (for example, driver age older than 100). In a second time, we need to look at the lapse reasons. In table 2.2, we put the lapse reasons.

The lapse by company is clearly to remove. However, the default of payment must be taken with care since it might represent an insured decision. We consider that it results from a too high premium, the customer can't afford. So we choose to keep those policies in our study (it was checked with line managers). Finally, the lapse motive will not be used in the GLM regression because this variable is endogeneous and so can not be known for prediction purposes.

Lapse motive	Renew	Lapse
	455615	0
Company decision	0	233
Insured decision	0	38134
Payment default	0	66596
Total	81.3 %	18.6 %

Table 2.2: Lapse motive

See 1.1.3 for a descriptive analysis of Portugal data.

GLM analysis

We investigate two models: the model with continuous variables and the model with categorical variables. In appendix B.2.1, the full backward selection can be found for both approaches. We put here only the final fitting result.

See below the summary table with coefficients values, standard errors, z-statistics and p-value. To see the effect of one variable, say the driver age (`age`) we will plot additional graphics, because as the variable is crossed with the price ratio, the interpretation is hard.

```
Call: glm(formula = did_lapse ~ age_policy + priceratio * (gender + age + age_vehicle + premium_before),
    family = binomial("logit"), data = workdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3538  -0.6757  -0.6045  -0.5145   2.5791

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.7737301  0.1907634  -4.056 4.99e-05 ***
age_policy               -0.0076121  0.0006013 -12.659  < 2e-16 ***
priceratio               -0.4459493  0.1903878  -2.342 0.019164 *
genderMALE                0.7540763  0.1159249   6.505 7.78e-11 ***
age                      -0.0352652  0.0031278 -11.275  < 2e-16 ***
age_vehicle              -0.0246664  0.0064425  -3.829 0.000129 ***
premium_before           -0.0017755  0.0001777  -9.989  < 2e-16 ***
priceratio:genderMALE    -0.6734653  0.1157564  -5.818 5.96e-09 ***
priceratio:age            0.0181797  0.0031110   5.844 5.11e-09 ***
priceratio:age_vehicle    0.0538313  0.0064235   8.380  < 2e-16 ***
priceratio:premium_before 0.0025015  0.0001790  13.973  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 539837  on 56034  degrees of freedom
    Residual deviance: 531382  on 56033  degrees of freedom   -   AIC: 53140
```

We also plot (but do not report) 1000 fitted probabilities against the price ratio $(p_i, \hat{\pi}_i)_{1 \leq i \leq 1000}$. However there is no particular pattern, so it seems hard to explain the lapse just with the price ratio variable.

We also test different link functions, but clog-log function did not converge, so the result is not reported. In terms of residuals deviance, the gain is not big: the probit link function is the best choice.

```
Analysis of Deviance Table

Model 1: did_lapse ~ age_policy + priceratio * (gender + age + age_vehicle + premium_before) - logit
Model 2: did_lapse ~ age_policy + priceratio * (gender + age + age_vehicle + premium_before) - probit
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1    560333     531382
2    560333     531339  0   43.354
```

Lapse rate prediction

As we are interested in deriving a portfolio elasticity based on individuals specificities, we compute an average lapse probability function of the price ratio p:

$$\hat{\pi}_{g,n}(p) = \frac{1}{n}\sum_{i=1}^{n} g^{-1}\left(x_i^T\hat{\beta}_{-p} + z_i^T\hat{\beta}_{+p} \times p\right),$$

where $\hat{\beta} = (\hat{\beta}_{-p}, \hat{\beta}_{+p})$ is the fitted parameter [*], x_i price-independent explanatory variables, z_i price-dependent explanatory variables and g the link function. Beware, it is a functiont of the price ratio p.

On the figure 2.3, we plot the average lapse function $\hat{\pi}_{g,n}$ for two link functions. For a global increase of 10% on the whole portfolio, the lapse rate could increase by 1.89 or 1.82 points respectively for logit and probit link function. Therefore we have clearly a price elasticity below 1. Those predictions seem unreliable.

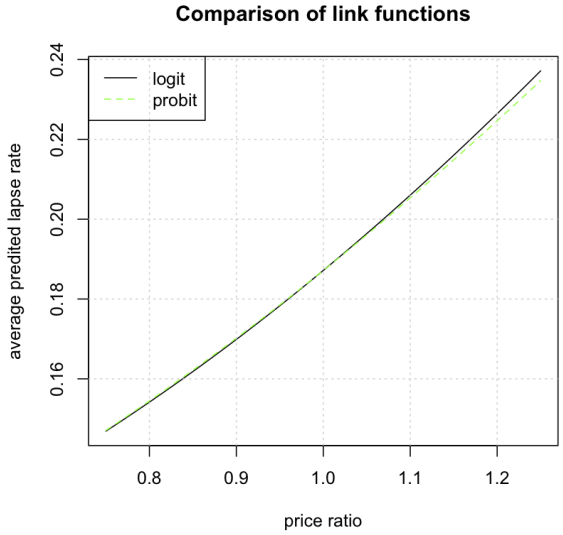


Figure 2.3: Average lapse function for different link functions

We also plot the effect of a given explanatory variable on the average lapse function, see figure B.7. Those plots emphasize the different behaviors in the portfolio, young vs. old drivers, male vs. female or young vs. old policies.

[*]. separated between coefficients for price-independent variable and price-dependent variable.

We refine the GLM analysis by using the categorical data. Compared to the full model with unsignificant variables, we observe that deviance residuals are better centered and less extreme in absolute value with the final model.

```
Call: glm(formula = did_lapse ~ agegroup3 + priceratio * (gender +  agevehgroup) + priceratio:agepolgroup2,
    family = binomial("logit"), data = workdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4640  -0.6572  -0.6197  -0.5657   2.2263

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -2.140869   0.115413 -18.550  < 2e-16 ***
agegroup3(25,99]              -0.399750   0.019230 -20.788  < 2e-16 ***
priceratio                     0.929192   0.114767   8.096 5.67e-16 ***
genderMALE                     0.625276   0.114278   5.472 4.46e-08 ***
agevehgroup(5,10]             -0.788156   0.103377  -7.624 2.46e-14 ***
agevehgroup(10,15]            -0.264146   0.117944  -2.240  0.02512 *
agevehgroup(15,99]            -0.373172   0.126790  -2.943  0.00325 **
priceratio:genderMALE         -0.576867   0.114071  -5.057 4.26e-07 ***
priceratio:agevehgroup(5,10]   0.960637   0.102957   9.331  < 2e-16 ***
priceratio:agevehgroup(10,15]  0.564691   0.117371   4.811 1.50e-06 ***
priceratio:agevehgroup(15,99]  0.722763   0.126233   5.726 1.03e-08 ***
priceratio:agepolgroup2(4,49] -0.234641   0.007087 -33.108  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 539782  on 560267  degrees of freedom
Residual deviance: 536859  on 560256  degrees of freedom
AIC: 536883
```

We do not report the fit for the other two link functions, because results are similar to above. Graphs to see the effect of one explanatory variable are put in appendix, figure B.8. Only for the vehicle age, the shape of the predicted lapse function is really different between continuous and categorical explanatory variables. The most unsensitive population are policies for young cars, but as the car gets old, the insured lapses more and more (especially for 10-year-old cars and older). Finally the effect of premium level paid last year by the client has the most important impact. As we do not have the cover type, it is a very good proxy for it.

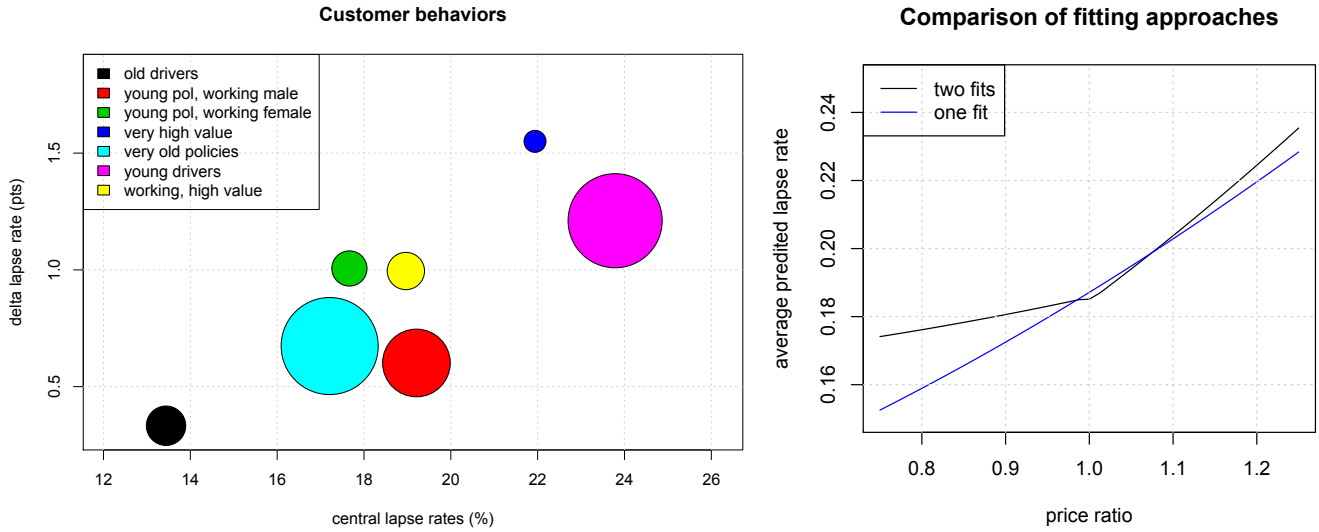
Sub-population study

A good and valuable output of the GLM analysis is that we can derive customer segmentations. By sorting individuals regarding their fitted lapse probabilites $\hat{\pi}_i$, we can group them into homogeneous group. We make 7 groups:

1. people older than 60 year old, a policy age between 4 and 8 years and a last premium amount less than 500 euros,
2. male between 35 and 60 years old with a policy age between 0 and 4 years,
3. female between 35 and 60 years old with a policy age between 0 and 4 years,
4. policies with premium amount above 1500 euros,
5. policies older than 8 years,
6. people between 20 and 35 years old,
7. people between 35 and 60 years old with a premium amount between 500 and 1500 euros.

Then, to better see the heterogeneity between groups, we plot the central lapse rates ($\hat{\pi}_{g,n}(1)$) and the "delta lapse rate" to a price incrase ($\Delta = \hat{\pi}_{g,n}(1.05) - \hat{\pi}_{g,n}(1)$)*.

On the figure 2.4a, we can observe the big differences between the seven sub-population, especially between young drivers and old drivers. However we are still thinking, we underestimate the price elasticity: considering very high premium (blue point), 1.5 pts for a 5% increase seems very little.



(a) Client behaviors

(b) 2 fits vs. 1 fit

Figure 2.4: Client behaviors and average behavior

To deal with this problem, we split the data according to their price ratio: a population experiencing a price increase, those experiencing a price decrease. The fit summaries can be found in appendix.

On figure 2.4b, one can observe the big differences between the two approaches: one GLM fit for the whole portfolio vs. two GLM fits. Obviously, there is a break at a price ratio of 1, for the black curve, and the slope is higher for a price increase rather than for a price decrease. This conclusion is very informative on the market insurance, and is a strong difference to other types of market.

In conclusion to this analysis, we must admit the price elasticity in non-life insurance is a complex topic. We answer the basic questions on this topic, but many questions remain because of the few data available. To further improve our understanding of the problem, we must have the claim history and the market prices.

*. the size of the circle corresponds to the proportion of the sub-population in the whole portfolio.

2.2.2 Québec

Let us continue with Québec data, see 1.2.2 for details. For the GLM analysis, we have to remove the lines with missing variables or the lines with aberrant variable value (for example, vehicle age older than 50).

In a second time, we exclude endogeneous variable such as the indicator variable for a change of wording by the insured and the indicator for a drop of cover (see the beginnning of section 2.2). To unbias the GLM fit, we remove also those policies where an intervention of the broker is indicated.

GLM analysis

We first work on the 2007 dataset. In appendix B.2.2, the full backward selection of the explanatory variables is available. Below, we only put a short summary of the GLM fit:

```
glm(formula = did_cancel ~ prev_prem_group2 + pol_age_group2 + resp_claim_2 +
multi_veh_dsc + pricefactor * (house_pol + price_group2 + cover2) +
pricefactor:(drivage_group2 + veh_age_group3), family = binomial(), data = workdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8569  -0.4280  -0.3379  -0.2525   3.0708

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -2.98228    0.22024 -13.541  < 2e-16 ***
prev_prem_group2(1e+03,2e+03]      0.29618    0.03021   9.803  < 2e-16 ***
prev_prem_group2(2e+03,Inf]        0.47842    0.09044   5.290 1.22e-07 ***
pol_age_group2(4,Inf]             -0.32606    0.01769 -18.436  < 2e-16 ***
resp_claim_2Y                      0.12788    0.04705   2.718  0.00657 **
multi_veh_dscY                    -0.27296    0.01865 -14.639  < 2e-16 ***
pricefactor                        1.11083    0.23002   4.829 1.37e-06 ***
house_polY                        -1.25777    0.14541  -8.650  < 2e-16 ***
price_group2(15,25]               -0.54633    0.21062  -2.594  0.00949 **
price_group2(25,99]               -0.58014    0.22254  -2.607  0.00914 **
cover2TPL+opt                      0.81110    0.14208   5.709 1.14e-08 ***
pricefactor:house_polY             0.36830    0.14671   2.510  0.01206 *
pricefactor:price_group2(15,25]    0.60226    0.21874   2.753  0.00590 **
pricefactor:price_group2(25,99]    0.69869    0.23060   3.030  0.00245 **
pricefactor:cover2TPL+opt         -1.03533    0.14518  -7.131 9.95e-13 ***
pricefactor:drivage_group2(30,55] -0.22602    0.02424  -9.324  < 2e-16 ***
pricefactor:drivage_group2(55,99] -0.37122    0.02669 -13.910  < 2e-16 ***
pricefactor:veh_age_group3(5,Inf]  0.09327    0.02329   4.005 6.20e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 118078  on 239927  degrees of freedom
Residual deviance: 113018  on 239910  degrees of freedom
AIC: 113054
```

Sub-population study

As in the previous sub-section, variable effect on the lapse is analyzed through the average lapse function $\hat{\pi}_{g,n}$ assessed for different groups, such young vs. old drivers.

All the figures are plotted in appendix: figure B.9 and figure B.10. From them, we can distinguish customer behaviors, summarised in the following table:

Sluggish customers	Sensitive customers
TPL contracts, old policies	All-risk cover, young policies
old drivers, no responsible claim	young drivers, experienced a claim
cross selling house, multi-vehicle	single contract, one vehicle
old car, low-value premium	new car, high-value premium
low pricing group	high pricing group

The analysis of the GLM fitted probabilities let us to identify segments of policies with similar behaviors relative to price. The process of segmentation also includes marketing aspect, so the conclusions are much more usable. We get the following segmentation:

(*black*) - young drivers with a low pricing group,

(*blue*) - young drivers with a high pricing group,

(*red*) - old drivers with full cross-selling (household and multi-vehicle),

(*green*) - old drivers with a household policy,

(*yellow*) - old drivers with a multi-vehicle discount,

(*azure*) - old drivers with no cross-selling,

(*grey*) - working class with all risk cover and responsible claims (in last 2 years),

(*orange*) - working class with all risk cover without responsible claims and young car,

(*turquoise*) - working class with all risk cover without responsible claims and old car,

(*pink*) - working class with third-part liability cover and possibly add-on cover.

On figure 2.5 below, we plot for each population the predicted central lapse rates and the delta lapse rate (for a 5%-price increase).

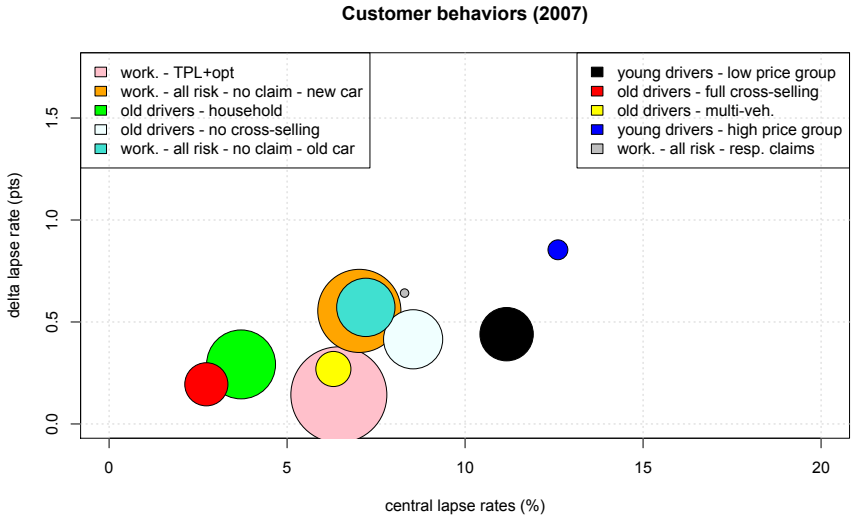


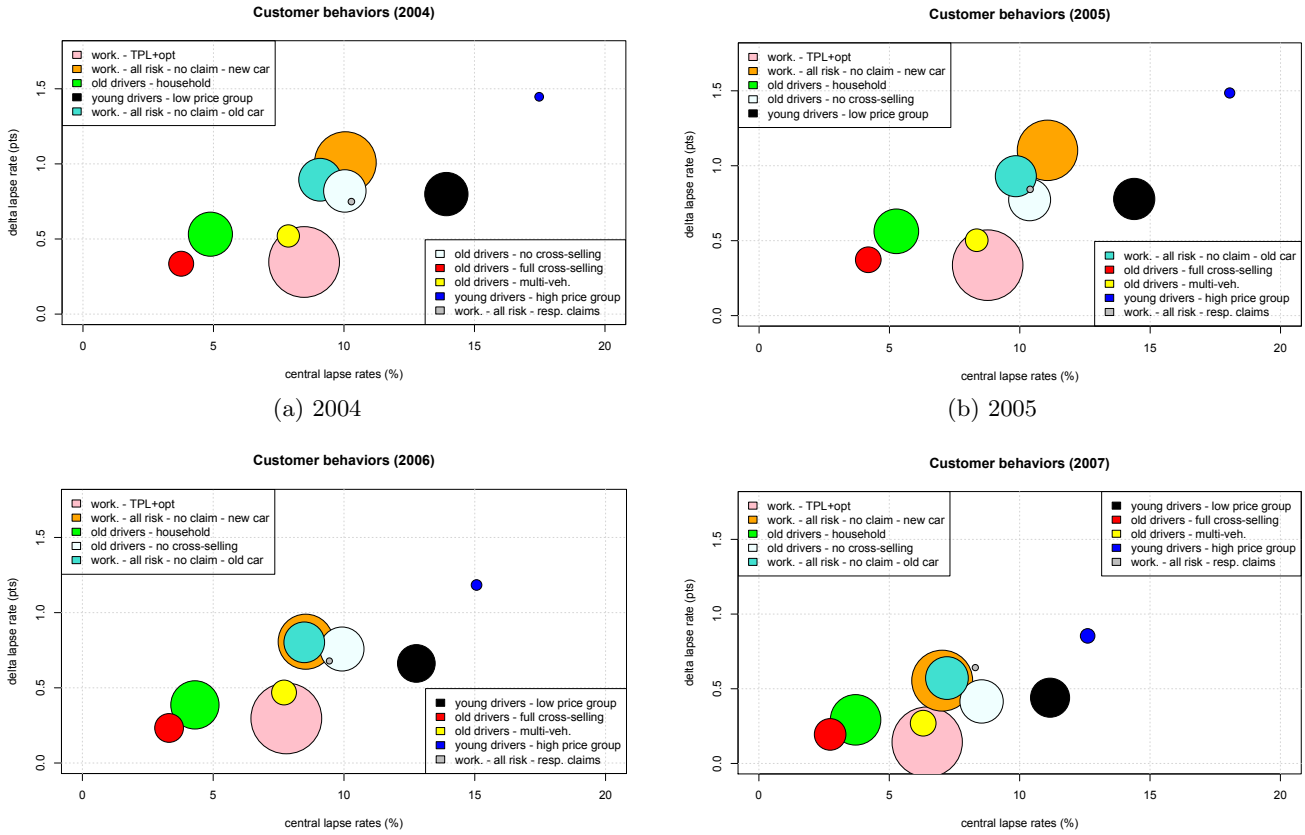
Figure 2.5: Customer behaviors

Without surprises, the most price-sensitive segment are young drivers, especially those with a high pricing group, i.e. with a valuable car. Among old drivers, there are big differences depending if the customer has many policies in AXA. Finally the "working class" (between 30 and 55 years old) cannot be well segmented, despite having a responsible claim has a significant impact on the central lapse rate.

Note that there is a strong difference with the Portugal data on the effect of old vehicle. In Portugal it is a factor of price-sensitiveness, while it has the opposite impact for Québec population. This is explained mainly by the fact, Québec data, unlike Portugal data, does not contain any lapses before renewal, in particular those customers who change vehicle and take the opportunity to shop around for another insurer.

The segmentation seems intuitive but as for Portugal data, the value of the additional lapse rate Δ are very low and so not very realistic. We split data between price increase and price decrease. Unfortunately, fitted results are less realistic than the current model, see figure B.10 in appendix.

In the current subsection, we work only with the 2007 dataset. If we work other dataset (2004, 2005, 2006), we have generally similar conclusions: the same set of significant variables, the effect of one variable on the lapse. We put all the regression summaries in appendix B.2.2 (one-variable plots have been omitted since they are almost the sames).



(a) 2004 (b) 2005

(c) 2006 (d) 2007

Figure 2.6: Heterogeneity of customer behaviors

However, there is one big difference between the 2007 results and other results. The price sensitiveness is much higher for other years. We can see it on figure 2.6 with results by populations.

As the time goes, the customers become less sensitive to price (increase) for all sub-populations and lapse less. This is especially true for 2006 and 2007 (figures 2.6c, 2.6d). Furthermore, sub-populations *orange*, *turquoise*, *azure* and *grey* stay pack together but moves from point (10%, 1pts) to (8%, 0.5pts). The most decrease in price sensitiveness is for the subpopulation *blue* (i.e. young drivers with high pricing group) going from (18%, 1.5pts) to (13%, 1pts).

We think a major reason for this phenomenon is the decreasing market trend at that time in Québec province :

Year	Market Premium	Loss Ratio (%)
2004	619.14	57.71
2005	619.94	58.33
2006	604.41	59.64
2007	592.20	59.63

As all major insurers decrease their premium (in average), customers (especially those of AXA) see their premium level decreasing. So, the market environment put the customer in a state of "sluggishness". This is reflected in figure 2.6.

We do not report here but we also try to use in the GLM regression prior explanatory variables, say 2004 variables in the GLM regression of 2005. Unfortunately, it does not really improve the model. In an attempt to take into account the dynamic, we fit a GLM three times a year, i.e. 12 fits. And then we try to model the GLM coefficients with time series. Again, it was not satisfactory.

GLM predictions vs. observed lapses The final test and challenge we can do with these data and the GLM models is to try to predict next year lapses. That is to say we use the individual lapse function

$$\hat{\pi}(p) = \frac{1}{n}\sum_{i=1}^{n} g^{-1}\left(x_i^T\hat{\beta}_{-p} + z_i^T\hat{\beta}_{+p} \times p\right),$$

where $\hat{\beta} = (\hat{\beta}_{-p}, \hat{\beta}_{+p})$ is the fitted parameter *, x_i price-independent explanatory variables, z_i price-dependent explanatory variables and g the logit function.

Assumed $\hat{\beta}$ is calibrated on 2004 data, we compute $\hat{\pi}(p_{2005})$ for all the individuals of 2005 present in the 2004 data. To compare the fitted probabilities with the binary response of 2005, we simply take the mean by sub-populations $j \in \{1, \ldots, 10\}$: $\hat{\pi}_j(p_{2005})$ vs. $r_j(2005)$.

In the table 2.3, we compare the predicted lapse rate and the observred lapse rate ($\sigma(\hat{\pi})$ denotes the standard error of $\hat{\pi}$). Most of the time we overestimate the true lapse rate, e.g. population *grey*. So we get a conservative picture of the lapses. Let us note for population *blue*, there is a kind of lag on the lapse estimate. This suggest there is a dynamic of GLM parameters. Unfortunately, we have a too few historic to verify this hypothesis with classic time series.

∗. separated between coefficients for price-independent variables and price-dependent variables.

Pop.	$r_j(2005)$	$\hat{\pi}_j(p_{2005})$	$\sigma(\hat{\pi}_j)$	$r_j(2006)$	$\hat{\pi}_j(p_{2006})$	$\sigma(\hat{\pi}_j)$	$r_j(2007)$	$\hat{\pi}_j(p_{2007})$	$\sigma(\hat{\pi}_j)$
black	13.57	13.36	0.33	12.13	13.00	0.314	10.98	11.69	0.303
blue	16.11	17.12	0.456	12.75	16.07	0.413	10.68	13.42	0.394
red	3.30	3.58	0.106	3.11	3.99	0.112	2.45	3.22	0.0973
green	4.85	4.65	0.133	4.014	4.90	0.129	3.50	4.15	0.121
yellow	7.60	7.61	0.208	7.75	7.97	0.200	6.49	7.52	0.205
azure	9.66	9.70	0.238	9.44	9.79	0.227	8.49	9.55	0.240
grey	9.61	10.55	0.494	8.88	9.55	0.424	8.57	8.77	0.462
orange	10.06	10.11	0.225	7.96	10.02	0.207	6.73	7.96	0.182
turquoise	9.05	8.79	0.202	8.15	9.00	0.197	7.11	7.98	0.198
pink	8.32	8.35	0.203	7.61	8.45	0.205	6.32	7.58	0.198
		2005			2006			2007	

Table 2.3: Accuracy of GLM predicted lapse rate (%)

In conclusion to the Québec analysis, we see a good step further in understanding the price elasticity in non-life insurance. The additional explanatory variables, we had compared to Portugal data, really enhance the prediction of the lapse rates. However the lapse values seem still dubious, despite our effort to identify distinct customer behaviors. Once we get the 10 sub-populations, we fit a GLM on each segment, but the predictions were closed to the unique GLM approach. Futhermore, the longer historic emphasizes a dynamic of price elasticity closely linked to the insurance market cycle. So the next German dataset with market prices should improve further our understanding of the problem.

2.2.3 Germany

We finish this chapter with the Germany data, see 1.3.2 for details. Note that market data are only available on TPL agent and TPL broker datasets. As for other datasets, we need to do a cleaning process to remove missing or aberrant records. In the GLM analysis, we should only consider exogeneous variables (see the beginning of section 2.2).

But excluding records where a rebate (ranging from 5% to 80% in some cases) is granted means to remove two thirds of the dataset... This is not satisfactory. To convince us it is not the right solution we fit a GLM without those records and a GLM with the full database (but without the rebate variable), the predictions were not very different.

The final solution, we choose, is to keep the records with rebates but to always include the rebate variable as explanatory variable in the GLM. We also need to take into account the granted rebates in the predictions. So we force to a null rebate in the predictions.

At our disposal, we have quite a large (but not the whole) database with different channels and cover types. Unless precised otherwise, the default dataset is the 2004 dataset. In table 2.4, we put the relative portfolio size by channel distribution and by cover type.

	TPL	PC	FC
Agent	11.104	20.324	30.026
Broker	5.155	9.573	11.728
Direct	2.625	5.284	4.176

Table 2.4: Proportions of portfolio by channel distribution and by cover type

Dataset split

The biggest part of policies are full comprehensive (FC) contracts sold through tied-agents (30%) and then other distribution channels for that cover. To split or not to split, that is the question?! Do we need (or not) to subdivide the 2008 dataset into the 9 nine subsets above before fitting GLMs? We did some tests with different subdivisions of the dataset.

We report in table the average lapse prediction $\hat{\pi}(p)$ for three different values of price ratio ($p = 0.95, 1, 1.05$) where $\hat{\pi}(p)$ is defined as

$$\hat{\pi}(p) = \frac{1}{n}\sum_{i=1}^{n} g^{-1}\left(x_i^T\hat{\beta}_{-p} + z_i^T\hat{\beta}_{+p} \times p\right),$$

with x_i's and z_i's are explanatory variables. Note that z_i's might depend on the price ratio p, so we need to update the covariate accordingly *.

*. e.g. the `diff2tech` variable depends on p since it is the difference between the proposed premium and the technical premium. We also force the rebate (one variable of the x_i's) to zero.

In the following, we select by a backward approach the significant explanatory variables for different datasets and report the prediction $\hat{\pi}$'s.

Premium change	-5%	0%	+5%	-5%	0%	+5%
Agent	6.55	7.65	9.18	6.67	7.94	9.59
Broker	8.75	11.10	13.94	8.67	10.81	13.52
Direct	10.43	11.60	12.72	9.00	11.39	14.44
Channel	One fit by channel			One fit for all channels		

Table 2.5: Predicted lapse rates for the TPL cover

On table 2.5, we can see that using the whole TPL dataset or the three subsets has a huge impact on lapse rate predictions. There is no obvious orders between those predictions.

Premium change	-5%	0%	+5%	-5%	0%	+5%
TPL	10.43	11.05	12.76	10.89	11.52	12.10
PC	10.02	11.25	12.48	10.67	11.39	12.17
FC	11.94	13.01	14.16	12.94	12.97	13.03
Cover	One fit by cover			One fit for all covers		

Table 2.6: Predicted lapse rates for the direct channel

On table 2.6, we observe that the predictions are quite close between the unique model for the whole dataset and the three distinct models. But, the predictions for the three separated models are higher than those for the unique model when considering a premium increase. Hence we are tempted to conclude that separate GLMs are better.

The opposite is observed for the TPL cover, thus the predictions are probably underestimated. However in table 2.5, we do not use the market data for the three-GLM approach. In table 2.7, we put the results when using market data.

Premium change	-5%	0%	+5%
Agent	5.63	6.73	8.31
Broker	NA	NA	NA
Direct	10.10	11.96	14.48

Table 2.7: Using market data

Those results are in line with the unique GLM approach of table 2.5. So in the following we will consider a different GLM for the nine subpopulations. In appendix B.2.3, we put the regression summaries for the nine regressions.

Channel and cover effects

On figure 2.7, we plot the central lapse rates and the delta lapse rates for the nine lapse rates. Unsurprisingly, broker and direct lines have the biggest lapse rates and the highest deltas. Moreover, in terms of lapse levels, theses two liens are equivalent but in terms of price sensitiveness, the broker deltas are clearly above.

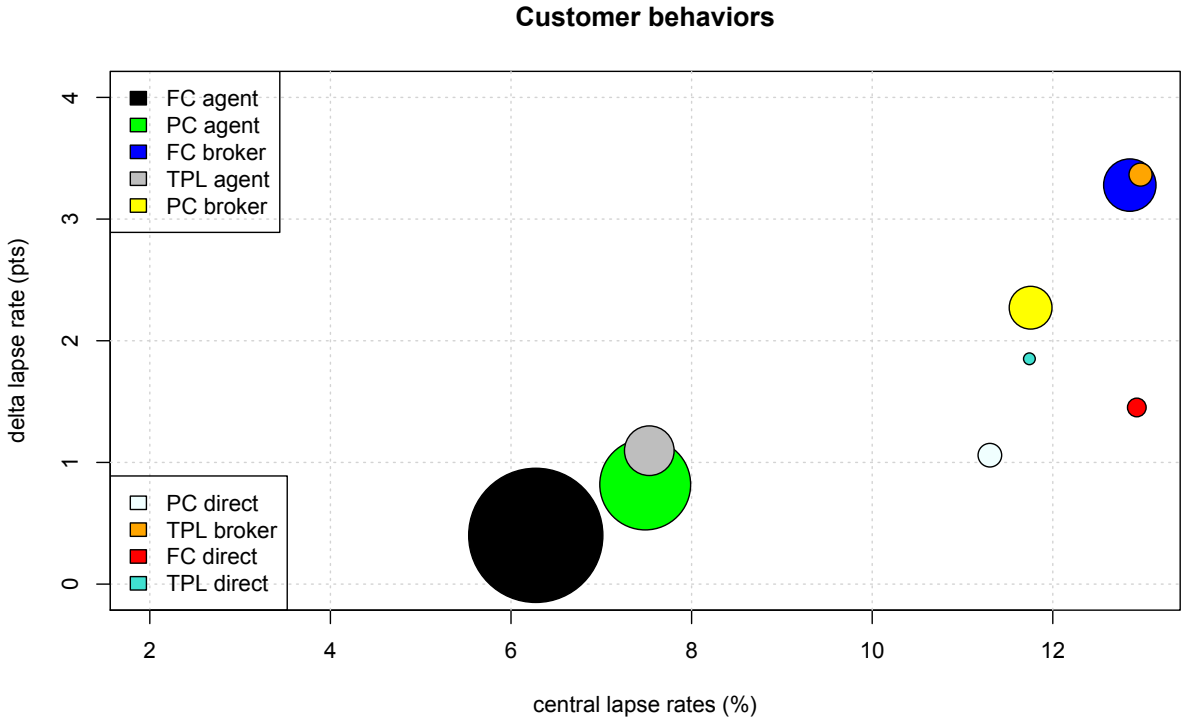


Figure 2.7: Comparison of distribution channels and cover types

Moreover, the better is the cover the lower is the price sensitiveness. When non tied-agent channel, TPL contracts are the targets of very active customers, those that shop around the most. But for tied-agent channel, the most price-sensitive population is PC contracts. As anticipated the most sluggish population are customers with FC policies.

We try to split this population into 4 groups to see the deviations from the average behavior of the FC agent *black* bubble. Four groups have been tested: (i) policies with no cross-sold products, (ii) policies with cross-sold products and policyholder working in the public sector, (iii-iv) policies with cross-sold products and policyholders working in private companies (splitting between "old" and "young" policyholders). It reveals that most FC policies lapse differently (x-axis differentiation) but do *not* react differently with respect to prices (y-axis differentiation).

Explanatory variables

In term of explanatory variables, we send the reader to the appendix B.2.3 for the regression details. In summary, the most relevant explanatory variables across the nine regressions are the

difference with technical premium (`diff2tech`), the policyholder age (`polholderage`) and the number of AXA contracts in household (`householdnbAXA`).

Furthermore, when the database is sufficiently large, the region (`region2`), the claim number (`nbclaim0608percust`, `nbclaim0708percust` and `nbclaim08percust`) and the bonus evolution (`bonusevol`). Note that the market variables (`diff2top10vip` and `diff2top10direct`) are also significant.

The table-figure 2.8 focuses on the effect of policyholder living region. We list the central lapse rates and the deltas in table 2.8a for two distribution channels. Combining with the map (figure 2.8b), we see that state-cities such as Hamburg, Berlin have a higher lapse rates and deltas. This can emphasize a different level of competition in big cities (than in rural areas) where it is easier to see competitor offers.

Due to the small differences on lapse rates (around 13%) and deltas (around 3%) for the broker channel, one can also conclude that brokers really do what they are pay for. The agent channel appears more volatile both in terms of lapse levels and deltas, probably due to a different level of competition in the German regions.

Num	Region	Central rate	Delta
1	Schleswig-H.	6.73 / 12.67	0.93 / 3.26
2	Hamburg	9.79 / 12.57	1.35 / 3.23
3	Niedersachsen	6.62 / 12.76	0.94 / 3.31
4	Bremen	8.92 / 13.46	1.27 / 3.32
5	Nordrhein-W.	8.03 / 13.28	1.11 / 3.28
6	Hessen	7.21 / 13.2	1.04 / 3.37
7	Rheinland-P.	6.52 / 12.76	0.93 / 3.32
8	Baden-W.	7.01 / 13.25	1.03 / 3.46
9	Bayern	6.69 / 13.18	1.03 / 3.52
10	Saarland	7.69 / 12.6	1.45 / 3.39
11	Berlin	8.91 / 12.74	1.11 / 3.11
12	Brandenburg	8.70 / 12.27	1.32 / 3.29
13	Mecklenburg-V.	8.90 / 12.7	1.42 / 3.40
14	Sachsen-A.	9.62 / 12.59	1.55 / 3.38
15	Sachsen	9.70 / 12.66	1.49 / 3.32
16	Thüringen	9.42 / 12.23	1.45 / 3.33

(a) Statistics for TPL agent/broker

(b) Map of Germany

Figure 2.8: Customer behaviors in the German market

The lapse is also higher in eastern Germany (states 11 to 15) than in western Germany (states 5 to 9). We do not have an answer to explain this phenomenon except that AXA might be cheaper than competitors (especially Allianz and HUK) in those states.

Fitted vs. observed values

As we want to challenge the GLM approach, we compare the observed lapse rates by region r and the average of the fitted lapse rates for region r. Note that the average fitted lapse rate and the predicted central rates are different concepts: the first one is defined as

$$\tilde{\pi} = \frac{1}{n}\sum_{i=1}^{n} g^{-1}\left(x_i^T\hat{\beta}_{-p} + z_i^T\hat{\beta}_{+p} \times p_i\right),$$

while the second one is

$$\hat{\pi}(1) = \frac{1}{n}\sum_{i=1}^{n} g^{-1}\left(x_i^T\hat{\beta}_{-p} + z_i^T\hat{\beta}_{+p} \times 1\right).$$

In table 2.8, we report the observed lapse rates π_r and the fitted lapse rates $\tilde{\pi}_r$. For the agent channel, the fitted values are close to the observed ones. But this is not the case for the broker and the direct channel: fitted values are very poor.

Num.	Region	Observed	Fitted	Observed	Fitted	Observed	Fitted
1	Schleswig-H.	6.53	6.67	8.93	10.08	10.79	11.12
2	Hamburg	8.60	9.65	9.27	11.02	8.00	10.98
3	Niedersachsen	6.85	6.61	12.66	11.12	11.49	11.21
4	Bremen	7.32	8.77	10.80	12.21	11.18	11.76
5	Nordrhein-W.	8.08	8.11	11.34	11.67	13.03	11.27
6	Hessen	6.80	7.20	11.03	11.63	11.63	10.97
7	Rheinland-P.	6.72	6.53	11.14	11.00	10.39	10.93
8	Baden-W.	6.91	7.02	10.51	10.53	10.28	11.24
9	Bayern	6.77	6.75	10.45	10.70	11.36	12.12
10	Saarland	9.67	9.74	10.57	10.24	11.15	12.16
11	Berlin	9.44	8.98	12.44	11.7	13.16	12.24
12	Brandenburg	9.25	8.73	14.04	10.98	13.37	13.03
13	Mecklenburg-V.	8.24	8.94	12.52	12.79	13.6	13.98
14	Sachsen-A.	9.84	9.68	15.1	13.05	16.17	13.54
15	Sachsen	9.13	9.72	12.81	13.19	15.37	14.07
16	Thüringen	9.41	9.47	13.12	12.61	12.78	13.74
		Agent channel		Broker channel		Direct channel	

Table 2.8: Lapse rates by region for TPL cover

We suspect that the gap of significant explanatory variables between the agent channel and other channels explain the difference. And there are few significant variables for the broker channel, because the datasets are small.

This is confirmed by the table B.30 (in appendix B.2.3), which contains observed-fitted lapse rates for FC cover. The fitted lapse rates are in line with the observed lapse rates for all channels. There is no wrong fitted lapse rates for broker and direct channels as there are in table 2.8.

Asymetric information and adverse selection testing

Definition Asymmetry of information occurs when two agents (e.g. a buyer and a seller of insurance policies) do not have access to the same information. Thus one of agents takes advantage of additional information in the deal. Typically, two problems can result from this information asymmetry: adverse selection and moral hazard. In insurance context, moral hazard can be observed in certain cases for high risk individuals, who will take more risks than if they were not insured.

Adverse selection is a different situation where the buyer of insurance coverage has a better understanding and knowledge of the risk he will transfer to the insurer than the insurer. So the buyer would choose a deductible according to its own risk. Hence high-risk individuals will have the tendency to choose lower deductibles.

The topic is of interest in customer behaviors, since a premium increase in hard market cycle phase (i.e. an increasing premium trend) can lead to higher loss ratio. Indeed if we brutally increase the price for all the policies by 10%, only high-risk individuals will renew their contracts (in an extreme case). Therefore the claim cost will increase per unit of sold insurance cover.

Deductible model In this report, we follow the framework of Dionne et al. (2001), which uses GLMs to test the evidence of adverse selection *. The approach is derived from the pioneer work of Rothschild & Stiglitz (1976), who models the insurance market with individuals choosing a "menu" (a couple of price and deductible) from the insurer offer set. Within this model, high-risk individuals choose contracts with more comprehensive coverage.

With this mind, Dionne et al. (2001) want to quantify the asymmetry of information benefic to the insured. To test this effect, we must verify the independence of between the consumer (deductible) choice and the endogeneous variables Y such as the observed claim number. Let Z be the deductible choice (an univariate discrete variable in $\{0, 1, \ldots, K\}$).

Z is modelled with an ordered logit model.

$$P(Z_i \leq k/X_i, Y_i) = g(\theta_k + X_i^T \beta + Y_i^T \gamma),$$

for individual i and deductible k, with g a distribution function (typically the logistic distribution) and X_i exogeneous explanatory variables as opposed to endegeneous variables Y_i. The parameters of this model equation are β and γ, the regression coefficients and θ_k the treshold parameter.

The treshold parameter is linked with the response variable Z by the equation

$$Z = k \Leftrightarrow \theta_{k-1} < U \leq \theta_k,$$

where U is a latent variable. So the trick to go from a binary model to a polytomous model is to have different intercept coefficients θ_k's for the different categorical value k.

∗. Similar works on this topic also consider the GLMs, see Chiappori & Salanié (2000) and Dardanoni & Donni (2008).

Alternative models could be multinomial models, (see, e.g., Fahrmeir & Tutz (1994)) where Z would be a multivariate binary vectors. In this case, the probability are modelled by

$$P(Z_i = e_k / X_i, Y_i) = \pi_k$$

where e_k is the unitary vector $(0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^K$ with a one at the k position. So there is vector of probabilities $(\pi_1, \ldots, \pi_K)$ linked to a linear predictor $X_i^T\beta + Y_i^T\gamma$ to estimate rather than multiple intercepts.

On explanatory variables To avoid drawing wrong conclusions about the adverse selection, Dionne et al. (2001) suggest to have the following variables into the endogeneous part Y:
– the observed number of claims,
– the expected number of claims,
where the expected number of claims is computed with a GLM and a Poisson or a quasi-Poisson family (see table 2.1).

In the application, many exogeneous explanatory variables will be tested for X, including the policyholder age, the vehicle age, the car class, the bonus-malus SF and the bonus evolution.

Application We test the information asymmetry on two datasets: FC agent and FC broker sets. Let us note that we cannot test it on TPL covers, since there is no deductible for this cover. We process in two steps, first we fit the deductible regression and then use the deductible probabilities in the lapse regression.

The numerical applications reveal that it is more relevant to regroup some deductible values which are too few in the dataset. Typically, the deductible is valued in {0, 150, 300, 500, 600, 1000, 2000, 2500}. As 300 euros is the standard deductible, very high deductibles are rarely chosen. So we regroup deductible values greater than 500 together, see table 2.9. Smaller deductible values might reveal high-risk individuals, so we decide to keep those values.

Deductible (€)	0	150	300	500	0	150	300	500
Proportion (%)	5.17	10.29	70.85	13.68	4.78	7.85	68.21	17.46
	Agent channel				Broker channel			

Table 2.9: Frequency table for FC deductibles values

In appendix B.2.3 , we put the regression summaries of the claim number and the deductible choice. The main significant variables for the claim number regression are the policyholder age, the difference between the driver and the policyholder, the SF class for FC cover.

Note that the intercept coefficients θ_k are denoted by `y>=150`, `y>=300` and `y>=500` in the summary. While for the deductible regresion, the main significant variables are the policyholder age, the SF class for FC cover and the estimated claim number coming from the GLM regression (i.e. fitted values of the first regression).

Looking at the coefficient p-values of the deductible regression, we can detect the presence of information asymmetry if the observed claim number is significant (at a 5% level). So we observe that this variable is significant for the FC agent dataset. However, the coefficient sign is positive, which implies that policies experiencing the highest number of claims are those with highest deductibles. This is a little counter-intuitive. However the sign of the expected claim number is negative, so it is hard to draw solid conclusions.

For the FC broker dataset, both coefficients have a negative sign, but the observed claim number is not significant. So we remove this variable for the following analysis. Nevertheless, we will use the deductible choice probability in the lapse regression for both channels.

From this analysis, we get the individual deductible probability $P(Z_i = k)$ with $k = \{0, 150, 300, 500\}$. As the purpose of information asymmetry testing is to see a non-standard behavior of the customers. We will use as explanatory variables the following probabilities $P(Z_i = 0)$, $P(Z_i = 150)$ and $P(Z_i = 500)$ in the lapse regression.

In table 2.10, we put the overall lapse rate predictions. At this aggregate level, taking into account the information asymmetry is not very convincing. The difference between the base and the "enhanced" approaches is very small.

Premium change	0%	+5%	0%	+5%
Base fit	6.274	6.675	12.853	16.132
With deduc. prob.	6.268	6.666	12.911	16.193
	Agent channel		Broker channel	

Table 2.10: Overall lapse predictions - Asymmetry of information impact

From the Germany analysis, we conclude that having the market variable to model insurance lapse deeply increases the accuracy of the lapse rate prediction. And in some cases (such as direct channel), it can unbias the original predictions. The lapse values seem correct for non tied-agent segments, but still remain underestimated for the tied-agent channel. For the Germany data, we do not spend time to identify customer segment, we only look at the region segment. But one can identify customer segments using the significant explanatory variables. Finally, we test the use of information asymmetry, but it proves to be inefficient at our aggregate output level.

2.3 Pros and cons of the GLM methodology

This section summarizes the advantages and the drawbacks of the GLM methodology when modelling lapse rates in non-life insurance.

2.3.1 Advantages

GLM is a classic and well-known method in actuarial science. This fact motivates our use to model lapse rate. Since it is a classic, fitting method and variable selection use state-of-art algorithms. So there is absolutely no problem in applying GLMs for daily use and estimation are also robust.

The goal of this memoir is to estimate an aggregate lapse rate curve function of the price ratio, which takes into account the individual characteristics of each policies of the portfolio. The GLM methodology fulfills this objective. Furthermore, using the predicted lapse rate values of GLMs, it is easy to identify customer segments, which react very differently to premium changes.

Finally, the back-fit of the GLMs on the identified population is good. So at an aggregate level or a customer segment, the GLM methodology provides a fair estimate of lapse rate and price sensitiveness for reasonable premium changes. But at a policy level, we think the predictions should be treated carefully.

2.3.2 Drawbacks

The GLM lapse rate predictions strongly depends on the data, so if the database present a small range of premium change, the predictions will be reliable only for a relatively small range of price change. We think, for high price change, the delta lapse rates are underestimated.

Moreover, as seen with the Québec data, the GLM follows the market cycle dynamic, i.e. in a decreasing trend, the lapse rate level decreases despite the customers do not change. It is not surprising since the GLM is a static setting which estimates the spot price elasticity. We also notice that the standard errors seem too small.

Despite some efforts to catch the dynamics, a too short historic renders the use of time serie modelling impossible. But we think this is a first step to model the price elasticity dynamic.

Finally, on the data we have, the use of information asymmetry testing does not reveal to be very successful. At the aggregate level, it does not provide new insight on our topic. But it will probably be useful at individual level for pricing and customer segmentation.

Chapter 3

Generalized Additive Models

The Generalized Additive Models (GAM) were introduced in the 90's by Hastie & Tibshirani (1990) by unifying the generalized linear models and additive models. So the generalized additive models combines two flexible and powerful methods: (i) the exponential family which can deal with many distribution for the response variable and (ii) additive models which relax the linearity assumption of the predictor. This chapter is divided into three sections: (i) model presentation, (ii) case studies and (iii) conclusions.

3.1 Model presentation

In this section, we present the Generalized Additive Models in two steps: from linear to additive models and then from additive to generalized additive models. Smoothing and fitting algorithms are then briefly presented. This section is divided into two parts: (i) theoretical description of GAMs, (ii) explanations on binary models and model selection.

3.1.1 Theoretical presentation

From linear to generalized additive models

Assuming observations X_i and response variables Y_i are i.i.d. random variables having the same distribution of generic random variables X and Y respectively. In a linear model, the model equation is

$$Y = X\Theta + \mathcal{E}$$

where Y as always stands for the response variable, X the design matrix and $\mathcal{E}$ the random noise. The parameter vector Θ has an easy interpretation, but it is not the most flexible model to model the relation between the response and the explanatory variables. One candidate to extend the linear model is the additive models defined by

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \mathcal{E},$$

with f_j smooth function of the jth explanatory variable X_j and $\mathcal{E}$ is assumed to be a centered random variable with variance σ^2.

The extension to Generalized Additive Models (GAM) is very similar to the previous chapter. A GAM is characterized by three components:

1. a random component: Y_i follows a distribution of the exponential family $\mathcal{F}_{exp}(\theta_i, \phi_i, a, b, c)$ *,
2. a systematic component: the covariate vector X_i produces a linear predictor $\eta_i = \alpha + \sum_{j=1}^{p} f_j(X_{ij})$,
3. a link function $g : \mathbb{R} \mapsto S$ which is monotone, differentiable and inversible, such that $E(Y_i) = g^{-1}(\eta_i)$,

for $i \in \{1, \ldots, n\}$, where θ_i is the shape parameter, ϕ_i the dispersion parameter, a, b, c three functions (characterizing the distribution) and f_j's smooth functions.

Note that the linear model (and GLM) is a special case of additive model (and GAM) with $f_j(x) = \beta_j x$. But much more complicated function can be used. The next subsection presents possible smooth functions we can use for f_j.

Smoothing for univariate data

In this sub-section, we present briefly some classic smoothing procedures. Probably the simplest method to get a smooth function is to regress a polynom on the whole data. Assuming observations are denoted by $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, a multiple regression model

$$Y = \alpha_0 + \alpha_1 X + \cdots + \alpha_p X^p,$$

does the job. Let us work on the `cars` data containing braking distances of a vehicle for different speeds. On the figure 3.1, we plot the fitted values for $p = 1, 2, 3$ †

Using $f(x) = \sum_i \alpha_i x^i$ is clearly not flexible and a better tool has to be found. One way to be less rigide in the smooth function is to subdivide the interval $[\min(x), \max(x)]$ into K segments. And then we can compute the average of the response variable Y on each segment $[c_k, c_{k+1}[$. This is called the bin smoother in the literature. As shown on Hastie & Tibshirani (1990) figure 2.1, this smoother is rather unsmooth.

Another way to find a smooth value at x, we can use points about x, in a symmetric neighborhood $N_S(x)$. Typically, we use the k nearest point at the left and k nearest at the right of x to compute the average of y_i's. We have

$$s(y|x) = \frac{1}{\text{Card}N_S(x)} \sum_{i \in N_S(x)} y_i,$$

where the cardinal $\text{Card}N_S(x)$ does not necessarily equal to $2k + 1$ if x is near the boundaries. Again we do not show the result and refers the reader to Hastie & Tibshirani (1990) figure 2.1. This method, called the running mean, takes into account better the specificities of the data. However we lose the smoothness of previous approches.

∗. See appendix A.1

†. The intercept α_0 was removed to avoid predicting non null values for a null speed.

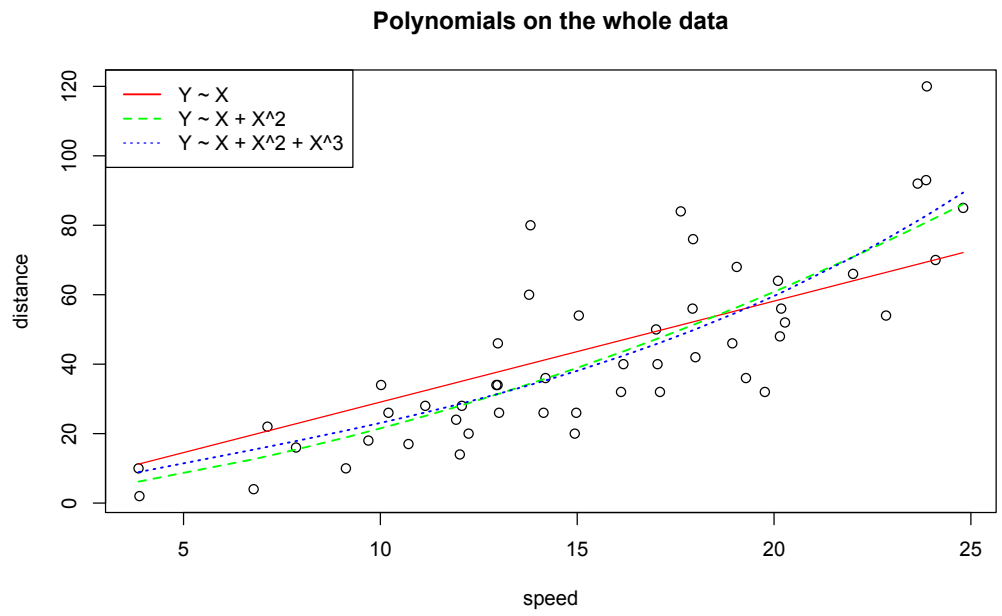


Figure 3.1: Polynom regression

An extension of this approach is to fit the linear model $y = \mu + \alpha x$ on the points (x_i, y_i) in the neighborhood (for $i \in N_S(x)$). That is to say we have a serie of intercepts μ and slopes α for all observations. We called this method the running line, which generalizes the running mean, where α is forced to 0.

Another enhancement is to weight the points in the regression (for x_i) inversely relative to the distance to x_i. Generally we use the tricube weight function

$$w(z) = (1 - |z|^3)^3 \mathbb{1}_{|z|<1}.$$

So the weight for x_j when computing the smooth value of x_i is $w(z_j)$ with $z_j = \frac{|x_i - x_j|}{b}$ and b the bandwith. Introduced by Cleveland (1979), this method is known as LOcally WEigthed Smoothing Scatterplots (LOWESS). Other weight function can be used as long as it is a symmetric, decreasing from 0, strictly positive on $]-1, 1[$ and null elsewhere.

Summarising the LOWESS approach, we have w a weight function, d the degree of polynoms (1 for running line, 0 for running mean), f the span defined as the proportion of points to use in the regression*. On the figure 3.2, we plot different LOWESS methods for 2 spans. Clearly the righthand side plot is smoother than the lefthand one. For a given span, as one can expected the running mean method is more robust, unlike the running square which goes up and down easily.

A popular method for smoothing is the Kernel smoothing. Choosing a Kernel function k, the associated smoother is

$$s(y|x) = \frac{\sum_{i=1}^n k\left(\frac{x-x_i}{b}\right) y_i}{\sum_{i=1}^n k\left(\frac{x-x_i}{b}\right)},$$

where k denotes the Kernel function, similar to a weight function and b the bandwidth.

∗. closely linked to the bandwidth.

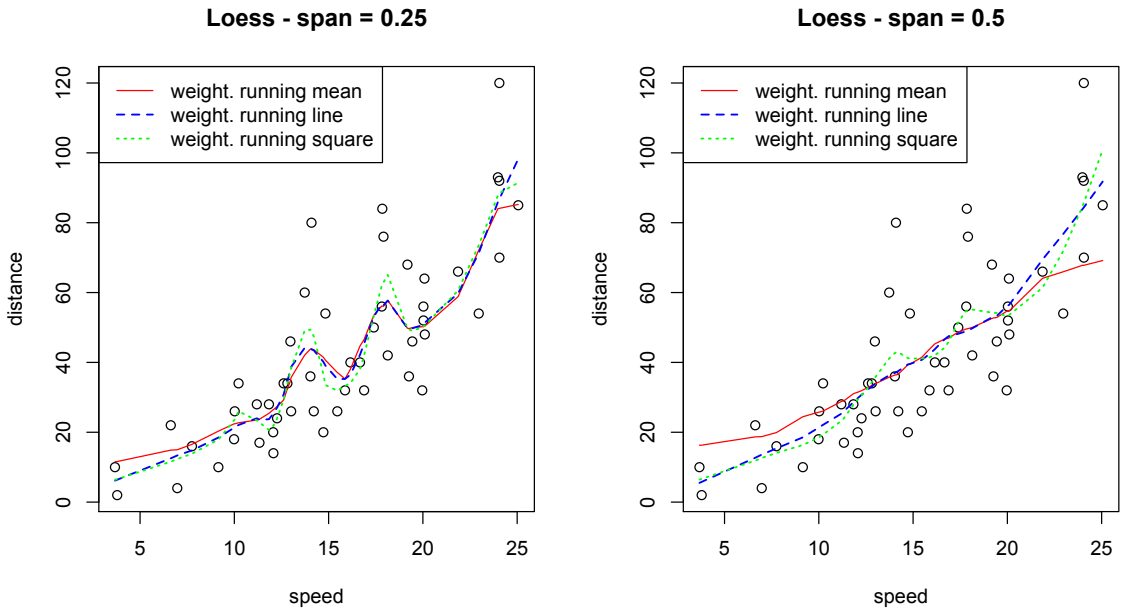
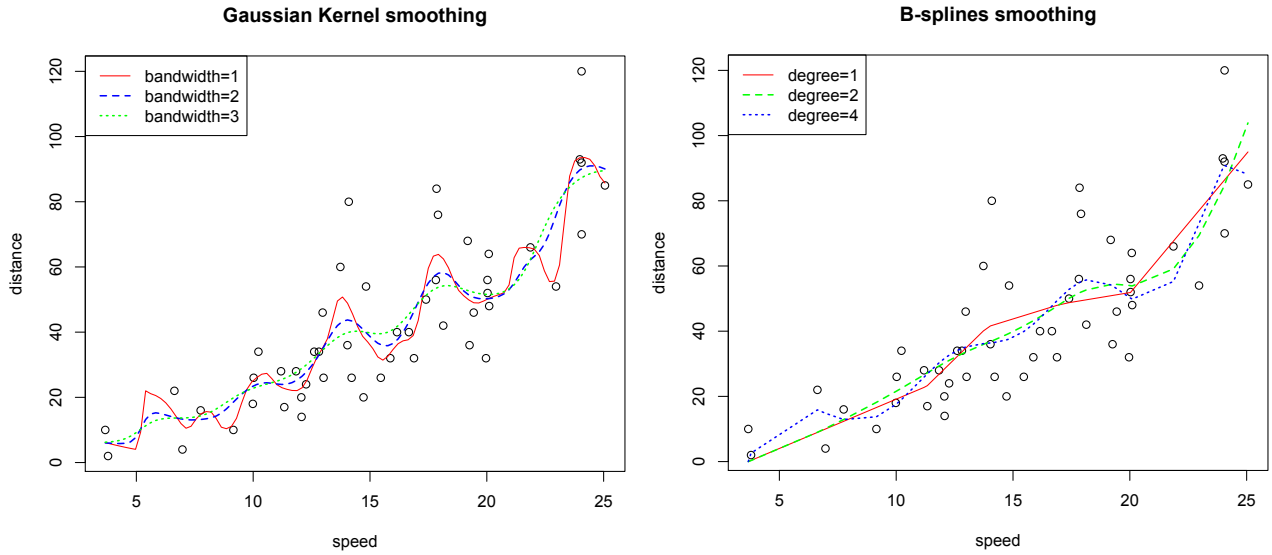


Figure 3.2: LOWESS

As suggested in Venables & Ripley (2002), Kernel smoothing can be seen as a local weighted running mean approach. However, the power of the Kernel approach is to use different Kernel functions. Common Kernel functions are the standard normal density function or the Epanechnikov function (a bisquare function).



(a) Kernel smoothing

(b) B-spline smoothing

Figure 3.3: Kernel and B-spline smoothing

As you can see on figure 3.3a, not surprisingly increasing the bandwidth increases the smoothness of the fitted curve.

The last and recent tool to fit a smooth curve is to use spline functions. The approach consists in splitting the interval in K knots $(t_1, \ldots, t_K)$ and fit a polynom on each segment, while imposing smooth conditions at the knots. One intuitive spline function is the polynom of third degree, since the smooth conditions ($f' = f'' = 0$) are easily written done.

A crucial fact is that using K polynoms can be represented by a banded matrix, a band for each segment. Using the matrix representation emphasizes that we use a basis of functions to approximate the function f. Many polynomial basis can be used (e.g. Bernstein polynoms for Bézier curves).

One popular basis is the B-spline basis. They are defined recursively starting polynoms of degree 0 defined by $B_{i,0}(t) = \mathbb{1}_{t_i \leq t < t_{i+1}}$ * and higher order $B_{i,d}$ obtained by convex combination (with t_i's increasing knots) of $(B_{i,d-1})_i$'s. To fit the data, we then minimize a penalized least square and use some quantiles as the knots (for instance quartiles).

On the figure 3.3b, we use the quantiles 20%, 40%, 60% and 80% as interior knots and three different degrees. As one can expect, B-splines with high degrees better fit the data. To conclude with this smoother presentation, all smoothers presented here are linear smoothers, since they can be written by $\hat{y} = Sy$ with S the smoother matrix (depending on observations x).

Fitting algorithms for GAM

In the previous subsection, we present a long list of smooth procedures. But we do not explain how to use it in an additive model. All smoothers have a smoothing parameter λ, (the polynom degree, the bandwidth or the span). By penalized least square we can fit a smoother S_λ for a given λ. A first problem is how to choose a criterion on which to optimize λ (hence to have an automatic selection). A second problem is to find a reliable estimate of the parameters α and smooths coefficients given a smoothing value λ.

We take the problem in the reverse way. Assuming a value of λ, we present a fitting method to calibre the model. In Hastie & Tibshirani (1990), they propose a local averaging generalized Fisher scoring method. However Wood (2008) propose a most recent and reliable method: a PIRLS method.

The Penalized Iteratively Reweighted Least Square method (PIRLS) is unsurprisingly an iterative method aims to minimize the penalized deviance

$$D(f_1, \ldots, f_p) + \sum_{j=1}^{p} \lambda_j \int f_j''(x_j)^2 dx_j.$$

The second term penalizes the wiggly behavior of smooth functions. Given a set of basis functions b_{jk}, we can express f_j as $f_j(x) = \sum_{k=1}^{K_j} \beta_{jk} b_{jk}(x)$. So the GAM can be represented as a GLM with $eta_i = \widetilde{X}_i \beta$ with $\widetilde{X}_i$ containing the basis functions evaluated at the covariate values and β containing α and the β_{jk}'s.

∗. See theorem 1.5 of Steihaug (2007).

Hence the penalized deviance can be rewritten as

$$\widetilde{D}(\beta) = D(\beta) + \sum_j \lambda_j \beta^T S_j \beta,$$

where S_j contains known coefficients and zero's and $D(\beta)$ the deviance defined the "GLM" version of the GAM. The PIRLS has the following scheme

- Initiate μ_i^0 typically with y_i.
- Iterate while no change in deviance $\widetilde{D}(\beta^k)$
  - compute the weight $w_i^k = \frac{1}{g'(\mu_i^{k-1})}\sqrt{\frac{\omega_i}{V(\mu_i^{k-1})}}$,
  - evaluate the pseudo data $z_i^k = g'(\mu_i^{k-1})(y_i - \mu_i^{k-1}) + \eta_i^{k-1}$ with $\eta_i^{k-1} = g(\mu_i^{k-1})$,
  - minimize over β the least square objective

$$||yW(z - \widetilde{X}\beta)||^2 y + \sum_j \lambda_j \beta^T S_j \beta,$$

    with $W = \text{diag}(w_1, \ldots, w_n)$. We get β^k.
  - prepare next estimate with $\eta^k = \widetilde{X}\beta^k$ and $\mu_i = g^{-1}(\eta_i^k)$.

Now we have a method PIRLS that for a λ gives the corresponding $\hat{\beta}(\lambda)$. We must find a criterion to select the vector λ. In the literature, there are many criteria to select the smoothing parameters:

- Restricted Maximum Likelihood REML,
- Maximum Likelihood ML,
- Generalized Cross Validation GCV,
- Generalized Approximate Cross Validation GACV.

These methods defer from one another if the smoothing parameter is treated as a random effect or not. So we either maximize a quantity linked to the likelihood (ML/REML) or minimize a prediction error (GCV/GACV).

Expressions of log-likelihoods ML and REML can be found in Wood (2010) with equation (4) and (5). Their expression use the deviance of the model, the satured deviance and a third-term penalizing the wiggliness of the smooth function f_j. The optimization procedure consists in using a Newton method for the optimization of the parameter λ where in each iteration a PIRLS is used (to find $\beta(\lambda)$). So this is a nested optimization where outer iteration optimizes over λ and the inner iterations optimized over β *.

As already said, an alternative approach seeks in minimizing the prediction error. The predictive error seems to be counter-intuitive but the usual justification is a leave-one-out argument. The leave-one-out consists in computing n deviances D_{-i} where D_{-i} is the deviance without the ith observation. The deviance cross validation is just a sum of the D_{-i}'s. In practice we do not fit n times the model (clearly too expensive!) but an approximate is used to compute the GCV or GACV (see Wood (2008) for details). Then again, a nested optimization procedure using the PIRLS scheme is used.

∗. See page 8 of Wood (2010) for the overview of the algorithm and the rest of the paper for details on the computation of the objective, the gradient and the Hessian functions.

On figure 3.4, we plot smooth functions for the `cars` dara with different smoothing selection and basis functions.

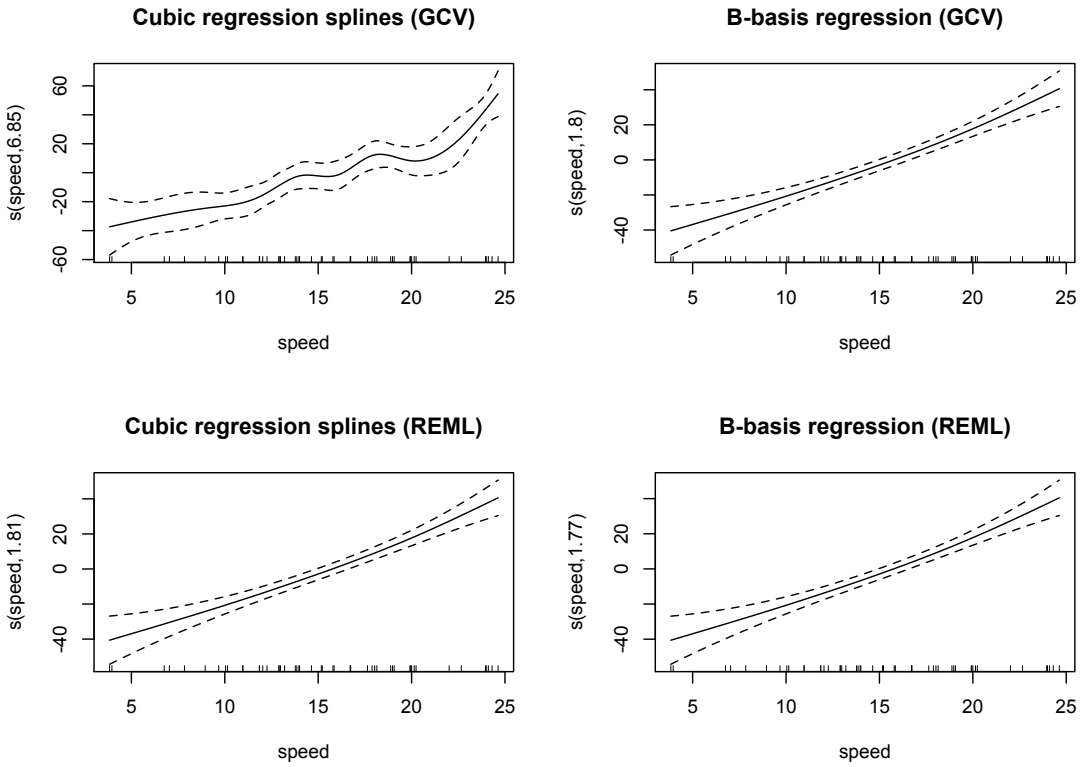


Figure 3.4: Additive model tests

For the REML criterion, the function basis has no influence while the GCV leads to dramatic different estimation of the smooth function f. Let us note that the estimated degrees of freedom are also very closed, except for the "top-left" method. In the following, we will use the REML criterion which seems more stable. It is also the recommended criterion by S. Wood.

3.1.2 Binary regression and model selection

As for GLMs, the binary regression means that we assume Y_i follows a Bernoulli distribution $\mathcal{B}(\pi_i)$. So we have the model equation is

$$\pi_i = g^{-1}(\eta_i),$$

where g is the link function and η_i the predictor. Unlike the GLM where the predictor was linear, for GAMs the predictor is a sum of smooth functions:

$$\alpha_0 + \sum_{j=1}^{p} f_j(X_j)y \quad \text{or} \quad \alpha_0 + \sum_{i=1}^{p_1} \alpha_i X_i + \sum_{j=1}^{p_2} f_j(X_j),$$

the latter being a semi-parametric approach.

As suggested in Hastie & Tibshirani (1995), the purpose to use linear terms can be motivated to avoid too much smooth terms which can noise one another. For instance, if a covariate represents the date or the time of events, it is "often" better to consider it as an increasing or decreasing trend with a single parameter α_i.

As for GLMs, we are able to compute confidence intervals using the Gaussian asymptotic approximation of the estimators. So the variable selection for GAMs is similar to those of GLMs. The true improvement is the higher degree of flexibility to model the effect of one explanatory variables on the response.

The procedure for variable selection is similar to the backward approach of GLMs, but taking into account a term has to be dropped only if no smooth function of it and no linear function of it is relevant. So a poor significance of a variable modelled by a smooth function might be significant when modelled by a single linear term.

We will use the rules of Wood (2001) to drop a term:
(a) Is the estimated degrees of freedom for the term close to 1?
(b) Does the plotted confidence interval band for the term include zero everywhere?
(c) Does the GCV score drop (or the REML score jump) when the term is dropped?
If the answer is "yes" to all questions (a, b, c), then we should drop the term. If only question (a) answer is "yes", then we should try a linear term. Otherwise there is no general rule to apply. For all the computation of GAM, we will use the recommended **mgcv** package written by S. Wood.

3.2 Case studies

This section focuses on the GAM analysis of the three datasets.

3.2.1 Portugal

The GLM analysis of Portugal data was an attempt to model price elasticity with a very limited dataset. Let us see what the generalized additive model can explain what the GLM cannot. We recall we have only few variables on the Portugal dataset, namely price ratio, driver age, policy age, the vehicle age and the last-year premium *.

GAM analysis

First we estimate a GAM by modelling all the terms by a smooth function. And then we apply the rules of previous section to remove, to linearize or to categorize the explanatory variables. Between models, we keep the model with the highest likelihood (i.e. the minimum of the REML score).

The final model is summarized below, while other fits can be found in the appendix B.3.1.

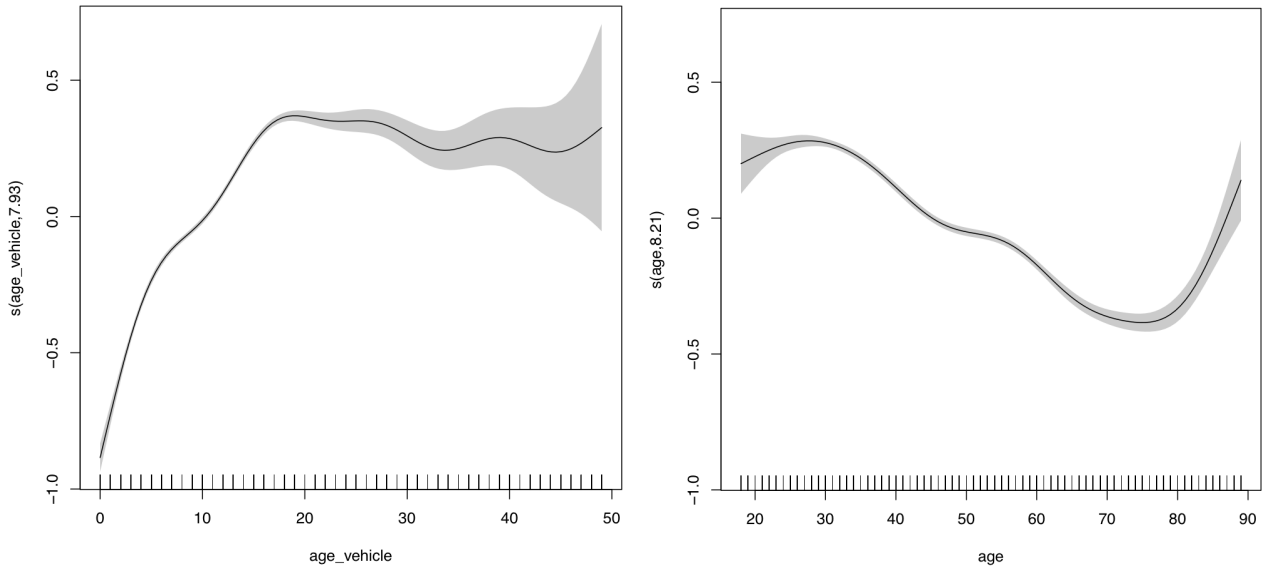
```
Family: binomial - Link function: logit

Formula: did_lapse ~ s(age) + s(age_vehicle) + agepolgroup + s(priceratio, premium_before)

Parametric coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.432685   0.005291 -270.75   <2e-16 ***
agepolgroup(4,8]   -0.167088   0.008709  -19.19   <2e-16 ***
agepolgroup(8,12]  -0.191395   0.010606  -18.05   <2e-16 ***
agepolgroup(12,49] -0.172698   0.013638  -12.66   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                               edf Ref.df Chi.sq p-value
s(age)                       8.213  8.797   2921  <2e-16 ***
s(age_vehicle)               7.927  8.508   4343  <2e-16 ***
s(priceratio,premium_before) 28.337 28.961   7581  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0272   Deviance explained = 2.78%
REML score = 2.6134e+05  Scale est. = 1         n = 557693
```

The summary is composed of two parts: a part for the linear terms (`Parametric coefficients`) and another for smooth terms (`Approximate significance of smooth terms`).

*. the gender was partially tested and then rejected because it is not significant.

As we can see with the term, the only term that we retain to link with price ratio is the last year premium (`premium_before`). The term `s(priceratio,premium_before)` corresponds to a bivariate smooth function $f(x, y)$. Note that this term has a high estimate degree of freedoms (28.337) compared to univariate smooth function (degree around 8). Other variables such as the driver age (`age`) if linked with the price ratio do not add higher significance.

On the following figures 3.5a and 3.5b, we plot the smooth functions for the driver age and the vehicle age. Beware the y-axis corresponds to the linear predictor scale. The solid line corresponds to the estimated function, while the shaded area is the standard error bandwidth around the smooth function. The plots show that the link between the lapse and those variables are not linear. The corresponding GLM approach would consist to categorize the variable so we approximate the smooth function by segments.



(a) Vehicle age

(b) Driver age

Figure 3.5: Smooth function plots

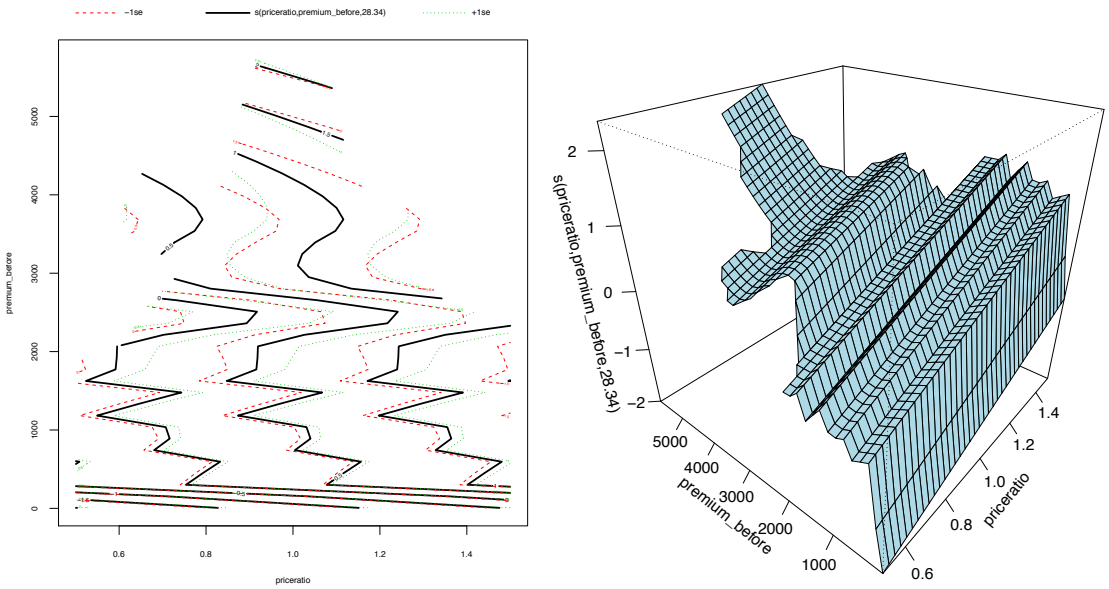
The plot of the smooth function for the price ratio and the last year premium reveals that the non-linear relationship between the response variable and the explanatory variables, see figure 3.6.

However for a given premium level, the function is linear. We try to model the term $f(x, y)$ where x corresponds to the price ratio and y to the premium level, by a simpler term $xf(y)$. Unfortunately, despite faster to compute, the model was worse in terms of REML score and deviance explained. The fitted probabilities were also quite different, so we keep the bivariate smooth function.

A first comparison with the GLM approach is the average predicted lapse function defined as

$$\hat{\pi}_{g,n}(p) = \frac{1}{n}\sum_{i=1}^{n} g^{-1}\left(x_i^T\hat{\beta} + \sum_{j=1}^{d}\hat{f}_j(z_i)\right),$$

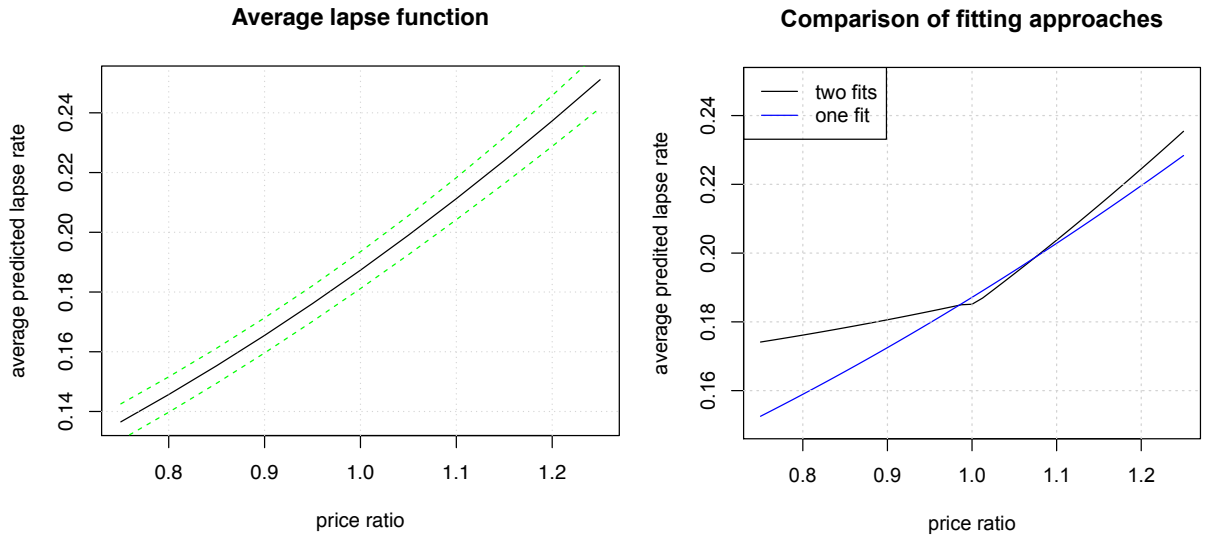
where x_i denotes the categorical variables (modelled linearly), z_i the continous variables modelled by a smooth function and g the link function (i.e. logit).



(a) Vehicle age (b) Driver age

Figure 3.6: Bivariate smooth function for price-linked variables

On the figure 3.7 below, we can see the difference between the GLM and the GAM approaches. The average lapse function is steeper with GAMs. The central lapse rates (for price ratio 1) are roughly the same around 18.7%.



(a) GAM (b) GLM

Figure 3.7: GAM vs. GLM

In a second time, we can compare the effect of explanatory variables on the lapse. All the graphs of one-variable effect have been put in appendix B.11. We observe that the explanatory variables have the same effect, such as the younger is the driver the more he lapses, etc...

Sub-population study

Finally we compare the behavior of customers by the populations we identify in the previous chapter. We recall here the populations:

1. people older than 60 year old, a policy age between 4 and 8 years and a last premium amount less than 500 euros,
2. male between 35 and 60 years old with a policy age between 0 and 4 years,
3. female between 35 and 60 years old with a policy age between 0 and 4 years,
4. policies with premium amount above 1500 euros,
5. policies older than 8 years,
6. people between 20 and 35 years old,
7. people between 35 and 60 years old with a premium amount between 500 and 1500 euros.

Then we plot for all populations the predicted lapse rate against the "delta lapse rate" a price increase (5%) * on figure 3.8. Comparing figure 3.8 with figure 2.4a, we note the delta lapse rates are higher for all population except high-value customers (in blue). However as for GLM, we think we underestimate the price-sensitivity of customers.

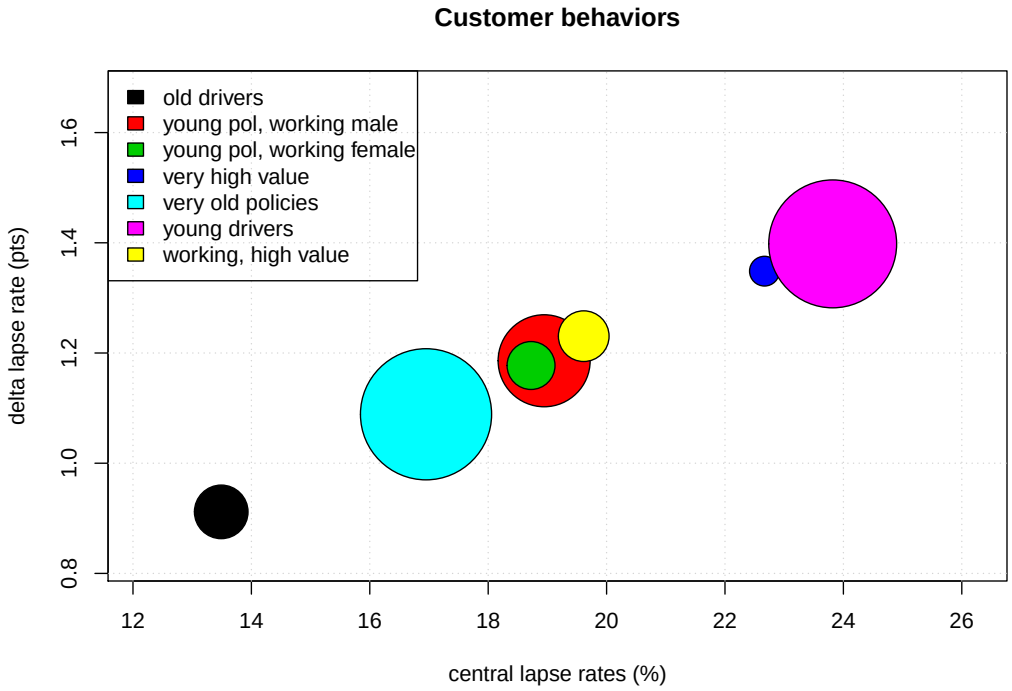


Figure 3.8: Client behaviors

In conclusion to the GAM analysis, we can say that additive modelling let us to model complex relationship within explanatory variables and between explanatory variables and the response variable. Despite this gain in modelling, the Portugal dataset has too few variables in order to have a very good predictive power.

∗. the size of the circle corresponds to the proportion of the sub-population in the whole portfolio.

3.2.2 Québec

The GLM analysis of Québec data reveals that many explanatory variables impact the choice of the customer to lapse or to renew. The central assumption of linearity contraints a bit the relationship between the explanatory variables and the response variable. We will test the potential benefit to use a generalized additive model. Let us recall that the main drawback of the Québec GLM analysis is an uncapicity to take into account the dynamic aspect. However the GAM do not add solution in this direction.

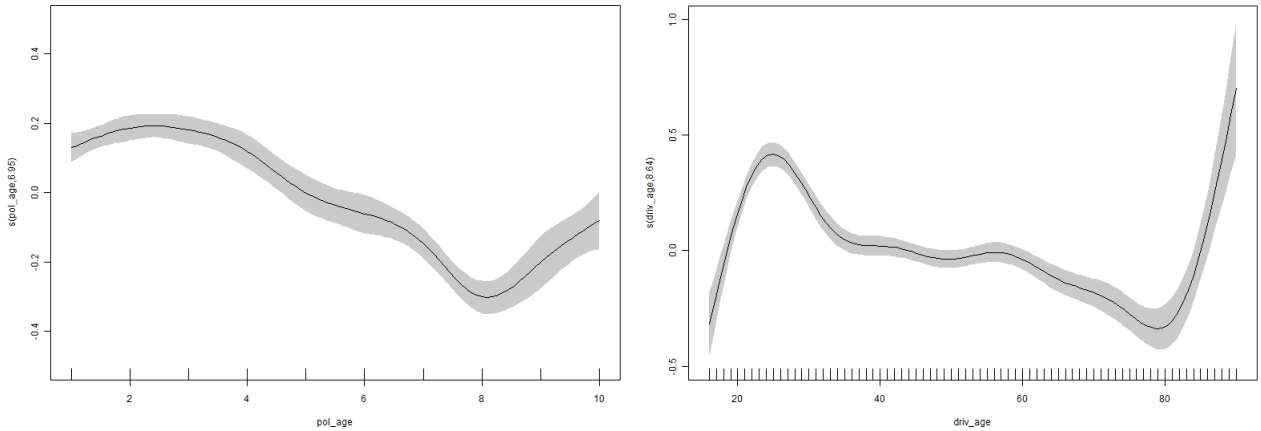
GAM analysis

Nevertheless, we carry out the GAM analysis for Québec data. First we estimate a GAM by modelling all the terms by a smooth function. And then we apply the rules of previous section to remove, to linearize or to categorize the explanatory variables, as in the Portugal analysis. We put below the final model, but the full models with and without price ratio interaction can be found in appendix B.3.2.

```
Family: binomial - Link function: logit

Formula:
did_cancel ~ claim_1 + house_pol + pricefactor:(multi_veh_dsc + cover)
    + prev_prem_group2 + pricefactor * (veh_age_group4) + s(pol_age) + s(driv_age)

Parametric coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -4.36024    0.11690 -37.299  < 2e-16 ***
claim_1                           -0.08108    0.02725  -2.975  0.00293 **
house_polY                        -0.89234    0.02195 -40.645  < 2e-16 ***
prev_prem_group2(1e+03,2e+03]      0.24977    0.03065   8.151 3.62e-16 ***
prev_prem_group2(2e+03,Inf]        0.48359    0.09285   5.208 1.90e-07 ***
pricefactor                        2.26238    0.11945  18.940  < 2e-16 ***
veh_age_group4(10,15]              1.12287    0.23519   4.774 1.80e-06 ***
veh_age_group4(15,Inf]             1.79452    0.26276   6.830 8.52e-12 ***
pricefactor:multi_veh_dscY        -0.28079    0.02065 -13.597  < 2e-16 ***
pricefactor:coverTPL              -0.21333    0.03116  -6.846 7.59e-12 ***
pricefactor:coverTPL+opt          -0.17159    0.02457  -6.983 2.90e-12 ***
pricefactor:veh_age_group4(10,15] -1.17859    0.24348  -4.841 1.29e-06 ***
pricefactor:veh_age_group4(15,Inf] -1.85951    0.27591  -6.740 1.59e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
             edf Ref.df Chi.sq p-value
s(pol_age)  6.951  8.026  315.9  <2e-16 ***
s(driv_age) 8.643  8.957  320.9  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0219   Deviance explained = 4.21%
REML score =  49772  Scale est. = 1         n = 202919
```

Let us note that the categorical variable such as the cover type cannot be smoothed, so compared to Portugal many variables of the Québec dataset must be modelled linearly. As we can see with the summary, we keep no term smoothed with the price ratio. However, some linear terms are crossed with price ratio, e.g. cover.

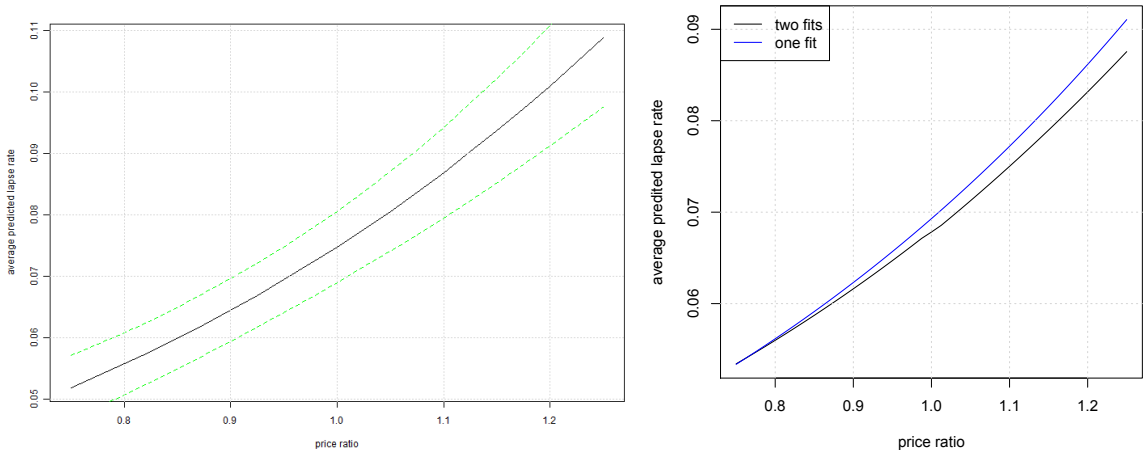
We plot on figures 3.9a and 3.9b. The policy age and the driver age are clearly not linear, which justifies the additive approach. Note that we try the model where the policy age and the driver age are modelled jointly by a smooth function. However in addition to being long to fit, the model was worse in terms of likelihood.



(a) Policy age (b) Driver age

Figure 3.9: Smooth function plots

Now let us take a look at the average predicted lapse function.



(a) GAM (b) GLM

Figure 3.10: GAM vs. GLM - average lapse function

From figure 3.10, we observe that the GAM approach provides a steeper lapse function compared to the GLM approach. The central lapse rate for the whole population is about 7.47% and the additional lapse rate for a 5% price increase is 0.6% (3.4% for a 25% price increase).

This was for an aggregate level for a price change considering other variables fixed. On figure B.12 in appendix, we plot the effect of each variable individually on the average lapse function. As for GLM, the most significant variables are cross-selling variables such as household or multi-vehicle discount, both in terms of the lapse level and the price sensitivity.

Sub-population study

Another criterion to compare the GLM and the GAM approach is to compare the (predicted) price-sensitiveness of customer subpopulations. In order to have comparable results, we use exactly the same segmentation:

(*black*) - young drivers with a low pricing group,

(*blue*) - young drivers with a high pricing group,

(*red*) - old drivers with full cross-selling (household and multi-vehicle),

(*green*) - old drivers with a household policy,

(*yellow*) - old drivers with a multi-vehicle discount,

(*azure*) - old drivers with no cross-selling,

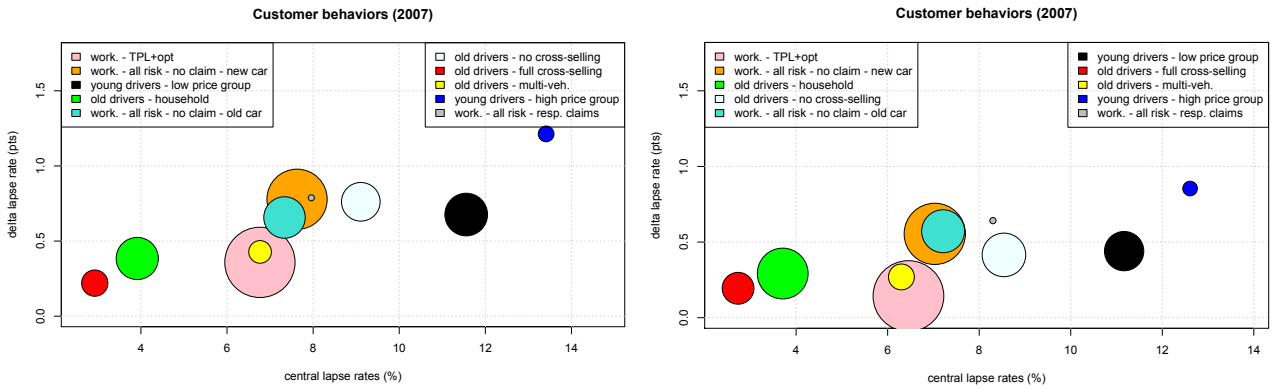
(*grey*) - working class with all risk cover and responsible claims (in last 2 years),

(*orange*) - working class with all risk cover without responsible claims and young car,

(*turquoise*) - working class with all risk cover without responsible claims and old car,

(*pink*) - working class with third-part liability cover and possibly add-on cover.

As we can see on figure 3.11, we observe additional lapse rates (ordinate) are always greater for the GAM model than for the GLM model. That's a first good point, because we thought we underestimate the effect of price increase with the GLMs.



(a) GAM (b) GLM

Figure 3.11: GAM vs. GLM - client behaviors

In terms of lapse levels, we roughly have the conclusions, working class with no claim and an old car (population *turquoise*) lapse 6 times more than the working class insured all-risk cover and experienced a responsible claim (population *grey*).

As for GLMs, we want to backfit the model, i.e. to predict next year lapse rates (by suppopulation). So we fit a GAM model on the 2006 data. As for the 2007 dataset, only few variables are modelled with smooth functions, namely the policy age and the driver age. In appendix B.3.2, we provide the fit summaries of the backward approach.

We also plot the two smooth functions in appendix. The figures B.13 is very similar to the one for the 2007 dataset, while for the policy age, we keep a bivariate term with the price ratio. Furthermore, we compare the average predicted lapse function between the two years, the 2006 fit is clearly steeper (see figure B.15). The same thing was observed for the GLM fit.

Similarly to the previous dataset, we make predictions for the 10 identified populations. The graph has been put in appendix also. From figure B.16, we again conclude that the price-sensitiveness estimated with GAMs is higher than the one fitted with GLMs.

Pop.	$r_j(2007)$	$\hat{\pi}_j^{GAM}(p_{2007})$	$\sigma^{GAM}(\hat{\pi}_j)$	$r_j(2007)$	$\hat{\pi}_j^{GLM}(p_{2007})$	$\sigma^{GLM}(\hat{\pi}_j)$
black	10.975	12.087	0.510	10.975	11.687	0.303
blue	10.684	13.667	0.610	10.684	13.421	0.394
red	2.446	3.498	0.171	2.446	3.220	0.097
green	3.495	4.363	0.219	3.495	4.150	0.121
yellow	6.487	7.905	0.354	6.487	7.524	0.205
azure	8.486	9.946	0.443	8.486	9.547	0.240
grey	8.568	9.009	0.471	8.568	8.773	0.462
orange	6.729	8.349	0.319	6.729	7.960	0.182
turquoise	7.105	8.358	0.313	7.105	7.981	0.198
pink	6.319	7.942	0.332	6.319	7.584	0.198
		GAM			GLM	

Table 3.1: GAM and GLM predicted lapse rate (%)

Finally we present the most interesting part: the backfit. In the above table 3.1, we see that the prediction from the GAM model are always higher than the GLM prediction, which are also always higher than the observed lapse rates. So using a GAM makes the prediction more conservative. As for the GLM, the GAM reveals the strong differences among populations, so additive models can also be used to segment customers.

In conclusion, the GAM analysis of the Québec data reveals to be a further degree of complexity compared to GLM. This higher sophistication permits to get a more cautious view of the price-sensitiveness of the customers. This is a positive aspect of the GAMs, but the variable selection is longer and the fitting time also. So we may wonder if this additional complexity does come with a too high cost.

3.2.3 Germany

The GLM analysis of the Germany dataset reveals that the channel distribution strongly impacts the GLM outputs. Especially, the lapse gap between tied-agent and other channels is far stronger than we could expect. However the price sensitiviness gap measured by the lapse deltas is also high. Let us see if it is still true with GAM results.

GAM analysis

On each channel and cover, we first estimate a GAM by modelling all the terms by a smooth function. And then we apply the Wood's rules to remove, to linearize or to categorize the explanatory variables. In appendix B.3.3, we provide the full list of regression summaries for the nine sub-datasets. Below, we list some comments:

(i) TPL cover

Agent: Using the market variables, we finally keep four non linear terms (`diff2tech`, `diff2top10vip`, `diff2top10direct`, `typeclassTPL`) all modelled jointly with the price ratio. We test to model these terms independently of price ratio, but it was worse in terms of REML scores.

Broker: We finally keep two non linear terms (`diff2tech` and `vehiclage`). Only the first term is modelled jointly with the price ratio, because the second term has a linear effect with the price ratio.

Direct: Due to the small dataset, it was hard to fit. We finally restrict the price ratio to be a smooth term of small order. This dataset also reveals some weird results with a negative elasticity for small premium increase, that the market variables could not deal with.

(ii) PC cover

Agent: Despite many attempts, only the price ratio (alone) has a real benefit to be modelled non linearly. This dataset is sufficiently big to make a lot of explanatory variables significant. And so we believe a big part of price sensitiveness is explained by linear terms.

Broker: As for the TPL covers, the variables are modelled non linearly (`diff2tech` and `vehiclage`), both joinly with the price ratio. The high estimated degrees of freedoms emphasizes this non linearity.

Direct: Only the `diff2tech` term is modelled through a smooth function, jointly with the price ratio.

(iii) FC cover

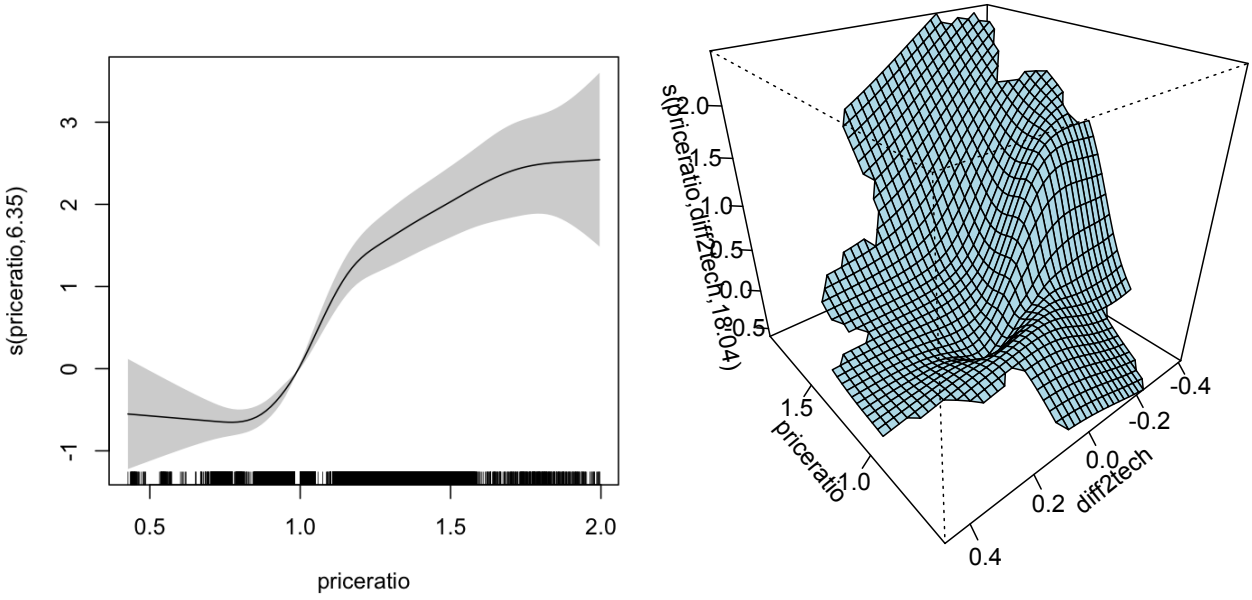
Agent: Three terms (`diff2tech`, `polholderage`, `typeclassFC`) are smoothed together with the price ratio. Again, the estimated degrees of freedom are high, especially for the `diff2tech` variable. So this is a real benefit compared to the GLMs.

Broker: Four terms (`diff2tech`, `vehiclage`, `polholderage`, `typeclassFC`) are modelled non linearly. This is astonishing because we retrive the difference with technical premium and the vehicle age as non linear terms. There might be a process made by brokers to target old vehicles and/or to detect a strong difference with technical premium. So the brokers have a major impact on the lapse decision.

Direct: Only two terms are modelled non linearly (`diff2tech`, `polholderage`): the estimated degree of freedom for the policyholder age variable is high. This may be linked to the close relationship between the motor (technical) premium and the policyholder age.

The regression coefficient analysis reveals some trends between channel distribution. Notably, the broker channel results are sensitive to the difference with technical premium and the vehicle age variables. There is also a datasize effect, since the datasets gradually increase in sizes from TPL to PC to FC covers. Obviously, the more we have data, the more we are confident with the regression.

On figure 3.12, we plot two smooth functions from two different GAM regressions *. The figure 3.12a represents the smooth function for the price ratio variable of the PC-agent regression: let us note the sharp increase for price ratio around 1.



(a) PC agent - price ratio smooth function

(b) FC broker - bivariate smooth function

Figure 3.12: GAM smooth functions

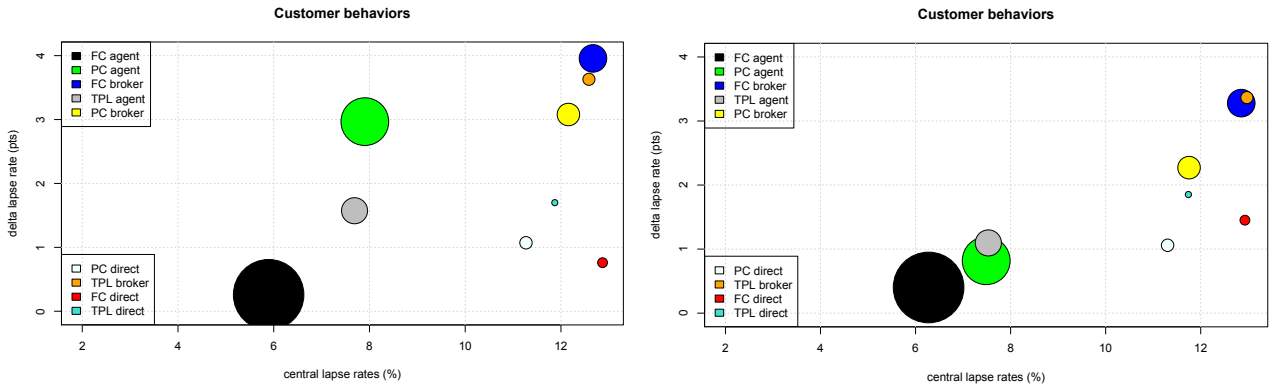
The figure 3.12b is the plot of the bivariate smooth function of the price ratio and the difference to technical premium variable for FC broker dataset. Note that there is a hollow in the curve around a price ratio of 1 and a zero difference with technical premium. In there, the price elasticity of the lapse decision is negative. Fortunately, this derived business inconsistency is small and located. If we had market variables for this dataset, it could be interesting to check this hollow vanish.

Sub-population study

On figure 3.13, we plot the traditional bubble plot to compare the differences between GAMs and GLMs on the different distribution channels and cover types. We observe that GAM delta predictions are higher than GLM ones in most cases. This is especially true for PC agent or FC broker: there is a high jump upward. Two channel-covers have a lower delta with GAMs: the FC

∗. The bandwith represents the standard error bandwidth around the smooth function.

direct case, a case where the dataset is small (so the GAM model selection was hard) and the FC agent case where the difference is limited.



(a) GAM (b) GLM

Figure 3.13: GAM vs. GLM - comparison of distribution channels and cover types

Study by region

Now we compare the GLM and the GAM results by looking at the central lapse rates and the deltas for the 16 regions. In the four-dimensional table 3.2, we put the results for the 16 regions of TPL agent and TPL broker channels.

We can observe that generally GAM predictions are higher than GLM ones, both for central lapse rates and deltas. This may be unconvenient for lapse rates to be overestimated, but we do not think it is a problem for deltas, because price sensitiveness is hard to catch and to estimate. So a conservative picture of deltas (additional lapse rate for 5% premium increase) is good.

As for GLMs, there are strong differences between agent and broker channels (9% against 13%). The central lapse rates are very high for brokers, which reflect the competition enforced by brokers. Deltas are relatively of the same order (2% against 3%).

In table 3.3, we compare the lapse rate fitted values for GAMs and GLMs by region. For the agent channel, the fitted values are close to the observed ones, as for GLMs. But for smaller datasets (i.e. broker and direct), the fitted values do not catch the specificity of the region. Typically, the observed lapse rate in Hamburg is 8.27%, while GLM and GAM fit a lapse rate of 12.02% and 12.11% respectively. The same bias can be observed for Saarland with a lapse rate of 14.57%, while fitted values are 11.24% and 11.29%.

This fact is explained by the regrouping process we did on the region variable. As a categorical variable with 16 values is hardly significant for all its categories, we group the regions according to their lapse rates. But for relative small datasets, we probably mix regions with quite different lapse rates.

Region	Central rate (%)		Delta (pts)		Central rate (%)		Delta (pts)	
Schleswig-H.	6.73	6.92	0.93	1.32	12.67	12.20	3.26	3.50
Hamburg	9.79	10.05	1.35	1.88	12.57	12.21	3.23	3.61
Niedersachsen	6.62	6.77	0.94	1.33	12.76	12.31	3.31	3.58
Bremen	8.92	9.22	1.27	1.78	13.46	13.23	3.32	3.78
Nordrhein-W.	8.03	8.18	1.11	1.57	13.28	12.93	3.28	3.66
Hessen	7.21	7.39	1.04	1.48	13.20	12.83	3.37	3.72
Rheinland-P.	6.52	6.66	0.93	1.38	12.76	12.50	3.32	3.51
Baden-W.	7.01	7.16	1.03	1.48	13.25	12.54	3.46	3.66
	GLM	GAM	GLM	GAM	GLM	GAM	GLM	GAM
Bayern	6.69	6.82	1.03	1.50	13.18	12.77	3.52	3.73
Saarland	7.69	9.92	1.45	2.07	12.60	12.32	3.39	3.76
Berlin	8.91	9.08	1.11	1.66	12.74	12.37	3.11	3.20
Brandenburg	8.70	8.89	1.32	1.89	12.27	12.00	3.29	3.55
Mecklenburg-V.	8.90	9.11	1.42	2.06	12.70	12.47	3.40	3.67
Sachsen-A.	9.62	9.82	1.55	2.14	12.59	12.50	3.38	3.61
Sachsen	9.70	9.89	1.49	2.09	12.66	12.51	3.32	3.60
Thüringen	9.42	9.54	1.45	2.06	12.23	12.02	3.33	3.67
	TPL agent				TPL broker			

Table 3.2: Customer behaviors in the German market

Region	Observed	GLM	GAM	Observed	GLM	GAM	Observed	GLM	GAM
Schleswig-H.	6.53	6.67	6.66	8.93	10.08	10.32	10.79	11.12	11.11
Hamburg	8.60	9.65	9.62	9.27	11.02	11.11	8.00	10.98	11.07
Niedersachsen	6.85	6.61	6.60	12.66	12.12	12.17	11.49	11.21	11.15
Bremen	7.32	8.77	8.76	10.80	12.21	12.42	11.18	11.76	11.31
Nordrhein-W.	8.08	8.11	8.11	11.34	11.67	11.79	13.03	11.27	11.25
Hessen	6.80	7.20	7.21	11.03	11.63	11.68	11.63	10.97	10.97
Rheinland-P.	6.72	6.53	6.53	11.14	11.00	11.12	10.39	10.93	10.92
Baden-W.	6.91	7.02	7.03	10.51	10.53	10.55	10.28	11.24	11.22
	Agent			Broker			Direct		
Bayern	6.77	6.75	6.75	10.45	10.70	10.79	11.36	12.12	12.18
Saarland	9.67	9.74	9.78	10.57	10.24	10.29	11.15	12.16	12.12
Berlin	9.44	8.98	9.00	12.44	11.70	11.65	13.16	12.24	11.26
Brandenburg	9.25	8.73	8.76	14.04	10.98	11.02	13.37	13.03	12.06
Mecklenburg-V.	8.24	8.94	8.91	12.52	12.79	11.05	13.60	13.98	13.98
Sachsen-A.	9.84	9.68	9.69	15.10	13.05	13.25	16.17	13.54	13.59
Sachsen	9.13	9.72	9.63	12.81	13.19	13.37	15.37	14.07	14.00
Thüringen	9.41	9.47	9.37	13.12	12.61	12.74	12.78	13.74	13.71

Table 3.3: Observed vs. fitted lapse rates for TPL cover

3.3 Pros and cons of the GAM methodology

This section summarizes the advantages and the drawbacks of the GAM methodology when modelling lapse rates in non-life insurance.

3.3.1 Advantages

GAM are less known tools than GLMs in actuarial, but they are used in reserving or claim modelling. GAMs introduced in the 90's are well known models, and the corresponding fitting procedures use state-of-the-art algorithms. But there are various ways to do model selections: prediction errors vs. likelihoods. In this memoir, we follow the advices of Wood's rules on the restricted maximum likelihood. We tested other statistical quantities and the impact was limited or inexistant.

As for GLMs, we meet the objective of this memoir with GAMs, since the overall estimated price elasticity takes into account the individual features of each policy. When data has enough variables, it is easy to identify customer segments with GAMs.

The additional complexity coming with additive modelling compared to GLMs permit to really fit the data. Especially for broker lines in Québec and Germany, we get a more cautious view of the price sensitiveness.

3.3.2 Drawbacks

For small datasets, GAM predictions may lead to irrelevant results, so we need to be careful when using the GAM methodology. However this was already noted for GLMs: with a small range of price change it is hard to extrapolate without having inconsistent results.

GAMs are generally longer to fit than GLMs and they require better computers with a lot of RAMs to be fitted. This is a limitation for GAMs to be used easily by everyone. Furthermore, some user judgement is needed to select the final appropriate model for GAMs. With Wood's rules, it is hard for newcomers to choose between two GAM models with the same "score" (i.e. likelihood or prediction errors).

Finally, we do not test the use of information asymmetry since GLMs reveal it was very useful at our aggregate level. But it will probably be useful at individual level for pricing and customer segmentation.

Chapter 4

Survival Regression Models

In this chapter, we use Survival Regression Models (SRM) to explain the dynamic price behavior. Using survival regression models to explain insurance lapses is recent but not new. Brockett et al. (2008) compares a classic logistic GLM with a Cox proportional hazard model.

The need for a dynamic model is obvious, because none of the previous models (GLMs and GAMs) can be used in a dynamic framework. The attempt to use time serie modelling on the GLM coefficients reveals to be rather inefficient. The following sections will provide good arguments in favor of survival regression models.

The chapter is divided into three sections: (i) model presentation, (ii) the application Québec data, (iii) the conclusions on the methodology.

4.1 Model presentation

In this section, we present the classic survival models, namely the non-parametric methods and the parametric regression models. Finally, we present the Cox PH model, a semi-parametric regression model for survival data.

In all this chapter, we stop to consider the lapse rate as a random variable independent of time, but rather the lapse rate of policies of age t is

$$r_t = P(T < t + 1 | T \geq t),$$

where T denotes the life span of the policy. Therefore in the following, we model directly the variable T with a survival model.

Survival models is a long-time studied topic, with roots in biology and failure time in mechanics. In actuarial science, it is also a well known topic for life insurance pricing since the 19th century. Survival models require specific tools, we present below:

- the survival function is $S_T(t) = P(T > t) = 1 - F_T(t)$,
- the hazard rate is a positive function $\lambda_T(t)$ such that $\lambda_T(t)dt = P(t \leq T < t+dt) = -\frac{S'_T(t)}{S_T(t)}dt$.

From the hazard rate, we can retrieve the survival function with

$$S_T(t) = e^{-\int_0^t \lambda_T(s)ds}.$$

So both quantities characterize the distribution of the variable T. In the following subsections, we present the non parametric estimation of the survival function S, the parametric regression model, Cox proportional hazard (PH) model and more advanced models *.

4.1.1 Non-parametric estimation

Kaplan & Meier (1958) provides a nonparametric estimator of the survival function defined by a step function. Let t_i be the exit times of the population and d_i the indicator of noncensored deaths. Then we have

$$\widehat{S}(t) = \prod_{t_{(i)} \leq t} \frac{n_i - d_i}{n_i},$$

where n_i stands for the cardinal of the population at risk for the ith period, d_i the number of "deaths" for the ith period and $t_{(i)}$ the ith ordered statistic. There exists a formula for the variance of this estimator (Greenwood's formula):

$$\widehat{\sigma}\left(\widehat{S}(t)\right) = \widehat{S}(t) \sum_{t_{(i)} \leq t} \frac{n_i}{n_i(n_i - d_i)}.$$

Using the normal approximation for binomial events and a log-minus-log transformation, confidence intervals can be computed

$$\left[\widehat{S}(t)^{\exp\left(y \pm \frac{\widehat{\sigma}(\widehat{S}(t))}{\widehat{S}(t)(1-\widehat{S}(t))}\hat{z}_{1-\frac{\alpha}{2}}\right)}\right],$$

where $z_{1-\frac{\alpha}{2}}$ the quantile of the standard normal distribution.

In the **survival** package, the Kaplan-Meier estimator is available in the `survfit` function as well as the confidence intervals. In the literature (but not presented here), the Nelson-Aalen estimator for the hazard rate function (and so for the survival function) is a competitive non parametric estimator.

4.1.2 Parametric regression model

Definition Let us continue with parametric models. There are three widely used distributions for T: the Weibull, the loglogistic and the lognormal distribution. Each of them can be characterized equivalently on T and $\ln(T)$.

The table 4.1 summarizes the relationship between the distributions of T and $\ln(T)$. Note that with a shape $\alpha = 1$, the distribution of Weibull is the exponential distribution. We pass from T to Y with $\sigma = 1/\alpha$, $\mu = -\ln(\lambda)$ and $y = \ln(t)$.

*. We remove the index T for simplicity.

T	Weibull	loglogistic	lognormal
$S_T(t)$	$e^{-(\lambda t)^\alpha}$	$\frac{1}{1+(\lambda t)^\alpha}$	$1-\Phi(\alpha \ln(\lambda t))$
$Y=\ln(T)$	extreme (min.) value	logistic	normal
$S_Y(y)$	$e^{-e^{\frac{y-\mu}{\sigma}}}$	$\frac{1}{1+e^{\frac{y-\mu}{\sigma}}}$	$1-\Phi(\frac{y-\mu}{\sigma})$

Table 4.1: Survival distributions

Link with GLMs Hidden in those expressions, we have the three link functions for binomial GLM.

- logit link: $g(\pi)=\ln\left(\frac{\pi}{1-\pi}\right)$,
- probit link: $g(\pi)=\Phi^{-1}(\pi)$,
- complementary log-log link: $g(\pi)=\ln(-\ln(1-\pi))$,

whose inverse are the distribution function of standard distributions. Let us note Z the variable characterized by g^{-1}. Then we have $Y=\mu+\sigma Z$.

To take into account, explanatory variables x_i of individual i, we will change the location parameter μ:

$$Y_i=\mu+x_i^T\beta+\sigma Z,$$

with β an unknown coefficient. This implies that

$$S_T(t)=S_{T_0}(e^{x_i^T\beta}t),$$

where T_0 is a baseline distribution (i.e. one of the distributions in table 4.1). From the last equation, we get the name of that type of model : accelerated / decelerated failure time model, since the coefficient $e^{x_i^T\beta}$ changes the scale of time implying a decrease / increase of the survival function.

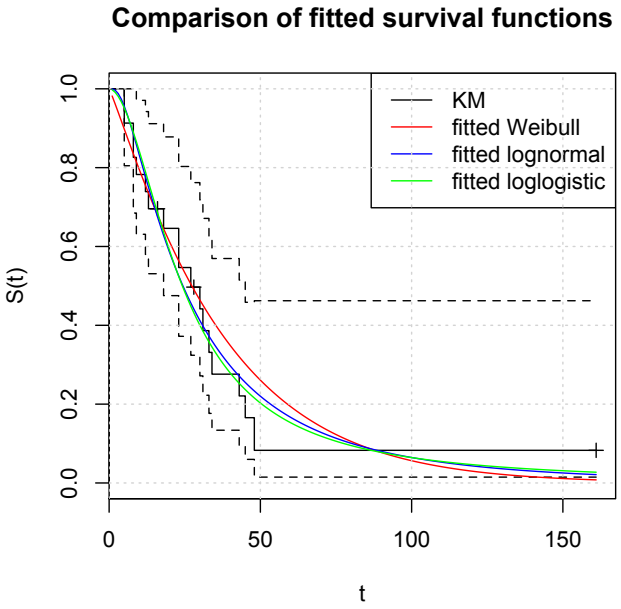
The estimation of the accelerated failure time model is done simply by maximising the loglikelihood. From the asymptotic normal behavior of maximum likelihood estimators, we can deduce confidence interval, hypothesis test for the β's components. Therefore a p-value is available for each coefficient of the regression, which help us to keep only the most significant variable. Hence we can adopt a backward or forward variable selection as for GLMs.

Base example Let us study a simple example. The dataset is `aml` from Therneau & Lumley (2009), a survival in "patients with Acute Myelogenous Leukemia. The question at the time was whether the standard course of chemotherapy should be extended ('maintainance') for additional cycles." The set consist of three variables, a survival time, a censoring status and a dummy variable indicating the "maintainance" of the chemotherapy.

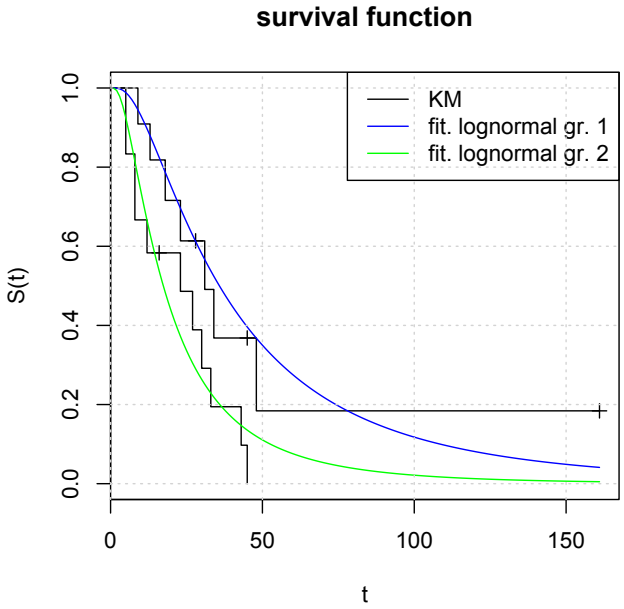
```
> head(aml)
  time status         x
1    9      1 Maintained
2   13      1 Maintained
3   13      0 Maintained
4   18      1 Maintained
5   23      1 Maintained
6   28      0 Maintained
```

First, we fit the three above distributions without using the covariate x indicating wether the chemotherapy was maintained. On figure 4.1a, we observe that all fits are quite good for small t but rather inadequate for large t. This is partly due to the fact that there are few individuals living a very long time. The standard error for the Kaplan-Meier estimation is really huge.

The model assumptions can be checked graphically (see (Tableman & Kim 2005, Chap. 6)). In appendix 4.1 , for each distribution we put the following plots: the qq-plot, the pp-plot, the survival function comparison and the deviance residuals against the fitted values. It is easy that for all distribution, the qqplot reveals extreme quantiles are not well explained by the model.



(a) Fitted survival functions



(b) Lognormal survival fit

Figure 4.1: Survival functions for AML data

In terms of loglikelihood, the three distributions have similar scores: -83.2, -80.7, -80.6 respectively for Weibull, lognormal and loglogistic.

Adding the covariate for the treatment maintainance makes the regression models much better. This is especially true for the survival functions or the qqplots (in appendix). In terms of likelihood, -80.5, -78.9, -79.4, the worst model is still the Weibull model.

We plot on figure 4.1b the survival function for the lognormal model. The group 2 corresponds to individuals where the chemotherapy was not maintained. We can see that their survival function is always lower than for group 1. The lognormal distribution captures this effect, especially for time below 50.

4.1.3 Cox PH model

From parametric regressions The Cox proportional hazard (PH) model can be seen as an extension of the accelerated failure time model. Let us recall that in the accelerated failure time model, the hazard function has the following form

$$h(t|x_i) = e^{x_i^T\beta} \times \begin{cases} \alpha\lambda^\alpha t^{\alpha-1} & \text{if Weibull} \\ \frac{\phi(t)}{1-\Phi(t)} & \text{if lognormal} \\ \frac{\alpha\lambda^\alpha t^{\alpha-1}}{1+(\lambda t)^\alpha} & \text{if loglogistic} \end{cases},$$

where ϕ and Φ denote respectively the density and the distribution function of the standard normal distribution.

To a semi-parametric regression A natural extension is therefore to consider models with hazard function

$$h(ty|x_i) = \theta(x_i, \beta) \times h_0(t)$$

with θ models the effect of covariates on the response (with an unknown parameter β) and h_0 is an arbitrary function. The name comes from the fact that $h(ty|x_i)/h(ty|x_j) = \theta(x_i, \beta)/\theta(x_j, \beta)$ is constant with respect to time t, so the hazard functions are "proportional".

Due to Efron (1977), the model can be interpreted by the following example: the survival of an individual i is represented by a time-varying coin. If the individual i is at risk during time interval $[t, t+\epsilon[$, we flip his coin with probability of heads equal to $h(t|x_i)$. In the Cox model, we assume $\theta(x_i(t), \beta) = e^{x_i^T\beta}$.

Estimation Given the risk set $R(t_j)$ (i.e. population at risk), the probability that an individual i_j failed at time t_j is

$$p_{i_j} = \frac{\theta(x_{i_j}(t), \beta)}{\sum\limits_{l \in R(t_j)} \theta(x_l, \beta)},$$

conditionnally on the failure times $t_1 < \cdots < t_d$, since the h_0's cancel. So we can deduce a partial loglikelihood is $\sum_{j=1}^d \ln(p_{i_j})$. By maximising this partial likelihood, we get an estimate $\hat{\beta}$, then by differentiating we get the information matrix at the estimate.

Assuming λ_0 is non-null at failure times $t_{(i)}$, $\lambda_0(t)$ has the form $\lambda_i\delta_{t,t_{(i)}}$. The author of the model considers the following form for $\lambda_0(t_{(i)})$

$$\frac{\pi_{(i)}e^{-\tilde{z}_{(i)}^T\hat{\beta}}}{1-\pi_{(i)}+\pi_{(i)}e^{-\tilde{z}_{(i)}^T\hat{\beta}}},$$

where $\tilde{z}_{(i)}$ is an arbitrary covariate constant at time $t_{(i)}$. Maximising the likelihood for π's, Cox (1972) derives a likelihood equation, that can be solved numerically.

However, the estimation of the baseline hazard function has been improved to a step function by Kalbfleisch & Prentice (1973), Breslow (1974) and Efron (1977).

The estimation process can summarized in two steps:

1. maximise the partial likelihood to find $\hat{\beta}$ using an iterative scheme (such as Newton),
2. given $\hat{\beta}$, estimation of h_0 at failure times t_i's using a nonparametric approach.

Thus, the Cox model belongs to the family of semiparametric regression model.

Summary The Cox PH model assumes the following equation for the survival function

$$S(t|yx_i) = S_0(t)^{e^{x_i^T \beta}},$$

where S_0 is the baseline survival function, x_i the covariate vector of individual i and β the regression coefficient. The baseline survival function S_0 is derived from the non parametric estimator of the hazard rate h_0, after having estimated the regression coefficient β by maximising the partial likelihood.

Extensions There have been various extensions of the Cox model. To increase a step further the flexibility of the Cox model, one extension, the stratified Cox model, considers different baseline hazard function $h_{0,j}$'s. This is intended to model a categorical variables with different baseline. So for that variable, the fit will be closer to the data. Obviously, the goal of this extension is not to transform all values of categorical variables into a stratum j.

Another extension is a refinement of the Cox PH model, which allows to have a time dependent coefficient $x_{it}(t)$. So the transformation on the hazard equation is relatively simple:

$$h(ty|yx_{it}) = (\theta(x_{it}, \gamma) \times \theta(x_i, \delta)) \times h_0(t),$$

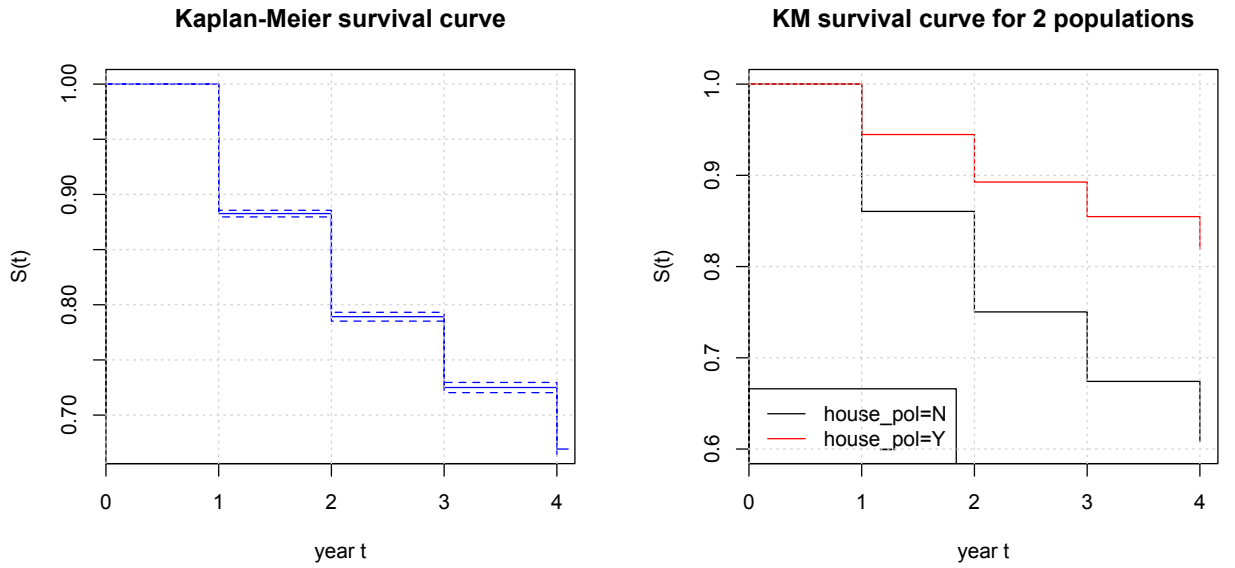
where θ being the exponential function generally, i.e. $\theta(x_{it}, \gamma) = e^{x_{it}^T \beta}$. So this extension only affects the estimation of the regression coefficients (from β to (γ, δ)). Note that this is reacher than the classic approach, we can use lag-time variables. Statistical inference * can be done to derive standard errors, hypothesis test,... With this extension, the Proportional Hazard assumption is no longer satisfied.†However, it is still compatible with the stratified Cox model.

∗. See Martinussen & Scheike (2006).

4.2 Case study - Québec

4.2.1 Non-parametric approach

Unless specified otherwise, we consider in the dataset the Québec policies facing their first renewal in 2004 during four years. We plot below the Kaplan-Meier survival curve for the whole population (figure 4.2a) and for two subpopulations (figure 4.2a).



(a) Whole population

(b) Cross-selling effect

Figure 4.2: Heterogeneity of customer behaviors

With these basic plots, we retrieve the effect of one variable of the lapses found in previous chapters. Two other plots can be found in appendix B.4.2. We can mix explanatory variables in the non-parametric approach to subdivide the dataset, but we do not get any further explanations about lapses.

4.2.2 Parametric regression

In a second step, we consider a full parametric regression for the three distributions presented previously, namely the Weibull, the lognormal and the loglogistic distributions. We consider two models for each distribution a simple intercept-only model (i.e. having no covariates $x_i^T \beta = 0$) and a model with covariates crossed with priceratio.

In appendix B.4.2, we give the regression summaries for these parametric regression. We use a backward approach to select the most significant variables. In table 4.2, we show the goodness-of-fit statistics for all the fitted models. Statistics subscripted with a zero denote the statistics for the intercept-only model, otherwise, it is for the with-covariate model.

	$\ln(\mathcal{L}_0)$	AIC_0	$\ln(\mathcal{L})$	AIC
Weibull	-37111.4	74224.8	-35360.3	70750.6
loglogistic	-36890.8	73783.7	-34956.7	69945.3
lognormal	-36295.3	72592.6	-34553.9	69141.8

Table 4.2: Goodness of fit

From those critertion, the lognormal model appears to be the best for all criteria. However the graphs to check model assumptions reveal that all the three distributions are inadequate, see appendix B.4.2.

On figure B.24, we compare the fitted survival functions for those models. The intercept-only models are particularly poor because we clearly overestimate the survival function for year 1, 2 and 3 and underestimate for year 4. Including explanatory variables enhances the model to fit the data with respect to the survival function.

With qq-plots B.25 and B.26, the inadequacy of the regression models is obvious (blue lines). However on the qq-plots, we also show the Ordinary Least Square estimate of the parameters. So this suggests the maximum likelihood estimators are not adapted for our datasets. One think that could improve the fit is to use the slope of the OLS method (i.e. the scale parameter). Unfortunately, it does not improve the adequacy with the survival function.

A last attempt we did for the parametric regression models is to extend the dataset. We consider the policies of age between 1 and 4 years old in 2004. With this, the resulting dataset is larger (see table 4.3).

Policy age	2004	2005	2006	2007
1	12117	0	0	0
2	9292	9663	0	0
3	7405	7843	7905	0
4	15416	6326	6375	11205
5	0	15136	5731	9731
6	0	0	13770	9594
7	0	0	0	29441

Table 4.3: Policy ages

In appendix B.4.2, we plot the usual figures: Kaplan Meier (fig B.27 and B.28), qq-plots (fig B.29 and B.30) and survival functions B.28. We do not report here any figures, because adding more policies worsen the model fit. We also test the use of the Ordinary-Least-Square estimate for the shape parameter. But again it does not solve the bad fitting.

In conclusion to this section, we arrive to the fact that parametric regression models despite their simplicity are not adequate for our problem. They are simplest model where explanatory variables can be used. However the parametric hazard rates assumed for each distribution is probably the weakness of this approach.

4.2.3 Cox regression

Baseline estimator

As presented in the previous section, the Cox model considers a hazard function which is the product of two components $h(ty|x_i) = e^{x_i^T \beta} \times h_0(t)$. The first term corresponds to an individual specific adjustment and the second to a baseline hazard function. In this subsection, we investigate briefly the differences between two estimators: Breslow (1974) and Efron (1977).

We work on a 4-generation Québec dataset, i.e. in 2004, we have policies of age below or equal to 4 years old, and in following years their renewal. See table 4.3.

To better judge the effect of the baseline estimator, we do not incorporate any explanatory variables: only an intercept will be estimated. We recall the estimator of h_0 is a non parametric estimator. As the Kaplan-Meier estimator, the survival function is a (decreasing) step function where the jumps occur at observed death times.

As we can see on figure 4.3, the Efron baseline estimator is always below the Breslow estimator. Efron's estimator in the case of covariate is nothing else than the Kaplan-Meier estimator. In the following, we keep the Efron estimator, the closest to the Kaplan-Meier estimator and the most conservative estimator of policy survival.

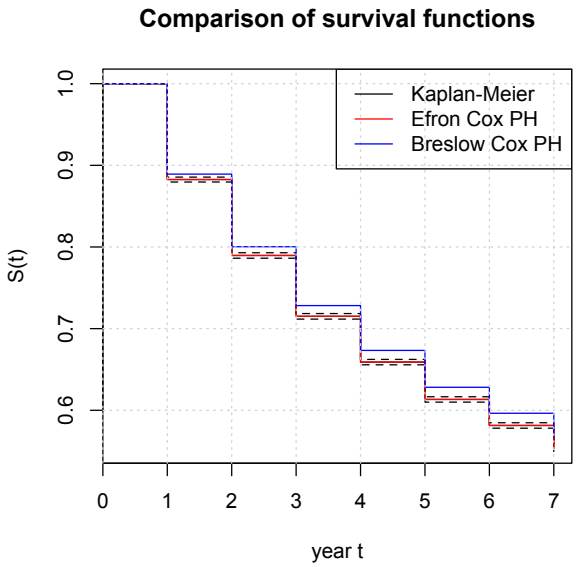


Figure 4.3: Baseline hazard estimators

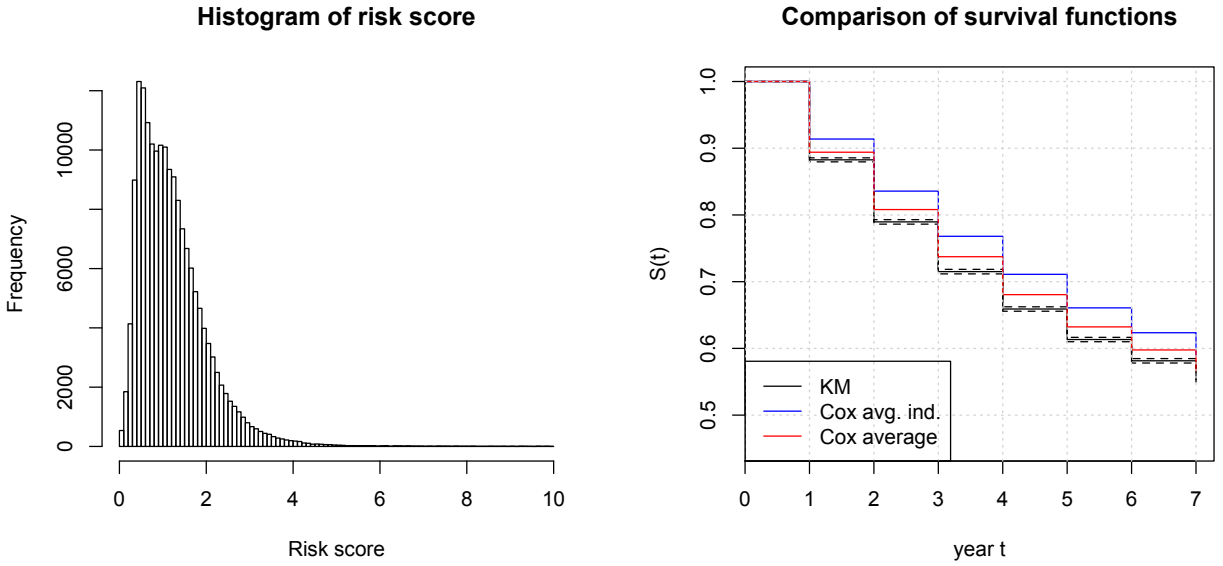
Cox model

Now we add all explanatory variables and their interaction with the price ratio. Again we use a backward approach to select the most significant variables. In appendix B.4.2, we put the regression summary. The most significant variables are similar to other regression models: the dummy variables indicating a household policy or a TPL cover.

By taking the exponential of the estimated coefficient for a given covariate, we have the effect on the hazard rate. If the exponentiated coefficient is greater than 1, thus fixing other covariate values, the variable increases the hazard rate (so decreases the survival function), whereas a value below than 1 decreases the hazard rate (so increases the survival function).

For the dummy variable indicating a household policy (alone), the exponentiated coefficient is 0.35351. Fixing other covariates, the hazard rate is reduced by 65% if an insured has a household policy compared to a single product customer. Other coefficients can be interpreted in a similar way.

The `coxph` function provides an estimate $\hat{\beta}_j$ of each coefficient β_j as well as p-values, standard errors and other useful statistics. However, this function (itself) does not provide an estimate of the survival function for each policy but rather the risk score $e^{x_i^T \hat{\beta}}$. To get the fitted survival functions for each policy, we need to use the `survfit` function *.



(a) Risk score

(b) Fitted survival function

Figure 4.4: Survival functions

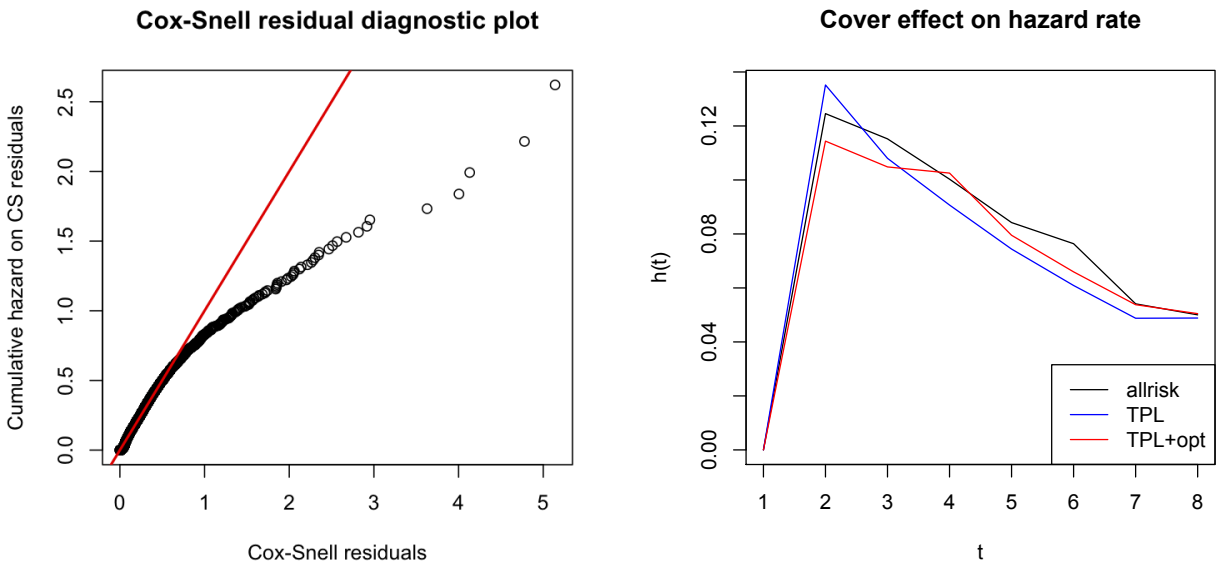
On figure 4.4a, we plot the histogram of risk scores. This is a few skewed distribution, even after keeping the risk score below 10. A lot of policies (around 46%) have a score below 1 (i.e. a survival function greater than the baseline function), but few policies have very high risk scores (14% of policies have score above 2).

On figure 4.4b, we plot the Kaplan-Meier survival function, the survival function for an ideal average policy (blue line) † and the average of all survival functions (red line). As one could expect from the 4.4a figure, the ideal average policy is not a good proxy for the average of individual policies. That's why in the following, we will focus on the "real" average of policies rather than the "average" policy.

Before going into further the conclusions, we will test the validity of the Cox model assumptions. The Cox-Snell residuals are designed to assess the overall fit a Cox model. If we have a positive random variable T with survival function S_T, we have $S_T(T)$ is uniformly distributed in $[0, 1]$ (hence $H_T(T) = -\ln(S_T(T))$ is exponentially distributed with parameter 1).

∗. Hence using this function on each policy is quite long on modern computers.

†. a policy having each covariate at the average value. For categorical variable, this means after computing the design matrix (of zeros and ones), we take the proportions of all variable values except the "first" value.



(a) Cox-Snell residuals

(b) Hazard rate

Figure 4.5: Residual plots

From this feature, the Cox-Snell residuals are defined as hazard rates $r_{CS,i} = H(T_i|yx_i) = H_0(T_i)e^{x_i^T\beta}$. The Cox-Snell plot displays $(r_{CS,i}, H_{CS}(r_{CS,i}))_i$, where H_{CS} is an estimator of the cumulative hazard rate of the (random) residuals $r_{CS,i}$. Under the assumption H is the true hazard function, the points should be closed to the 45° line.

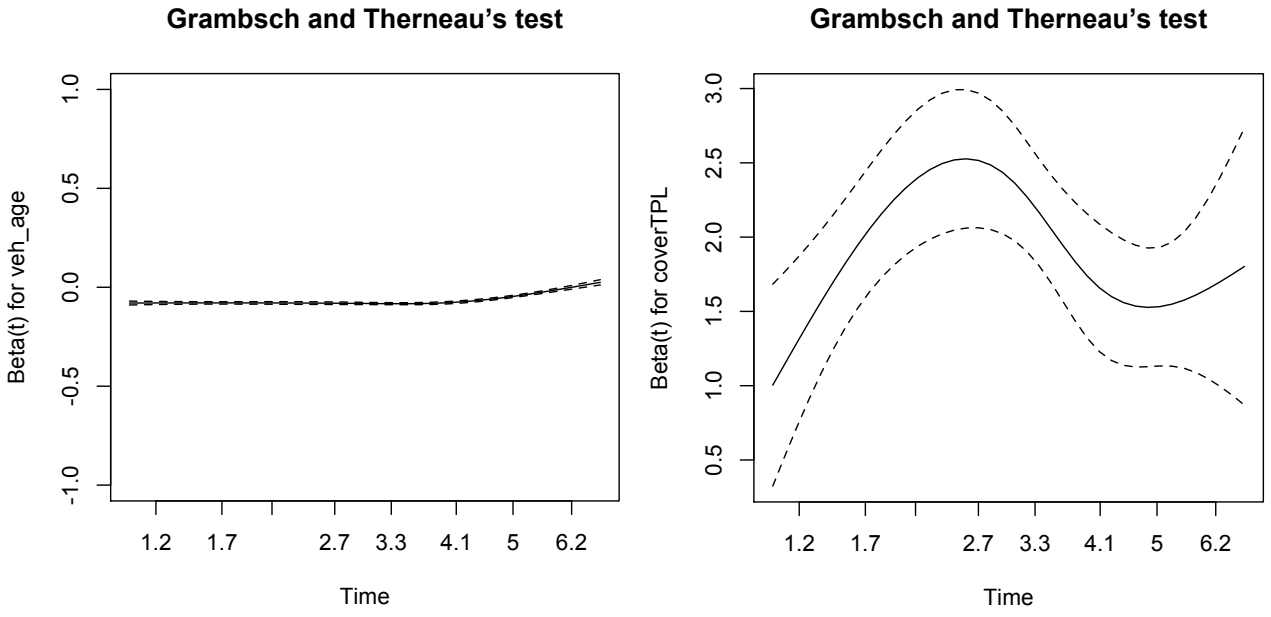
Another diagnostic plot is the hazard rate plot by value of categorical variables. It consists in plotting $(t, h(t|x_i))_t$ for different values of a covariate x_i. On figure 4.5b, we observe the effect of the cover type on the hazard rates. The Cox model assumes that the hazard ratio between two covariates is contant with respect to times. This cannot hold for the cover type variable since the hazard rates on figure 4.5b cross over time.

Other hazard rate graphs can be found in appendix B.4.2. From these two graphs, we conclude the Cox model is not well adapted because the proportional hazard assumption is not met for all variables. There exists other graphs such as the deviance residuals against an explanatory variables, but their purpose is to detect outlier or observation influence.

To test the proportional hazard assumption, Grambsch & Therneau (1994) considers a time-varying coefficient approach: $\beta(t) = \beta + \theta g(t)$ with $g(t)$ an unknown function. Using the Schoenfeld residuals covariance matrix, they can derive an estimate $\hat{\beta}$ of the function $t \to \beta(t)$. Note that the Schoenfeld residuals * are designed to detect outliers in explanatory variables. So the diagnostic test consists in plotting the $\hat{\beta}$ function for the different covariates.

∗. For j covariate index and the ith individual, the Schoenfeld residuals are defined as $r_{S,j}(i) = x_{i,j} - \frac{\sum_m x_{m,j} e^{x_m^T\hat{\beta}}}{\sum_m e^{x_m^T\hat{\beta}}}$, where $x_{i,j}$ is the jth covariate value.

On figure 4.6, we plot the Grambsch and Therneau's test for two covariates, namely the vehicle age and the dummy variable indicating TPL cover. On figure 4.6a, the PH assumption seems validated. However, the figure 4.6b strongly rejects the PH assumption, since the beta function is everything but linear. In appendix B.4.2, the figure B.32 shows other examples of PH violation or validation.



(a) Vehicle age coefficient

(b) Cover TPL coefficient

Figure 4.6: Grambsch and Therneau's test

Several options are available to tackle this problem. According to Tableman & Kim (2005), we can either stratifies some variables (i.e. use a different baseline function h_0 for the different variable value) or use a time dependent variable. The model resulting from the last option is called an extended Cox model.

Extended Cox model

The extended Cox model consists in modelling the hazard rate by $h(t|x_i) = e^{x_i(t)^T\beta}h_0(t)$. The value of the covariate $x_i(t)$ depends on time but is deterministic. In terms of data, this additional complexity adds flexibility in the explanatory variables we can use.

i	t_0	t_i	y_i	$x_{1,i}$	$x_{2,i}$	$x_{3,i}$
1	3	4	0	4	34	28
2	1	5	0	12	17	74
3	3	7	0	3	28	64
4	0	1	1	14	8	18
5	0	2	0	12	18	40
6	0	3	1	2	24	31
		⋮				

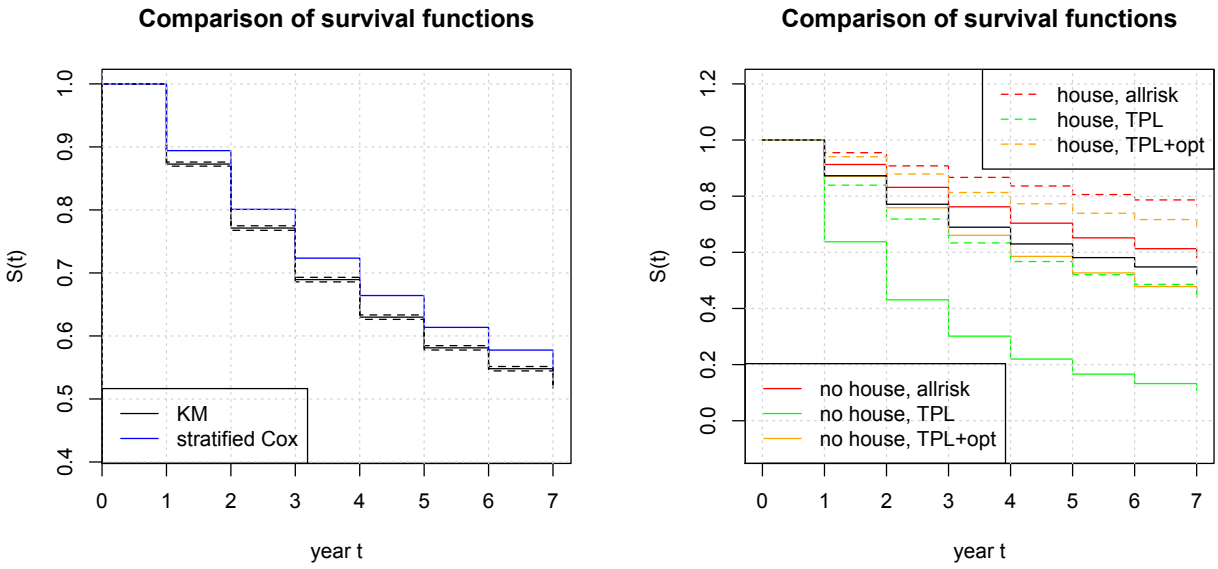
$\Rightarrow$

i	t_{i-1}	t_i	y_i	x_{1,i,t_i}	x_{2,i,t_i}	x_{3,i,t_i}
1	3	4	0	4	34	28
2	1	2	0	9	17	71
2	2	3	0	10	17	72
2	3	4	0	11	17	73
2	4	5	0	12	17	74
3	3	4	0	5	25	61
		⋮				

This flexibility however increases the datasize since variables are needed by interval. The supra example clearly presents the expanding effect happening to policy number 2.

A second strong difference between this sub-section and the previous one is the use of a stratified Cox regression. Two categorical variables for which the proportional hazard assumption is not verified will be stratified: the cover type and the household cross-selling dummy variable.

Then we use a backward selection for the other explanatory variables. In appendix B.4.2, we put the regression summaries. Except two variables (`cover` and `household`) are no longer present, almost the same variables remain significant.



(a) Fitted survival function

(b) Stratified survival function

Figure 4.7: Survival functions

On figure 4.7a, we plot the Kaplan-Meier estimator and the average of all individual survival functions predicted by the stratified Cox model. As with the Cox model, the predicted survival probability are a little bit overestimated.

Unlike the previous figure, figure 4.7b focuses on the heterogeneity of customer behaviors in the studied insured portfolio. The step functions are the predicted survival function for an ideal "average" customer in each stratification or group. As detected by the GLM, the most price sensitive population is the individuals with the "basic" TPL cover and having no household policy.

The black line on figure 4.7b indicates the Kaplan Meier survival estimator. For two out three cases, having a household policy is a sufficient condition to get a small decreasing survival function. However these effect are softer for the average of the whole portfolio then the "average" individual. As shown on figure B.33 in appendix, the survival curves decreases slower. Note that the order is not respected.

The difference with the previous sub-section is that we stratify explanatory variables for which the PH assumption was not satisfied, namely the `cover` and `housepol` variables. So we cannot test these variables with the Grambsch and Therneau's test. However, for other explanatory variables, the test can be used. As reported in appendix B.4.2, the figures B.34a and B.34b do not show any strong violation of the PH assumption.

Average behavior As for previous chapters, we are interesting in computing an average lapse function. From the Bayes' rule, the lapse probability is written as

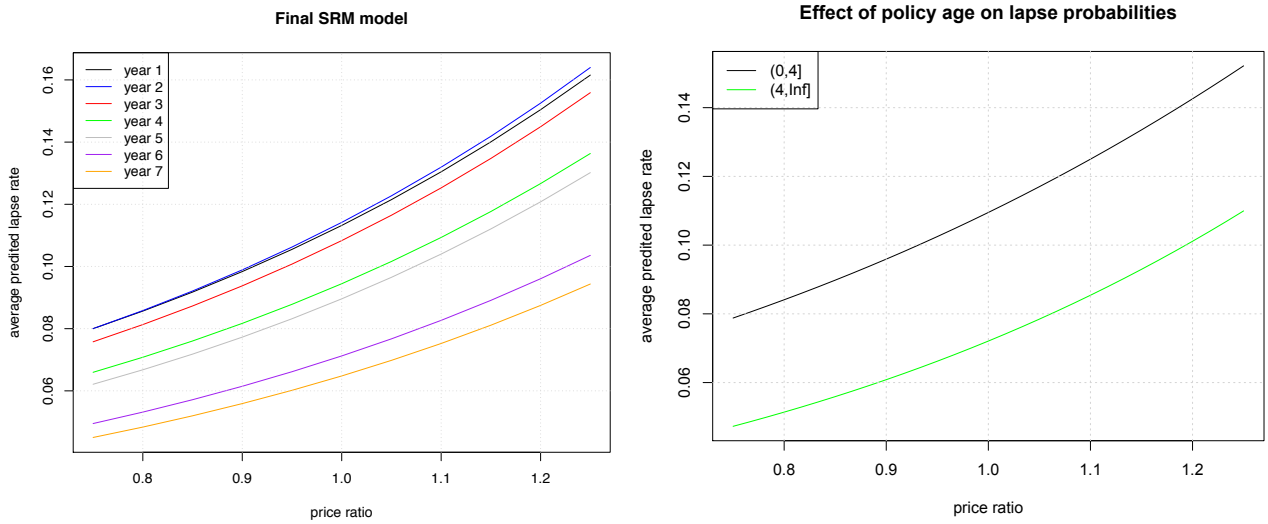
$$\hat{\pi}_t(x_{it}) \triangleq P(T < t+1/T \geq t, x_{it}) = \frac{\hat{S}(t|x_{it}) - \hat{S}(t+1|x_{it})}{\hat{S}(t|x_{it})}, \tag{4.1}$$

for a policy i with characteristic x_{it}. Note that even if a policy i is terminated in year t_0, we have a predicted survival function $\hat{S}(t|x_{it})$ for $t \in [0, T_{max}]$ with $T_{max} = 7$ the maximum policy age.

To compute the average lapse function at time t, we take the average over the whole population and a given value of the price ratio,

$$\hat{\pi}_t(p) = \frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_t(x_{it}),$$

where p denotes the price ratio, one of the explanatory variables x_{it}. On the figure 4.8a, we plot the average lapse function from the extended cox model estimated at different time t. As the time increases, the lapse probability decreases, except between year 1 and 2. Note that this aging effect was already observed when using GLMs (see figure 4.8b) and GAMs. So it is a good point that the SRM does not contradict this fact.



(a) SRM (b) GLM

Figure 4.8: Model comparison - average lapse function

In appendix B.4.2, we give the regression summary. By observing the sign of each coefficient, we can deduce the impact of each explanatory variable. We conclude the same effects as for GLMs and GAMs, e.g. the higher is the previous premium group, the greater is the lapse rate.

Segmentation The most interesting thing is now to study the subpopulations identified in the GLM analysis. We recall below the subpopulations:

(*black*) - young drivers with a low pricing group,

(*blue*) - young drivers with a high pricing group,

(*red*) - old drivers with full cross-selling (household and multi-vehicle),

(*green*) - old drivers with a household policy,

(*yellow*) - old drivers with a multi-vehicle discount,

(*azure*) - old drivers with no cross-selling,

(*grey*) - working class with all risk cover and responsible claims (in last 2 years),

(*orange*) - working class with all risk cover without responsible claims and young car,

(*turquoise*) - working class with all risk cover without responsible claims and old car,

(*pink*) - working class with third-part liability cover and possibly add-on cover.

We want to compare $\hat{\pi}^m_{t,j}(p)$ for $p \in \{1, 1.05\}$ for different population j at time t^* with different models $m \in \{\text{SRM, GAM, GLM}\}$. As in previous chapters, the delta lapse rate is defined as $\Delta^m_{t,j} = \hat{\pi}^m_{t,j}(1.05) - \hat{\pi}^m_{t,j}(1)$.

Nevertheless, there is a strong difference between static GLMs and GAMs and the dynamic SRM. The first ones estimate the regression coefficients and/or functions using one year of data in order to provide a lapse function for a given year (e.g. $t = 2007$), while the second one uses four years of data and provides a lapse function for a given policy age (e.g. $t = 2$).

This is completely different because a given year, there are multiple policy age and vice-versa. So to solve this issue, we decided to use for each population the average policy age as the time to compute the lapse rate in equation 4.1.

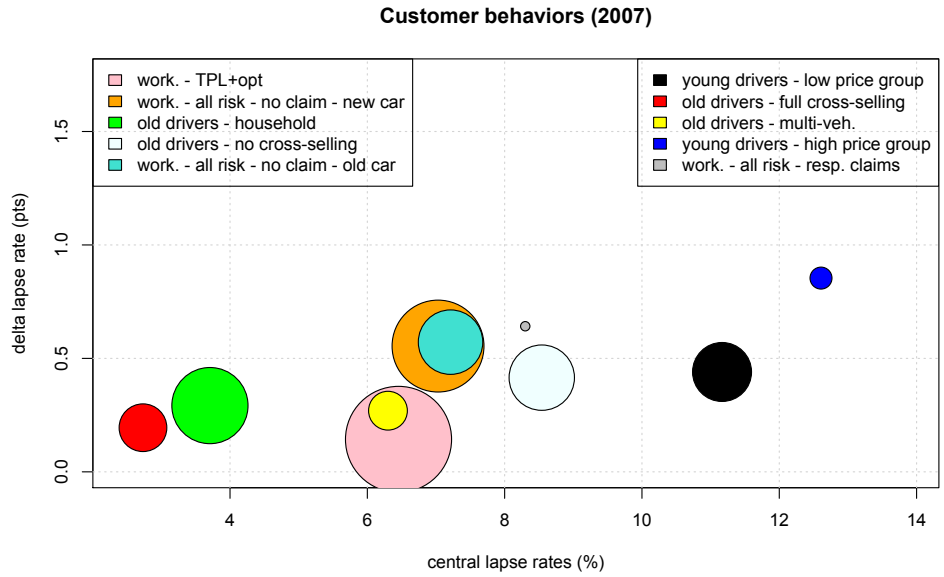
On figure 4.9, we plot the bubble plots for the three models, namely GLM, GAM and SRM. Beware the y-axis has the same scale for all graphs but not the x-axis. We can observe that the SRM provides less different estimates of lapse rate and delta lapse rate for the 10 groups.

On figure 4.9c, three main groups appear: high-value clients (*red* and *green*), working class or other old drivers (*pink*, *yellow*, *turquoise*, *orange*, *grey*) and young drivers (*black*, *blue*); whereas on figures 4.9a and 4.9b almost all subpopulations have distinct lapse rates and delta's. This is maybe due to the approximation on the policy age t used to compute $\hat{\pi}^{\text{SRM}}_{t,j}(p)$, where t must be an integer [†]. Hence, there is an approximation.

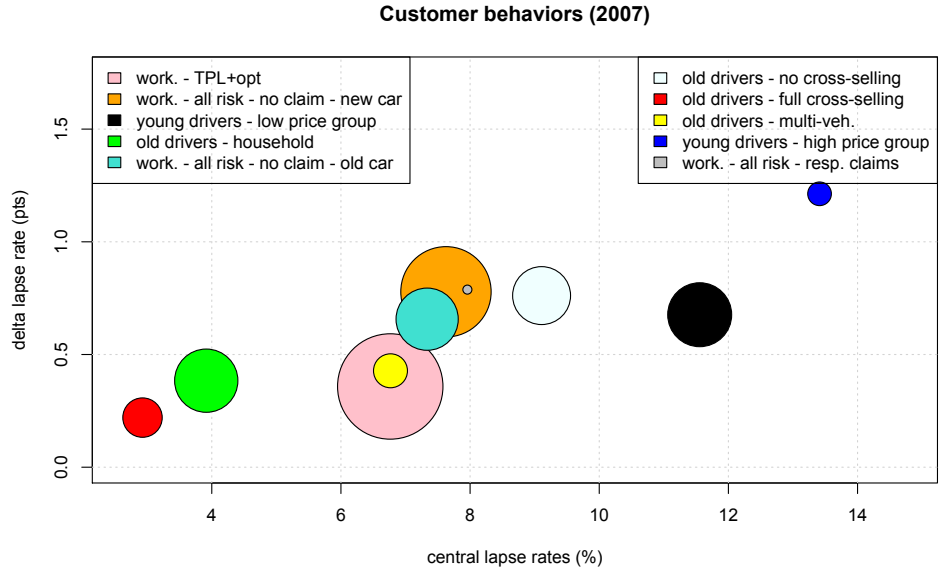
However, the ten populations are arranged in the same order in all models, so we can conclude that the most price-sensitive and the less loyal population is the *blue* population, young drivers. At the opposite, the most loyal population and the less price-sensitive population is the *red* population, old drivers with full cross-selling. Finally, we notice that with the SRM methodology customers are less loyal to the insurer, i.e. higher central lapse rates.

∗. which denotes the year or the policy age.

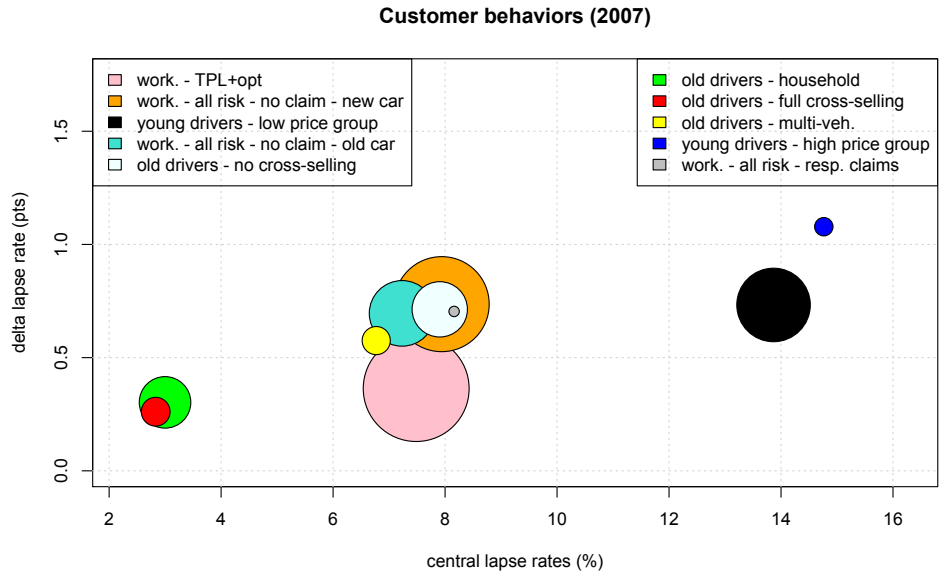
†. since the observed times were all integers.



(a) GLM



(b) GAM



(c) SRM

Figure 4.9: Model comparison - client behaviors

Back-testing On table 4.4, we put the lapse rate predictions by the three models $\hat{\pi}^m_{t,j}(p_{2007})$ against the observed lapse rates $r_j(2007)$ for each population j. Note that $\hat{\pi}^{\text{GAM}}_{2006,j}(p_{2007})$ is the lapse rates predicted by the GAM (estimated on 2006 data) for the (observed) price ratio p_{2007} and population j.

Pop.	$r_j(2007)$	$\hat{\pi}^{\text{SRM}}_{t_{2006}+1,j}(p_{2007})$	$\hat{\pi}^{\text{GAM}}_{2006,j}(p_{2007})$	$\hat{\pi}^{\text{GLM}}_{2006,j}(p_{2007})$
black	10.975	12.838	12.087	11.687
blue	10.684	13.642	13.667	13.421
red	2.446	2.578	3.498	3.220
green	3.495	2.754	4.363	4.150
yellow	6.487	6.234	7.905	7.524
azure	8.486	7.395	9.946	9.547
grey	8.568	7.948	9.009	8.773
orange	6.729	7.544	8.349	7.960
turquoise	7.105	6.490	8.358	7.981
pink	6.319	7.181	7.942	7.584
		SRM	GAM	GLM

Table 4.4: SRM, GAM and GLM predicted lapse rates (%)

Unlike static models, the SRM does not always overestimate the lapse rates. For the most sensitive populations (*black*, *blue*), the lapse rates is really overestimated while for other populations (*azure*, *grey*, *turquoise*), it seems to be underestimated. Unfortunately, we cannot benchmark the delta lapse rates: we can only guess. So the analysis of subpopulations stops here.

4.3 Pros and cons of the SRM methodology

This section summarizes the advantages and the drawbacks of the SRM methodology when modelling lapse rates in non-life insurance.

4.3.1 Advantages

Survival regression models are widely known and used in life insurance, but their use in non-life insurance is relatively limited. And for the price elasticity topic, it is even harder to see an application of such model. Only Brockett et al. (2008) seems to be the only one to use such model.

However in terms of estimation process and variable selection, it is not hard to find a rich literature in biology or in social sciences. As for GLMs and GAMs, the parametric and Cox regression models have good algorithms and software implementations.

Since we can have as many explanatory variables as we want, SRMs meet the requirement of this memoir to use individual characteristics to derive an aggregate elasticity. Compared to GLMs and GAMs, they take into account the dynamic aspects of a policy life. The (extended) Cox model allow to use dynamic explanatory variables, which lead to relevant results.

4.3.2 Drawbacks

Despite its easy use, the (extended) Cox model is not the ultimate methodology since the coefficients of the explanatory variables, say for price ratio, cannot evolve through the time. There is no possibilities to include a dynamic on the regression coefficients. The model of Fahrmeir (1994) has been tested to deal with such issue: this model uses a latent approach to incorporate a dynamic on the regression coefficients. They use the Kalman filter (an estimation procedure) to fit the model. However this model reveals to be unreliable in the coefficient estimate and heavily depends on the initial parameter values. So we discard this model.

It was not possible with our data to use a full dynamic approach, but the survival regression models should be useful the full lifetime of a policy where termination can occur for different reasons: insured lapse, company lapse, disappearance of the risk. In addition in a dynamic framework, we could model the cashflows: the incoming premiums and the outcoming claims. The survival approach is the only method that can deal with such matters.

Conclusion

Being dependent on the market's environment, price elasticity forecasts require rigorous attention to detail to prevent the risk of erroneous conclusions. Not surprisingly, a data cleaning process has also found to be essential prior to regression fitting. In short, some explanatory variables supplied significantly affect the results attained. Omitting these variables in the data can, in itself, lead to unreliable findings.

These must-have variables include distribution channels, market premium proxies, rebate levels, coverage types, driver age, and cross-selling indicators. As the Portugal dataset only provides the driver age, this example leads to inconclusive results. Whatever the model we use, the old versus young segmentation alone, cannot in itself substantiate the lapse reasons.

In the Quebec dataset, the coverage type, and the cross-selling indicators were added to the regression fit. This enabled us to refine our analysis and to zero in on customer segments. Having or not having a household policy with AXA was thus proven to be a driving factor in renewing or allowing a contract to lapse.

In the German dataset, the price sensitiveness fit was significantly enhanced along with our ability to fine tune the results thanks to the inclusion of distribution channels, a market proxy, and a rebate level. Disposing of market variables proved to make testing market scenarios possible (e.g. -5%, +5%). Being able to provide such forecasts is highly valuable in taking pricing actions.

Generalized Linear Models (GLM) of McCullagh & Nelder (1989), are widely known and respected methods in non-life insurance, especially in pricing policies. They are the first regression models that have dealt with a binary response. As a base tool, they serve as the benchmark model. We must remember, nonetheless, that GLMs are generally too approximate for they tend to underestimate the price sensitiveness of customers.

With the gradual addition of explanatory variables, we have seen an increased accuracy of the lapse rate predictions. Additionnaly, the market variables, along with the technical premium have enabled us to gain a better understanding of the portfolio elasticity. We are consequently more confident with the accuracy of our forecasts reached in studies of the caliber of the Germany study.

Generalized Additive Models (GAM) of Hastie & Tibshirani (1990) are a generalization of GLMs, in the sense that they allow for non linear terms in the predictor. Like GLMs, the quality of the findings attained is directly related to the data provided. Exploiting limited variables in the Portugal dataset produced approximate results, whereas, dealing with an extensive set of variables in Germany database, lead to proven results.

Applying GAMs despite their additional complexity can be justified in cases where GLMs fail to provide realistic lapse predictions and substantial datasets. It should also be noted that GAMs can model interactions between explanatory variables. Not restricted to linear terms, they consequently provide us with a more adaptive tool. Caution should however be exercised, as they may overfit the data when applied to limited datasets. This could then imply business inconsistency.

GLMs and GAMs are static models, however, and this is a major drawback. One option could have been to use time serie models on regression coefficients. This was impossible with our datasets due to a limited number of years. Generalized Linear Mixed Models (GLMM), where the linear predictor becomes the sum of a (unknown deterministic) fixed term and a random term, are a natural extension of GLMs, when it is necessary to deal with heterogeneity across time.

Among others, Frees (2004) presents GLMMs in the context of longitudinal and panel data. Since a panel data model cannot deal with right-censoring (when a policy is terminated), we need survival models. Despite discarding GLMMs for dynamic lapse modelling, we try to use the GLMMs to model endogeneous effects such as dropping coverage for a random term. Unfortunately, this has been inefficient.

The Survival Regression Models of Cox (1972) were applied to eliminate the inherent limits of the static regression models previously used. They, by their nature, take into account the lapse that is the modeled variable's dynamic aspects. GLMs and GAMs clearly demonstrate that renewing a policy for the 1st time is not motivated by the same factors as renewing one for the 10th time.

As explained, the full power survival models could not be applied to the Quebec and German datasets as these only gave the lapse by the insured. With other policy termination factors, it would be feasible to model the complete life cycle of a policy. With a full picture integrating cash flow, claims, and premiums, risk could also be better evaluated, and risk managers could be better equipped. Further advanced models such as Fahrmeir (1994), mentioned in chapter 4, exist, yet, at present, they remain difficult to put into place due to the fitting process involved.

In this memoir we have explored the wide range of existing statistical regression models. Each of the three models presented are informative forecasting tools. To sum up, survival regression models allow the modeling of a policy's life cycle. On the other hand, GLMs and GAMs provide more proven results. Once again, our research has further demonstrated that the quality of data used in actuarial studies unequivocally affects the findings reached.

The conclusions drawn from customer price sensitiveness studies should in any respect be weighed carefully. Charging higher premiums to loyal customers could seem unfair in light of the fact that those same customers usually have a better claims history. By the same token, relying on the market context with its inherent uncertainty to predict price sensitiveness could be misleading.

In summary, insurers must have a well informed overview of the market, the customer base, and a keen awareness of the pros and cons of potential pricing adjustments. The models presented herein serve as decision-making support tools and reinforce business acumen.

Bibliography

Bland, R., Carter, T., Coughlan, D., Kelsey, R., Anderson, D., Cooper, S. & Jones, S. (1997), Workshop - customer selection and retention, *in* 'General Insurance Convention & ASTIN Colloquium'. 9

Breslow, N. (1974), 'Covariance analysis of censored data', *Biometrics* **30**, 89–99. 75, 79

Brockett, P. L., Golden, L. L., Guillen, M., Nielsen, J. P., Parner, J. & Perez-Marin, A. M. (2008), 'Survival analysis of a household portfolio insurance policies: How much time do you have to stop total customer defection?', *Journal of Risk and Insurance* **75**(3), 713–737. 9, 71, 88

Chiappori, P.-A. & Salanié, B. (2000), 'Testing for asymmetric information in insurance markets', *Journal of Political Economy* **108**(1), 56–78. 44

Clark, D. R. & Thayer, C. A. (2004), 'A primer on the exponential family of distributions', *2004 call paper program on generalized linear models* . 94

Cleveland, W. S. (1979), 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association* . 51

Cox, D. R. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society: Series B* . 75, 90

Cummins, J. D. & Venard, B. (2007), *Handbook of international insurance*, Springer. 11, 14, 16

Dardanoni, V. & Donni, P. L. (2008), Testing for asymmetric information in insurance markets with unobservable types. HEDG working paper. 44

Dionne, G., Gouriéroux, C. & Vanasse, C. (2001), 'Testing for evidence of adverse selection in the automobile insurance market: A comment', *Journal of Political Economy* **109**(2), 444–453. 44, 45

Dionne, G., Pinquet, J., Maurice, M. & Vanasse, C. (2009), Incentive mechanisms for safe driving: a comparative analysis with dynamic data. working paper. 14

Dreyer, V. (2000), Study the profitability of a customer, Master's thesis, ULP - magistère d'actuariat. Confidential memoir - AXA Insurance U.K. 21

Efron, B. (1977), 'The efficiency of cox's likelihood function for censored data', *Journal of the American Statistical Association* **72**(359), 557–565. 75, 79

Fahrmeir, L. (1994), 'Dynamic modelling and penalized likelihood estimation for discrete time survival data', *Biometrika* **81**(2), 317–330. 88, 90

Fahrmeir, L. & Tutz, G. (1994), *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer. 45

Fox, J. (2010), Logit and probit models, Technical report, York SPIDA. 25

Frees, E. W. (2004), *Longitudinal and Panel data*, Cambridge University Press. 90

Grambsch, P. & Therneau, T. (1994), 'Proportional hazard tests and diagnostics based on weighted residuals', *Biometrika* **81**, 515–526. 81

Guillen, M., Parner, J., Densgsoe, C. & Perez-Marin, A. M. (2003), *Using Logistic Regression Models to Predict and Understand Why Customers Leave an Insurance Company*, Vol. 6 of *Innovative Intelligence* Shapiro & Jain (2003), chapter 13. 9

Hamel, S. (2007), Prédiction de l'acte de résiliation de l'assuré et optimisation de la performance en assurance automobile particulier, Master's thesis, ENSAE. Mémoire confidentiel - AXA France. 21

Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman and Hall. 49, 50, 53, 89

Hastie, T. J. & Tibshirani, R. J. (1995), 'Generalized additive models', *to appear in Encyclopedia of Statistical Sciences* . 56

Kalbfleisch, J. D. & Prentice, R. L. (1973), 'Marginal likelihoods based on cox's regression and life model', *Biometrika* **60**, 267–278. 75

Kaplan, E. L. & Meier, P. (1958), 'Nonparametric estimation from incomplete observations', *Journal of the American Statistical Association* **53**(282), 457–481. 72

Kelsey, R., Anderson, D., Beauchamp, R., Black, S., Bland, R., Klauke, P. & Senator, I. (1998), Workshop - price/demand elasticity, *in* 'General Insurance Convention & ASTIN Colloquium'. 9

Martinussen, T. & Scheike, T. H. (2006), *Dynamic Regression models for survival data*, Springer. 76

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, 2nd edn, Chapman and Hall. 21, 24, 26, 89

Nelder, J. A. & Wedderburn, R. W. M. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society* . 21

Ohlsson, E. & Johansson, B. (2010), *Non-Life Insurance Pricing with Generalized Linear Models*, Springer. 21

R Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org* 9

Rabehi, H. (2007), Study of multi-risk household, Master's thesis, ISUP. Mémoire confidentiel - AXA France. 21

Rothschild, M. & Stiglitz, J. E. (1976), 'Equilibrium in competitive insurance markets: an essay on the economics of imperfect information', *The Quarterly Journal of Economics* **90**(4), 630–649. 44

Sergent, V. (2004), Etude de la sensibilité de l'assuré au prix en assurance auto de particuliers, Master's thesis, ISUP. Mémoire confidentiel - AXA France. 21

Shapiro, A. F. & Jain, L. C. (2003), *Intelligent and Other Computational Techniques in Insurance*, World Scientific Publishing. 9, 92, 93

Steihaug, T. (2007), Splines and b-splines: an introduction, Technical report, University of Oslo. 53

Tableman, M. & Kim, J. S. (2005), *Survival Analysis using S: Analysis of time-to-event data*, Chapman and Hall. 74, 82

Therneau, T. & Lumley, T. (2009), *survival: Survival analysis, including penalised likelihood.* R package version 2.35-8.
URL: *http://CRAN.R-project.org/package=survival* 73

Turner, H. (2008), Introduction to generalized linear models, Technical report, Vienna University of Economics and Business. 27

Venables, W. N. & Ripley, B. D. (2002), *Modern Applied Statistics with S*, 4th edn, Springer. 22, 52

Wood, S. N. (2001), 'mgcv: Gams and generalized ridge regression for r', *R News* **1**, 20–25. 56

Wood, S. N. (2008), 'Fast stable direct fitting and smoothness selection for generalized additive models', *Journal of the Royal Statistical Society: Series B* **70**(3). 53, 54

Wood, S. N. (2010), 'Fast stable reml and ml estimation of semiparametric glms', *Journal of the Royal Statistical Society: Series B* . 54

Yeo, A. C. & Smith, K. A. (2003), *An integretated Data Mining Approach to Premium Pricing for the Automobile Insurance Industry*, Vol. 6 of *Innovative Intelligence* Shapiro & Jain (2003), chapter 5. 9

Appendix A

Statistics

A.1 Exponential family

A.1.1 Characterization

Clark & Thayer (2004) defines the exponential family by the following density or mass probability function

$$f(x) = e^{d(\theta)e(x)+g(\theta)+h(x)},$$

where d, e, g and h are known functions and θ the vector of paremeters. Let us note that the support of the distribution can be $\mathbb{R}$ or $\mathbb{R}_+$ or $\mathbb{N}$. This form for the exponential family is called the natural form.

When we deal with generalized linear models, we use the natural form of the exponential family, which is

$$f(x, \theta, \phi) = e^{\frac{\theta x - b(\theta)}{a(\phi)} + c(x,\phi)},$$

where a, b, c are known functions and θ, ϕ* denote the parameters. This form is derived from the previous by setting $d(\theta) = \theta$, $e(x) = x$ and adding a dispersion parameter ϕ.

Let μ be the mean of the variable of an exponential family distribution. We have $\mu = \tau(\theta)$ since ϕ is only a dispersion parameter. The mean value form of the exponential family is

$$f(x) = e^{\frac{\tau^{-1}(\mu)x - b(\tau^{-1}(\mu))}{a(\phi)} + c(x,\phi)}.$$

A.1.2 Properties

For the exponential family, we have

$$E(X) = \mu = b'(\theta) \quad \text{and} \quad Var(X) = a(\phi)V(\mu) = a(\phi)b''(\theta),$$

∗. the canonic and the dispersion parameters.

where V is the unit variance function. The skewness is given by

$$\gamma_3(X) = \frac{dV}{d\mu}(\mu)\sqrt{\frac{a(\phi)}{V(\mu)}} = \frac{b^{(3)}(\theta)a(\phi)^2}{Var(Y)^{3/2}},$$

while the kurtosis is

$$\gamma_4(X) = 3 + \left[\frac{d^2V}{d\mu^2}(\mu)V(\mu) + \left(\frac{dV}{d\mu}(\mu)\right)^2\right]\frac{a(\phi)}{V(\mu)} = 3 + \frac{b^{(4)}(\theta)a(\phi)^3}{Var(Y)^2}.$$

The property of uniqueness is the fact that the variance function V uniquely identifies the distribution.

A.1.3 Special cases

The exponential family of distributions in fact contains the most frequently used distributions. Here are the corresponding parameters, listed in a table:

Law	Distribution	θ	ϕ	Expectation	Variance
Normal $\mathcal{N}(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2	$\mu=\theta$	1
Gamma $\mathcal{G}(\alpha,\beta)$	$\frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)}e^{-\beta x}$	$-\frac{\beta}{\alpha}=\frac{1}{\mu}$	$\frac{1}{\alpha}$	$\mu=-\frac{1}{\theta}$	μ^2
Inverse Normal $\mathcal{I}(\mu,\lambda)$	$\sqrt{\frac{\lambda}{2\pi x^3}}e^{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}}$	$-\frac{1}{2\mu^2}$	$\frac{1}{\lambda}$	$\mu=(-2\theta)^{-\frac{1}{2}}$	μ^3
Bernoulli $\mathcal{B}(\mu)$	$\mu^x(1-\mu)^{1-x}$	$\log(\frac{\mu}{1-\mu})$	1	$\mu=\frac{e^\theta}{1+e^\theta}$	$\mu(1-\mu)$
Poisson $\mathcal{P}(\mu)$	$\frac{\mu^x}{x!}e^{-\mu}$	$\log(\mu)$	1	$\mu=e^\theta$	μ
Overdispersed Poisson $\mathcal{P}(\phi,\mu)$	$\frac{\mu^{\frac{x}{\phi}}}{\frac{x}{\phi}!}e^{-\mu}$	$\log(\mu)$	ϕ	ϕe^θ	$\phi\mu$

Appendix B

Additional tables and graphics

B.1 Data presentation

B.1.1 Descriptive analysis for Portugal data

	(0,100]	(100,200]	(200,300]	(300,400]	(400,500]
Lapse rate (%)	14	14	20	25	22.0
Prop. of total (%)	14	26	43	11	5.1

Table B.1: Proposed premium - premium_after

Pearson's Chi-squared test for Proposed premium

data: mytable X-squared = 4895.49, df = 4, p-value < 2.2e-16

	(0,100]	(100,200]	(200,300]	(300,400]	(400,500]
Lapse rate (%)	14	14	20	25	22.1
Prop. of total (%)	14	27	44	11	4.8

Table B.2: Last paid premium - premium_before

Pearson's Chi-squared test for Last paid premium

data: mytable X-squared = 4775.193, df = 4, p-value < 2.2e-16

	FEMALE	MALE
Lapse rate (%)	18	19
Prop. of total (%)	20	80

Table B.3: Gender

Pearson's Chi-squared test with Yates' continuity correction for Gender

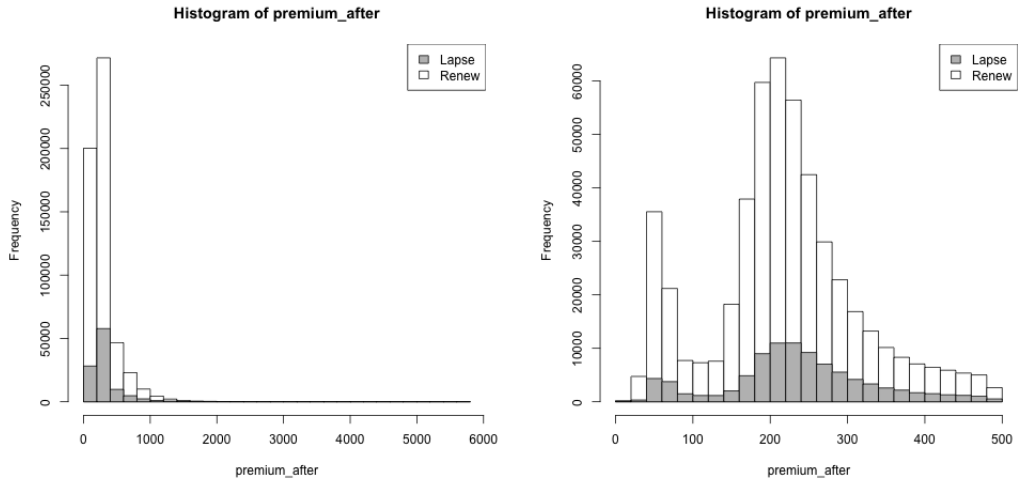


Figure B.1: Proposed premium

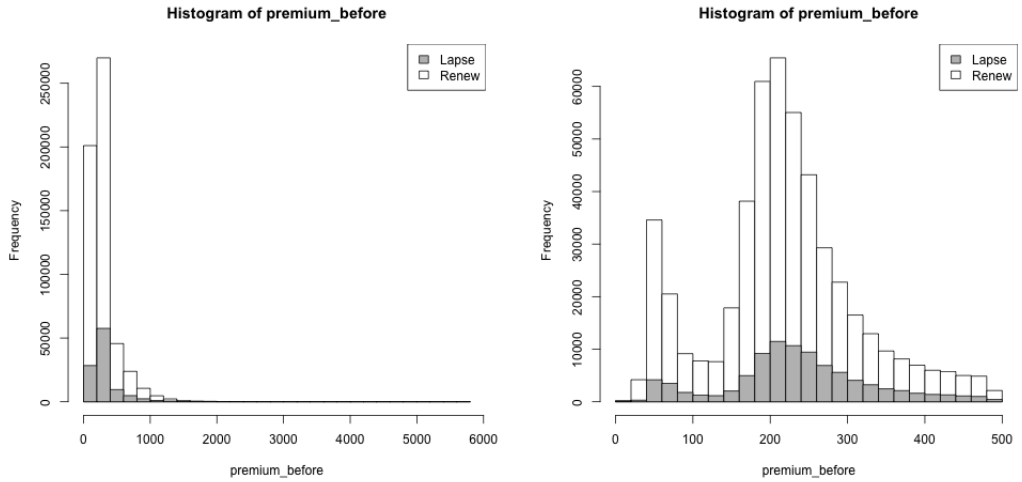


Figure B.2: Last paid premium

data: mytable X-squared = 23.7255, df = 1, p-value = 1.111e-06

	(32.5,47.5]	(47.5,62.5]	(62.5,77.5]	(77.5,92.5]
Lapse rate (%)	20	17	14	14.6
Prop. of total (%)	38	42	17	2.8

Table B.4: Driver age - age

Pearson's Chi-squared test for Driver age

data: mytable X-squared = 1500.512, df = 3, p-value < 2.2e-16

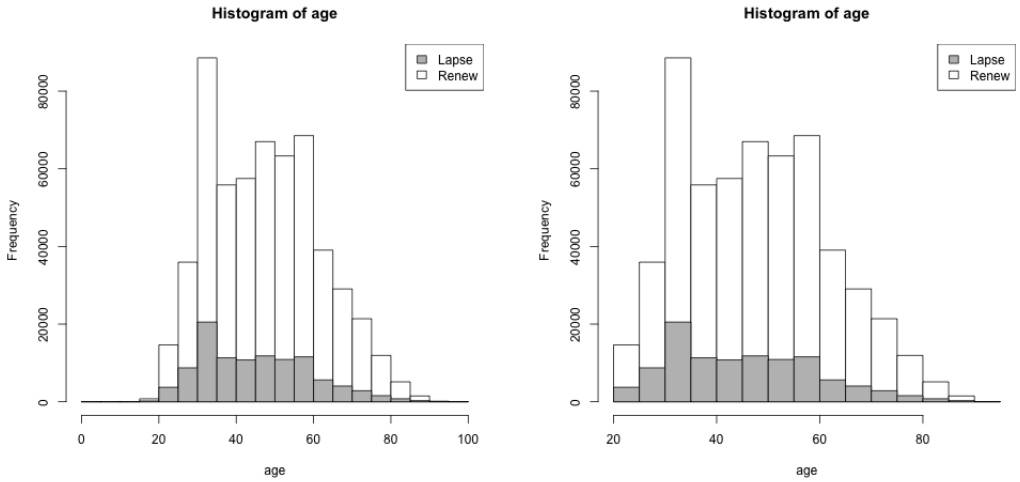


Figure B.3: Driver age

	(2.5,5.5]	(5.5,8.5]	(8.5,11.5]	(11.5,14.5]	(14.5,17.5]
Lapse rate (%)	21	17	18	16.6	17.5
Prop. of total (%)	38	33	22	3.6	2.3

Table B.5: Policy age - age_policy

Pearson's Chi-squared test for Policy age

data: mytable X-squared = 896.5751, df = 4, p-value < 2.2e-16

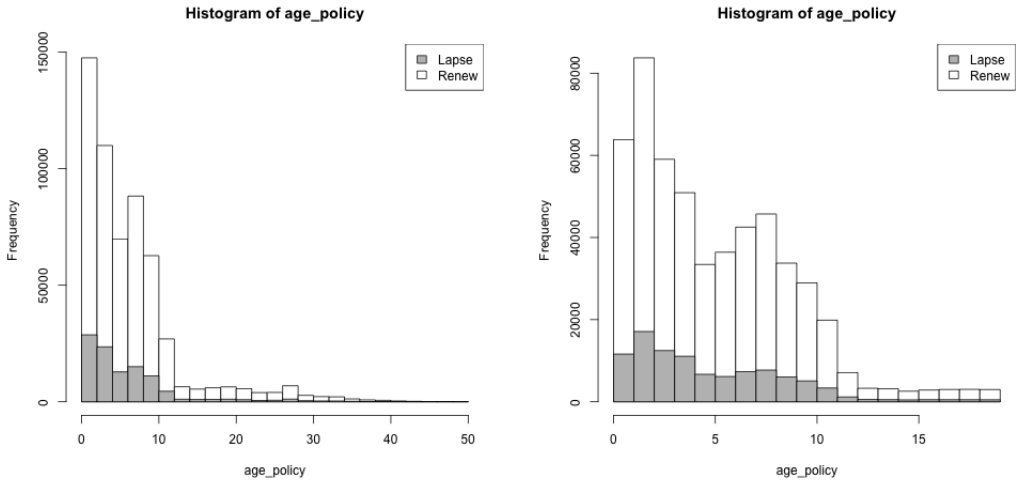


Figure B.4: Policy age

Pearson's Chi-squared test for Vehicle age

	(2.5,5.5]	(5.5,8.5]	(8.5,11.5]	(11.5,14.5]	(14.5,17.5]	(17.5,20.5]	(20.5,26.5]
Lapse rate (%)	17	18	19	20	21	21.1	39.3
Prop. of total (%)	15	21	21	16	14	8.4	4.4

Table B.6: Vehicle age - age_vehicle

data: mytable X-squared = 632.7147, df = 7, p-value < 2.2e-16

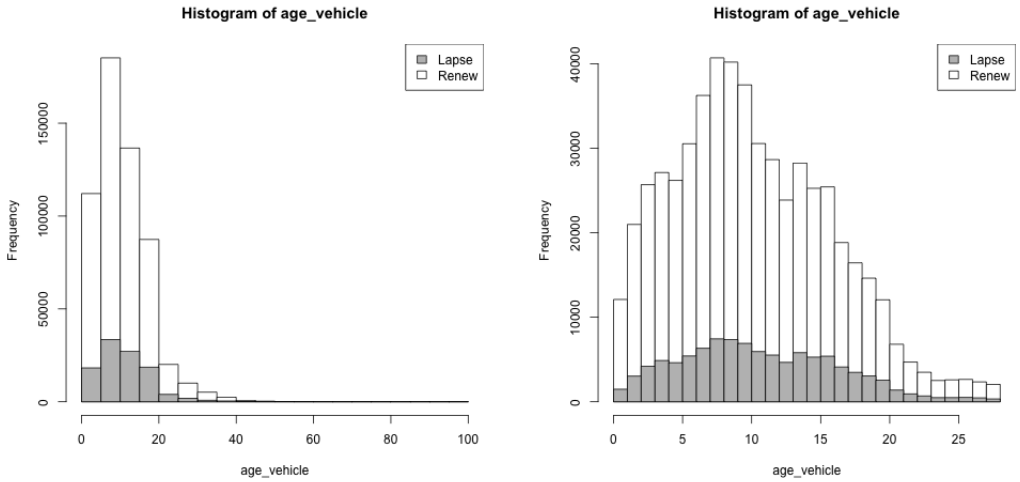


Figure B.5: Vehicle age

	(0.925,0.955]	(0.955,0.985]	(0.985,1.02]	(1.02,1.04]	(1.04,1.08]
Lapse rate (%)	18.8	20	18	18	21.9
Prop. of total (%)	8.5	20	40	29	2.5

Table B.7: Price ratio

Pearson's Chi-squared test for Price ratio

data: mytable X-squared = 484.1171, df = 4, p-value < 2.2e-16

B.1.2 Descriptive analysis for QuÈbec data

	(0,1]	(1,2]	(2,27]		N	Y
prop. size (%)	59.317196	20.524838	20.157966	prop. size (%)	59.165306	40.834694
lapse rate (%)	6.700996	5.398152	4.708123	lapse rate (%)	7.121054	4.453746

Table B.8: Vehicle number group - veh_num_group / Multi-vehicle discount - multi_veh_dsc

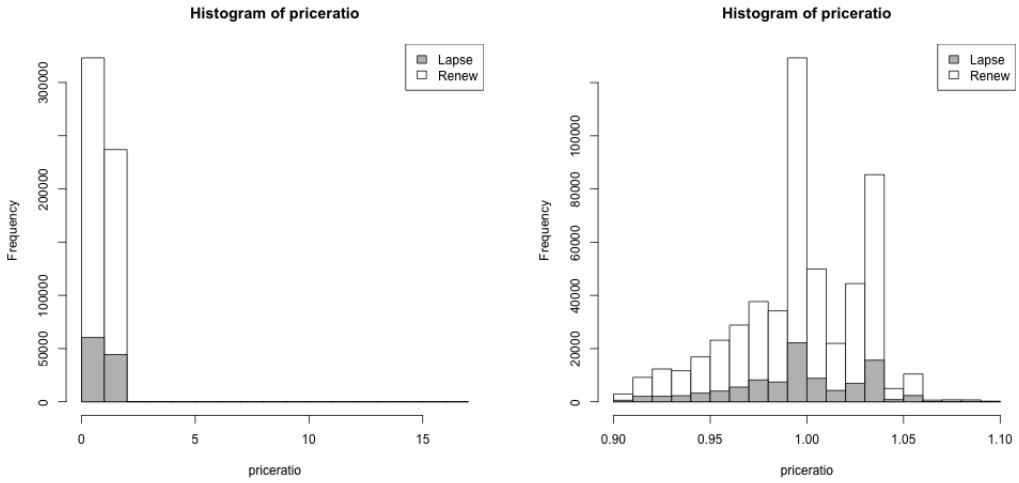


Figure B.6: Price ratio

	F	M		other	transfer
prop. size (%)	43.370755	56.629245	prop. size (%)	32.241395	67.758605
lapse rate (%)	5.958236	6.088259	lapse rate (%)	7.298461	5.429187

Table B.9: Gender / Billing mode - bill_mode

	(-1,5]	(5,10]	(10,15]	(15,91]
prop. size (%)	39.477699	28.809350	20.437710	11.059257
lapse rate (%)	5.927667	6.235154	6.190484	5.641583

Table B.10: Vehicle age group - veh_age_group

	N	Y		0	1
prop. size (%)	56.460330	43.539670	prop. size (%)	97.998725	2.001275
lapse rate (%)	8.242083	3.165756	lapse rate (%)	6.121663	1.634696

Table B.11: Have house policy at AXA - house_pol / Drop cover - drop_cover

	(0,2]	(2,6]	(6,11]
prop. size (%)	24.30206	29.241986	46.437930
lapse rate (%)	8.10588	6.927178	4.376424

Table B.12: Policy age group - pol_age_group

	allrisk	TPL	TPL+opt
prop. size (%)	61.713052	16.749288	21.537660
lapse rate (%)	5.856643	6.702675	6.012276

Table B.13: Cover

	N	Y
prop. size (%)	57.035175	42.964825
lapse rate (%)	7.086167	4.632299

Table B.14: Have multi-vehicle - multi_veh

	(15,25]	(25,35]	(35,45]	(45,55]	(55,65]	(65,75]	(75,99]
prop. size (%)	9.325041	10.662340	18.585654	22.501744	19.268658	12.568142	7.085416
lapse rate (%)	9.826734	8.168441	6.270319	5.548467	4.996448	4.382587	4.466431

Table B.15: Driver age group - drivage_group

	0	1	2+		0	1	2+
prop. size (%)	90.575813	8.717815	0.7063717	prop. size	90.789127	8.510510	0.7003629
lapse rate (%)	6.013755	6.176527	6.5689981	lapse rate	5.980873	6.558406	6.2440419

Table B.16: Last year claim group - claim_1_group / 2-year-ago claim group - claim_2_group

	0	1	2+		0	1	2+
prop. size (%)	97.435897	2.499341	0.06476187	prop. size	97.332078	2.604161	0.0637604
lapse rate (%)	5.992572	7.426205	11.34020619	lapse rate	5.990047	7.601590	5.7591623

Table B.17: Last year responsible claim group - respclaim_1_group / 2-year-ago responsible claim group - respclaim_2_group

	(0,250]	(250,500]	(500,750]	(750,1e+03]	(1e+03,2e+03]	(2e+03,Inf]
prop. size (%)	26.859150	32.468395	25.25780	9.332719	5.705387	0.3758859
lapse rate (%)	5.266036	4.997841	6.23827	8.251958	10.303669	0.3758859

Table B.18: Proposed premium group - next_prem_group

	(0,250]	(250,500]	(500,750]	(750,1e+03]	(1e+03,2e+03]	(2e+03,Inf]
prop. size (%)	24.30840	30.424724	26.933926	10.819571	6.981262	0.5321155
lapse rate (%)	5.42311	5.143735	5.992588	7.657894	9.209583	11.8569636

Table B.19: Last year premium group - prev_prem_group

	(0,15]	(15,25]	(25,99]
prop. size (%)	12.625226	45.41376	41.961016
lapse rate (%)	5.830249	6.08052	6.039873

Table B.20: Pricing group - price_group2

	(0,0.75]	(0.75,1]	(1,1.25]	(1.25,Inf]
prop. size (%)	6.615057	73.908312	18.127981	1.347982
lapse rate (%)	3.936213	5.983315	6.586992	11.515602

Table B.21: Price ratio - pricechange

B.1.3 Descriptive analysis for Germany data

The tables below are computed for the 2008 direct business only. Similar conclusions can be drawn for other distribution channels.

polage	(0,1]	(1,2]	(2,7]	(7,34]
prop. size	24.97228	16.78816	34.38068	23.85888
lapse rate	17.43146	15.27215	11.25925	8.781606

cover	fully compr.	partial compr.	TPL
prop. size	36.16025	37.60943	26.23033
lapse rate	14.2557	12.64208	12.79180

product	altern. garag	eco	VIP
prop. size	22.25926	75.67923	2.061505
lapse rate	14.41161	12.88283	14.90649

Table B.22: Policy information

bonusevol	down	stable	up
prop. size	33.31746	62.91899	3.763551
lapse rate	16.68709	11.52702	12.02169

typeclassTPL	9	10	11	12	13	14	15	16	17
prop. size	0.1161	0.090	0.07230	0.3735	2.790	10.38	10.95	15.06	18.70
lapse rate	12.26	10.84	7.575	8.797	10.87	12.06	12.43	13.18	13.26

typeclassTPL	18	19	20	21	22	23	24	25	
prop. size	17.95	13.14	4.990	2.904	1.580	0.7230	0.1500	0.007668	
lapse rate	12.31	13.35	13.43	12.82	14.62	15.75	8.029	14.28	

typeclassFC	9	10	11	12	13	14	15	16	17
prop. size	0.1161	7.853	7.441	8.935	6.038	7.639	7.127	10.1	10.82
lapse rate	12.26	9.737	10.61	11.59	12.20	12.16	12.71	13.18	13.27

typeclassFC	18	19	20	21	22	23	24	25	
prop. size	9.524	8.682	4.554	3.816	2.0	1.647	1.385	0.8271	
lapse rate	13.94	14.67	14.55	14.75	13.62	14.42	14.22	14.03	

Table B.23: Bonus information

polholderage	(-1,37]	(37,44]	(44,54]	(54,85]		
prop. size	26.82628	26.34495	22.08789	24.74089		
lapse rate	16.32178	13.39251	12.24752	10.74608		
maritalstatus	0	1	2	3		
prop. size	94.79958	0.8025974	1.854942	2.542883		
lapse rate	13.34483	12.30552	9.241126	13.70536		
diffdriverPH		all drivers > 24	commercial	learner 17	only partner	same-young drivers
prop. size	0.077108	7.885427	0.3707987	0.2812174	45.06622	46.31923
lapse rate	8.82353	13.07161	10.70336	14.11290	13.91943	12.68361
jobgroup	medical	normal	public			
prop. size	49.47725	50.50687	0.01587518			
lapse rate	12.79994	13.72219	7.142857			

Table B.24: Policyholder information

mileage	(1,9]	(9,12]	(12,15]	(15,35]
prop. size	33.85136	26.16386	20.03272	19.95206
lapse rate	13.17202	13.29512	13.34354	13.35193
finance	lease	none		
prop. size	0.9150905	99.0849		
lapse rate	15.24164	13.24659		
caruse	commercial	private	unknown	
prop. size	0.04195582	76.89935	23.05869	
lapse rate	13.51351	14.46266	9.269732	
carclass	(0,2]	(2,3]	(3,5]	
prop. size	59.24966	30.20909	10.54125	
lapse rate	12.84	13.55510	12.51374	
vehiclage	(0,6]	(6,10]	(10,13]	(13,18]
prop. size	26.05579	31.01423	21.84650	21.08348
lapse rate	15.49827	13.55724	12.71630	10.67061

Table B.25: Car information

housepol		flat owner	house with axa	no property	not with axa
prop. size	0.443371	19.36204	1.054565	51.01715	28.12287
lapse rate	10.48593	10.69400	14.40860	13.61606	14.39861
householdnbAXA	(0,2]	(2,3]	(3,15]		
prop. size	93.31812	4.045908	2.635970		
lapse rate	13.37177	12.14927	11.36951		
isinsuredinhealth	0	1			
prop. size	98.95787	1.042128			
lapse rate	13.28194	12.09150			
isinsuredinlife	0	1			
prop. size	97.52296	2.477040			
lapse rate	13.27831	12.92392			
isinsuredinaccident	0	1			
prop. size	99.48007	0.5199287			
lapse rate	13.27612	12.00873			

Table B.26: Car information

claimamount	(-1,435]	(435,1570]	(1570,48600]			
prop. size	60.00149	19.99926	19.99926			
lapse rate	13.46154	14.25647	14.07035			
nbclaim08percust	0	1	2	3	4	[5 - 13]
prop. size	70.59108	25.28847	3.599345	0.43549	0.07444	0.017444
lapse rate	13.75165	13.36473	16.02896	12.82051	20	100
nbclaim0708percust	0	1	2	3	4	[5, 14]
prop. size	38.32353	47.5322	11.28936	2.1365	0.55832	0.15191
lapse rate	12.45144	14.27565	15.13353	16.89895	15.33333	21.3583
nbclaim0608percust	(-1,1]	(1,2]	(2,27]			
prop. size	77.20539	17.00663	5.787985			
lapse rate	13.30151	15.01423	15.88424			

Table B.27: Claim information

lastprem	(0,500]	(500,1e+03]	(1e+03,5e+03]			
prop. size	72.4981	24.44247	3.059425			
lapse rate	10.47965	13.00631	14.85349			
finalprem	(0,500]	(500,1e+03]	(1e+03,5e+03]			
prop. size	72.3931	24.58395	3.022953			
lapse rate	10.42744	13.12656	15.06524			
priceratio	(0,0.75]	(0.75,0.95]	(0.95,1]	(1,1.05]	(1.05,1.25]	(1.25,5]
prop. size	0.2369838	2.148523	88.97731	1.460095	4.935531	2.241554
lapse rate	12.80992	12.09359	10.83894	14.01744	16.07143	13.99447
top10vip	(53.9,227]	(227,282]	(282,349]	(349,472]	(472,2460]	
prop. size	20	20	20	20	20	
lapse rate	10.15625	11.76215	13.02083	14.14931	14.90885	
top10eco	(39.7,185]	(185,229]	(229,282]	(282,382]	(382,1930]	
prop. size	20	20	20	20	20	
lapse rate	10.63922	11.61430	12.71939	13.7565	15.30126	
top10direct	(39.7,185]	(185,229]	(229,282]	(282,382]	(382,1933]	
prop. size	20	20	20	20	20	
lapse rate	10.63922	11.61430	12.71939	13.7565	15.30126	
paymentfreq	1	2	4	12		
prop. size	55.14356	16.12464	19.55255	9.179253		
lapse rate	13.67469	13.25598	12.61961	12.19271		
directdebit	0	1				
prop. size	12.63806	87.36194				
lapse rate	13.0821	13.29162				

Table B.28: Premium information

cumulrebate	(-1,4]	(4,9]	(9,14]	(14,90]
prop. size	70.40656	20.76536	5.695849	3.132224
lapse rate	12.82658	13.49583	12.13695	9.303952
techrebate	(-1,4]	(4,9]	(9,14]	(14,90]
prop. size	98.99646	0.5905102	0.2202088	0.1928197
lapse rate	12.80641	15.21336	12.43781	10.79545

Table B.29: Agent information

B.2 GLM analyses

B.2.1 GLM analysis for Portugal data

Continuous variables

Here follows the backward selection when all variables are continuous.

```
> summary(resglm25)

Call:
glm(formula = did_lapse ~ priceratio * (gender + age + age_policy + age_vehicle + premium_before),
   family = binomial(), data = workdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4472  -0.6757  -0.6045  -0.5145   2.5865

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.7666660  0.1917413  -3.998 6.38e-05 ***
priceratio                -0.4531173  0.1914016  -2.367 0.017915 *
genderMALE                 0.7566560  0.1161393   6.515 7.27e-11 ***
age                       -0.0350644  0.0031744 -11.046  < 2e-16 ***
age_policy                -0.0107819  0.0094287  -1.144 0.252820
age_vehicle               -0.0246413  0.0064303  -3.832 0.000127 ***
premium_before            -0.0017808  0.0001781  -9.999  < 2e-16 ***
priceratio:genderMALE     -0.6760655  0.1159748  -5.829 5.56e-09 ***
priceratio:age             0.0179810  0.0031567   5.696 1.23e-08 ***
priceratio:age_policy      0.0031629  0.0093887   0.337 0.736204
priceratio:age_vehicle     0.0538058  0.0064112   8.392  < 2e-16 ***
priceratio:premium_before  0.0025069  0.0001794  13.973  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 539837  on 560343  degrees of freedom
Residual deviance: 531382  on 560332  degrees of freedom
AIC: 531406

Number of Fisher Scoring iterations: 4

> summary(resglm26)

Call:
glm(formula = did_lapse ~ age_policy + priceratio * (gender + age + age_vehicle + premium_before),
    family = binomial(), data = workdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3538  -0.6757  -0.6045  -0.5145   2.5791

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.7737301  0.1907634  -4.056 4.99e-05 ***
age_policy                -0.0076121  0.0006013 -12.659  < 2e-16 ***
priceratio                -0.4459493  0.1903878  -2.342 0.019164 *
genderMALE                 0.7540763  0.1159249   6.505 7.78e-11 ***
age                       -0.0352652  0.0031278 -11.275  < 2e-16 ***
age_vehicle               -0.0246664  0.0064425  -3.829 0.000129 ***
premium_before            -0.0017755  0.0001777  -9.989  < 2e-16 ***
priceratio:genderMALE     -0.6734653  0.1157564  -5.818 5.96e-09 ***
priceratio:age             0.0181797  0.0031110   5.844 5.11e-09 ***
priceratio:age_vehicle     0.0538313  0.0064235   8.380  < 2e-16 ***
priceratio:premium_before  0.0025015  0.0001790  13.973  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 539837  on 560343  degrees of freedom
Residual deviance: 531382  on 560333  degrees of freedom
AIC: 531404

Number of Fisher Scoring iterations: 4
```

Analysis of link functions:

```
> summary(resglm27)

Call:
glm(formula = did_lapse ~ age_policy + priceratio * (gender + age + age_vehicle + premium_before),
    family = binomial("probit"), data = workdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5676  -0.6765  -0.6055  -0.5121   2.6552

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.5683747  0.1090992  -5.210 1.89e-07 ***
age_policy               -0.0041449  0.0003329 -12.453  < 2e-16 ***
priceratio               -0.1909652  0.1089723  -1.752  0.07970 .
genderMALE                0.4313946  0.0669263   6.446 1.15e-10 ***
age                      -0.0187913  0.0017734 -10.596  < 2e-16 ***
age_vehicle              -0.0108735  0.0036373  -2.989  0.00279 **
premium_before           -0.0010077  0.0001022  -9.859  < 2e-16 ***
priceratio:genderMALE    -0.3855923  0.0669052  -5.763 8.25e-09 ***
priceratio:age            0.0092317  0.0017656   5.229 1.71e-07 ***
priceratio:age_vehicle    0.0276823  0.0036275   7.631 2.33e-14 ***
priceratio:premium_before  0.0014363  0.0001031  13.934  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 539837  on 560343  degrees of freedom
Residual deviance: 531339  on 560333  degrees of freedom
AIC: 531361

Number of Fisher Scoring iterations: 5

> summary(resglm28)

Call:
glm(formula = did_lapse ~ age_policy + priceratio * (gender + age + age_vehicle + premium_before),
    family = binomial("cloglog"), data = workdata)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-8.490e+00  -8.490e+00  -2.107e-08  -2.107e-08   8.490e+00

Coefficients:
                           Estimate Std. Error    z value Pr(>|z|)
(Intercept)              -1.965e+14  4.785e+06  -41071229   <2e-16 ***
age_policy               -1.395e+13  1.475e+04 -946234735   <2e-16 ***
priceratio                7.060e+14  4.784e+06  147577169   <2e-16 ***
genderMALE                1.076e+15  3.104e+06  346751583   <2e-16 ***
age                      -6.826e+12  7.938e+04  -86001767   <2e-16 ***
age_vehicle              -5.303e+13  1.595e+05 -332499024   <2e-16 ***
premium_before           -1.827e+12  4.419e+03 -413476726   <2e-16 ***
priceratio:genderMALE    -1.002e+15  3.108e+06 -322572510   <2e-16 ***
priceratio:age           -2.081e+13  7.913e+04 -263027784   <2e-16 ***
priceratio:age_vehicle    9.127e+13  1.591e+05  573841804   <2e-16 ***
priceratio:premium_before  2.556e+12  4.465e+03  572450712   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance:   539837  on 560343  degrees of freedom
Residual deviance: 14549742  on 560333  degrees of freedom
AIC: 14549764

Number of Fisher Scoring iterations: 25

> summary(resglm26)

Call:
glm(formula = did_lapse ~ age_policy + priceratio * (gender + age + age_vehicle + premium_before),
    family = binomial("logit"), data = workdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3538  -0.6757  -0.6045  -0.5145   2.5791

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.7737301  0.1907634  -4.056 4.99e-05 ***
age_policy               -0.0076121  0.0006013 -12.659  < 2e-16 ***
priceratio               -0.4459493  0.1903878  -2.342 0.019164 *
genderMALE                0.7540763  0.1159249   6.505 7.78e-11 ***
age                      -0.0352652  0.0031278 -11.275  < 2e-16 ***
age_vehicle              -0.0246664  0.0064425  -3.829 0.000129 ***
premium_before           -0.0017755  0.0001777  -9.989  < 2e-16 ***
priceratio:genderMALE    -0.6734653  0.1157564  -5.818 5.96e-09 ***
priceratio:age            0.0181797  0.0031110   5.844 5.11e-09 ***
priceratio:age_vehicle    0.0538313  0.0064235   8.380  < 2e-16 ***
priceratio:premium_before  0.0025015  0.0001790  13.973  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

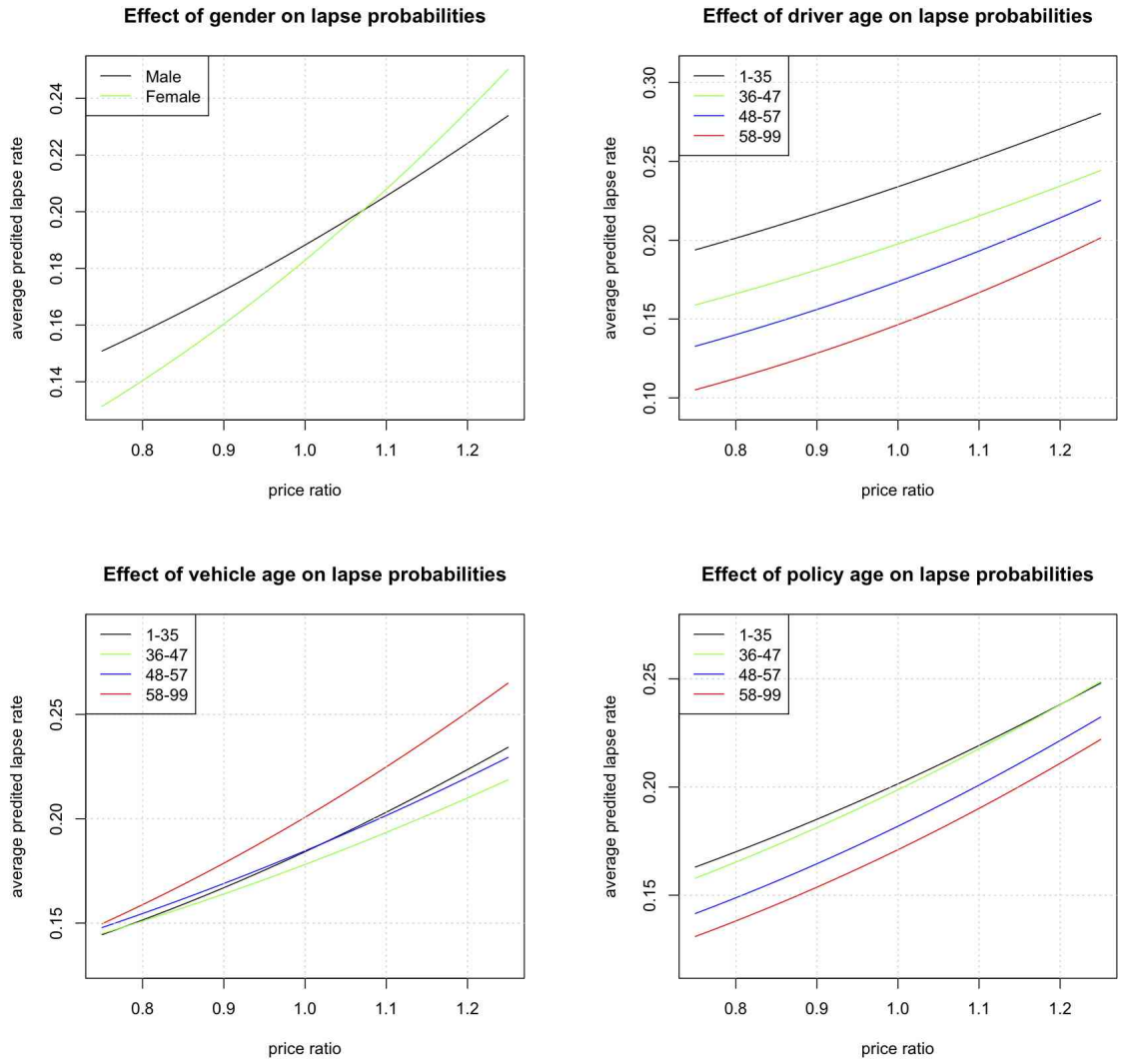


Figure B.7: One-variable effect on lapse for continuous variables

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 539837  on 560343  degrees of freedom
Residual deviance: 531382  on 560333  degrees of freedom
AIC: 531404

Number of Fisher Scoring iterations: 4

>
> anova(resglm26, resglm27, test="Chisq")
Analysis of Deviance Table

Model 1: did_lapse ~ age_policy + priceratio * (gender + age + age_vehicle + premium_before) - logit
Model 2: did_lapse ~ age_policy + priceratio * (gender + age + age_vehicle + premium_before) - probit
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1    560333     531382
2    560333     531339  0   43.354
>
```

Categorical variables

Here follows the backward selection when variables are categorical.

```
> summary(resglm38)

Call:
glm(formula = did_lapse ~ priceratio * (gender + agegroup + agepolgroup +
    agevehgroup + prembeforegroup), family = binomial(), data = workdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3082  -0.6770  -0.6028  -0.5189   2.5276

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                              -2.289692   0.723096  -3.167 0.001543 **
priceratio                                0.904127   0.742198   1.218 0.223157
genderMALE                                0.709440   0.116143   6.108 1.01e-09 ***
agegroup(20,35]                           0.151167   0.716661   0.211 0.832940
agegroup(35,60]                          -0.464333   0.716779  -0.648 0.517111
agegroup(60,99]                          -1.187186   0.727779  -1.631 0.102839
agepolgroup(4,8]                          0.046926   0.098402   0.477 0.633449
agepolgroup(8,12]                        -0.227676   0.113902  -1.999 0.045623 *
agepolgroup(12,49]                       -0.237745   0.211287  -1.125 0.260496
agevehgroup(5,10]                        -0.706848   0.106784  -6.619 3.61e-11 ***
agevehgroup(10,15]                       -0.184412   0.121030  -1.524 0.127587
agevehgroup(15,99]                       -0.231443   0.130382  -1.775 0.075880 .
prembeforegroup(500,1e+03]               -0.370542   0.138315  -2.679 0.007385 **
prembeforegroup(1e+03,1.5e+03]           -1.093366   0.314019  -3.482 0.000498 ***
prembeforegroup(1.5e+03,1e+04]           -0.667078   0.639235  -1.044 0.296690
priceratio:genderMALE                    -0.638175   0.115970  -5.503 3.74e-08 ***
priceratio:agegroup(20,35]               -0.263852   0.735989  -0.359 0.719969
priceratio:agegroup(35,60]               -0.006562   0.736058  -0.009 0.992886
priceratio:agegroup(60,99]                0.378138   0.746560   0.507 0.612500
priceratio:agepolgroup(4,8]              -0.183033   0.097695  -1.874 0.060996 .
priceratio:agepolgroup(8,12]              0.076816   0.113106   0.679 0.497041
priceratio:agepolgroup(12,49]             0.021355   0.210124   0.102 0.919048
priceratio:agevehgroup(5,10]              0.978501   0.106478   9.190  < 2e-16 ***
priceratio:agevehgroup(10,15]             0.638095   0.120390   5.300 1.16e-07 ***
priceratio:agevehgroup(15,99]             0.803768   0.129724   6.196 5.79e-10 ***
priceratio:prembeforegroup(500,1e+03]     0.644801   0.139476   4.623 3.78e-06 ***
priceratio:prembeforegroup(1e+03,1.5e+03] 1.502788   0.315483   4.763 1.90e-06 ***
priceratio:prembeforegroup(1.5e+03,1e+04] 1.112602   0.644471   1.726 0.084279 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 539782  on 560267  degrees of freedom
Residual deviance: 532308  on 560240  degrees of freedom
  (76 observations deleted due to missingness)
AIC: 532364

Number of Fisher Scoring iterations: 4

> summary(resglm46)

Call:
glm(formula = did_lapse ~ agepolgroup2 + priceratio:agegroup4 +
    priceratio * (gender + agevehgroup2 + prembeforegroup2),
    family = binomial(), data = workdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1587  -0.6633  -0.6060  -0.5193   2.8747

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                              -2.522477   0.120852 -20.873  < 2e-16 ***
agepolgroup2(4,49]                       -0.153793   0.007270 -21.154  < 2e-16 ***
priceratio                                1.018771   0.120903   8.426  < 2e-16 ***
genderMALE                                0.681454   0.117045   5.822 5.81e-09 ***
agevehgroup2(5,10]                       -0.684290   0.106741  -6.411 1.45e-10 ***
agevehgroup2(10,99]                      -0.262674   0.101038  -2.600  0.00933 **
prembeforegroup2(500,1e+03]              -0.295837   0.137011  -2.159  0.03083 *
prembeforegroup2(1e+03,1e+04]            -0.923435   0.283603  -3.256  0.00113 **
priceratio:agegroup4(35,60]              -0.352247   0.008083 -43.579  < 2e-16 ***
priceratio:agegroup4(60,99]              -0.674209   0.011248 -59.938  < 2e-16 ***
priceratio:genderMALE                    -0.607070   0.116885  -5.194 2.06e-07 ***
priceratio:agevehgroup2(5,10]             0.956935   0.106426   8.992  < 2e-16 ***
priceratio:agevehgroup2(10,99]            0.766736   0.100552   7.625 2.44e-14 ***
priceratio:prembeforegroup2(500,1e+03]    0.569856   0.138151   4.125 3.71e-05 ***
priceratio:prembeforegroup2(1e+03,1e+04]  1.340304   0.285123   4.701 2.59e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 539782  on 560267  degrees of freedom
```

```
Residual deviance: 532580  on 560253  degrees of freedom
  (76 observations deleted due to missingness)
AIC: 532610

Number of Fisher Scoring iterations: 4

>
```

Below, we put the GLM fit summaries respectively for the price-increase population and the price-decrease population.

```
> summary(resglm50up)

Call:
glm(formula = did_lapse ~ agepolgroup2 + priceratio:(agegroup4 +
    prembeforegroup2) + priceratio * (gender + agevehgroup2),
    family = binomial(), data = workdata[idxup, ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9921  -0.6652  -0.6045  -0.5266   3.0007

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -2.604269   0.138423 -18.814  < 2e-16 ***
agepolgroup2(4,49]                     -0.127231   0.007764 -16.388  < 2e-16 ***
priceratio                              1.077198   0.136502   7.891 2.99e-15 ***
genderMALE                              0.822231   0.138906   5.919 3.23e-09 ***
agevehgroup2(5,10]                     -0.945460   0.128975  -7.331 2.29e-13 ***
agevehgroup2(10,99]                    -0.753005   0.124929  -6.027 1.67e-09 ***
priceratio:agegroup4(35,60]            -0.330512   0.008643 -38.241  < 2e-16 ***
priceratio:agegroup4(60,99]            -0.646742   0.011841 -54.618  < 2e-16 ***
priceratio:prembeforegroup2(500,1e+03]  0.295764   0.013132  22.523  < 2e-16 ***
priceratio:prembeforegroup2(1e+03,1e+04]  0.442775   0.027677  15.998  < 2e-16 ***
priceratio:genderMALE                  -0.743046   0.136960  -5.425 5.79e-08 ***
priceratio:agevehgroup2(5,10]           1.195475   0.126936   9.418  < 2e-16 ***
priceratio:agevehgroup2(10,99]          1.221063   0.122737   9.949  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 472406  on 490261  degrees of freedom
Residual deviance: 466542  on 490249  degrees of freedom
AIC: 466568

Number of Fisher Scoring iterations: 4

> summary(resglm50down)

Call:
glm(formula = did_lapse ~ agepolgroup2 + gender + priceratio:(agegroup4 +
    prembeforegroup2) + priceratio * (agevehgroup2), family = binomial(),
    data = workdata[idxdown, ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9758  -0.6617  -0.5996  -0.5145   2.3583

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -2.964607   0.206029 -14.389  < 2e-16 ***
agepolgroup2(4,49]                     -0.161592   0.008509 -18.991  < 2e-16 ***
genderMALE                              0.099735   0.009984   9.990  < 2e-16 ***
priceratio                              1.463632   0.212330   6.893 5.45e-12 ***
agevehgroup2(5,10]                      0.785047   0.260462   3.014  0.00258 **
agevehgroup2(10,99]                     1.901443   0.245635   7.741 9.87e-15 ***
priceratio:agegroup4(35,60]            -0.392557   0.009551 -41.100  < 2e-16 ***
priceratio:agegroup4(60,99]            -0.713592   0.013853 -51.511  < 2e-16 ***
priceratio:prembeforegroup2(500,1e+03]  0.289150   0.013990  20.669  < 2e-16 ***
priceratio:prembeforegroup2(1e+03,1e+04]  0.413289   0.028901  14.300  < 2e-16 ***
priceratio:agevehgroup2(5,10]          -0.529244   0.267397  -1.979  0.04779 *
priceratio:agevehgroup2(10,99]         -1.413007   0.252110  -5.605 2.09e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 406081  on 425550  degrees of freedom
Residual deviance: 400537  on 425539  degrees of freedom
AIC: 400561

Number of Fisher Scoring iterations: 4
```

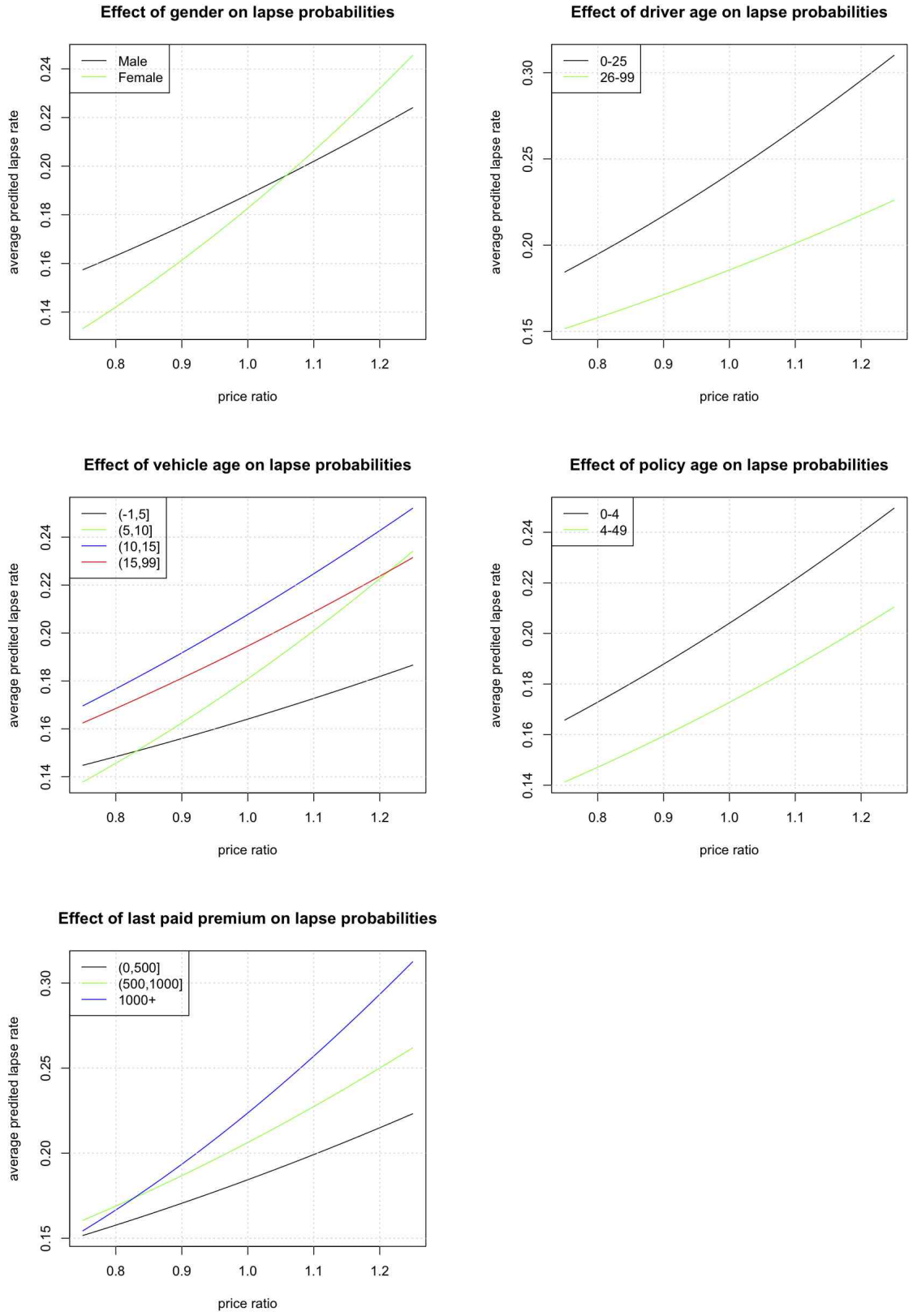


Figure B.8: One-variable effect on lapse for categorical variables

B.2.2 GLM analysis for Québec data

Year 2007

Here follows the backward selection when all variables are continuous.

```
> summary(resglm01)

Call:
glm(formula = did_cancel ~ pricefactor * (pol_age_group + multi_veh_dsc +
    house_pol + veh_age_group + price_group2 + cover2 + gender +
    drivage_group + prev_prem_group + respclaim_1_group + respclaim_2_group),
    family = binomial(), data = workdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8740  -0.4268  -0.3318  -0.2488   3.0148

Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -2.73354    0.40746  -6.709 1.96e-11 ***
pricefactor                             0.45475    0.42096   1.080 0.280016
pol_age_group(2,6]                      0.04676    0.15541   0.301 0.763519
pol_age_group(6,11]                    -0.44718    0.16686  -2.680 0.007362 **
multi_veh_dscY                         -0.19659    0.15624  -1.258 0.208302
house_polY                             -1.16914    0.15312  -7.635 2.25e-14 ***
veh_age_group(5,10]                    -0.15624    0.17714  -0.882 0.377759
veh_age_group(10,15]                    0.03098    0.26806   0.116 0.907980
veh_age_group(15,91]                    0.16869    0.35498   0.475 0.634629
price_group2(15,25]                    -0.28099    0.28616  -0.982 0.326141
price_group2(25,99]                    -0.31968    0.32556  -0.982 0.326129
cover2TPL+opt                           0.59909    0.17247   3.474 0.000513 ***
genderM                                 0.05151    0.13145   0.392 0.695144
drivage_group(25,35]                   -0.10856    0.20932  -0.519 0.604034
drivage_group(35,45]                    0.09019    0.21660   0.416 0.677113
drivage_group(45,55]                   -0.30168    0.21703  -1.390 0.164516
drivage_group(55,65]                   -0.06805    0.24201  -0.281 0.778583
drivage_group(65,75]                   -0.22715    0.27263  -0.833 0.404744
drivage_group(75,99]                   -0.81497    0.30644  -2.659 0.007826 **
prev_prem_group(1e+03,2e+03]           -0.36103    0.29977  -1.204 0.228457
prev_prem_group(250,500]               -0.69695    0.23202  -3.004 0.002665 **
prev_prem_group(2e+03,Inf]              0.10988    0.51690   0.213 0.831655
prev_prem_group(500,750]               -0.75412    0.26599  -2.835 0.004580 **
prev_prem_group(750,1e+03]             -0.51114    0.29284  -1.745 0.080903 .
respclaim_1_group1                      0.34025    0.21743   1.565 0.117608
respclaim_1_group2+                     1.57437    0.67674   2.326 0.019997 *
respclaim_2_group1                      0.51399    0.22532   2.281 0.022541 *
respclaim_2_group2+                     1.35116    1.67507   0.807 0.419880
pricefactor:pol_age_group(2,6]         -0.13094    0.15926  -0.822 0.410971
pricefactor:pol_age_group(6,11]         0.08384    0.16895   0.496 0.619727
pricefactor:multi_veh_dscY             -0.04643    0.15887  -0.292 0.770082
pricefactor:house_polY                  0.28769    0.15461   1.861 0.062776 .
pricefactor:veh_age_group(5,10]         0.34609    0.18001   1.923 0.054532 .
pricefactor:veh_age_group(10,15]        0.23223    0.27462   0.846 0.397761
pricefactor:veh_age_group(15,91]        0.12160    0.36585   0.332 0.739599
pricefactor:price_group2(15,25]         0.23024    0.29613   0.777 0.436866
pricefactor:price_group2(25,99]         0.19926    0.33556   0.594 0.552638
pricefactor:cover2TPL+opt              -0.67688    0.17614  -3.843 0.000122 ***
pricefactor:genderM                    -0.08546    0.13377  -0.639 0.522913
pricefactor:drivage_group(25,35]        0.11203    0.21757   0.515 0.606623
pricefactor:drivage_group(35,45]       -0.24265    0.22262  -1.090 0.275709
pricefactor:drivage_group(45,55]        0.13568    0.22301   0.608 0.542909
pricefactor:drivage_group(55,65]       -0.12258    0.24797  -0.494 0.621074
pricefactor:drivage_group(65,75]       -0.06388    0.27604  -0.231 0.816994
pricefactor:drivage_group(75,99]        0.51930    0.30831   1.684 0.092117 .
pricefactor:prev_prem_group(1e+03,2e+03]  1.17909    0.30795   3.829 0.000129 ***
pricefactor:prev_prem_group(250,500]    0.93871    0.23971   3.916 9.00e-05 ***
pricefactor:prev_prem_group(2e+03,Inf]  0.88518    0.54871   1.613 0.106699
pricefactor:prev_prem_group(500,750]    1.27567    0.27307   4.672 2.99e-06 ***
pricefactor:prev_prem_group(750,1e+03]  1.24924    0.29997   4.165 3.12e-05 ***
pricefactor:respclaim_1_group1         -0.40788    0.20242  -2.015 0.043907 *
pricefactor:respclaim_1_group2+        -1.28800    0.53702  -2.398 0.016467 *
pricefactor:respclaim_2_group1         -0.44224    0.22503  -1.965 0.049380 *
pricefactor:respclaim_2_group2+        -1.83594    1.82804  -1.004 0.315225
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 118078  on 239927  degrees of freedom
Residual deviance: 112618  on 239874  degrees of freedom
  (394 observations deleted due to missingness)
AIC: 112726

Number of Fisher Scoring iterations: 6
```

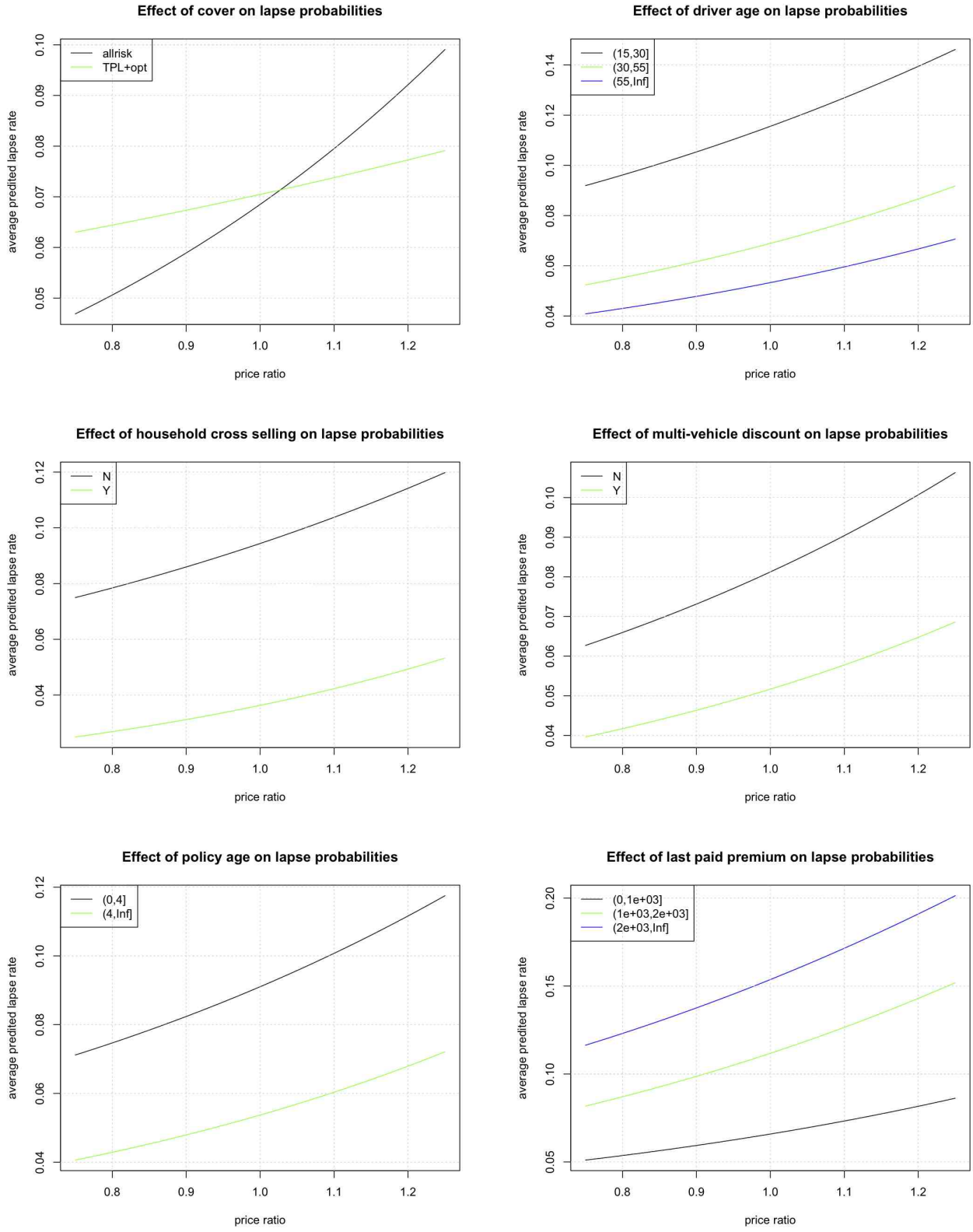


Figure B.9: One-variable effect on lapse (2007 data)

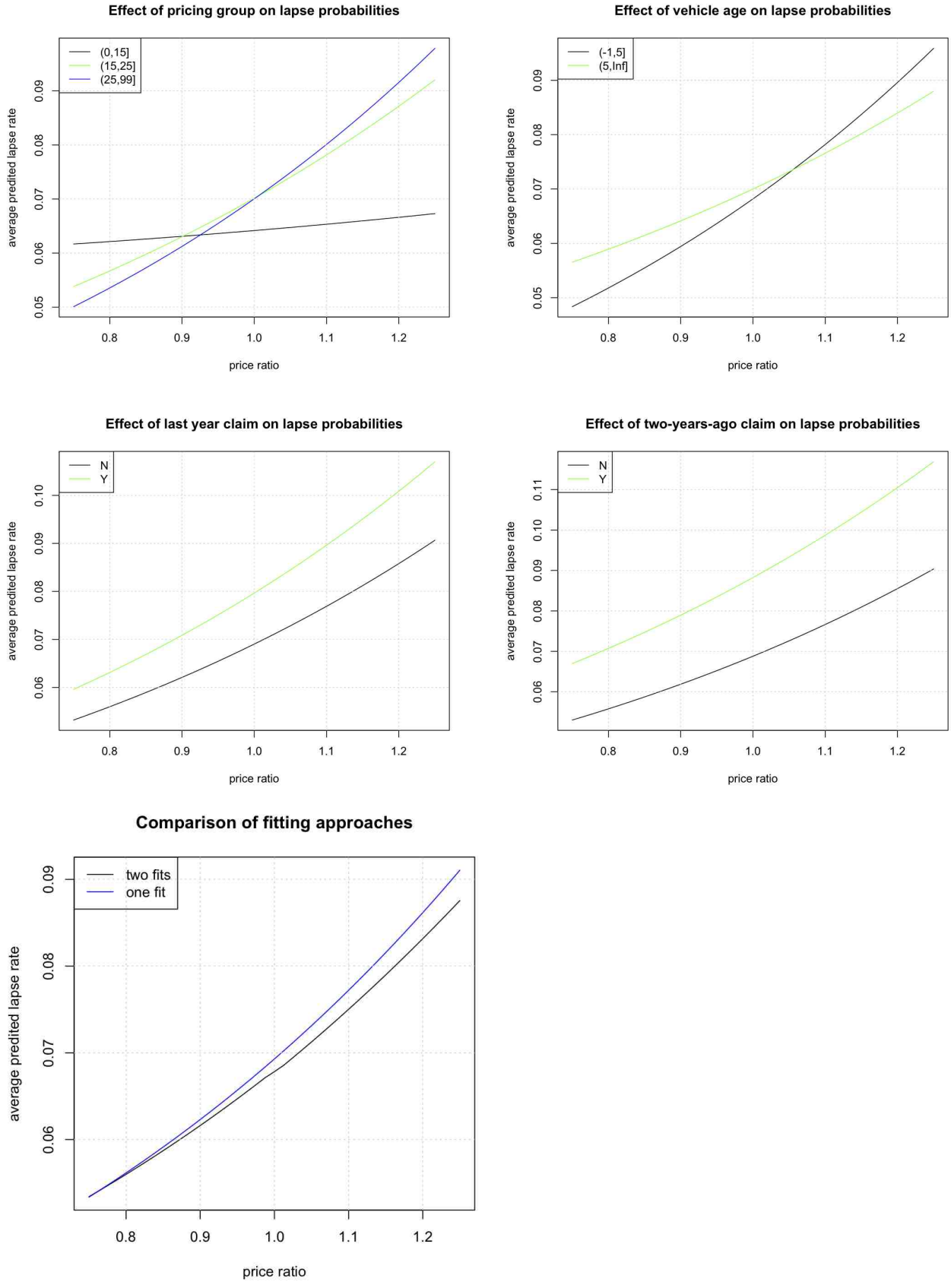


Figure B.10: One-variable effect on lapse (2007 data) (continued)

Year 2006

Please find below the final variable selection of the regression.

```
Call: glm(formula = did_cancel ~ drivage_group2 + house_pol + pol_age_group2 +    pricefactor * (cover2 + resp_claim_1 + resp_claim_2) + pricefactor:(m
    veh_age_group2 + prev_prem_group3), family = binomial(), data = workdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3965  -0.4565  -0.3635  -0.2677   2.8822

Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -3.66656    0.09274 -39.536  < 2e-16 ***
drivage_group2(30,55]                -0.17948    0.02270  -7.908 2.62e-15 ***
drivage_group2(55,Inf]               -0.27242    0.02510 -10.855  < 2e-16 ***
house_polY                           -0.89993    0.01923 -46.787  < 2e-16 ***
pol_age_group2(4,Inf]                -0.36730    0.01679 -21.872  < 2e-16 ***
pricefactor                           1.74083    0.09530  18.267  < 2e-16 ***
cover2TPL+opt                         0.94588    0.12981   7.287 3.18e-13 ***
resp_claim_1Y                         0.41288    0.20864   1.979  0.04783 *
resp_claim_2Y                         0.63793    0.21949   2.906  0.00366 **
pricefactor:cover2TPL+opt            -1.05538    0.13364  -7.897 2.86e-15 ***
pricefactor:resp_claim_1Y            -0.45033    0.20191  -2.230  0.02572 *
pricefactor:resp_claim_2Y            -0.64133    0.22308  -2.875  0.00404 **
pricefactor:multi_veh_dscY           -0.18801    0.01811 -10.380  < 2e-16 ***
pricefactor:veh_age_group2(5,10]      0.17581    0.02175   8.082 6.37e-16 ***
pricefactor:veh_age_group2(10,Inf]    0.12624    0.02796   4.515 6.35e-06 ***
pricefactor:prev_prem_group3(500,1e+03]  0.35641    0.02256  15.795  < 2e-16 ***
pricefactor:prev_prem_group3(1e+03,Inf]  0.64826    0.03262  19.875  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 126871  on 232510  degrees of freedom
Residual deviance: 120983  on 232494  degrees of freedom
AIC: 121017

Number of Fisher Scoring iterations: 6
```

Year 2005

Please find below the final variable selection of the regression.

```
Call: glm(formula = did_cancel ~ drivage_group2 + pol_age_group2 + pricefactor * (house_pol + price_group2 + cover2 + resp_claim_2) +
    pricefactor:(multi_veh_dsc + veh_age_group3 + resp_claim_1 + prev_prem_group3), family = binomial(), data = workdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6252  -0.4856  -0.3922  -0.2998   2.8428

Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -3.03286    0.17184 -17.649  < 2e-16 ***
drivage_group2(30,55]                -0.20989    0.01897 -11.064  < 2e-16 ***
drivage_group2(55,Inf]               -0.39234    0.02159 -18.171  < 2e-16 ***
pol_age_group2(4,Inf]                -0.31619    0.01431 -22.095  < 2e-16 ***
pricefactor                           1.25063    0.18089   6.914 4.73e-12 ***
house_polY                           -1.10539    0.12250  -9.024  < 2e-16 ***
price_group2(15,25]                  -0.49213    0.16477  -2.987 0.002819 **
price_group2(25,99]                  -0.44995    0.17391  -2.587 0.009676 **
cover2TPL+opt                         0.80632    0.11695   6.895 5.39e-12 ***
resp_claim_2Y                         0.59906    0.19666   3.046 0.002317 **
pricefactor:house_polY                0.36616    0.12482   2.934 0.003352 **
pricefactor:price_group2(15,25]       0.52303    0.17316   3.020 0.002524 **
pricefactor:price_group2(25,99]       0.51810    0.18197   2.847 0.004411 **
pricefactor:cover2TPL+opt            -0.95515    0.12061  -7.919 2.39e-15 ***
pricefactor:resp_claim_2Y            -0.66833    0.20085  -3.328 0.000876 ***
pricefactor:multi_veh_dscY           -0.14885    0.01548  -9.614  < 2e-16 ***
pricefactor:veh_age_group3(5,Inf]     0.08699    0.01933   4.501 6.78e-06 ***
pricefactor:resp_claim_1Y            -0.21944    0.03996  -5.491 4.00e-08 ***
pricefactor:prev_prem_group3(500,1e+03]  0.36928    0.01983  18.619  < 2e-16 ***
pricefactor:prev_prem_group3(1e+03,Inf]  0.61236    0.02857  21.437  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 169763  on 278590  degrees of freedom
Residual deviance: 162503  on 278571  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 162543

Number of Fisher Scoring iterations: 5
```

Year 2004

Please find below the final variable selection of the regression.

```
glm(formula = did_cancel ~ drivage_group2 + pricefactor * (pol_age_group2 + house_pol + cover2 + resp_claim_1 + resp_claim_2) + pricefactor:(price_grou
    prev_prem_group3 + multi_veh_dsc + gender + veh_age_group3), family = binomial(), data = workdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5240  -0.4795  -0.3890  -0.2928   3.0141

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                           -3.58257    0.09460 -37.869  < 2e-16 ***
drivage_group2(30,55]                 -0.22323    0.01912 -11.677  < 2e-16 ***
drivage_group2(55,Inf]                -0.39959    0.02207 -18.109  < 2e-16 ***
pricefactor                            1.62437    0.09703  16.740  < 2e-16 ***
pol_age_group2(4,Inf]                 -0.47006    0.11160  -4.212 2.53e-05 ***
house_polY                            -1.03525    0.12285  -8.427  < 2e-16 ***
cover2TPL+opt                          0.95824    0.11282   8.494  < 2e-16 ***
resp_claim_1Y                          0.40308    0.18659   2.160 0.030753 *
resp_claim_2Y                          0.75472    0.19971   3.779 0.000157 ***
pricefactor:pol_age_group2(4,Inf]      0.26288    0.11177   2.352 0.018671 *
pricefactor:house_polY                 0.24882    0.12250   2.031 0.042231 *
pricefactor:cover2TPL+opt             -1.02438    0.11420  -8.970  < 2e-16 ***
pricefactor:resp_claim_1Y             -0.51115    0.17705  -2.887 0.003888 **
pricefactor:resp_claim_2Y             -0.75389    0.19900  -3.788 0.000152 ***
pricefactor:price_group2(15,25]        0.05876    0.02094   2.806 0.005011 **
pricefactor:price_group2(25,99]        0.13252    0.02757   4.807 1.53e-06 ***
pricefactor:prev_prem_group3(500,1e+03]  0.40126    0.02006  19.999  < 2e-16 ***
pricefactor:prev_prem_group3(1e+03,Inf]  0.60923    0.02933  20.772  < 2e-16 ***
pricefactor:multi_veh_dscY            -0.19931    0.01560 -12.779  < 2e-16 ***
pricefactor:genderM                   -0.05170    0.01436  -3.601 0.000317 ***
pricefactor:veh_age_group3(5,Inf]      0.14495    0.01952   7.427 1.11e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 162900  on 271827  degrees of freedom
Residual deviance: 155657  on 271807  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 155699

Number of Fisher Scoring iterations: 5
```

B.2.3 GLM analysis for Germany data

Lapse regression

Below we list the regression summaries for the Germany data.

TPL direct channel

```
Call: glm(formula = lapse ~ diff2tech + product2 + claimamount + cumulrebate2 +
    isinsuredinhealth + diff2top10vip + priceratio:(diff2tech +
    isinsuredinhealth + polholderage + jobgroup2), family = binomial("logit"),
    data = idata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3193  -0.5411  -0.4924  -0.4371   2.8131
```

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.329e+00  9.469e-02 -14.038  < 2e-16 ***
diff2tech                      1.024e+01  3.087e+00   3.315 0.000915 ***
product2eco                   -1.529e-01  5.000e-02  -3.059 0.002221 **
claimamount                    3.712e-05  1.488e-05   2.495 0.012581 *
cumulrebate2_10+              -4.972e-01  1.358e-01  -3.661 0.000251 ***
isinsuredinhealth             -8.712e+00  3.157e+00  -2.759 0.005794 **
diff2top10vip                 -2.417e-01  8.437e-02  -2.865 0.004169 **
diff2tech:priceratio          -1.280e+01  3.145e+00  -4.072 4.67e-05 ***
isinsuredinhealth:priceratio   8.740e+00  3.127e+00   2.795 0.005186 **
priceratio:polholderage       -1.238e-02  1.702e-03  -7.275 3.46e-13 ***
priceratio:jobgroup2public    -1.080e-01  4.067e-02  -2.656 0.007917 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 17943  on 24234  degrees of freedom
Residual deviance: 17730  on 24224  degrees of freedom
  (113 observations deleted due to missingness)
AIC: 17752

Number of Fisher Scoring iterations: 5
```

TPL broker channel

```
Call: glm(formula = lapse ~ isinsuredinhealth + gender + polage + bonusevol +
    cumulrebate2 + priceratio:(lastprem_group2 + diff2tech +
    paymentfreq + directdebit + isinsuredinhealth + vehiclage +
    householdnbAXA + polholderage + bonusevol), family = binomial("logit"),
    data = idata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1980  -0.5436  -0.4768  -0.4004   2.7510

Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -3.760450   0.420864  -8.935  < 2e-16 ***
isinsuredinhealth                  -2.975226   1.274174  -2.335 0.019542 *
gender                             -0.079972   0.027322  -2.927 0.003423 **
polage                             -0.017645   0.005243  -3.365 0.000765 ***
bonusevolstable                    -3.410175   0.715415  -4.767 1.87e-06 ***
bonusevolup                         1.401355   0.585020   2.395 0.016602 *
cumulrebate2_10-20                 -0.368879   0.029987 -12.301  < 2e-16 ***
cumulrebate2_25+                   -0.789049   0.064987 -12.142  < 2e-16 ***
priceratio:lastprem_group2(0,500]   2.825986   0.433422   6.520 7.02e-11 ***
priceratio:lastprem_group2(500,5e+03]  2.973426   0.444911   6.683 2.34e-11 ***
priceratio:diff2tech               -1.763109   0.271805  -6.487 8.78e-11 ***
priceratio:paymentfreq             -0.031802   0.003662  -8.685  < 2e-16 ***
priceratio:directdebit              0.081609   0.032096   2.543 0.011001 *
isinsuredinhealth:priceratio        2.690558   1.229736   2.188 0.028676 *
priceratio:vehiclage               -0.031511   0.003129 -10.069  < 2e-16 ***
priceratio:householdnbAXA          -0.048904   0.011522  -4.244 2.19e-05 ***
priceratio:polholderage            -0.005533   0.001041  -5.313 1.08e-07 ***
bonusevolstable:priceratio          3.175966   0.707735   4.488 7.21e-06 ***
bonusevolup:priceratio             -1.639430   0.516407  -3.175 0.001500 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 43971  on 59780  degrees of freedom
Residual deviance: 42996  on 59762  degrees of freedom
  (2451 observations deleted due to missingness)
AIC: 43034

Number of Fisher Scoring iterations: 5
```

TPL agent channel

```
Call: glm(formula = lapse ~ diff2tech + diff2top10vip + product2 +
    region2 + cumulrebate3 + nbclaim0608percust + isinsuredinhealth +
    isinsuredinlife + vehiclage + householdnbAXA + polholderage +
    maritalstatus2 + jobgroup2 + gender + typeclassTPL + bonusevol2 +
    priceratio:(diff2tech + diff2top10vip + diff2top10direct +
        paymentfreq + nbclaim08percust + nbclaim0608percust +
        nbclaim0708percust + isinsuredinaccident + householdnbAXA +
        gender + typeclassTPL + bonusevol2), family = binomial("logit"),
    data = idata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3426  -0.4098  -0.3465  -0.2769   3.0779

Coefficients:
```

```
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.4688261  0.1096307 -13.398  < 2e-16 ***
diff2tech                      7.8703759  1.4738254   5.340 9.29e-08 ***
diff2top10vip                 -1.2285174  0.3513642  -3.496 0.000472 ***
product2eco                   -0.3178553  0.0351763  -9.036  < 2e-16 ***
product2VIP                   -0.6104159  0.0429973 -14.197  < 2e-16 ***
region2_02-04-11               0.2551451  0.0429930   5.935 2.95e-09 ***
region2_05                     0.1901793  0.0278148   6.837 8.07e-12 ***
region2_08-09                  0.0556719  0.0260395   2.138 0.032518 *
region2_10                     0.4514134  0.0907537   4.974 6.56e-07 ***
region2_12-13                  0.3451752  0.0407160   8.478  < 2e-16 ***
region2_14-15-16               0.4415889  0.0374315  11.797  < 2e-16 ***
cumulrebate3                   0.1292981  0.0237389   5.447 5.13e-08 ***
nbclaim0608percust             0.2430882  0.0863587   2.815 0.004880 **
isinsuredinhealth             -0.2055506  0.0738987  -2.782 0.005411 **
isinsuredinlife               -0.0921472  0.0404841  -2.276 0.022838 *
vehiclage                     -0.0377862  0.0025184 -15.004  < 2e-16 ***
householdnbAXA                -0.1597084  0.0348957  -4.577 4.72e-06 ***
polholderage                  -0.0142757  0.0008121 -17.578  < 2e-16 ***
maritalstatus2b               -0.2667296  0.0760361  -3.508 0.000452 ***
maritalstatus2d               -0.1017938  0.0340552  -2.989 0.002798 **
jobgroup2public               -0.1189075  0.0215773  -5.511 3.57e-08 ***
gender                        -0.8174742  0.1755621  -4.656 3.22e-06 ***
typeclassTPL                  -0.0940907  0.0324866  -2.896 0.003776 **
bonusevol2up-down              3.5677414  0.6106271   5.843 5.13e-09 ***
diff2tech:priceratio          -8.3070757  1.4777530  -5.621 1.89e-08 ***
diff2top10vip:priceratio       1.2629656  0.3803163   3.321 0.000898 ***
priceratio:diff2top10direct   -0.8715317  0.1895395  -4.598 4.26e-06 ***
priceratio:paymentfreq        -0.0344649  0.0026126 -13.192  < 2e-16 ***
priceratio:nbclaim08percust   -0.1013398  0.0384519  -2.635 0.008401 **
nbclaim0608percust:priceratio -0.2291441  0.0915974  -2.502 0.012362 *
priceratio:nbclaim0708percust  0.1556257  0.0367502   4.235 2.29e-05 ***
priceratio:isinsuredinaccident -0.1484186  0.0506585  -2.930 0.003392 **
householdnbAXA:priceratio      0.0817327  0.0348671   2.344 0.019072 *
gender:priceratio              0.7452469  0.1764711   4.223 2.41e-05 ***
typeclassTPL:priceratio        0.1238598  0.0324918   3.812 0.000138 ***
bonusevol2up-down:priceratio  -3.3975682  0.6090299  -5.579 2.42e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 91145  on 188123  degrees of freedom
Residual deviance: 88147  on 188088  degrees of freedom
  (8175 observations deleted due to missingness)
AIC: 88219

Number of Fisher Scoring iterations: 6
```

PC direct channel

```
Call: glm(formula = lapse ~ region2 + nbclaim08percust + nbclaim0708percust +
    vehiclage + cumulrebate2 + polholderage + jobgroup2 + polage +
    typeclassPC + priceratio:(diff2tech + paymentfreq + directdebit +
    nbclaim08percust + nbclaim0608percust + nbclaim0708percust +
    gender), family = binomial("logit"), data = idata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6072  -0.5554  -0.4753  -0.3818   2.6485

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.256199   0.160158  -7.844 4.38e-15 ***
region2_02-06                 -0.226159   0.054341  -4.162 3.16e-05 ***
region2_07-08-09-10-11         0.122042   0.040287   3.029 0.002451 **
region2_12-13-15               0.256615   0.070120   3.660 0.000253 ***
region2_14-16                  0.492771   0.065816   7.487 7.04e-14 ***
nbclaim08percust              -1.987397   0.524222  -3.791 0.000150 ***
nbclaim0708percust             1.447482   0.450107   3.216 0.001301 **
vehiclage                     -0.022334   0.004833  -4.622 3.81e-06 ***
cumulrebate2_10+              -0.125790   0.060858  -2.067 0.038739 *
polholderage                  -0.007928   0.001498  -5.291 1.22e-07 ***
jobgroup2public               -0.132496   0.035014  -3.784 0.000154 ***
polage                        -0.067280   0.005880 -11.442  < 2e-16 ***
typeclassPC                    0.020726   0.005134   4.037 5.42e-05 ***
priceratio:diff2tech          -2.424138   0.352605  -6.875 6.20e-12 ***
priceratio:paymentfreq        -0.027927   0.005840  -4.782 1.74e-06 ***
priceratio:directdebit        -0.154461   0.051364  -3.007 0.002637 **
nbclaim08percust:priceratio    1.966404   0.517571   3.799 0.000145 ***
priceratio:nbclaim0608percust  0.138503   0.046313   2.991 0.002784 **
nbclaim0708percust:priceratio -1.454186   0.453706  -3.205 0.001350 **
priceratio:gender             -0.108971   0.035543  -3.066 0.002170 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24733  on 33355  degrees of freedom
Residual deviance: 24045  on 33336  degrees of freedom
  (851 observations deleted due to missingness)
```

```
AIC: 24085

Number of Fisher Scoring iterations: 5
```

PC broker channel

```
Call: glm(formula = lapse ~ lastprem_group2 + diff2tech + paymentfreq +
    region2 + cumulrebate2 + vehiclage + householdnbAXA + polholderage +
    diffdriverPH + jobgroup2 + polage + bonusevol2 + priceratio:(diff2tech +
    paymentfreq + directdebit + region2 + nbclaim0608percust +
    typeclassPC + bonusevol2), family = binomial("logit"), data = idata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7358  -0.5273  -0.4469  -0.3613   2.9528

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -2.326e+00  1.016e-01 -22.900  < 2e-16 ***
lastprem_group2(500,5e+03]       1.333e-01  2.975e-02   4.480 7.47e-06 ***
diff2tech                        8.930e+00  2.130e+00   4.193 2.76e-05 ***
paymentfreq                      7.122e-02  3.073e-02   2.318  0.02046 *
region2_03-04-06-07-10          -2.363e+00  5.185e-01  -4.557 5.20e-06 ***
region2_05-09                   -2.504e+00  4.583e-01  -5.464 4.64e-08 ***
region2_08                      -2.728e+00  5.482e-01  -4.977 6.46e-07 ***
region2_11-12-13-16             -2.644e+00  5.531e-01  -4.781 1.75e-06 ***
region2_14-15                   -3.040e+00  6.034e-01  -5.039 4.69e-07 ***
cumulrebate2_10-20              -3.320e-01  2.421e-02 -13.717  < 2e-16 ***
cumulrebate2_25+                -5.431e-01  5.135e-02 -10.576  < 2e-16 ***
vehiclage                       -1.949e-02  2.905e-03  -6.711 1.93e-11 ***
householdnbAXA                  -6.688e-02  8.602e-03  -7.775 7.57e-15 ***
polholderage                    -7.051e-03  8.457e-04  -8.338  < 2e-16 ***
diffdriverPHall drivers > 24     4.492e-01  5.680e-02   7.908 2.61e-15 ***
diffdriverPHcommercial           4.694e-01  1.432e-01   3.278  0.00105 **
diffdriverPHlearner 17           7.936e-01  1.806e-01   4.395 1.11e-05 ***
diffdriverPHonly partner         5.178e-01  5.040e-02  10.274  < 2e-16 ***
diffdriverPHsame                 4.713e-01  5.185e-02   9.090  < 2e-16 ***
diffdriverPHyoung drivers        6.743e-01  6.079e-02  11.091  < 2e-16 ***
jobgroup2public                 -1.646e-01  2.642e-02  -6.229 4.68e-10 ***
polage                          -2.124e-02  3.718e-03  -5.714 1.11e-08 ***
bonusevol2up-down                2.854e+00  4.495e-01   6.350 2.15e-10 ***
diff2tech:priceratio            -1.061e+01  2.051e+00  -5.170 2.34e-07 ***
paymentfreq:priceratio          -8.886e-02  3.047e-02  -2.916  0.00355 **
priceratio:directdebit           6.062e-02  2.806e-02   2.161  0.03072 *
region2_03-04-06-07-10:priceratio  2.538e+00  5.164e-01   4.915 8.86e-07 ***
region2_05-09:priceratio         2.861e+00  4.563e-01   6.271 3.59e-10 ***
region2_08:priceratio            2.892e+00  5.468e-01   5.289 1.23e-07 ***
region2_11-12-13-16:priceratio   3.208e+00  5.504e-01   5.829 5.57e-09 ***
region2_14-15:priceratio         3.777e+00  6.013e-01   6.282 3.34e-10 ***
priceratio:nbclaim0608percust    7.081e-02  1.306e-02   5.424 5.83e-08 ***
priceratio:typeclassPC           9.804e-03  3.172e-03   3.090  0.00200 **
bonusevol2up-down:priceratio    -2.709e+00  4.398e-01  -6.160 7.27e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 61944  on 88899  degrees of freedom
Residual deviance: 59864  on 88866  degrees of freedom
  (9532 observations deleted due to missingness)
AIC: 59932

Number of Fisher Scoring iterations: 5
```

PC agent channel

```
Call: glm(formula = lapse ~ lastprem_group + region2 + cumulrebate2 +
    nbclaim0608percust + isinsuredinaccident + housepol + vehiclage +
    householdnbAXA + polholderage + maritalstatus2 + diffdriverPH7 +
    jobgroup2 + gender + polage + typeclassPC + priceratio:(diff2tech +
    paymentfreq + isinsuredinlife + housepol + bonusevol), family = binomial("logit"),
    data = idata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2313  -0.4054  -0.3307  -0.2402   3.2781

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.4834163  0.0808352 -18.351  < 2e-16 ***
lastprem_group(1e+03,5e+03]      0.2756085  0.0573667   4.804 1.55e-06 ***
lastprem_group(500,1e+03]        0.1230814  0.0199186   6.179 6.44e-10 ***
region2_02-04-11                 0.3028182  0.0308358   9.820  < 2e-16 ***
region2_05                       0.1759933  0.0173652  10.135  < 2e-16 ***
region2_10                       0.2878711  0.0677199   4.251 2.13e-05 ***
```

```
region2_12-13                       0.4521249  0.0307907  14.684  < 2e-16 ***
region2_14-15-16                    0.4607204  0.0271870  16.946  < 2e-16 ***
cumulrebate2_25+                   -0.2349028  0.0440478  -5.333 9.67e-08 ***
cumulrebate2_5-20                  -0.0636038  0.0147217  -4.320 1.56e-05 ***
nbclaim0608percust                  0.0738475  0.0073792  10.007  < 2e-16 ***
isinsuredinaccident                -0.0675953  0.0292559  -2.310 0.020861 *
housepolflat owner                 -2.1528313  0.3698894  -5.820 5.88e-09 ***
housepolhouse with axa             -2.2298853  0.3314990  -6.727 1.74e-11 ***
housepolno property                -1.7573732  0.2219465  -7.918 2.41e-15 ***
housepolnot with axa               -2.0179795  0.2676815  -7.539 4.75e-14 ***
vehiclage                          -0.0267052  0.0019401 -13.765  < 2e-16 ***
householdnbAXA                     -0.0788257  0.0032649 -24.143  < 2e-16 ***
polholderage                       -0.0108440  0.0005959 -18.199  < 2e-16 ***
maritalstatus2b                    -0.2156059  0.0513198  -4.201 2.65e-05 ***
maritalstatus2d                    -0.1452715  0.0219042  -6.632 3.31e-11 ***
diffdriverPH7learner 17             0.4722532  0.1273644   3.708 0.000209 ***
diffdriverPH7only partner           0.1028589  0.0155482   6.616 3.70e-11 ***
diffdriverPH7young drivers          0.1571513  0.0260498   6.033 1.61e-09 ***
jobgroup2public                    -0.1800172  0.0153234 -11.748  < 2e-16 ***
gender                             -0.0770005  0.0150856  -5.104 3.32e-07 ***
polage                             -0.0247578  0.0011942 -20.732  < 2e-16 ***
typeclassPC                         0.0135060  0.0021534   6.272 3.57e-10 ***
priceratio:diff2tech               -0.9813502  0.1699192  -5.775 7.68e-09 ***
priceratio:paymentfreq             -0.0150664  0.0019221  -7.839 4.56e-15 ***
priceratio:isinsuredinlife         -0.0562095  0.0250614  -2.243 0.024905 *
housepolflat owner:priceratio       2.0545687  0.3673471   5.593 2.23e-08 ***
housepolhouse with axa:priceratio   1.9719108  0.3287744   5.998 2.00e-09 ***
housepolno property:priceratio      1.6871427  0.2178579   7.744 9.62e-15 ***
housepolnot with axa:priceratio     1.9774989  0.2642932   7.482 7.31e-14 ***
priceratio:bonusevolstable         -0.1690751  0.0200085  -8.450  < 2e-16 ***
priceratio:bonusevolup             -0.5876312  0.0690713  -8.508  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 167680  on 365672  degrees of freedom
Residual deviance: 160477  on 365636  degrees of freedom
  (13700 observations deleted due to missingness)
AIC: 160551

Number of Fisher Scoring iterations: 6
```

FC direct channel

```
Call: glm(formula = lapse ~ region2 + householdnbAXA + gender + polage +
    priceratio:(paymentfreq + directdebit + diff2tech + nbclaim0608percust +
        polholderage + jobgroup2 + typeclassFC), family = binomial("logit"),
    data = idata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9999  -0.5914  -0.5087  -0.4097   2.5622

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -1.117554   0.145889  -7.660 1.86e-14 ***
region2_03-05-09              0.281153   0.046924   5.992 2.08e-09 ***
region2_07-08                 0.171349   0.055933   3.063 0.002188 **
region2_10-11-12              0.528628   0.066725   7.922 2.33e-15 ***
region2_13-14-15-16           0.615216   0.062533   9.838  < 2e-16 ***
householdnbAXA               -0.046267   0.018119  -2.553 0.010665 *
gender                       -0.152176   0.035834  -4.247 2.17e-05 ***
polage                       -0.067047   0.005869 -11.425  < 2e-16 ***
priceratio:paymentfreq       -0.021333   0.005695  -3.746 0.000180 ***
priceratio:directdebit       -0.155778   0.051573  -3.021 0.002523 **
priceratio:diff2tech         -2.909523   0.341014  -8.532  < 2e-16 ***
priceratio:nbclaim0608percust  0.059747   0.019089   3.130 0.001749 **
priceratio:polholderage      -0.013141   0.001412  -9.306  < 2e-16 ***
priceratio:jobgroup2public   -0.084517   0.034102  -2.478 0.013198 *
priceratio:typeclassFC        0.016423   0.005011   3.277 0.001048 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 25559  on 31726  degrees of freedom
Residual deviance: 24892  on 31712  degrees of freedom
  (841 observations deleted due to missingness)
AIC: 24922

Number of Fisher Scoring iterations: 5
```

FC broker channel

```
Call: glm(formula = lapse ~ lastprem_group + diff2tech + directdebit +
    region2 + cumulrebate2 + housepol + vehiclage + householdnbAXA +
```

```
    polholderage + typeclassFC + bonusevol2 + priceratio:(directdebit +
    diff2tech + nbclaim08percust + nbclaim0608percust + jobgroup2 +
    gender + typeclassFC + bonusevol2), family = binomial("logit"),
    data = idata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7490  -0.5337  -0.4521  -0.3701   2.9438

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -2.109e+00  8.517e-02 -24.756  < 2e-16 ***
lastprem_group(1e+03,5e+03]  2.667e-01  5.171e-02   5.157 2.51e-07 ***
lastprem_group(500,1e+03]    1.785e-01  2.277e-02   7.839 4.54e-15 ***
diff2tech                    9.930e+00  2.220e+00   4.472 7.74e-06 ***
directdebit                 -1.215e+00  3.377e-01  -3.597 0.000322 ***
region2_02-04-05-11          3.434e-01  6.114e-02   5.616 1.95e-08 ***
region2_03-09-10             3.288e-01  2.736e-02  12.017  < 2e-16 ***
region2_04-05-06-07          1.202e-01  2.678e-02   4.489 7.17e-06 ***
region2_12-13                4.811e-01  4.310e-02  11.163  < 2e-16 ***
region2_14-15-16             6.013e-01  3.415e-02  17.610  < 2e-16 ***
cumulrebate2_10-20          -3.653e-01  2.017e-02 -18.110  < 2e-16 ***
cumulrebate2_25-50          -5.021e-01  3.666e-02 -13.697  < 2e-16 ***
cumulrebate2_50+            -1.375e+00  1.956e-01  -7.030 2.06e-12 ***
housepolflat owner           5.525e-01  5.190e-02  10.646  < 2e-16 ***
housepolhouse with axa       4.016e-01  5.888e-02   6.820 9.11e-12 ***
housepolno property          5.024e-01  4.035e-02  12.451  < 2e-16 ***
housepolnot with axa         5.435e-01  4.030e-02  13.487  < 2e-16 ***
vehiclage                    9.570e-03  2.817e-03   3.397 0.000681 ***
householdnbAXA              -5.244e-02  6.344e-03  -8.266  < 2e-16 ***
polholderage                -9.749e-03  6.915e-04 -14.099  < 2e-16 ***
typeclassFC                 -6.511e-02  1.962e-02  -3.319 0.000904 ***
bonusevol2up-down            2.000e+00  3.145e-01   6.358 2.04e-10 ***
directdebit:priceratio       1.197e+00  3.351e-01   3.572 0.000354 ***
diff2tech:priceratio        -1.281e+01  2.171e+00  -5.899 3.65e-09 ***
priceratio:nbclaim08percust -5.405e-02  2.472e-02  -2.187 0.028745 *
priceratio:nbclaim0608percust  6.115e-02  1.208e-02   5.062 4.14e-07 ***
priceratio:jobgroup2public  -9.569e-02  2.051e-02  -4.664 3.10e-06 ***
priceratio:gender           -4.744e-02  1.955e-02  -2.427 0.015241 *
typeclassFC:priceratio       6.733e-02  1.915e-02   3.516 0.000438 ***
bonusevol2up-down:priceratio -1.979e+00  3.116e-01  -6.352 2.13e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 84866  on 119129  degrees of freedom
Residual deviance: 82034  on 119100  degrees of freedom
  (26555 observations deleted due to missingness)
AIC: 82094

Number of Fisher Scoring iterations: 5
```

FC agent channel

```
Call: glm(formula = lapse ~ lastprem_group + glasscover + region2 +
    cumulrebate3 + nbclaim08percust + nbclaim0608percust + householdnbAXA +
    polholderage + maritalstatus2 + jobgroup2 + gender + polage +
    priceratio:(diff2tech + directdebit + product2 + isinsuredinhealth +
        isinsuredinlife + isinsuredinaccident + householdnbAXA +
        diffdriverPH7 + bonusevol2), family = binomial("logit"),
    data = idata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8899  -0.3725  -0.2859  -0.2099   3.2505

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.2444367  0.0609177 -20.428  < 2e-16 ***
lastprem_group(1e+03,5e+03]      0.3142438  0.0419218   7.496 6.58e-14 ***
lastprem_group(500,1e+03]        0.2030701  0.0165221  12.291  < 2e-16 ***
glasscover                      -0.1239641  0.0203622  -6.088 1.14e-09 ***
region2_06-07-08-09             -0.1455102  0.0154264  -9.433  < 2e-16 ***
region2_10-11                    0.2345939  0.0367294   6.387 1.69e-10 ***
region2_12-13                    0.2600399  0.0295366   8.804  < 2e-16 ***
region2_14-15-16                 0.2843734  0.0260932  10.898  < 2e-16 ***
cumulrebate3                    -0.0330928  0.0161422  -2.050 0.040356 *
nbclaim08percust                -0.0465684  0.0158363  -2.941 0.003276 **
nbclaim0608percust               0.0661664  0.0072763   9.093  < 2e-16 ***
householdnbAXA                  -0.1798880  0.0228308  -7.879 3.30e-15 ***
polholderage                    -0.0132590  0.0005775 -22.959  < 2e-16 ***
maritalstatus2b                 -0.1961024  0.0510359  -3.842 0.000122 ***
maritalstatus2d                 -0.1116925  0.0194238  -5.750 8.91e-09 ***
jobgroup2public                 -0.1639196  0.0143147 -11.451  < 2e-16 ***
gender                          -0.0525636  0.0147925  -3.553 0.000380 ***
polage                          -0.0231433  0.0009947 -23.268  < 2e-16 ***
priceratio:diff2tech            -1.4944324  0.1437874 -10.393  < 2e-16 ***
priceratio:directdebit          -0.0599526  0.0176559  -3.396 0.000685 ***
priceratio:product2eco          -0.0955695  0.0449387  -2.127 0.033448 *
priceratio:product2VIP          -0.4084230  0.0460922  -8.861  < 2e-16 ***
```

```
priceratio:isinsuredinhealth          -0.0746706  0.0320412  -2.330 0.019782 *
priceratio:isinsuredinlife            -0.0852991  0.0223092  -3.823 0.000132 ***
priceratio:isinsuredinaccident        -0.0836531  0.0247572  -3.379 0.000728 ***
householdnbAXA:priceratio              0.1004204  0.0228780   4.389 1.14e-05 ***
priceratio:diffdriverPH7learner 17     0.2381382  0.1133038   2.102 0.035574 *
priceratio:diffdriverPH7only partner   0.1015512  0.0147733   6.874 6.24e-12 ***
priceratio:diffdriverPH7young drivers  0.0873415  0.0314806   2.774 0.005529 **
priceratio:bonusevol2up-down           0.0500753  0.0165456   3.027 0.002474 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 180209  on 445742  degrees of freedom
Residual deviance: 170558  on 445713  degrees of freedom
  (78983 observations deleted due to missingness)
AIC: 170618

Number of Fisher Scoring iterations: 6
```

Backfit

Num.	Region	Observed	Fitted	Observed	Fitted	Observed	Fitted
1	Schleswig-H.	5.01	5.18	8.15	8.36	11.02	10.5
2	Hamburg	5.63	6.56	8.07	9.49	11.56	10.2
3	Niedersachsen	5.23	4.99	11.43	12.9	13.61	14.19
4	Bremen	5.21	5.06	9.68	10.68	11.52	10.4
5	Nordrhein-W.	5.36	5.11	10.25	10.51	14.12	13.86
6	Hessen	4.49	4.83	9.80	10.77	10.00	10.51
7	Rheinland-P.	4.58	4.14	10.74	10.22	13.39	12.96
8	Baden-W.	4.24	4.23	8.89	8.68	13.03	13.24
9	Bayern	4.90	4.45	12.32	12.4	14.7	14.78
10	Saarland	7.78	6.99	12.91	12.65	17.41	17.89
11	Berlin	7.59	7.53	13.03	13.12	15.27	15.36
12	Brandenburg	8.16	7.64	14.70	14.92	16.90	16.76
13	Mecklenburg-V.	8.02	7.88	15.30	15.44	19.22	19.37
14	Sachsen-A.	8.14	8.13	18.15	17.49	19.85	19.79
15	Sachsen	8.26	8.34	16.05	17.91	20.65	19.92
16	Thüringen	9.97	8.09	15.27	17.12	19.6	19.95
		Agent channel		Broker channel		Direct channel	

Table B.30: Lapse rates by region for FC cover

Information asymmetry

Claim number for FC agent channel

```
Call: glm(formula = nbclaim08FC ~ vehiclage + polholderage + diffdriverPH +
    jobgroup + gender + polage + typeclassFC + bonusevol, family = poisson("log"),
    data = idata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.1291  -0.5794  -0.4979  -0.4077  16.5866

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.3220232  0.0272575 -48.501  < 2e-16 ***
vehiclage                   -0.0359920  0.0011359 -31.685  < 2e-16 ***
polholderage                -0.0157506  0.0002576 -61.149  < 2e-16 ***
diffdriverPHall drivers > 24 -0.5343152  0.0154902 -34.494  < 2e-16 ***
```

```
diffdriverPHcommercial       -1.3454677  0.0178036 -75.573  < 2e-16 ***
diffdriverPHlearner 17       -0.3986107  0.0601646  -6.625 3.46e-11 ***
diffdriverPHonly partner     -0.7549764  0.0139778 -54.013  < 2e-16 ***
diffdriverPHsame             -0.9286748  0.0156502 -59.339  < 2e-16 ***
diffdriverPHyoung drivers    -0.3830581  0.0183989 -20.820  < 2e-16 ***
jobgroupnormal                0.1212390  0.0075581  16.041  < 2e-16 ***
jobgrouppublic                0.5913743  0.0198723  29.759  < 2e-16 ***
gender                        0.2166531  0.0026268  82.478  < 2e-16 ***
polage                       -0.0085524  0.0004025 -21.246  < 2e-16 ***
typeclassFC                   0.0350851  0.0009224  38.035  < 2e-16 ***
bonusevolstable               0.0407246  0.0084191   4.837 1.32e-06 ***
bonusevolup                   1.0170884  0.0138727  73.316  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 396684  on 524725  degrees of freedom
Residual deviance: 355448  on 524710  degrees of freedom
AIC: 500798

Number of Fisher Scoring iterations: 6
```

Deductible choice for FC agent channel

```
Logistic Regression Model

lrm(formula = deductibleFC3 ~ nbclaim08FC + ClaimNBhat + vehiclage +
    polholderage + polage + typeclassFC, data = idata, method = "lrm.fit",
    se.fit = TRUE)

Frequencies of Responses
     0    150    300    500
  6147  33000 408532  77047

       Obs  Max Deriv Model L.R.       d.f.          P          C
    524726      5e-09    30423.6          6          0       0.65
       Dxy      Gamma      Tau-a         R2      Brier
     0.299      0.305       0.11      0.075      0.011

             Coef     S.E.      Wald Z  P
y>=150        3.41899 0.0252019  135.66 0
y>=300        1.46877 0.0222785   65.93 0
y>=500       -3.07040 0.0226560 -135.52 0
nbclaim08FC   0.07929 0.0044965   17.63 0
ClaimNBhat   -0.26989 0.0239767  -11.26 0
vehiclage     0.03794 0.0009884   38.38 0
polholderage -0.01183 0.0002231  -53.03 0
polage       -0.03250 0.0003362  -96.68 0
typeclassFC   0.10546 0.0009525  110.72 0
```

Claim number for FC broker channel

```
Call: glm(formula = nbclaim08FC ~ isinsuredinaccident + housepol2 +
    vehiclage + polholderage + gender + typeclassFC + bonusevol,
    family = poisson("log"), data = idata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9287  -0.5187  -0.4582  -0.4022  16.8296

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.8498687  0.0447814 -41.309  < 2e-16 ***
isinsuredinaccident    0.2066575  0.0435667   4.743 2.10e-06 ***
housepol2flat owner   -0.2798096  0.0357390  -7.829 4.91e-15 ***
housepol2no property  -0.4191739  0.0210313 -19.931  < 2e-16 ***
housepol2not with axa -0.2670818  0.0198762 -13.437  < 2e-16 ***
vehiclage             -0.0394027  0.0022163 -17.779  < 2e-16 ***
polholderage          -0.0093380  0.0004227 -22.090  < 2e-16 ***
gender                 0.0691520  0.0053671  12.884  < 2e-16 ***
typeclassFC            0.0297784  0.0018224  16.340  < 2e-16 ***
bonusevolstable       -0.1680848  0.0160898 -10.447  < 2e-16 ***
bonusevolup            0.9357764  0.0258422  36.211  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 94040  on 145684  degrees of freedom
Residual deviance: 87067  on 145674  degrees of freedom
AIC: 122626

Number of Fisher Scoring iterations: 6
```

Deductible choice for FC broker channel

```
Logistic Regression Model

lrm(formula = deductibleFC3 ~ nbclaim08FC + ClaimNBhat + vehiclage +
    polholderage + maritalstatus + gender + polage + typeclassFC +
    bonusevol, data = idata, method = "lrm.fit", se.fit = TRUE)

Frequencies of Responses
     0    150    300    500
  1060   5318 101164  23302

Frequencies of Missing Values Due to Each Variable
deductibleFC3   nbclaim08FC    ClaimNBhat     vehiclage  polholderage
         2467             0             0             0             0
maritalstatus        gender        polage   typeclassFC     bonusevol
        12899             0             0             0             0

       Obs  Max Deriv Model L.R.       d.f.          P          C
    130844      1e-10     4016.1         10          0      0.611
       Dxy      Gamma      Tau-a         R2      Brier
     0.222       0.23      0.082      0.041      0.008

                 Coef      S.E.      Wald Z P
y>=150            4.547938 0.0663771  68.52 0.0000
y>=300            2.699288 0.0600989  44.91 0.0000
y>=500           -1.939587 0.0592427 -32.74 0.0000
nbclaim08FC      -0.034567 0.0184862  -1.87 0.0615
ClaimNBhat       -5.541685 0.2479716 -22.35 0.0000
vehiclage         0.008217 0.0022353   3.68 0.0002
polholderage     -0.012881 0.0005667 -22.73 0.0000
maritalstatus    -0.074755 0.0078458  -9.53 0.0000
gender           -0.055136 0.0140376  -3.93 0.0001
polage           -0.025597 0.0013145 -19.47 0.0000
typeclassFC       0.110312 0.0021592  51.09 0.0000
bonusevol=stable -0.344774 0.0160222 -21.52 0.0000
bonusevol=up      1.086692 0.0637438  17.05 0.0000
```

Lapse regression for FC agent channel

```
Call: glm(formula = lapse ~ lastprem_group + glasscover + region2 +
    cumulrebate3 + nbclaim08percust + nbclaim0608percust + householdnbAXA +
    polholderage + maritalstatus2 + jobgroup2 + gender + polage +
    deductibleFC3.0 + deductibleFC3.150 + deductibleFC3.500. +
    priceratio:(diff2tech + directdebit + product2 + isinsuredinhealth +
        isinsuredinlife + isinsuredinaccident + householdnbAXA +
        diffdriverPH7 + bonusevol2), family = binomial("logit"),
    data = idata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8387  -0.3729  -0.2861  -0.2095   3.2668

Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -7.842e-01  1.338e-01  -5.862 4.57e-09 ***
lastprem_group(1e+03,5e+03]         3.306e-01  4.336e-02   7.625 2.45e-14 ***
lastprem_group(500,1e+03]           2.029e-01  1.725e-02  11.763  < 2e-16 ***
glasscover                         -1.238e-01  2.036e-02  -6.080 1.20e-09 ***
region2_06-07-08-09                -1.457e-01  1.543e-02  -9.444  < 2e-16 ***
region2_10-11                       2.340e-01  3.673e-02   6.369 1.90e-10 ***
region2_12-13                       2.580e-01  2.955e-02   8.733  < 2e-16 ***
region2_14-15-16                    2.826e-01  2.612e-02  10.820  < 2e-16 ***
cumulrebate3                       -3.098e-02  1.619e-02  -1.914 0.055671 .
nbclaim08percust                   -4.511e-02  1.589e-02  -2.838 0.004538 **
nbclaim0608percust                  6.547e-02  7.290e-03   8.981  < 2e-16 ***
householdnbAXA                     -1.797e-01  2.285e-02  -7.865 3.68e-15 ***
polholderage                       -1.286e-02  6.118e-04 -21.015  < 2e-16 ***
maritalstatus2b                    -1.942e-01  5.105e-02  -3.804 0.000143 ***
maritalstatus2d                    -1.110e-01  1.943e-02  -5.713 1.11e-08 ***
jobgroup2public                    -1.648e-01  1.432e-02 -11.507  < 2e-16 ***
gender                             -5.086e-02  1.504e-02  -3.382 0.000719 ***
polage                             -2.043e-02  1.480e-03 -13.803  < 2e-16 ***
deductibleFC3.0                     5.662e+01  1.812e+01   3.125 0.001778 **
deductibleFC3.150                  -1.528e+01  4.415e+00  -3.462 0.000537 ***
deductibleFC3.500.                 -1.407e+00  3.722e-01  -3.780 0.000157 ***
priceratio:diff2tech               -1.491e+00  1.444e-01 -10.324  < 2e-16 ***
priceratio:directdebit             -6.006e-02  1.767e-02  -3.398 0.000678 ***
priceratio:product2eco             -9.930e-02  4.496e-02  -2.209 0.027192 *
priceratio:product2VIP             -4.090e-01  4.613e-02  -8.867  < 2e-16 ***
priceratio:isinsuredinhealth       -7.388e-02  3.206e-02  -2.305 0.021191 *
priceratio:isinsuredinlife         -8.624e-02  2.232e-02  -3.864 0.000111 ***
priceratio:isinsuredinaccident     -8.427e-02  2.477e-02  -3.402 0.000669 ***
householdnbAXA:priceratio           1.000e-01  2.290e-02   4.368 1.25e-05 ***
priceratio:diffdriverPH7learner 17  2.364e-01  1.133e-01   2.086 0.036981 *
priceratio:diffdriverPH7only partner 9.961e-02  1.482e-02   6.722 1.79e-11 ***
```

```
priceratio:diffdriverPH7young drivers  8.670e-02  3.173e-02   2.732 0.006291 **
priceratio:bonusevol2up-down           5.064e-02  1.677e-02   3.019 0.002532 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 180209  on 445742  degrees of freedom
Residual deviance: 170541  on 445710  degrees of freedom
  (78983 observations deleted due to missingness)
AIC: 170607

Number of Fisher Scoring iterations: 6
```

Lapse regression for FC broker channel

```
Call: glm(formula = lapse ~ lastprem_group + diff2tech + directdebit +
    region2 + cumulrebate2 + housepol + householdnbAXA + polholderage +
    typeclassFC + bonusevol2 + deductibleFC3.0 + deductibleFC3.150 +
    priceratio:(directdebit + diff2tech + nbclaim08percust +
        nbclaim0608percust + jobgroup2 + typeclassFC + bonusevol2),
    family = binomial("logit"), data = idata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7735  -0.5354  -0.4528  -0.3666   2.9938

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -6.860e-01  1.698e-01  -4.041 5.33e-05 ***
lastprem_group(1e+03,5e+03]   3.246e-01  5.236e-02   6.199 5.67e-10 ***
lastprem_group(500,1e+03]     1.865e-01  2.294e-02   8.132 4.22e-16 ***
diff2tech                     9.418e+00  2.226e+00   4.230 2.33e-05 ***
directdebit                  -1.183e+00  3.391e-01  -3.488 0.000486 ***
region2_02-04-05-11           3.256e-01  6.145e-02   5.298 1.17e-07 ***
region2_03-09-10              3.025e-01  2.761e-02  10.954  < 2e-16 ***
region2_04-05-06-07           1.002e-01  2.701e-02   3.711 0.000207 ***
region2_12-13                 4.593e-01  4.339e-02  10.585  < 2e-16 ***
region2_14-15-16              5.717e-01  3.437e-02  16.633  < 2e-16 ***
cumulrebate2_10-20           -3.727e-01  2.027e-02 -18.384  < 2e-16 ***
cumulrebate2_25-50           -5.010e-01  3.703e-02 -13.531  < 2e-16 ***
cumulrebate2_50+             -1.397e+00  1.961e-01  -7.123 1.05e-12 ***
housepolflat owner            5.240e-01  5.238e-02  10.005  < 2e-16 ***
housepolhouse with axa        4.104e-01  5.941e-02   6.907 4.96e-12 ***
housepolno property           4.541e-01  4.094e-02  11.094  < 2e-16 ***
housepolnot with axa          4.995e-01  4.088e-02  12.217  < 2e-16 ***
householdnbAXA               -3.277e-02  6.515e-03  -5.030 4.91e-07 ***
polholderage                 -5.265e-03  7.808e-04  -6.743 1.55e-11 ***
typeclassFC                  -1.143e-01  2.041e-02  -5.602 2.12e-08 ***
bonusevol2up-down             1.709e+00  3.169e-01   5.392 6.97e-08 ***
deductibleFC3.0               1.702e+02  5.687e+01   2.994 0.002758 **
deductibleFC3.150            -5.224e+01  1.297e+01  -4.027 5.65e-05 ***
directdebit:priceratio        1.154e+00  3.364e-01   3.431 0.000602 ***
diff2tech:priceratio         -1.222e+01  2.178e+00  -5.613 1.99e-08 ***
priceratio:nbclaim08percust  -5.638e-02  2.488e-02  -2.266 0.023472 *
priceratio:nbclaim0608percust  6.327e-02  1.217e-02   5.200 2.00e-07 ***
priceratio:jobgroup2public   -9.732e-02  2.049e-02  -4.749 2.04e-06 ***
typeclassFC:priceratio        6.677e-02  1.934e-02   3.453 0.000554 ***
bonusevol2up-down:priceratio -1.674e+00  3.142e-01  -5.329 9.90e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


    Null deviance: 83782  on 117481  degrees of freedom
Residual deviance: 80832  on 117452  degrees of freedom
  (28203 observations deleted due to missingness)
AIC: 80892

Number of Fisher Scoring iterations: 6
```

B.3 GAM analyses

B.3.1 GAM analysis for Portugal data

Here follows the summaries of the full GAM regression when no terms are crossed with the price ratio.

```
Family: binomial - Link function: logit

Formula:
did_lapse ~ s(premium_before) + s(priceratio) + s(age) + s(age_policy) + s(age_vehicle)

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.520535   0.003586    -424   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                    edf Ref.df Chi.sq p-value
s(premium_before) 8.939  8.999 4718.4  <2e-16 ***
s(priceratio)     8.324  8.873 1030.4  <2e-16 ***
s(age)            8.230  8.806 3013.9  <2e-16 ***
s(age_policy)     8.288  8.800  409.4  <2e-16 ***
s(age_vehicle)    7.528  8.212 4384.8  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0251   Deviance explained =  2.5%
REML score = 2.6205e+05  Scale est. = 1         n = 557693
```

Here follows the summaries of the full GAM regression when all terms are crossed with the price ratio.

```
Family: binomial - Link function: logit

Formula:
did_lapse ~ s(priceratio, premium_before) + s(priceratio, age)
+ s(priceratio, age_policy) + s(priceratio, age_vehicle)

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.528645   0.003612  -423.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                               edf Ref.df Chi.sq p-value
s(priceratio,premium_before) 28.31  28.96   7334  <2e-16 ***
s(priceratio,age)            25.08  27.36   3092  <2e-16 ***
s(priceratio,age_policy)     18.82  22.61    653  <2e-16 ***
s(priceratio,age_vehicle)    17.09  21.45   4370  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0279   Deviance explained = 2.85%
REML score = 2.6121e+05  Scale est. = 1         n = 557693
```

B.3.2 GAM analysis for QuÈbec data

2007 dataset

Here follows the summaries of the full GAM regression when no terms are crossed with the price ratio.

```
Family: binomial - Link function: logit
```

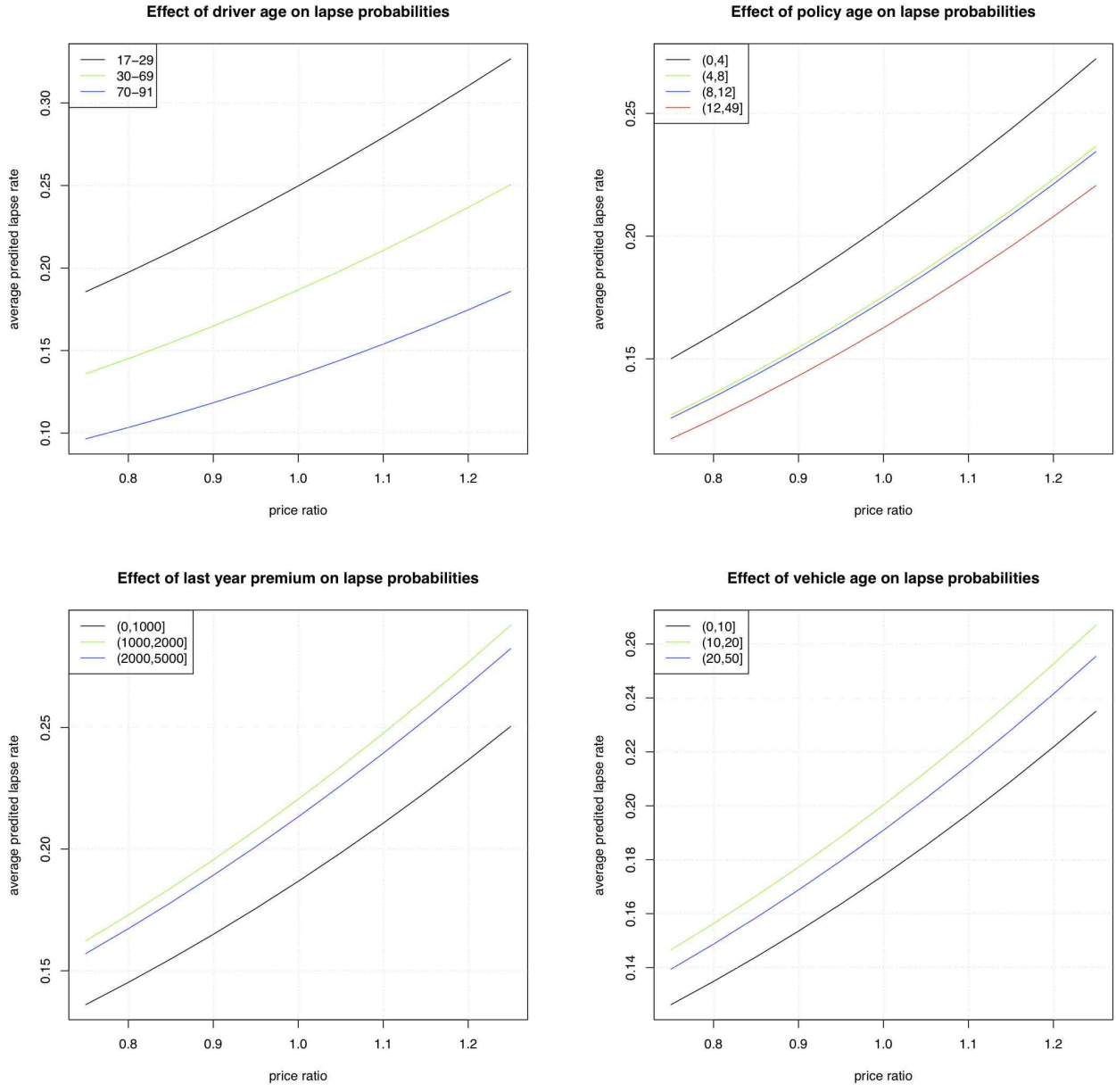


Figure B.11: One-variable effect on lapse

```
Formula:
did_cancel ~ house_pol + multi_veh_dsc + cover + gender + claim_1 +
    s(pricefactor) + s(pol_age) + s(veh_age) + s(price_group) + s(driv_age) + s(prev_prem)

Parametric coefficients:
               Estimate Std. Error  z value Pr(>|z|)
(Intercept)    -2.36015    0.01917 -123.105  < 2e-16 ***
house_polY     -0.87995    0.02049  -42.940  < 2e-16 ***
multi_veh_dscY -0.19401    0.01966   -9.869  < 2e-16 ***
coverTPL        0.28178    0.03954    7.127 1.02e-12 ***
coverTPL+opt   -0.01491    0.02429   -0.614 0.539288
genderM        -0.06270    0.01775   -3.533 0.000411 ***
claim_1        -0.13724    0.02628   -5.221 1.78e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                 edf Ref.df Chi.sq  p-value
s(pricefactor) 7.134  8.189 533.37  < 2e-16 ***
s(pol_age)     7.635  8.492 375.90  < 2e-16 ***
s(veh_age)     7.180  7.794 326.20  < 2e-16 ***
s(price_group) 2.948  3.777  34.20 5.11e-07 ***
s(driv_age)    8.132  8.678 191.03  < 2e-16 ***
s(prev_prem)   6.427  7.416 609.76  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0259   Deviance explained = 5.08%
REML score =  55607  Scale est. = 1         n = 238673
```

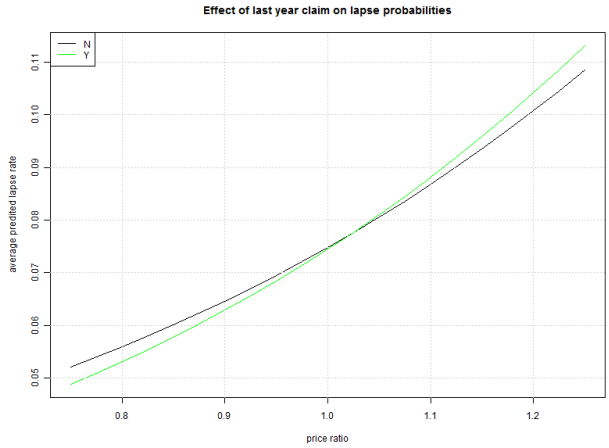
Here follows the summaries of the full GAM regression when all terms are crossed with the price ratio.

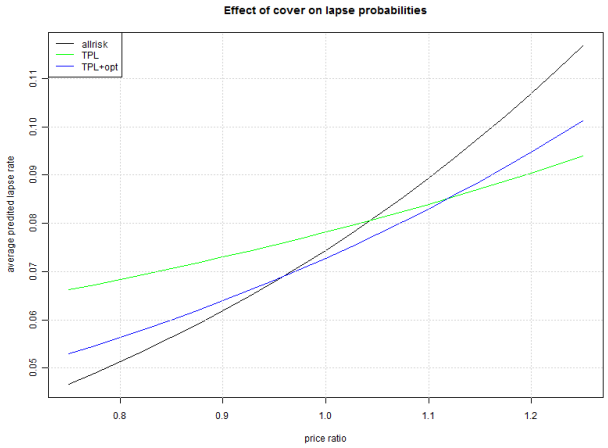
```
Family: binomial - Link function: logit

Formula:
did_cancel ~ claim_1 + house_pol + pricefactor:(multi_veh_dsc +
    cover) + s(pricefactor, prev_prem) + s(pricefactor, veh_age) +
    s(pricefactor, pol_age) + s(pricefactor, driv_age)

Parametric coefficients:
                              Estimate Std. Error  z value Pr(>|z|)
(Intercept)                  -2.49732    0.02088 -119.625  < 2e-16 ***
claim_1                      -0.11614    0.02737   -4.243 2.20e-05 ***
house_polY                   -0.88009    0.02201  -39.990  < 2e-16 ***
pricefactor:multi_veh_dscN    0.21060    0.02096   10.045  < 2e-16 ***
pricefactor:multi_veh_dscY    0.00000    0.00000       NA       NA
pricefactor:coverTPL          0.13864    0.04164    3.330  0.00087 ***
pricefactor:coverTPL+opt     -0.10306    0.02564   -4.020 5.83e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                           edf Ref.df Chi.sq p-value
s(pricefactor,prev_prem) 14.60  18.48  714.2  <2e-16 ***
s(pricefactor,veh_age)   21.68  25.31  455.1  <2e-16 ***
s(pricefactor,pol_age)   10.40  12.88  299.4  <2e-16 ***
s(pricefactor,driv_age)  17.09  21.31  219.3  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0257   Deviance explained = 4.91%
REML score =  49486  Scale est. = 1         n = 202919
```

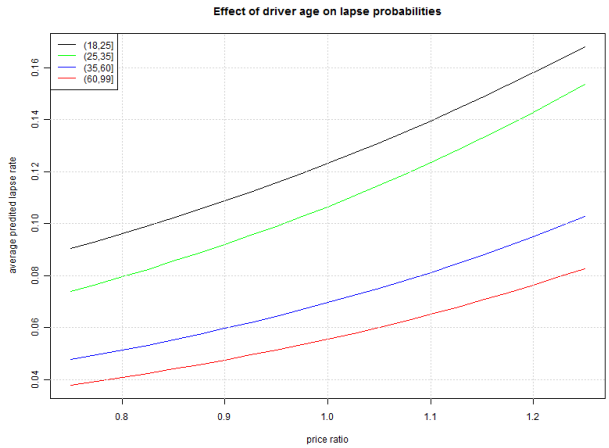
2006 dataset

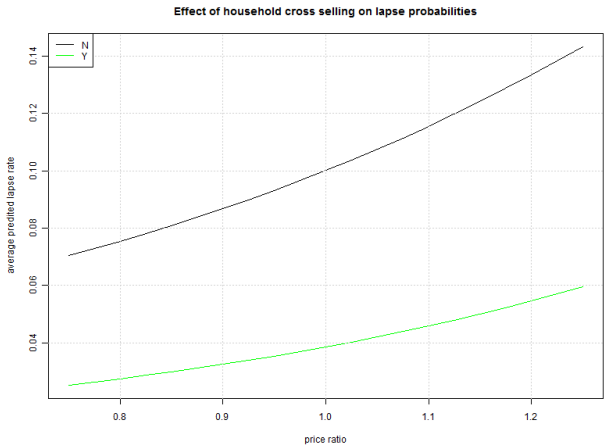
Here follows the summaries of the full GAM regression when no terms are crossed with the price ratio.

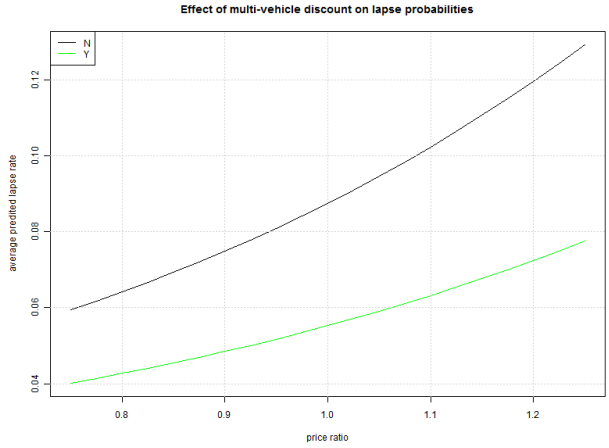
```
Family: binomial - Link function: logit

Formula:
did_cancel ~ house_pol + multi_veh_dsc + cover + gender + claim_1 +
    claim_2 + s(pricefactor) + s(pol_age) + s(veh_age) + s(price_group) +
    s(driv_age) + s(prev_prem)

Parametric coefficients:
                Estimate Std. Error  z value Pr(>|z|)
(Intercept)    -2.2129297  0.0195052 -113.453  < 2e-16 ***
```

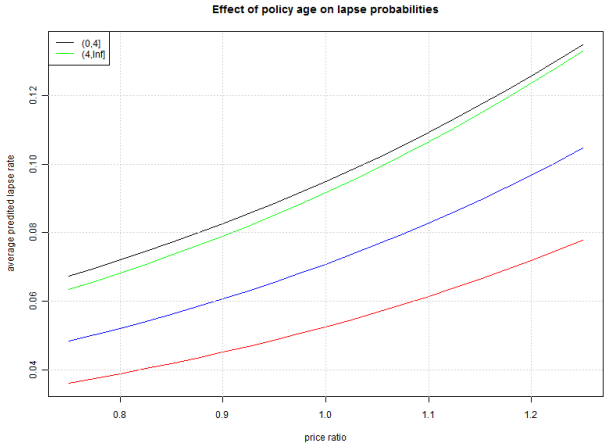

Effect of last year claim on lapse probabilities
N
Y
average predited lapse rate
price ratio

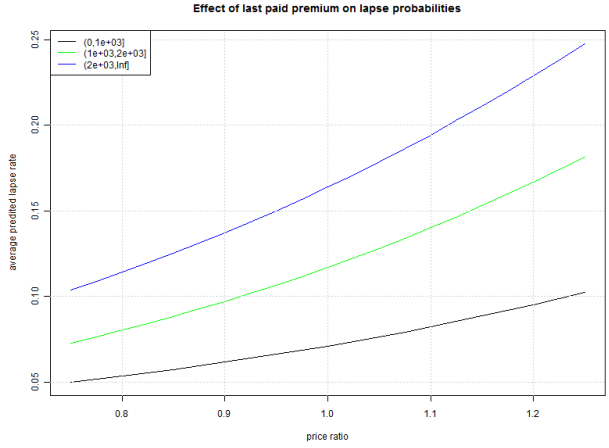

Effect of cover on lapse probabilities
allrisk
TPL
TPL+opt
average predited lapse rate
price ratio

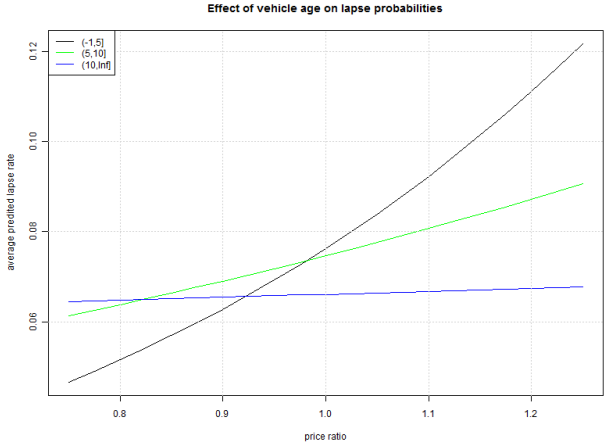

Effect of driver age on lapse probabilities
(18,25]
(25,35]
(35,60]
(60,99]
average predited lapse rate
price ratio


Effect of household cross selling on lapse probabilities
N
Y
average predited lapse rate
price ratio


Effect of multi-vehicle discount on lapse probabilities
N
Y
average predited lapse rate
price ratio


Effect of policy age on lapse probabilities
(0,4]
(4,Inf]
average predited lapse rate
price ratio


Effect of last paid premium on lapse probabilities
(0,1e+03]
(1e+03,2e+03]
(2e+03,Inf]
average predited lapse rate
price ratio


Effect of vehicle age on lapse probabilities
(-1,5]
(5,10]
(10,Inf]
average predited lapse rate
price ratio

```
house_polY     -0.8817779  0.0206811  -42.637  < 2e-16 ***
multi_veh_dscY -0.1139312  0.0196460   -5.799 6.66e-09 ***
coverTPL        0.3410253  0.0396561    8.600  < 2e-16 ***
coverTPL+opt    0.0362405  0.0235205    1.541 0.123365
genderM        -0.0743876  0.0177287   -4.196 2.72e-05 ***
claim_1        -0.0838275  0.0252417   -3.321 0.000897 ***
claim_2        -0.0006465  0.0245648   -0.026 0.979005
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                edf Ref.df Chi.sq  p-value
s(pricefactor) 6.705  7.838 606.05  < 2e-16 ***
s(pol_age)     7.980  8.730 419.12  < 2e-16 ***
s(veh_age)     6.860  7.657 275.06  < 2e-16 ***
s(price_group) 1.009  1.018  30.95 2.77e-08 ***
s(driv_age)    8.455  8.903 160.17  < 2e-16 ***
s(prev_prem)   6.716  7.644 657.99  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0289   Deviance explained = 5.02%
REML score =  53847  Scale est. = 1         n = 200662
```

Here follows the summaries of the full GAM regression when all terms are crossed with the price ratio.

```
Family: binomial - Link function: logit

Formula:
did_cancel ~ pricefactor * (gender + claim_1 + house_pol + claim_2 +
    multi_veh_dsc + cover) + s(pricefactor, prev_prem) + s(pricefactor, veh_age) +
    s(pricefactor, pol_age) + s(pricefactor, driv_age)

Parametric coefficients:
                           Estimate Std. Error  z value Pr(>|z|)
(Intercept)                -2.18845    0.01908 -114.716  < 2e-16 ***
pricefactor                 0.00000    0.00000       NA       NA
genderM                    -0.29946    0.18942   -1.581 0.113883
claim_1                     0.77562    0.24741    3.135 0.001719 **
house_polY                 -1.26136    0.23984   -5.259 1.45e-07 ***
claim_2                     0.07491    0.23737    0.316 0.752308
multi_veh_dscY              0.53759    0.20822    2.582 0.009827 **
coverTPL                    2.62170    0.24699   10.614  < 2e-16 ***
coverTPL+opt                0.79179    0.23055    3.434 0.000594 ***
pricefactor:genderM         0.22830    0.19582    1.166 0.243649
pricefactor:claim_1        -0.88217    0.25363   -3.478 0.000505 ***
pricefactor:house_polY      0.39052    0.24660    1.584 0.113278
pricefactor:claim_2        -0.08125    0.24707   -0.329 0.742272
pricefactor:multi_veh_dscY -0.68485    0.21473   -3.189 0.001426 **
pricefactor:coverTPL       -2.47206    0.25893   -9.547  < 2e-16 ***
pricefactor:coverTPL+opt   -0.83183    0.23833   -3.490 0.000483 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                          edf Ref.df Chi.sq p-value
s(pricefactor,prev_prem) 15.73  19.73  883.4  <2e-16 ***
s(pricefactor,veh_age)   12.86  16.70  374.2  <2e-16 ***
s(pricefactor,pol_age)   15.30  18.83  428.1  <2e-16 ***
s(pricefactor,driv_age)  18.23  22.45  162.6  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0298   Deviance explained = 5.18%
REML score =  53792  Scale est. = 1         n = 200662
```

Here follows the final GAM regression when all terms are crossed with the price ratio.

```
Family: binomial - Link function: logit

Formula:
did_cancel ~ house_pol + prev_prem_group2 + pricefactor * (claim_1 + cover) +
pricefactor:multi_veh_dsc + veh_age_group3 + s(pol_age, pricefactor) + s(driv_age)

Parametric coefficients:
                                 Estimate Std. Error  z value Pr(>|z|)
(Intercept)                      -2.12457    0.01826 -116.323  < 2e-16 ***
house_polY                       -0.88011    0.02063  -42.656  < 2e-16 ***
prev_prem_group2(1e+03,2e+03]     0.37826    0.02873   13.167  < 2e-16 ***
prev_prem_group2(2e+03,Inf]       0.41728    0.08778    4.753    2e-06 ***
pricefactor                       0.00000    0.00000       NA       NA
claim_1                           0.81699    0.24764    3.299 0.000970 ***
coverTPL                          2.44312    0.23343   10.466  < 2e-16 ***
coverTPL+opt                      0.60688    0.22839    2.657 0.007878 **
```

```
veh_age_group3(5,Inf]          0.05281    0.02015    2.620 0.008783 **
pricefactor:claim_1           -0.87599    0.25387   -3.451 0.000559 ***
pricefactor:coverTPL          -2.81112    0.24341  -11.549  < 2e-16 ***
pricefactor:coverTPL+opt      -0.77036    0.23606   -3.263 0.001101 **
pricefactor:multi_veh_dscY    -0.21726    0.01950  -11.140  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                           edf Ref.df Chi.sq p-value
s(pol_age,pricefactor) 14.712 17.919  961.2  <2e-16 ***
s(driv_age)             8.493  8.915  284.3  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0255   Deviance explained = 4.46%
REML score =  54129  Scale est. = 1         n = 200662
```

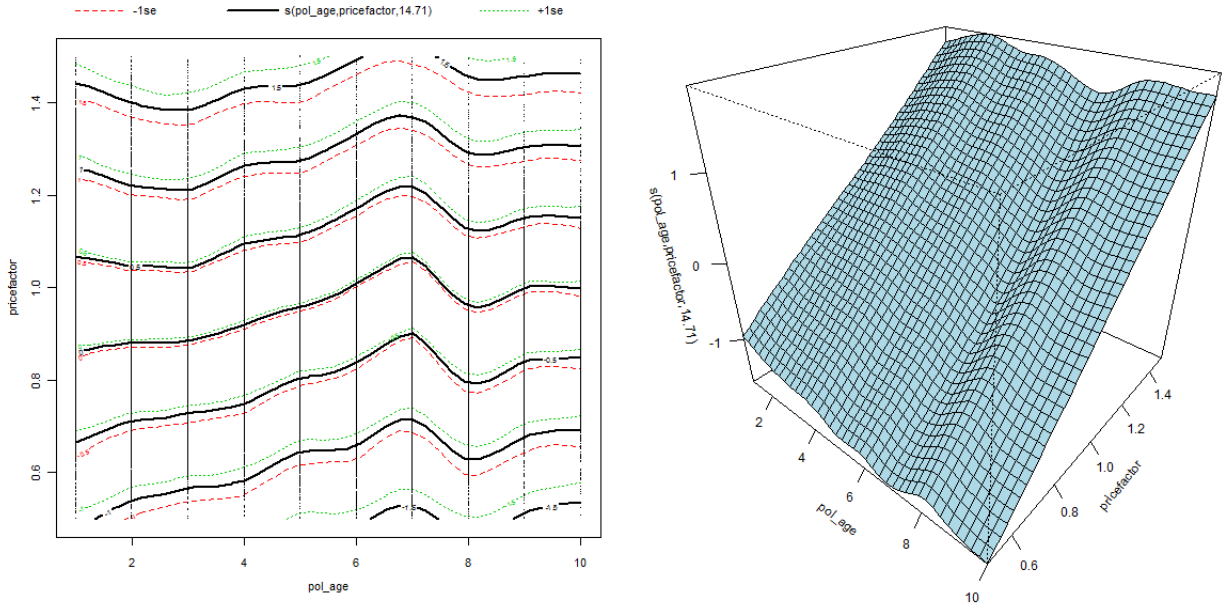


Figure B.13: Smooth function for the policy age

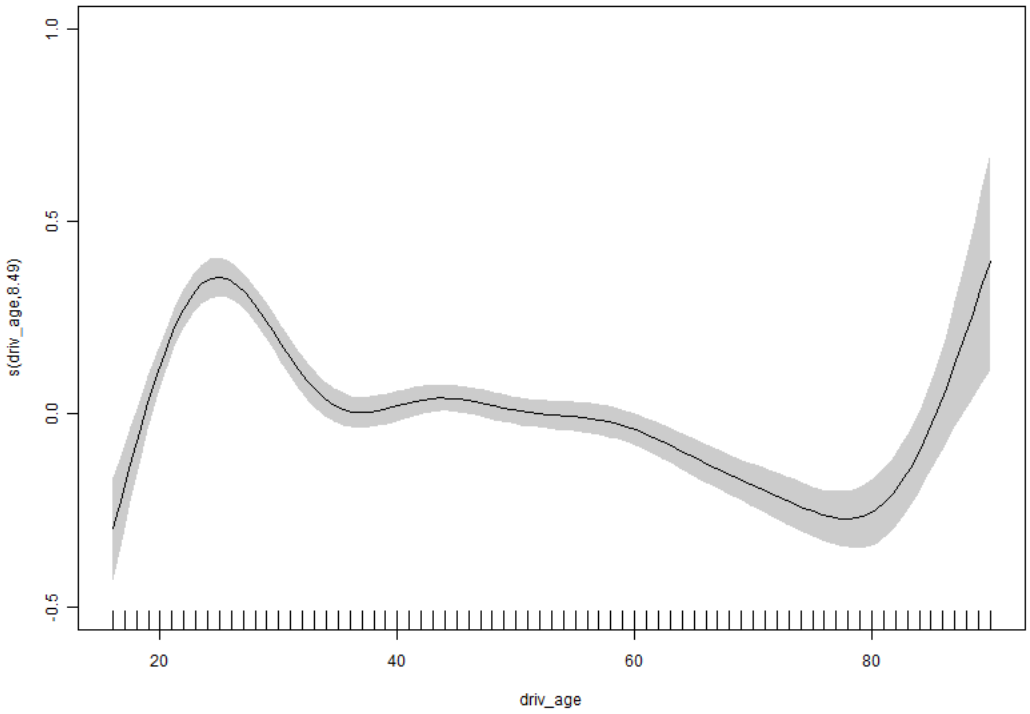
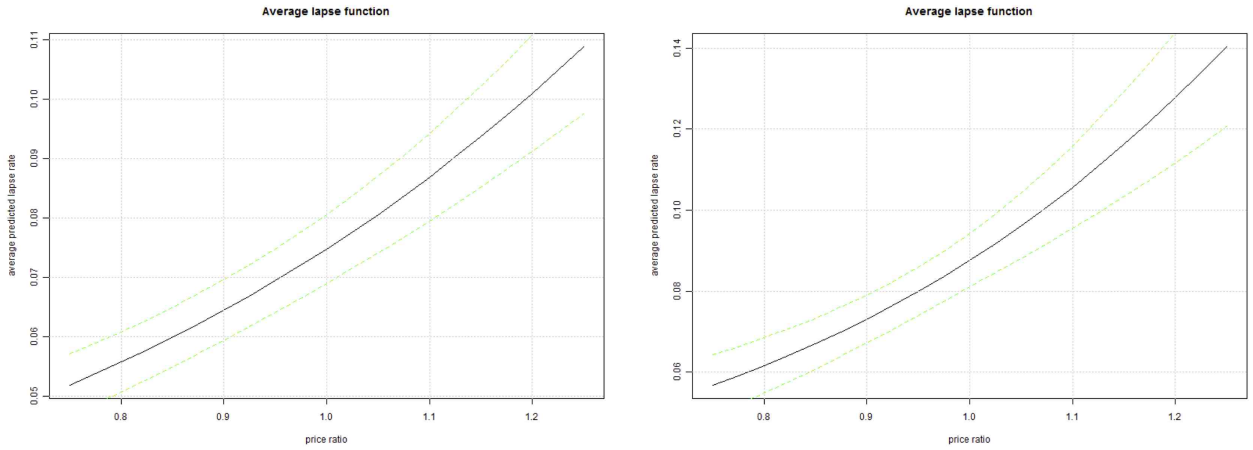
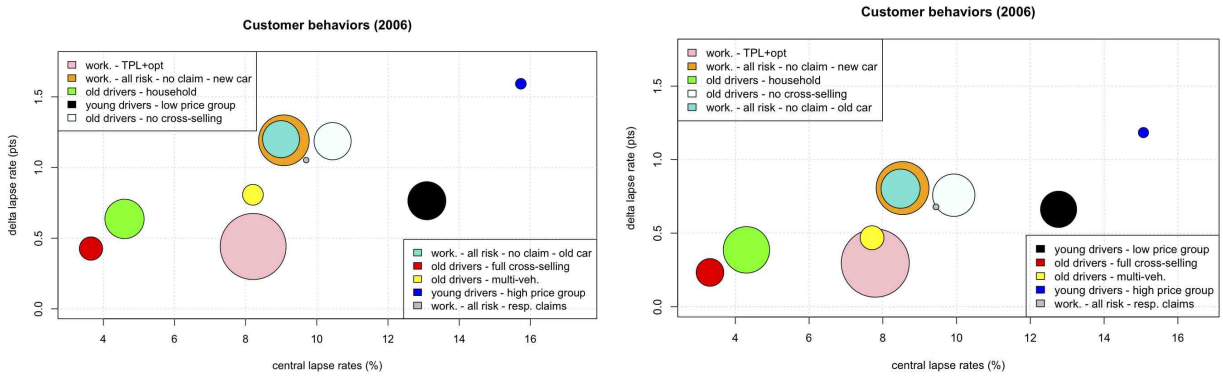


Figure B.14: Smooth function for the driver age



(a) 2007 (b) 2006

Figure B.15: 2007 vs. 2006 dataset



(a) GAM (b) GLM

Figure B.16: GAM vs. GLM

B.3.3 GAM analysis for Germany data

Below we list the regression summaries for the Germany data.

TPL direct channel

```
Family: binomial - Link function: logit

Formula: lapse ~ product2 + claimamount + cumulrebate2 + priceratio:(polholderage +
    diff2tech) + s(priceratio, k = 3) + s(diff2top10direct)

Parametric coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.489e+00  8.348e-02 -17.839  < 2e-16 ***
product2eco             -1.612e-01  5.045e-02  -3.196 0.001394 **
claimamount              4.061e-05  1.517e-05   2.677 0.007437 **
cumulrebate2_10+        -5.103e-01  1.358e-01  -3.758 0.000171 ***
priceratio:polholderage -1.148e-02  1.771e-03  -6.482 9.03e-11 ***
priceratio:diff2tech    -2.506e+00  3.626e-01  -6.912 4.76e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                     edf Ref.df Chi.sq p-value
s(priceratio)       1.771  1.948  4.331 0.10966
s(diff2top10direct) 2.593  3.325 12.863 0.00673 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.00784   Deviance explained = 1.09%
REML score = 8893.1  Scale est. = 1         n = 24194
```

TPL broker channel

```
Family: binomial - Link function: logit

Formula: lapse ~ isinsuredinhealth + gender + polage + cumulrebate2 +
    priceratio:(lastprem_group2 + paymentfreq + directdebit +
        isinsuredinhealth + householdnbAXA + polholderage) +
    s(vehiclage) + s(priceratio, diff2tech)

Parametric coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                          0.000000   0.000000      NA       NA
isinsuredinhealth                   -2.950964   1.274851  -2.315  0.02063 *
gender                              -0.087811   0.027580  -3.184  0.00145 **
polage                              -0.014217   0.005451  -2.608  0.00910 **
cumulrebate2_10-20                  -0.347806   0.030150 -11.536  < 2e-16 ***
cumulrebate2_25+                    -0.686823   0.070555  -9.735  < 2e-16 ***
priceratio:lastprem_group2(0,500]   -1.230270   0.074370 -16.543  < 2e-16 ***
priceratio:lastprem_group2(500,5e+03] -1.063100   0.078223 -13.591  < 2e-16 ***
priceratio:paymentfreq              -0.031337   0.003701  -8.468  < 2e-16 ***
priceratio:directdebit               0.064782   0.032448   1.996  0.04588 *
isinsuredinhealth:priceratio         2.656010   1.230556   2.158  0.03090 *
priceratio:householdnbAXA           -0.045711   0.011535  -3.963 7.40e-05 ***
priceratio:polholderage             -0.008972   0.001192  -7.525 5.28e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                       edf Ref.df Chi.sq p-value
s(vehiclage)         5.619  6.682  148.2  <2e-16 ***
s(priceratio,diff2tech) 19.134 23.879  612.5  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0187   Deviance explained = 2.46%
REML score =  21227  Scale est. = 1         n = 58645
```

TPL agent channel

```
Family: binomial - Link function: logit

Formula: lapse ~ product2 + region2 + cumulrebate3 + nbclaim0608percust +
```

```
    isinsuredinhealth + isinsuredinlife + vehiclage + householdnbAXA +
    polholderage + maritalstatus2 + jobgroup2 + gender + bonusevol2 +
    priceratio:(paymentfreq + nbclaim08percust + nbclaim0608percust +
        nbclaim0708percust + isinsuredinaccident + bonusevol2) +
    s(priceratio, diff2tech) + s(priceratio, diff2top10vip) +
    s(priceratio, diff2top10direct) + s(priceratio, typeclassTPL)

Parametric coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -0.9881832  0.0744176 -13.279  < 2e-16 ***
product2eco                  -0.2957239  0.0365839  -8.083 6.30e-16 ***
product2VIP                  -0.5888125  0.0439784 -13.389  < 2e-16 ***
region2_02-04-11              0.2474500  0.0432128   5.726 1.03e-08 ***
region2_05                    0.1820856  0.0279436   6.516 7.21e-11 ***
region2_08-09                 0.0627676  0.0260959   2.405 0.016161 *
region2_10                    0.4597820  0.0908178   5.063 4.13e-07 ***
region2_12-13                 0.3600178  0.0408722   8.808  < 2e-16 ***
region2_14-15-16              0.4440049  0.0377465  11.763  < 2e-16 ***
cumulrebate3                  0.1287561  0.0241245   5.337 9.44e-08 ***
nbclaim0608percust            0.2144964  0.0968126   2.216 0.026720 *
isinsuredinhealth            -0.2018414  0.0739308  -2.730 0.006331 **
isinsuredinlife              -0.0978298  0.0405763  -2.411 0.015908 *
vehiclage                    -0.0367641  0.0025963 -14.160  < 2e-16 ***
householdnbAXA               -0.0783881  0.0048668 -16.107  < 2e-16 ***
polholderage                 -0.0150938  0.0008334 -18.111  < 2e-16 ***
maritalstatus2b              -0.2629597  0.0760885  -3.456 0.000548 ***
maritalstatus2d              -0.1017553  0.0341228  -2.982 0.002863 **
jobgroup2public              -0.1161175  0.0217312  -5.343 9.12e-08 ***
gender                       -0.0790535  0.0209269  -3.778 0.000158 ***
bonusevol2up-down             7.4827223  1.0625789   7.042 1.89e-12 ***
priceratio:paymentfreq       -0.0343715  0.0026481 -12.980  < 2e-16 ***
priceratio:nbclaim08percust  -0.0893319  0.0393116  -2.272 0.023062 *
nbclaim0608percust:priceratio -0.2010502  0.1016136  -1.979 0.047864 *
priceratio:nbclaim0708percust  0.1538349  0.0369590   4.162 3.15e-05 ***
priceratio:isinsuredinaccident -0.1409923  0.0508941  -2.770 0.005600 **
bonusevol2up-down:priceratio  -7.2677291  1.0573222  -6.874 6.26e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                                edf Ref.df Chi.sq  p-value
s(priceratio,diff2tech)      12.440 16.687 113.56  < 2e-16 ***
s(priceratio,diff2top10vip)   8.901 12.069  29.36  0.00361 **
s(priceratio,diff2top10direct)  8.177 11.277  18.63  0.07569 .
s(priceratio,typeclassTPL)    4.160  5.687  43.91 5.43e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0176   Deviance explained = 3.46%
REML score =  44028  Scale est. = 1         n = 187733
```

PC direct channel

```
Family: binomial - Link function: logit

Formula: lapse ~ region2 + nbclaim08percust + nbclaim0708percust + cumulrebate2 +
    polholderage + jobgroup2 + polage + typeclassPC + priceratio:(paymentfreq +
    directdebit + vehiclage + nbclaim08percust + nbclaim0608percust +
    nbclaim0708percust + gender) + s(priceratio, diff2tech, k = 5)

Parametric coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -1.181041   0.161795  -7.300 2.89e-13 ***
region2_02-06              -0.223266   0.054336  -4.109 3.97e-05 ***
region2_07-08-09-10-11      0.122193   0.040309   3.031 0.002434 **
region2_12-13-15            0.258495   0.070136   3.686 0.000228 ***
region2_14-16               0.493010   0.065834   7.489 6.96e-14 ***
nbclaim08percust           -2.334435   0.554587  -4.209 2.56e-05 ***
nbclaim0708percust          1.592221   0.467221   3.408 0.000655 ***
cumulrebate2_10+           -0.126755   0.060875  -2.082 0.037324 *
polholderage               -0.008126   0.001512  -5.373 7.75e-08 ***
jobgroup2public            -0.134604   0.035090  -3.836 0.000125 ***
polage                     -0.067032   0.005893 -11.375  < 2e-16 ***
typeclassPC                 0.021070   0.005159   4.084 4.42e-05 ***
priceratio:paymentfreq     -0.028385   0.005844  -4.858 1.19e-06 ***
priceratio:directdebit     -0.158882   0.051641  -3.077 0.002093 **
priceratio:vehiclage       -0.022113   0.004819  -4.589 4.45e-06 ***
nbclaim08percust:priceratio  2.293884   0.545987   4.201 2.65e-05 ***
priceratio:nbclaim0608percust  0.139346   0.046146   3.020 0.002530 **
nbclaim0708percust:priceratio -1.601264   0.470395  -3.404 0.000664 ***
priceratio:gender          -0.107735   0.036464  -2.955 0.003131 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                       edf Ref.df Chi.sq  p-value
s(priceratio,diff2tech) 2.097  2.190  48.23 4.78e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
R-sq.(adj) =  0.0211   Deviance explained = 2.81%
REML score =  12068  Scale est. = 1         n = 33329
```

PC broker channel

```
Family: binomial - Link function: logit

Formula: lapse ~ lastprem_group2 + paymentfreq + cumulrebate2 + householdnbAXA +
    polholderage + diffdriverPH + jobgroup2 + bonusevol2 + priceratio:(paymentfreq +
    directdebit + region2 + nbclaim0608percust + bonusevol2 +
    typeclassPC + polage) + s(priceratio, diff2tech) + s(priceratio,
    vehiclage)

Parametric coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -2.3686039  0.0983407 -24.086  < 2e-16 ***
lastprem_group2(500,5e+03]      0.1604532  0.0308136   5.207 1.92e-07 ***
paymentfreq                     0.0817162  0.0322749   2.532  0.01135 *
cumulrebate2_10-20             -0.3331866  0.0242329 -13.749  < 2e-16 ***
cumulrebate2_25+               -0.5479358  0.0514444 -10.651  < 2e-16 ***
householdnbAXA                 -0.0664575  0.0086113  -7.718 1.19e-14 ***
polholderage                   -0.0070960  0.0008486  -8.362  < 2e-16 ***
diffdriverPHall drivers > 24    0.3999880  0.0579515   6.902 5.12e-12 ***
diffdriverPHcommercial          0.4229071  0.1437853   2.941  0.00327 **
diffdriverPHlearner 17          0.7418775  0.1810437   4.098 4.17e-05 ***
diffdriverPHonly partner        0.4703600  0.0517579   9.088  < 2e-16 ***
diffdriverPHsame                0.4253787  0.0530989   8.011 1.14e-15 ***
diffdriverPHyoung drivers       0.6337959  0.0620598  10.213  < 2e-16 ***
jobgroup2public                -0.1626077  0.0264531  -6.147 7.90e-10 ***
bonusevol2up-down               3.9282621  0.8495474   4.624 3.77e-06 ***
paymentfreq:priceratio         -0.0989592  0.0320204  -3.091  0.00200 **
priceratio:directdebit          0.0591160  0.0280854   2.105  0.03530 *
priceratio:region2_03-04-06-07-10  0.1722009  0.0540630   3.185  0.00145 **
priceratio:region2_05-09        0.3507387  0.0510888   6.865 6.64e-12 ***
priceratio:region2_08           0.1719197  0.0557036   3.086  0.00203 **
priceratio:region2_11-12-13-16  0.5598971  0.0573334   9.766  < 2e-16 ***
priceratio:region2_14-15        0.7452055  0.0601616  12.387  < 2e-16 ***
priceratio:nbclaim0608percust   0.0682477  0.0131804   5.178 2.24e-07 ***
bonusevol2up-down:priceratio   -3.7270938  0.8380135  -4.448 8.69e-06 ***
priceratio:typeclassPC          0.0062400  0.0032487   1.921  0.05476 .
priceratio:polage              -0.0217244  0.0037257  -5.831 5.51e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                          edf Ref.df Chi.sq  p-value
s(priceratio,diff2tech) 14.74 19.374 255.63  < 2e-16 ***
s(priceratio,vehiclage)  4.81  6.482  62.48 2.57e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.025   Deviance explained = 3.48%
REML score =  29987  Scale est. = 1         n = 88874
```

PC agent channel

```
Family: binomial - Link function: logit

Formula: lapse ~ lastprem_group + region2 + cumulrebate2 + nbclaim0608percust +
    isinsuredinaccident + housepol2 + vehiclage + householdnbAXA +
    polholderage + maritalstatus2 + diffdriverPH7 + jobgroup2 +
    gender + polage + typeclassPC + priceratio:(paymentfreq +
    isinsuredinlife + bonusevol + diff2tech) + s(priceratio)

Parametric coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.3190466  0.0769756 -17.136  < 2e-16 ***
lastprem_group(1e+03,5e+03]      0.3425350  0.0602800   5.682 1.33e-08 ***
lastprem_group(500,1e+03]        0.1450698  0.0202595   7.161 8.03e-13 ***
region2_02-04-11                 0.2975915  0.0308691   9.640  < 2e-16 ***
region2_05                       0.1692593  0.0174239   9.714  < 2e-16 ***
region2_10                       0.2911690  0.0677203   4.300 1.71e-05 ***
region2_12-13                    0.4461087  0.0308384  14.466  < 2e-16 ***
region2_14-15-16                 0.4604619  0.0272014  16.928  < 2e-16 ***
cumulrebate2_25+                -0.2349860  0.0439285  -5.349 8.83e-08 ***
cumulrebate2_5-20               -0.0653231  0.0147267  -4.436 9.18e-06 ***
nbclaim0608percust               0.0718586  0.0074045   9.705  < 2e-16 ***
isinsuredinaccident             -0.0681740  0.0292625  -2.330 0.019820 *
housepol2flat owner             -0.0628338  0.0292950  -2.145 0.031963 *
housepol2house with axa         -0.2178244  0.0258309  -8.433  < 2e-16 ***
housepol2no property            -0.0331257  0.0170782  -1.940 0.052423 .
vehiclage                       -0.0265574  0.0019453 -13.652  < 2e-16 ***
householdnbAXA                  -0.0787359  0.0032673 -24.098  < 2e-16 ***
```

```
polholderage                -0.0111607  0.0006061 -18.413  < 2e-16 ***
maritalstatus2b             -0.2184142  0.0513733  -4.252 2.12e-05 ***
maritalstatus2d             -0.1466518  0.0219140  -6.692 2.20e-11 ***
diffdriverPH7learner 17      0.4642937  0.1273389   3.646 0.000266 ***
diffdriverPH7only partner    0.0994980  0.0154306   6.448 1.13e-10 ***
diffdriverPH7young drivers   0.1579198  0.0259588   6.083 1.18e-09 ***
jobgroup2public             -0.1799238  0.0153295 -11.737  < 2e-16 ***
gender                      -0.0762593  0.0151031  -5.049 4.44e-07 ***
polage                      -0.0241535  0.0011833 -20.411  < 2e-16 ***
typeclassPC                  0.0110029  0.0022055   4.989 6.07e-07 ***
priceratio:paymentfreq      -0.0151536  0.0019261  -7.867 3.62e-15 ***
priceratio:isinsuredinlife  -0.0596722  0.0250983  -2.378 0.017429 *
priceratio:bonusevolstable  -0.3537506  0.0511614  -6.914 4.70e-12 ***
priceratio:bonusevolup      -1.4742191  0.1420716 -10.377  < 2e-16 ***
priceratio:diff2tech        -0.8063376  0.1718221  -4.693 2.69e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                edf Ref.df Chi.sq p-value
s(priceratio) 6.352  7.436  151.7  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.021   Deviance explained = 4.34%
REML score =  80233  Scale est. = 1         n = 365213
```

FC direct channel

```
Family: binomial - Link function: logit

Formula: lapse ~ region2 + polage + cumulrebate2 + householdnbAXA + typeclassFC +
    priceratio:(paymentfreq + diffdriverPH7 + jobgroup2 + gender) +
    s(priceratio, diff2tech) + s(priceratio, polholderage)

Parametric coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -1.925779   0.129231 -14.902  < 2e-16 ***
region2_03-05-09                    0.284929   0.046978   6.065 1.32e-09 ***
region2_07-08                       0.171518   0.055983   3.064  0.00219 **
region2_10-11-12                    0.529380   0.066802   7.925 2.29e-15 ***
region2_13-14-15-16                 0.614435   0.062596   9.816  < 2e-16 ***
polage                             -0.055197   0.005924  -9.318  < 2e-16 ***
cumulrebate2_15+                   -0.178099   0.102547  -1.737  0.08243 .
householdnbAXA                     -0.041041   0.018065  -2.272  0.02310 *
typeclassFC                         0.022004   0.005246   4.195 2.73e-05 ***
priceratio:paymentfreq             -0.024130   0.005683  -4.246 2.18e-05 ***
priceratio:diffdriverPH7young drivers  0.178967   0.073348   2.440  0.01469 *
priceratio:jobgroup2public         -0.078890   0.034651  -2.277  0.02280 *
priceratio:gender                  -0.171572   0.036450  -4.707 2.51e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                           edf Ref.df Chi.sq  p-value
s(priceratio,diff2tech)  5.062  7.122   77.5 5.22e-14 ***
s(priceratio,polholderage) 7.351  9.834  111.4  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0219   Deviance explained = 2.77%
REML score =  12483  Scale est. = 1         n = 31728
```

FC broker channel

```
Family: binomial - Link function: logit

Formula: lapse ~ lastprem_group + cumulrebate2 + paymentfreq + directdebit +
    region2 + nbclaim08percust + claimamount + isinsuredinlife +
    crossell + jobgroup2 + polage + typeclassFC + priceratio:(paymentfreq +
    directdebit + glasscover + nbclaim08percust + diffdriverPH) +
    s(priceratio, diff2tech) + s(priceratio, vehiclage) + s(priceratio,
    polholderage) + s(priceratio, typeclassFC)

Parametric coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                         0.000e+00  0.000e+00      NA       NA
lastprem_group(1e+03,5e+03]         2.597e-01  6.025e-02   4.310 1.64e-05 ***
lastprem_group(500,1e+03]           1.669e-01  2.506e-02   6.662 2.70e-11 ***
cumulrebate2_10-20                 -3.551e-01  2.199e-02 -16.146  < 2e-16 ***
cumulrebate2_25-50                 -4.734e-01  4.227e-02 -11.201  < 2e-16 ***
cumulrebate2_50+                   -1.579e+00  2.273e-01  -6.946 3.75e-12 ***
paymentfreq                         8.955e-02  4.533e-02   1.975 0.048216 *
```

```
directdebit                                -8.721e-01  4.066e-01  -2.145 0.031942 *
region2_02-04-05-11                         2.944e-01  6.533e-02   4.507 6.58e-06 ***
region2_03-09-10                            2.715e-01  2.937e-02   9.244  < 2e-16 ***
region2_04-05-06-07                         7.973e-02  2.879e-02   2.770 0.005613 **
region2_12-13                               4.319e-01  4.619e-02   9.351  < 2e-16 ***
region2_14-15-16                            5.175e-01  3.695e-02  14.004  < 2e-16 ***
nbclaim08percust                            7.657e-01  2.256e-01   3.395 0.000686 ***
claimamount                                 8.147e-06  4.592e-06   1.774 0.076029 .
isinsuredinlife                            -8.023e-02  4.456e-02  -1.801 0.071760 .
crossell                                   -1.785e-01  2.201e-02  -8.114 4.91e-16 ***
jobgroup2public                            -1.199e-01  2.230e-02  -5.379 7.50e-08 ***
polage                                     -2.386e-02  3.060e-03  -7.797 6.34e-15 ***
typeclassFC                                -1.168e-01  3.200e-03 -36.511  < 2e-16 ***
paymentfreq:priceratio                     -9.519e-02  4.506e-02  -2.112 0.034651 *
directdebit:priceratio                      8.670e-01  4.033e-01   2.150 0.031552 *
priceratio:glasscover                      -2.337e-01  5.172e-02  -4.519 6.21e-06 ***
nbclaim08percust:priceratio                -7.346e-01  2.192e-01  -3.351 0.000806 ***
priceratio:diffdriverPHall drivers > 24      1.643e-01  5.100e-02   3.222 0.001273 **
priceratio:diffdriverPHcommercial            1.966e-01  9.851e-02   1.995 0.045997 *
priceratio:diffdriverPHlearner 17            5.373e-01  1.427e-01   3.766 0.000166 ***
priceratio:diffdriverPHonly partner          3.090e-01  4.619e-02   6.691 2.22e-11 ***
priceratio:diffdriverPHsame                  2.143e-01  4.907e-02   4.368 1.26e-05 ***
priceratio:diffdriverPHyoung drivers         4.692e-01  6.267e-02   7.487 7.04e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                               edf Ref.df Chi.sq p-value
s(priceratio,diff2tech)     18.043 22.891 362.36  <2e-16 ***
s(priceratio,vehiclage)      6.965  9.263  99.98  <2e-16 ***
s(priceratio,polholderage) 10.754 13.949 166.35  <2e-16 ***
s(priceratio,typeclassFC)   7.824 10.431 787.37  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0269   Deviance explained = 3.75%
REML score =  35465  Scale est. = 1         n = 102955
```

FC agent channel

```
Family: binomial - Link function: logit

Formula: lapse ~ lastprem_group + region2 + nbclaim0608percust + claimamount +
    isinsuredinhealth + isinsuredinlife + isinsuredinaccident +
    householdnbAXA + polholderage + jobgroup2 + gender + polage +
    typeclassFC + cumulrebate3 + priceratio:(glasscover + diffdriverPH7 +
    nbclaim08percust + nbclaim0608percust + householdnbAXA +
    claimamount + vehiclage) + s(priceratio, diff2tech) + s(priceratio,
    polholderage) + s(priceratio, typeclassFC)

Parametric coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                              0.000e+00  0.000e+00      NA       NA
lastprem_group(1e+03,5e+03]              2.806e-01  4.488e-02   6.253 4.03e-10 ***
lastprem_group(500,1e+03]                1.481e-01  1.743e-02   8.494  < 2e-16 ***
region2_06-07-08-09                     -1.461e-01  1.538e-02  -9.499  < 2e-16 ***
region2_10-11                            2.604e-01  3.672e-02   7.090 1.34e-12 ***
region2_12-13                            2.700e-01  2.946e-02   9.163  < 2e-16 ***
region2_14-15-16                         2.941e-01  2.610e-02  11.266  < 2e-16 ***
nbclaim0608percust                      -2.187e-01  8.444e-02  -2.590 0.009597 **
claimamount                              8.512e-05  2.719e-05   3.130 0.001747 **
isinsuredinhealth                       -7.372e-02  3.176e-02  -2.321 0.020290 *
isinsuredinlife                         -1.071e-01  2.205e-02  -4.859 1.18e-06 ***
isinsuredinaccident                     -1.352e-01  2.380e-02  -5.679 1.36e-08 ***
householdnbAXA                          -1.782e-01  3.192e-02  -5.584 2.36e-08 ***
polholderage                            -2.129e-02  3.352e-03  -6.352 2.12e-10 ***
jobgroup2public                         -1.765e-01  1.427e-02 -12.365  < 2e-16 ***
gender                                  -6.032e-02  1.497e-02  -4.029 5.61e-05 ***
polage                                  -2.667e-02  9.904e-04 -26.932  < 2e-16 ***
typeclassFC                             -5.383e-02  1.003e-02  -5.368 7.98e-08 ***
cumulrebate3                            -1.607e-01  1.435e-02 -11.196  < 2e-16 ***
priceratio:glasscover                   -1.416e-01  2.023e-02  -7.002 2.52e-12 ***
priceratio:diffdriverPH7all drivers > 24 -1.029e-01  2.001e-02  -5.144 2.69e-07 ***
priceratio:nbclaim08percust             -4.524e-02  1.647e-02  -2.746 0.006030 **
nbclaim0608percust:priceratio            2.822e-01  8.415e-02   3.354 0.000797 ***
householdnbAXA:priceratio                8.816e-02  3.195e-02   2.760 0.005789 **
claimamount:priceratio                  -8.360e-05  2.661e-05  -3.142 0.001678 **
priceratio:vehiclage                     7.158e-03  2.233e-03   3.205 0.001351 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                              edf Ref.df Chi.sq  p-value
s(priceratio,diff2tech)    19.957 24.673 229.29  < 2e-16 ***
s(priceratio,polholderage)  8.392 11.411 105.73  < 2e-16 ***
s(priceratio,typeclassFC)   6.490  8.737  76.42 6.11e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
R-sq.(adj) =  0.0231   Deviance explained = 5.33%
REML score =  85970  Scale est. = 1         n = 450799
```

B.4 SRM analysis

B.4.1 AML example

Weibull

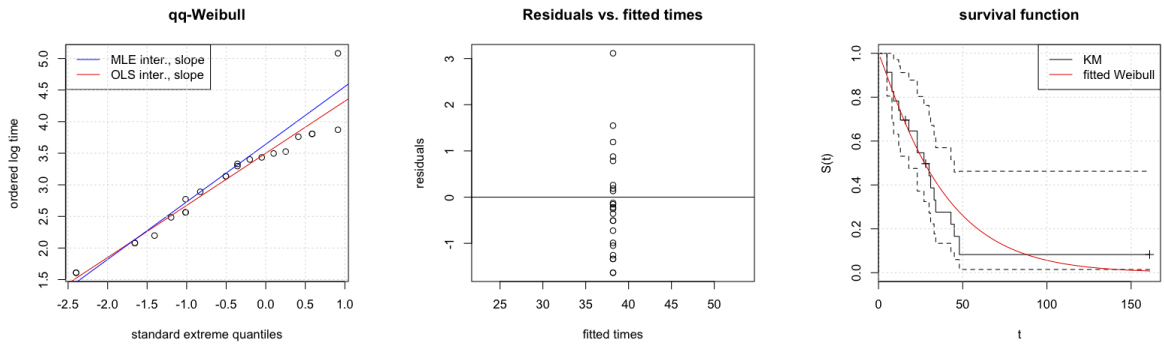


Figure B.17: Model assumptions check for Weibull distribution - intercept only

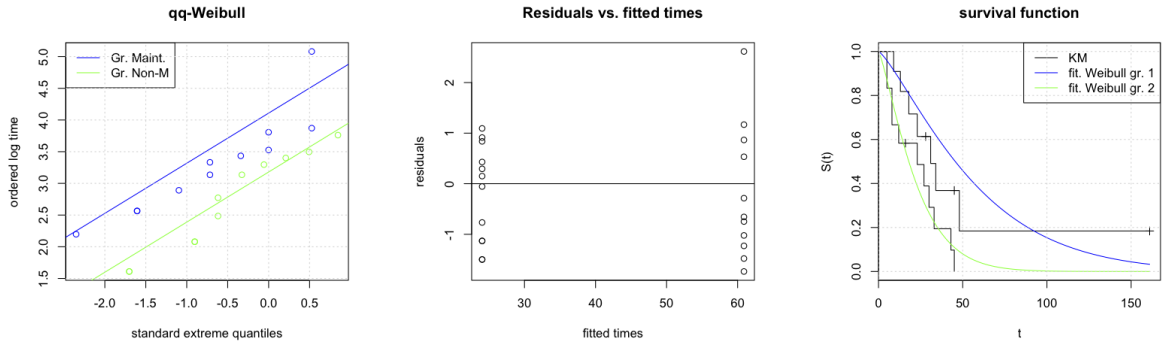


Figure B.18: Model assumptions check for Weibull distribution - with cov.

Lognormal

Loglogistic

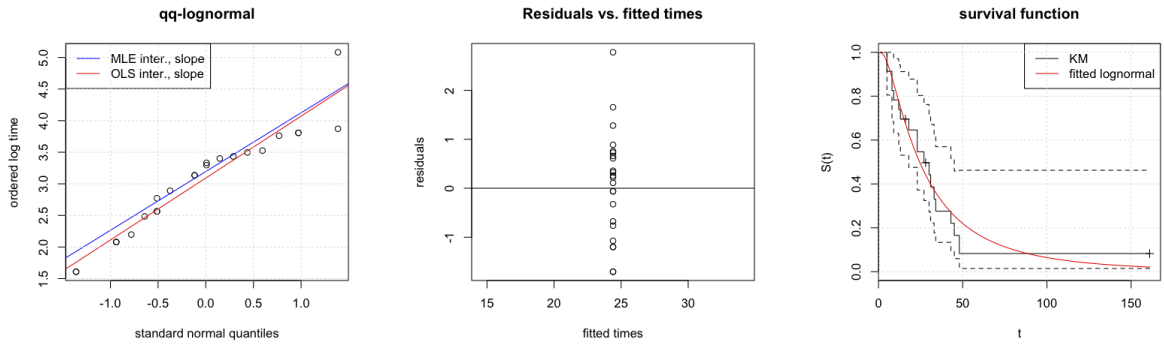


Figure B.19: Model assumptions check for lognormal distribution - intercept only

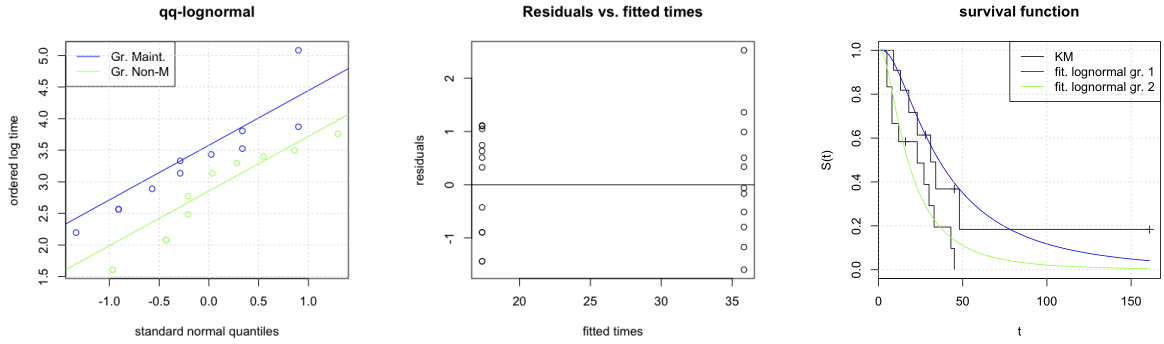


Figure B.20: Model assumptions check for lognormal distribution - with cov.

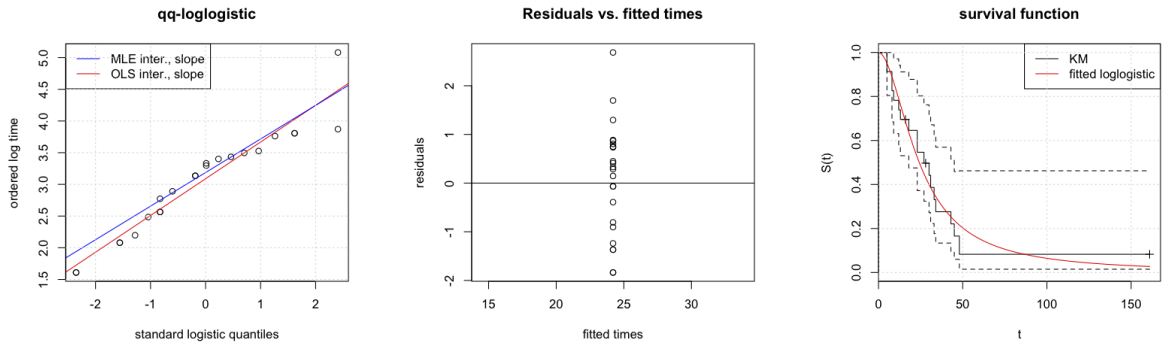


Figure B.21: Model assumptions check for loglogistic distribution - intercept only

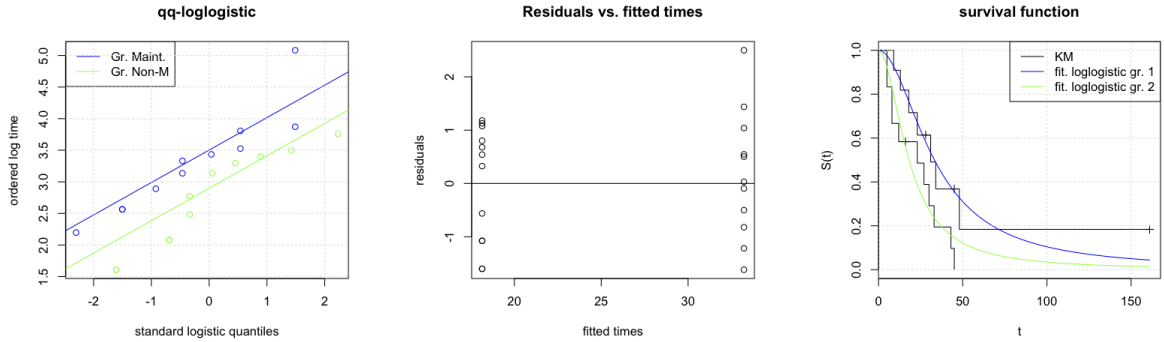
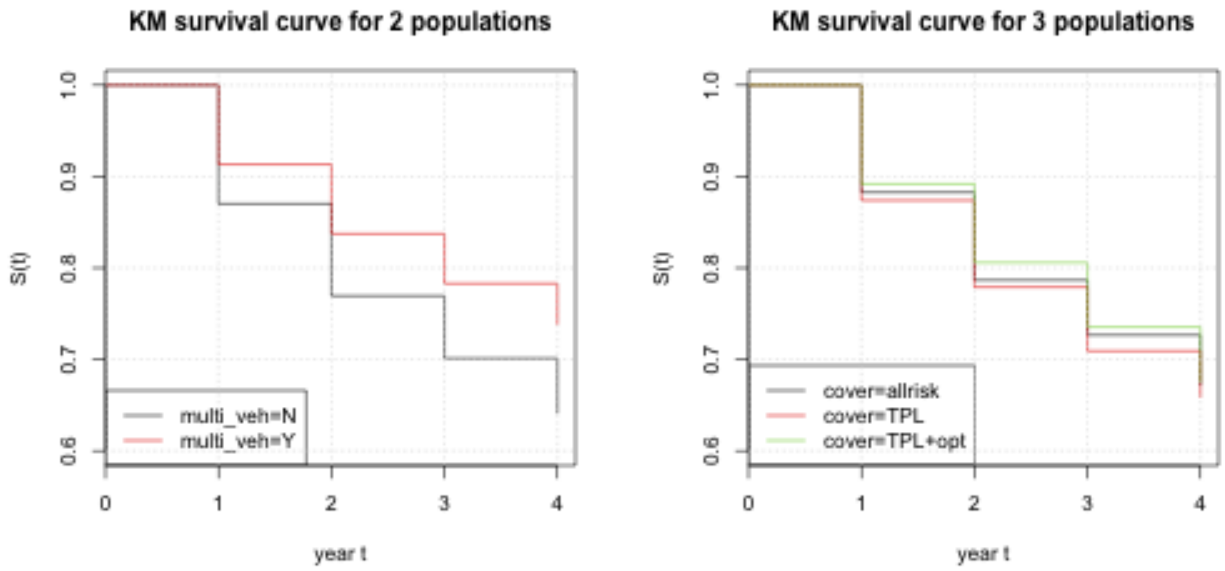


Figure B.22: Model assumptions check for loglogistic distribution - with cov.

B.4.2 Québec

Kaplan-Meier curves



(a) Multi-vehicle (b) Cover type

Figure B.23: Heterogeneity of customer behaviors

Parametric regression summaries

Here follows the regression summaries for the parametric regression methods.

Weibull Intercept-only model:

```
Call: survreg(formula = Surv(t_end, did_cancel) ~ 1, data = untruncfulldata,
    dist = "weibull")
             Value Std. Error   z p
(Intercept)  1.871    0.00822 228 0
Log(scale)  -0.441    0.00802 -55 0

Scale= 0.643

Weibull distribution
Loglik(model)= -37111.4   Loglik(intercept only)= -37111.4
Number of Newton-Raphson Iterations: 6
n= 44380
```

With-covariate model:

```
Call: survreg(formula = Surv(t_end, did_cancel) ~ gender + driv_age +
    house_pol + claim_1 + cover + priceratio:(gender + veh_age +
    price_group + house_pol + claim_2 + cover), data = untruncfulldata,
    dist = "weibull")
                      Value Std. Error      z         p
(Intercept)         1.89686   0.031410  60.39  0.00e+00
```

```
genderM                -0.09789   0.032852  -2.98  2.88e-03
driv_age                0.00678   0.000381  17.77  1.14e-70
house_polY              1.05559   0.062982  16.76  4.77e-63
claim_1                 0.10082   0.016926   5.96  2.58e-09
coverTPL               -0.54095   0.041557 -13.02  9.81e-39
coverTPL+opt           -0.16210   0.047751  -3.39  6.87e-04
genderF:priceratio     -1.73664   0.055860 -31.09 3.34e-212
genderM:priceratio     -1.69501   0.050405 -33.63 6.58e-248
priceratio:veh_age      0.04816   0.001946  24.75 3.36e-135
priceratio:price_group  0.03806   0.001494  25.47 4.19e-143
house_polY:priceratio  -0.55051   0.061499  -8.95  3.51e-19
priceratio:claim_2      0.06386   0.015919   4.01  6.03e-05
coverTPL:priceratio     0.57218   0.037608  15.21  2.84e-52
coverTPL+opt:priceratio 0.21524   0.046393   4.64  3.49e-06
Log(scale)             -0.50577   0.007850 -64.43  0.00e+00

Scale= 0.603

Weibull distribution
Loglik(model)= -35360.3   Loglik(intercept only)= -37111.4
Chisq= 3502.19 on 14 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 7
n= 44380
```

Loglogistic Intercept-only model:

```
Call: survreg(formula = Surv(t_end, did_cancel) ~ 1, data = untruncfulldata,
    dist = "loglogistic")
             Value Std. Error     z p
(Intercept)  1.676    0.00807 207.6 0
Log(scale)  -0.546    0.00779 -70.1 0

Scale= 0.579

Log logistic distribution
Loglik(model)= -36890.8   Loglik(intercept only)= -36890.8
Number of Newton-Raphson Iterations: 4
n= 44380
```

With-covariate model:

```
Call: survreg(formula = Surv(t_end, did_cancel) ~ veh_age + price_group +
    driv_age + house_pol + claim_1 + claim_2 + cover + priceratio:(gender +
    house_pol + claim_1 + claim_2 + cover), data = untruncfulldata,
    dist = "loglogistic")
                        Value Std. Error      z         p
(Intercept)           1.16878   0.076224  15.33  4.57e-53
veh_age               0.05060   0.002078  24.35 5.27e-131
price_group           0.03778   0.001493  25.30 3.31e-141
driv_age              0.00742   0.000398  18.67  8.27e-78
house_polY            0.81941   0.090439   9.06  1.30e-19
claim_1              -0.33168   0.097110  -3.42  6.37e-04
claim_2              -0.32316   0.083706  -3.86  1.13e-04
coverTPL             -1.27783   0.068513 -18.65  1.25e-77
coverTPL+opt         -0.68064   0.083302  -8.17  3.07e-16
priceratio:genderF   -1.28908   0.062108 -20.76  1.09e-95
priceratio:genderM   -1.36069   0.062299 -21.84 9.40e-106
house_polY:priceratio -0.29148  0.091892  -3.17  1.51e-03
claim_1:priceratio    0.47060   0.097554   4.82  1.41e-06
claim_2:priceratio    0.40361   0.086032   4.69  2.71e-06
coverTPL:priceratio   1.32434   0.069077  19.17  6.34e-82
coverTPL+opt:priceratio 0.75068 0.085648   8.76  1.87e-18
Log(scale)           -0.64412   0.007702 -83.63  0.00e+00

Scale= 0.525

Log logistic distribution
Loglik(model)= -34956.6   Loglik(intercept only)= -36890.8
Chisq= 3868.39 on 15 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 4
n= 44380
```

Lognormal Intercept-only model:

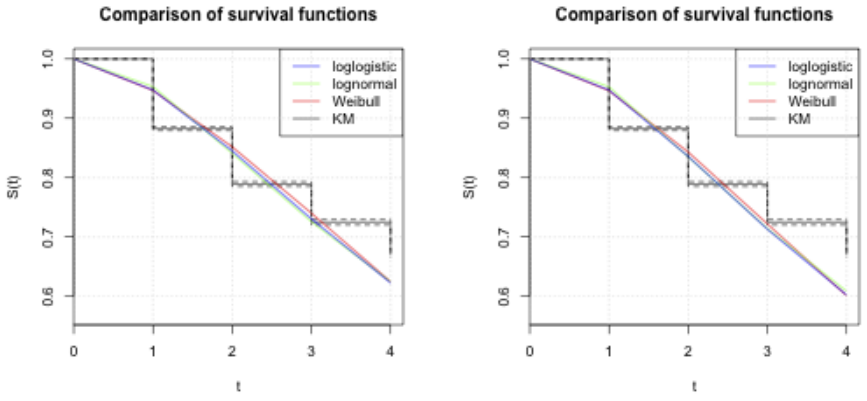
```
Call: survreg(formula = Surv(t_end, did_cancel) ~ 1, data = untruncfulldata,
    dist = "lognormal")
             Value Std. Error      z      p
(Intercept) 1.7091    0.00893 191.48 0.0000
Log(scale)  0.0177    0.00723   2.45 0.0144
```

```
Scale= 1.02

Log Normal distribution
Loglik(model)= -36295.3   Loglik(intercept only)= -36295.3
Number of Newton-Raphson Iterations: 4
n= 44380
```

With-covariate model:

```
Call: survreg(formula = Surv(t_end, did_cancel) ~ veh_age + price_group +
    driv_age + house_pol + claim_1 + claim_2 + cover + priceratio:(gender +
    price_group + house_pol + claim_1 + claim_2 + cover), data = untruncfulldata,
    dist = "lognormal")
                             Value Std. Error      z         p
(Intercept)                1.44829   0.116451  12.44  1.65e-35
veh_age                    0.04224   0.001853  22.79 5.62e-115
price_group                0.01663   0.004120   4.04  5.43e-05
driv_age                   0.00709   0.000395  17.96  3.97e-72
house_polY                 0.97778   0.086137  11.35  7.30e-30
claim_1                   -0.14789   0.095152  -1.55  1.20e-01
claim_2                   -0.20813   0.081780  -2.55  1.09e-02
coverTPL                  -0.93899   0.059124 -15.88  8.51e-57
coverTPL+opt              -0.41100   0.075242  -5.46  4.70e-08
priceratio:genderF        -1.28076   0.114515 -11.18  4.87e-29
priceratio:genderM        -1.34887   0.114917 -11.74  8.16e-32
price_group:priceratio     0.01410   0.004097   3.44  5.81e-04
house_polY:priceratio     -0.47775   0.087663  -5.45  5.04e-08
claim_1:priceratio         0.27876   0.095373   2.92  3.47e-03
claim_2:priceratio         0.27196   0.083794   3.25  1.17e-03
coverTPL:priceratio        0.96251   0.058611  16.42  1.33e-60
coverTPL+opt:priceratio    0.46814   0.076783   6.10  1.08e-09
Log(scale)                -0.07178   0.007165 -10.02  1.27e-23

Scale= 0.93

Log Normal distribution
Loglik(model)= -34553.9   Loglik(intercept only)= -36295.3
Chisq= 3482.88 on 16 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 4
n= 44380
```

Adequacy graphs (1 generation data)



(a) Intercept only (b) With covariates

Figure B.24: Heterogeneity of customer behaviors (1 generation data)

Adequacy graphs (4 generation data)

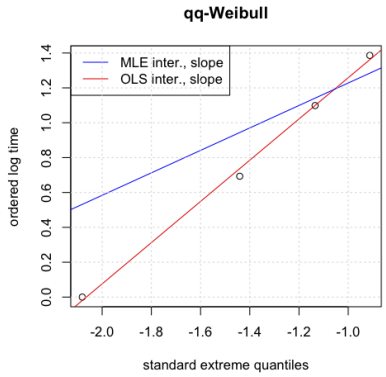


Figure B.25: Intercept-only models (1 generation data)

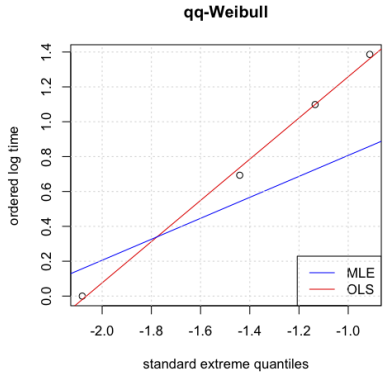
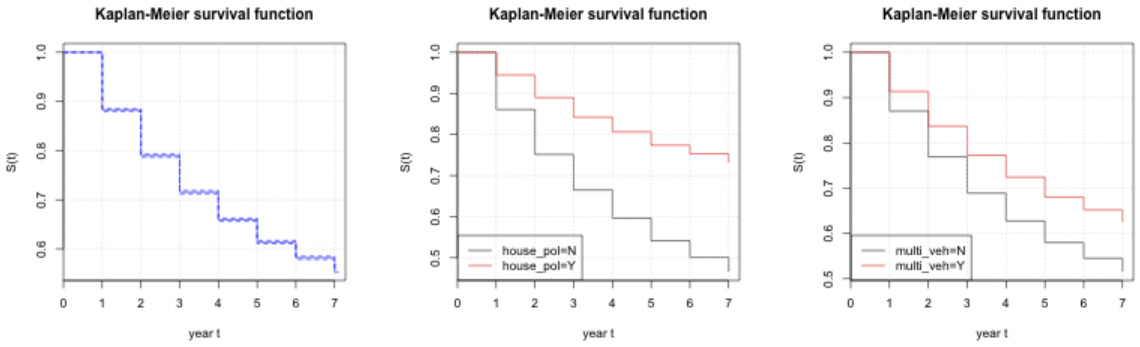
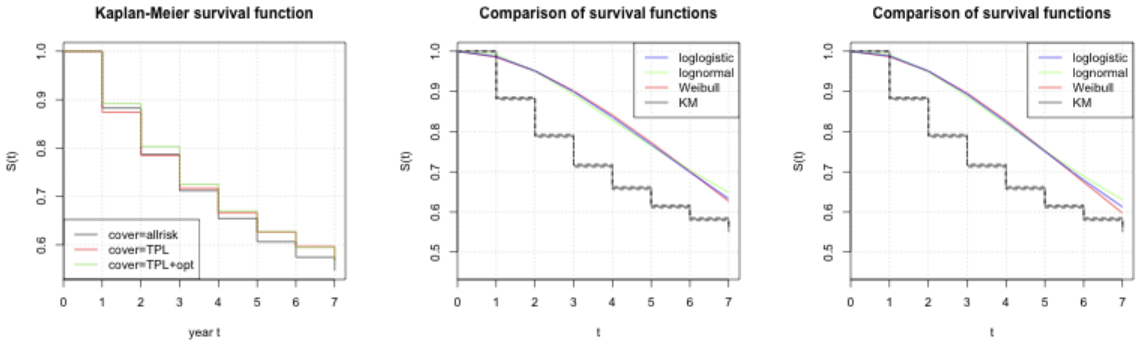


Figure B.26: Model with covariates (1 generation data)



(a) Whole population (b) Cross-selling (c) Multi-vehicle

Figure B.27: Heterogeneity of customer behaviors (4 generation data)



(a) Cover type (b) Intercept only (c) With covariate

Figure B.28: Heterogeneity / Survival functions (4 generation data)

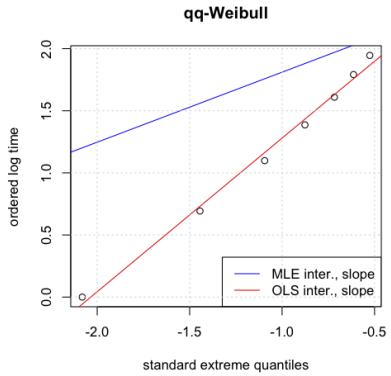


Figure B.29: Intercept-only (4 generation data)

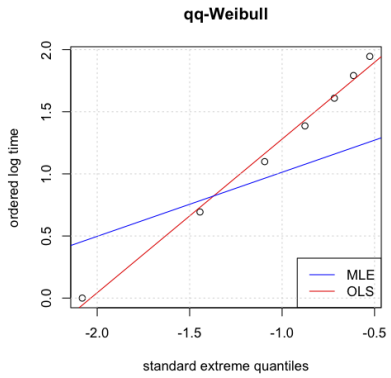


Figure B.30: Model with covariates (4 generation data)

Cox summaries

```
Call: coxph(formula = Surv(t_beg, t_end, did_cancel) ~ veh_age + price_group + driv_age + house_pol +
claim_1 + claim_2 + cover + prev_prem_group + priceratio:(gender + house_pol + claim_1 + claim_2 +
cover + prev_prem_group), data = lasttimedata, method = "efron")

  n= 176158

                                          coef  exp(coef)   se(coef)       z Pr(>|z|)
veh_age                             -0.0660426  0.9360910  0.0017890 -36.917  < 2e-16 ***
price_group                         -0.0731033  0.9295048  0.0013595 -53.771  < 2e-16 ***
driv_age                            -0.0077500  0.9922799  0.0003497 -22.162  < 2e-16 ***
house_polY                          -1.0398235  0.3535171  0.0816893 -12.729  < 2e-16 ***
claim_1                              0.3079118  1.3605810  0.0761343   4.044 5.25e-05 ***
claim_2                              0.2409094  1.2724057  0.0791048   3.045 0.002323 **
coverTPL                             1.8407229  6.3010916  0.1251767  14.705  < 2e-16 ***
coverTPL+opt                         0.6215845  1.8618759  0.0844531   7.360 1.84e-13 ***
prev_prem_group(500,1e+03]           0.3428904  1.4090143  0.0892352   3.843 0.000122 ***
prev_prem_group(1e+03,Inf]           1.3141232  3.7214865  0.0994147  13.219  < 2e-16 ***
priceratio:genderF                   2.1384384  8.4861750  0.0869452  24.595  < 2e-16 ***
priceratio:genderM                   2.1473092  8.5617896  0.0869610  24.693  < 2e-16 ***
house_polY:priceratio                0.2605641  1.2976620  0.0830592   3.137 0.001706 **
claim_1:priceratio                  -0.4627637  0.6295414  0.0751736  -6.156 7.46e-10 ***
claim_2:priceratio                  -0.3507617  0.7041515  0.0804754  -4.359 1.31e-05 ***
coverTPL:priceratio                 -1.8585603  0.1558969  0.1300347 -14.293  < 2e-16 ***
coverTPL+opt:priceratio             -0.5942489  0.5519770  0.0868030  -6.846 7.60e-12 ***
prev_prem_group(500,1e+03]:priceratio  0.1855884  1.2039266  0.0906994   2.046 0.040737 *
prev_prem_group(1e+03,Inf]:priceratio -0.3864228  0.6794832  0.1010785  -3.823 0.000132 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                                    exp(coef) exp(-coef) lower .95 upper .95
veh_age                                0.9361     1.0683    0.9328    0.9394
price_group                            0.9295     1.0758    0.9270    0.9320
driv_age                               0.9923     1.0078    0.9916    0.9930
house_polY                             0.3535     2.8287    0.3012    0.4149
claim_1                                1.3606     0.7350    1.1720    1.5795
claim_2                                1.2724     0.7859    1.0897    1.4858
coverTPL                               6.3011     0.1587    4.9302    8.0532
coverTPL+opt                           1.8619     0.5371    1.5778    2.1970
prev_prem_group(500,1e+03]             1.4090     0.7097    1.1829    1.6783
prev_prem_group(1e+03,Inf]             3.7215     0.2687    3.0626    4.5221
priceratio:genderF                     8.4862     0.1178    7.1566   10.0628
priceratio:genderM                     8.5618     0.1168    7.2201   10.1528
house_polY:priceratio                  1.2977     0.7706    1.1027    1.5271
claim_1:priceratio                     0.6295     1.5885    0.5433    0.7295
claim_2:priceratio                     0.7042     1.4201    0.6014    0.8245
coverTPL:priceratio                    0.1559     6.4145    0.1208    0.2012
coverTPL+opt:priceratio                0.5520     1.8117    0.4656    0.6543
prev_prem_group(500,1e+03]:priceratio  1.2039     0.8306    1.0079    1.4381
prev_prem_group(1e+03,Inf]:priceratio  0.6795     1.4717    0.5574    0.8284

Rsquare= 0.081   (max possible= 0.994 )
Likelihood ratio test= 14861  on 19 df,   p=0
Wald test            = 14468  on 19 df,   p=0
Score (logrank) test = 14576  on 19 df,   p=0
```

One variable-effect on hazard rate

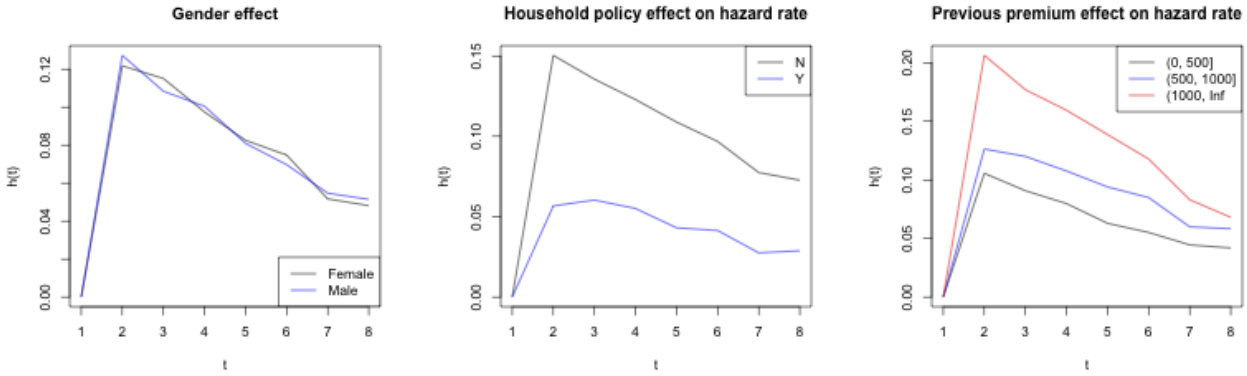
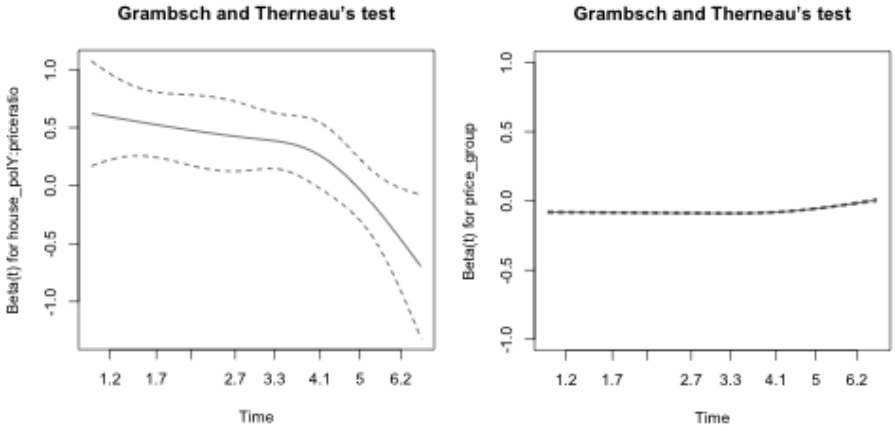


Figure B.31: Hazard rates

PH assumption test



(a) Vehicle age coefficient (b) Cover TPL coefficient

Figure B.32: Grambsch and Therneau's test

Extended Cox summaries

```
Call: coxph(formula = Surv(t_beg, t_end, did_cancel) ~ strata(house_pol) + strata(cover) + price_group +
driv_age + claim_2 + prev_prem_group + priceratio:(gender + veh_age + driv_age + claim_1 + claim_2 + cover),
data = allyeardata, method = "efron")

  n= 429038

                              coef exp(coef)  se(coef)       z Pr(>|z|)
price_group                -0.005611  0.994404  0.001088  -5.158 2.50e-07 ***
driv_age                   -0.014517  0.985588  0.002102  -6.907 4.95e-12 ***
claim_2                     0.168547  1.183584  0.081378   2.071 0.038343 *
prev_prem_group(500,1e+03]  0.372831  1.451840  0.015171  24.575  < 2e-16 ***
prev_prem_group(1e+03,Inf]  0.623761  1.865933  0.021016  29.680  < 2e-16 ***
priceratio:genderF          1.569859  4.805969  0.099857  15.721  < 2e-16 ***
priceratio:genderM          1.527194  4.605236  0.100217  15.239  < 2e-16 ***
priceratio:veh_age          0.008310  1.008344  0.001507   5.515 3.50e-08 ***
driv_age:priceratio         0.007515  1.007543  0.002144   3.504 0.000458 ***
priceratio:claim_1         -0.082710  0.920618  0.014417  -5.737 9.64e-09 ***
claim_2:priceratio         -0.175893  0.838708  0.082739  -2.126 0.033514 *
priceratio:coverTPL        -1.591663  0.203587  0.114096 -13.950  < 2e-16 ***
priceratio:coverTPL+opt    -0.531740  0.587581  0.087128  -6.103 1.04e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                           exp(coef) exp(-coef) lower .95 upper .95
price_group                   0.9944     1.0056    0.9923    0.9965
driv_age                      0.9856     1.0146    0.9815    0.9897
claim_2                       1.1836     0.8449    1.0091    1.3883
prev_prem_group(500,1e+03]    1.4518     0.6888    1.4093    1.4957
prev_prem_group(1e+03,Inf]    1.8659     0.5359    1.7906    1.9444
priceratio:genderF            4.8060     0.2081    3.9517    5.8449
priceratio:genderM            4.6052     0.2171    3.7840    5.6048
priceratio:veh_age            1.0083     0.9917    1.0054    1.0113
driv_age:priceratio           1.0075     0.9925    1.0033    1.0118
priceratio:claim_1            0.9206     1.0862    0.8950    0.9470
claim_2:priceratio            0.8387     1.1923    0.7132    0.9864
priceratio:coverTPL           0.2036     4.9119    0.1628    0.2546
priceratio:coverTPL+opt       0.5876     1.7019    0.4953    0.6970

Rsquare= 0.008   (max possible= 0.828 )
Likelihood ratio test= 3641  on 13 df,   p=0
Wald test            = 3847  on 13 df,   p=0
Score (logrank) test = 3638  on 13 df,   p=0
```

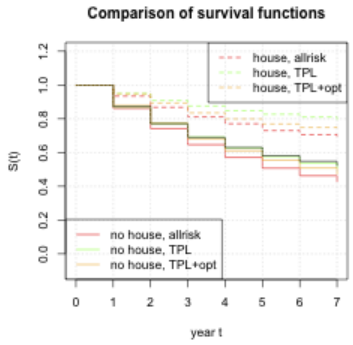
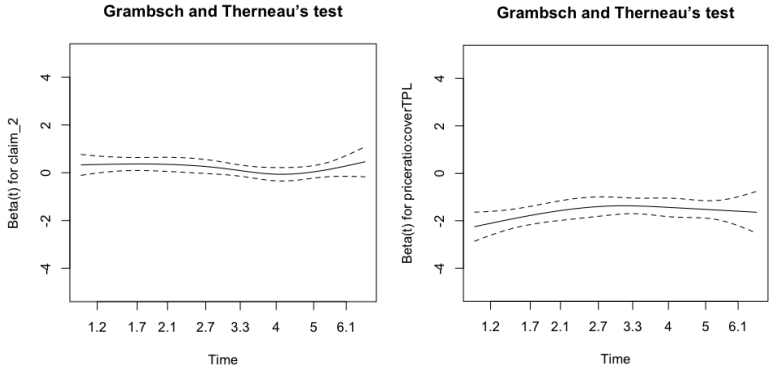


Figure B.33: Average stratified survival curves

One variable-effect on hazard rate

PH assumption test



(a) Vehicle age coefficient (b) Cover TPL coefficient

Figure B.34: Grambsch and Therneau's test

"Insurance never covers you against damages sustained by Chuck Norris, as it's classed as an Act of God!"
from http://www.chucknorrisfacts.com/.