The customer, the insurer and the market

Christophe Dutang

THE CUSTOMER, THE INSURER AND THE MARKET

Christophe DUTANG[1]
Université de Lyon – Université Lyon 1 [2]

Abstract:

Price elasticity studies analyze how customers react to price changes. In this paper, we focus on their effect on the renewal of non-life insurance contracts. 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? Regression models are used to explore the triangular relationship of the customer, the insurer and the market. We conclude that the latter cannot be ignored if we want to get reliable lapse predictions. Furthermore, we also investigate empirical evidence of adverse selection and study its potential impact on lapse decisions.

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

Résumé :

L'élasticité prix consiste à étudier l'effet d'un changement de prix sur le comportement d'un client. Dans cet article, nous nous intéressons à l'élasticité prix dans le cadre d'un renouvellement de contrat d'assurance non-vie. Chaque année, les assureurs font face au dilemme pour établir leur prix: soit augmenter les prix pour améliorer les marges soit diminuer les prix pour attaquer de nouvelles parts de marché? Des modèles de régression sont employés pour explorer la relation triangulaire assuré - assureur - marché. On en conclut que le marché ne peut être ignoré si on veut obtenir des résultats fiables de prédiction. De plus, nous nous testons la présence éventuelle d'anti-sélection et étudions son lien possible avec les résiliations.

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

[1] Corresponding author. Email address: dutangc@gmail.com. Address: ISFA, 50 Avenue Tony Garnier, F-69007 Lyon. Tel.: +33 4 37 28 74 46.
[2] ISFA, Laboratoire SAF, 50 Avenue Tony Garnier, F-69007 Lyon.

1. INTRODUCTION

In price elasticity studies, one analyzes how customers react to price changes. In this paper, we focus on its effect on the renewal of non-life insurance contracts. The methodologies developed 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 strike a compromise between these contradictory objectives. The price elasticity is therefore a factor to contend with in actuarial and marketing departments of every insurance company.

In order 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 sufficiently refined to identify customer segments. Consequently the aim of this paper is to determine the price sensitivity of non life insurance portfolios with respect to individual policy characteristics constituting the portfolio.

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

Price elasticity is not a new topic in actuarial literature. Two ASTIN[1] workshops (see Bland et al. (1997), Kelsey et al. (1998)) were held in the 90's to analyze customer retention and price/demand elasticity topics. Shapiro & Jain (2003) also devote two chapters of their book to price elasticity: Guillen et al. (2003) use logisitic regressions, whereas Yeo & Smith (2003) consider neural networks.

In the context of life insurance, the topic is more complex as the lapse can occur at any time, whereas for non-life policies, most lapses occur at renewal dates. There are some trigger effects due to contractual constraints: penalties are enforced when lapses occur at the beginning of the policy duration, while after that period, penalties no longer applies.

[1] ASTIN stands for Actuarial STudies In Non-Life insurance.

Another influencial feature is the profit benefit option of some life insurance policies allowing insurers to distribute part of benefits to customers in a given year. This benefit option stimulates customers to shop around for policies with higher profit benefits.

In terms of models, Kagraoka (2005), Atkins & Gallop (2007) use counting process to model surrenders of life insurance, while Kim (2005) uses a logistic regression to predict the lapse. Milhaud et al. (2011) point out relevant customer segments when using Classification And Regression Trees models (CART) and logistic regression. Finally, Loisel & Milhaud (2011) study the copycat behavior of insureds during correlation crises.

In non-life insurance, generalized linear models have been the main tool to analyze price-sensitivity, see Hamel (2007) and the references therein. However, generalized linear model outputs might underestimate the true price sensitivity. This could lead to irrelevant conclusions, and therefore gross premium optimization based on such results may lead to biased and sometimes irrelevant pricing decisions, see, e.g., (Hamel 2007, Part 5), (Bella & Barone 2004, Sect. 3).

What makes the present paper different from previous research on the topic is the fact that we tackle the issue of price elasticity from various points of view. Our contribution is to focus on price elasticity of different markets, to check the impact of distribution channels, to investigate the use of market proxies and to test for evidence of adverse selection. We have furthermore given ourselves the dual objective of comparing regression models as well as identifying the key variables needed.

In this paper, we only exploit private motor datasets, but the methodologies can be applied to other personal non-life insurance lines of business. After a brief introduction of generalized linear models in Section 2, Section 3 presents a naive application. Based on the dubious empirical results of Section 3, the Section 4 tries to correct the price-sensitivity predictions by including new variables. Section 5 looks for empirical evidence of asymmetry of information on our datasets. Section 6 discusses the use of other regression models, and Section 7 concludes. Unless otherwise specified, all numerical applications are carried out with the R statistical software, R Core Team (2012).

2. GLMS, A BRIEF INTRODUCTION

In this paper, we are interested in modelling the lapse of (individual) customers. Thus, our interest variable Y_i represents the lapse indicator of customer i, i.e. Y_i follows a Bernoulli variable with 1 indicating a lapse and 0 a renewal of the ith policy. Generalized

Linear Models are a natural choice for modellling Bernoulli events with explanatory variables. Therefore, the purpose of this section is to briefly present such models.

The Generalized Linear Models (GLM[1]) were introduced by Nelder & Wedderburn (1972) to deal with discrete and/or bounded response variables. A response variable on the whole space of real numbers $\mathbb{R}$ is too retrictive, while with GLMs the response variable space can be restricted to a discrete and/or bounded sets. They became widely popular with the book of McCullagh and Nelder, cf. McCullagh and Nelder (1989).

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 intensive use of GLMs by actuaries. There are even books dedicated to this topic, see, e.g., Ohlsson & Johansson (2010). Hence, GLMs seem to be the very first choice of models we can use to model price elasticity. 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 Theoretical presentation

In this section, we only consider fixed-effect models, i.e. statistical models where explanatory variables have deterministic values, unlike random-effect or mixed models. GLMs are an extension of classic linear models, so that linear models form a suitable starting point for discussion. Therefore, the first subsection shortly describes linear models. Then, we introduce GLMs in the second subsection.

2.1.1 Starting from the linear model

Let $X \in M_{np}(\mathbb{R})$ be the matrix where each row contains the value of the explanatory variables for a given individual and $Y \in \mathbb{R}^n$ 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. The Gaussian linear model assumptions are: (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) parameter identification: $rank(X) = p < n$. Then, the Gauss-Markov theorem

[1] Note that in this document, the term GLM will *never* be used for general linear model.

gives us the following results: (i) the least square estimator $\hat{\Theta}$ of Θ is $\hat{\Theta} = (X^T X)^{-1} X^T Y$ and an estimator $\hat{\sigma}^2 = \| Y - X\Theta \|_2^2 / (n-p)$ for σ^2, (ii) $\hat{\Theta}$ is a Gaussian vector independent of the random variable $\hat{\sigma}^2 \sim \chi^2_{n-p}$, (iii) $\hat{\Theta}$ is the unbiased estimator with minimum variance of Θ, such that $Var(\hat{\Theta}) = \sigma^2 (X^T X)^{-1}$ and $\hat{\sigma}^2$ is an unbiased estimator of σ^2.

Let us note that first four assumptions can be expressed into one single assumption $\mathcal{E} \sim \mathcal{N}(0, \sigma^2 I_n)$. But splitting the normality assumption will help us to identify the strong differences between linear models and GLMs. The term $X\Theta$ is generally referred to the linear predictor of Y.

Linear models include a wide range of statistical models, e.g. the simple linear regression $y_i = a + bx_i + \varepsilon_i$ is obtained with a 2-column matrix X having 1 in first column and $(x_i)_i$ in second column. Many properties can be derived for linear models, notably hypothesis tests, confidence intervals for parameter estimates as well as estimator convergence, see, e.g., Chapter 6 of Venables & Ripley (2002).

We now focus on the limitations of linear models resulting from the above assumptions. The following problems have been identified. When X contains near-colinear variables, the computation of the estimator $\hat{\Theta}$ will be numerically unstable. This would lead to an increase in the variance estimator[1]. Working with a constrained linear model is not an appropriate answer. In pratice, a solution is to test models with omitting one explanatory variable after another to check for near colinearity. Another stronger limitation lies in the fact that the response variance is assumed to be the same (σ^2) for all individuals. One way to deal with this issue is to transform the response variable by the nonlinear Box-Cox transformation. However, this response transformation can still be unsatifactory in certain cases. Finally, the strongest limitation is the assumed support of the response variable. By the normal assumption, Y must lies in the whole set $\mathbb{R}$, which excludes count variable (e.g. Poisson distribution) or positive variable (e.g. exponential distribution). To address this problem, we have to use a more general model than linear models.

As already mentioned, Y represents the lapse indicator of customers, i.e. Y follows a Bernoulli variable with 1 indicating a lapse. For Bernoulli variables, there are two main pitfalls when using (Gaussian) linear models. Since the value of $E(Y)$ is contained within the interval $[0,1]$, it seems natural the expected values $\hat{Y}$ should also lie in $[0,1]$. However, predicted values $\hat{\theta}X$ may fall out of this range for sufficiently large or small

[1] This would be one way to detect such isssue.

values of X. Furthermore, the normality hypothesis of the residuals is clearly not met: $Y - E(Y)$ 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 error distribution. This motivates to use an extended model that can deal with discrete-valued variables.

2.1.2 Toward generalized linear models

A Generalized Linear Model is characterized by three components:

1. a random component: Y_i follows a specific distribution of the exponential family $\mathcal{F}_{exp}(\theta_i, \phi_i, a, b, c)$ [1],
2. a systematic component: the covariate vector X_i provides a linear predictor[2] $\eta_i = X_i^T \beta$,
3. a link function: $g : \mathbb{R} \mapsto S$ which is monotone, differentiable and invertible, such that $E(Y_i) = g^{-1}(\eta_i)$,
4. for all individuals $i \in \{1, \dots, n\}$, where θ_i is the shape parameter, ϕ_i the dispersion parameter, a, b, c three functions and S a set of possible values of the expectation $E(Y_i)$. Let us note that we get back to linear models with a Gaussian distribution and an identity link function ($g(x) = x$). However, there are many other distributions and link functions. We say a link function to be canonical if $\theta_i = \eta_i$.

There are many applications of GLM in actuarial science, e.g., claim severity modelling with gamma or inverse normal distributions and claim frequency modelling with the Poisson distribution. Apart from the identity link function, the logarithm link function is the most commonly used link function in actuarial applications. In fact, with this link function, the explanatory variables have multiplicative effects on the observed variable and the observed variable stays positive, since $E(Y) = \prod_i e^{\beta_i x_i}$. For example, the effect of being a young driver and owning an expensive car on average loss could be the product of the two separate effects: the effect of being a young driver and the effect of owning an expensive car. The logarithm link function is a key element in most actuarial pricing models and is used for modelling the frequency and the severity of claims. This makes

[1] See, e.g., Subsection 2.2.2 of McCullagh & Nelder (1989) or more recently Clark & Thayer (2004).
[2] For GLMs, the name `linear predictor' is kept, despite η_i is not a linear predictor of Y_i.

possible to have a standard premium and multiplicative individual factors to adjust the premium.

2.2 Binary regression

Since the insurer choice by customers is modelled by a Bernoulli variable, we give further details on binary regression in this subsection.

2.2.1 Base model assumption

In binary regression, the response variable is either 1 or 0 for success and failure, respectively. We cannot parametrize two outcomes with more than one parameter. So, a Bernoulli distribution $\mathcal{B}(\pi_i)$ is assumed, i.e. $P(Y_i = 1) = \pi_i = 1 - P(Y_i = 0)$, with π_i the parameter. The mass probability function can be expressed as

$$f_{Y_i}(y) = \pi_i^y (1-\pi_i)^{1-y} = e^{y \log(\pi_i/(1-\pi_i)) + \log(1-\pi_i)},$$

which emphasizes the exponential family characteristic $\theta_i = \log(\pi_i / (1-\pi_i))$. Let us recall that the first two moments are $E(Y_i) = \pi_i$ and $Var(Y_i) = \pi_i(1-\pi_i) = V(\pi_i)$. Hence, 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 Bernoulli model is expressed as follows

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

Let us note that if some 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, this is only possible if all covariates are categorical. As indicated in McCullagh & Nelder (1989), the link function and the response variable can be reformulated in term of a latent variable approach. $\pi_i = P(Y_i = 1) = P(x_i^T \beta - \varepsilon_i > 0)$. If ε_i follows a normal distribution (resp. a logistic distribution), we have $\pi_i = \Phi(x_i^T \beta)$ ($\pi_i = F_{logistic}(x_i^T \beta)$).

Now, the log-likelihood is derived as

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

plus an omitted term not involving π_i, see, e.g., (McCullagh & Nelder, 1989, Chap. 4) for further details.

2.2.2 Link functions

Generally, the following three functions are considered as link functions 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.

All these three functions are the inverses of a distribution function, so other link functions can be obtained using inverses of other distribution function. Let us note that the first two links are symmetrical, while the last one is not.

In addition to being the canonical link function for which the fitting procedure is simplified, the logit link is generally preferred because of its simple interpretation as the logarithm of the odds ratio. Indeed, assume there is one explanatory variable X, the logit link model is $p/(1-p) = e^{\mu+\alpha X}$. If $\hat{\alpha} = 2$, increasing X by 1 will lead to increase the odds by $e^2 \approx 7.389$.

2.3 Variable selection and model adequacy

As fitting a GLM is quick in most standard software, then a relevant question is to check for its validity on the dataset used.

2.3.1 Model adequacy

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

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

where $\hat{\pi}$ is the estimate of the beta vector. The ``best'' model is the one having the lowest deviance. However, if all responses are binary data, the first term can be infinite. So in practice, we consider the deviance simply as

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

Furthermore, the deviance is used as a relative measure to compare two models. In most software, in particular in R, the GLM fitting function provides two deviances: the null deviance and the deviance. 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 only an intercept full of 1[1]. The

[1] It means all the heterogeneity of data comes from the random component.

(second) deviance is the deviance for the model $D(y,\hat{\pi})$ with the p explanatory variables. Note that if there are as many parameters as there are observations, then the deviance will be the best possible, but the model does not explain anything.

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, AIC aims to penalized overfitted models, i.e. models with too many parameters (compared to the length of the dataset). AIC is defined by

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

where k the number of parameters, i.e. the length of β. This criterion is a trade-off between further improvement in terms of log-likelihood with additional variables and the additional model cost of including new variables. To compare two models with different parameter numbers, we look for the one having the lowest AIC.

In a linear model, the analysis of residuals (which are assumed to be identical and independent Gaussian variables) may reveal that the model is unappropriate. Typically we can plot the fitted values against the fitted residuals. For GLMs, the analysis of residuals is much more complex, because we loose the normality assumption. Furthermore, for binary data, i.e. not binomial data, the plot of residuals exhibits straight lines, which are hard to interpret, see Appendix 8.1. We believe that the residual analysis is not appropriate for binary regressions.

2.3.2 Variable selection

From the normal asymptotic distribution of the maximum likelihood estimator, we can derive confidence intervals as well as hypothesis tests for coefficents. Therefore, a p-value is available for each coefficient of the regression, which helps us to keep only the most significant variable. However, as removing one variable impacts the significance of other variables, it can be hard to find the optimal set of explanatory variables.

There are mainly two approaches: either a forward selection, i.e. starting from the null model, we add the most significant variable at each step, or a backward elimination, i.e. starting from the full model, we 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 $\ln\mathcal{L}$ between two models, i.e. ratios of likelihood. Using an asymptotic distribution, either chi-square or Fisher-Snedecor distributions, a p-value can be used to remove or to keep an explanatory variable. Based on

this fact, statistical softwares generally provide a function for the backward and the forward selection using an automatic deviance analysis.

In conclusion, GLM is a well-known statistical method in actuarial science. This fact motivates its use to model lapse rate. Since it is a classic among statistical models, fitting method and variable selection use state-of-art algorithms providing robust estimators. So there is absolutely no problem in applying GLMs for a daily use. In the following section, we apply GLMs to explain the customer price-sensitivity.

3. SIMPLISTIC APPLICATIONS AND BIASED BUSINESS CONCLUSIONS

This section is intended to present quite naive GLM applications and to show how they can lead to inconclusive or even biased findings. First, we use a dataset with poor and limited data, and then a larger dataset with more comprehensive data. Finally, we summarize the issues encountered. It may seem obvious, but to study customer price-sensitivity, insurers need to collect the premium proposed to customers when renewing policy, especially for those who lapse.

For confidential reasons, the country names are not revealed, but we study two continental European insurance markets. In this part of the world, the insurance penetration rate is considered high, e.g., 8.6% in France, 7% in Germany, 7.6% in Italy, according to Cummins & Venard (2007). Thus, the insurance markets studied are mature as well as competition level is intense. Furthermore, data outputs presented in this paper have been perturbed, but original conclusions have been preserved.

3.1 An example of poor data

In this subsection, we work with a (representative) subset of a 1-year lapse history database in 2003. Each line of the dataset represents a policy for a given vehicle. With only few variables, we expect the data analysis to be difficult and the model outputs to be unreliable.

3.1.1 Descriptive analysis

To better understand interactions between lapses, the premium and other explanatory variables, we start with a short descriptive analysis. As a general comment, all variables in the dataset are dependent to the lapse variable according to a Chi-square test. At our disposal, we have the last year premium and the proposed premium. Computing the premium ratio, we observe that most of the portfolio experienced a price decrease, probably

due to the ageing and the market conditions. We expect to slightly underestimate the true price sensitivity of clients, since customers attention will be released.

Turning to customer variables, we focus on gender and driver age variables, reported in Table 1. As the age of the customer increases, the lapse rate decreases. So, the most sensitive clients seem to be the youngest clients. The gender[1] does not have any particular impact on the lapse (alone). However, the GLM analysis may reveal some links between the gender and lapses if the gender variable is crossed with other explanatory variables.

	(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: Driver age and Gender

We also have a categoric variable containing a lapse type with three possible values: lapse by insured, lapse by company and payment default. We observe a total lapse rate of 18%, of which 11% is a payment default, 6% a lapse by the customer, only 1% a lapse by the company. The lapse by company has to be removed, because those lapses generally result from the pruning strategy of insurers. However, default of payment must be taken with care since it might represent a hidden insured decision. It may result from a too high premium that the customer can't afford. Thus, we choose to keep those policies in our study. Note that the lapse motive cannot be used in the regression because its value is not known in advance, i.e. the lapse motive is endogeneous.

The last variables to explore are policy age and vehicle age. According to Table 2, some first conclusions can be derived. As the policy age increases, the remaining customers are more and more loyal, i.e. lapse rates decrease. Unlike the policy age, when the vehicle age increases, the lapse rate increases. One explanation may be that the customer may shop around for a new insurer when changing the vehicle.

	(1, 5]	(5,9]	(9,13]	(13,17]	(1,8]	(8,14]	(14,20]	(20,26]
Lapse rate (%)	21	17	18	16.9	17.6	19.4	21	39
Prop. of total (%)	38	33	22	7	36	37	22	4

Table 2: Policy age and vehicle age

[1] In a near 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.

3.1.2 GLM analysis

For the GLM analysis of this dataset, we use a backward selection. The explanatory variables are driver age, gender, policy age, vehicle age, the last year premium and the price ratio, i.e. ratio of the premium proposed and the premium paid last year. In order to have better fit and predictive power, all explanatory variables are crossed with the price ratio: crossing variable x_j with price ratio p consists in creating a dummy variable $x_{ji} \times p_i$ for all observations $1 \le i \le n$.

Note that variable x_j might be categorical, i.e. valued in $\{0,\dots,d\}$, which allows to zoom in on some particular features of individuals. The linear predictor for observation i is thus given by

$$\eta_i = \beta_0 \times 1 + (x_{1i},\dots,x_{ki})^T \beta_{-p} + (z_{1i},\dots,z_{ki})^T \beta_{+p} \times p_i,$$

where β_0 is the intercept, β_{-p} (resp. β_{+p}) the coefficient for price-noncross variables (resp. price-cross), x_i price-noncross variables, z_i price-cross variables and p_i the price ratio.

Yet not reported here, we test two models: (i) a GLM with original (continuous) variable and (ii) a GLM with categorized variables. We expect the second model with categorized data to be better. Using continuous variables limits the number of parameters: 1 parameter for a continuous variable and $d-1$ parameters for a categorical variable with d categories. Cutting the driver age, for example, into three values $]18,35]$, $]35,60]$ and $]60,99]$ enables to test for the significance of the different age classes.

The numerical application reveals that a GLM with categorical data is better in terms of deviance and AIC. Hence, we only report this model in Appendix 8.2, first column is the coefficient estimates $\hat{\beta}_0$, $\hat{\beta}_{-p}$ and $\hat{\beta}_{+p}$.

The GLM with continuous variables also has business inconsistent fitted coefficients, e.g. the coefficient for the price ratio was negative. This also argues in favor of the GLM with categorized variables. We also analyze (but do not report) different link functions to compare with the (default) logit link function. But the fit gives similar estimate for the coefficients $\hat{\beta}_0$, $\hat{\beta}_{-p}$ and $\hat{\beta}_{+p}$, as well as similar predictions.

To test our model, we want to make lapse rate predictions and to compare against observed lapse rates. From a GLM fit, we get the fitted probabilities $\hat{\pi}_i$ for $1 \le i \le n$. Plotting those probabilities against the observed price ratios does not help to understand the link between a premium increase/decrease and the predicted lapse rate. Recall that we are

interested in deriving a portfolio elasticity based on individual policy features, we choose to use an average lapse probability function defined as

$$\hat{\pi}_n(p) = \frac{1}{n}\sum_{i=1}^{n} g^{-1}\left(\hat{\beta}_0 + x_i(p)^T \hat{\beta}_{-p} + z_i(p)^T \hat{\beta}_{+p} \times p\right), \quad (1)$$

where $(\hat{\beta}_0, \hat{\beta}_{-p}, \hat{\beta}_{+p})$ are the fitted parameters, x_i price-noncross explanatory variables, z_i price-cross explanatory variables[1] and g the logit link function, i.e. $g^{-1}(x) = 1/(1+e^{-x})$. Note that this function applies a price ratio constant to all policies. For example, $\hat{\pi}_n(1)$ the average lapse rate, called central lapse rate, if the premium remains constant compared to last year for all our customers.

Computing this sum for different values of price ratio is quite heavy. We could have use a prediction for a new obsveration $(\tilde{x}, \tilde{y}, \tilde{p})$,

$$g^{-1}\left(\hat{\beta}_0 + \tilde{x}^T \hat{\beta}_{-p} + \tilde{y}^T \hat{\beta}_{+p} \times \tilde{p}\right),$$

where the covariate $(\tilde{x}, \tilde{y}, \tilde{p})$ corresponds to the average individual. But in our datasets, the ideal average individual is not the best representative of the average behavior. Equation (1) has the advantage to really take into account portfolio specificities, as well as the summation can be done over a subset of the overall data. In Table 3, we put the predicted lapse rates, i.e. $\hat{\pi}_n(1)$. We also present a measure of price sensitivity, the delta lapse rate defined as

$$\Delta_{1-}(\delta) = \hat{\pi}_n(1-\delta) - \hat{\pi}_n(1) \text{ and } \Delta_{1+}(\delta) = \hat{\pi}_n(1+\delta) - \hat{\pi}_n(1), \quad (2)$$

where δ represents a premium change, for example 5%. As mentioned in the introduction, this measure has many advantages compared to the price elasticity[2] ($e_r(p) = \frac{dr(p)}{dp} \times \frac{p}{r(p)}$) : it is easier to compute, more robust[3], easier to interpret.

	$\Delta_{1-}(5\%)$	$\hat{\pi}_n(1)$	$\Delta_{1+}(5\%)$
All	-0.745	14.714	0.772
Old drivers	-0.324	9.44	0.333
Young pol., working male	-0.585	15.208	0.601
Young drivers	-1.166	19.784	1.211

Table 3: Central lapse rates (%) and deltas (pts)

[1] Both x_i and y_i may depend on the price ratio, e.g. if x_i represents the difference between the proposed premium and a technical premium.

[2] It is the customer's sensitivity to price changes relative to their current price. A price elasticity of e means that an increase by 1% of p increase the lapse rate by e%.

[3] Price elasticity interpretation is based on a serie expansion around the point of computation. So, price elasticity is not adapted for large δ.

In Table 3, we report the predicted lapse rates and deltas for the whole dataset (first line) as well as for three subsets: old drivers, young policies and working male, young drivers. This first result exhibits the wide range of behaviors among a portfolio: young vs. old drivers. However, delta values seem unrealistic: a 5% premium increase will increase the lapse rate only by 0.772 pts. Based only on such predictions, one will certainly not hesitate to increase premium.

As this small dataset only provides the driver age, GLM outputs lead to inconclusive or dubious results. The old versus young segmentation alone cannot in itself substantiate the lapse reasons. We conclude that the number of explanatory variables are too few to get reliable findings with GLMs, and probably with any statistical models.

3.2 A larger database

In this subsection, we study another dataset from a different country in continental Europe in 2004. As for the other dataset, a record is a policy purchased by an individual, so an individual may have different records for the different covers he bought.

3.2.1 Descriptive analysis

This dataset is very rich and contains much more variables than the previous set. The full list is available in Appendix 8.3. In Table 4, we present some explanatory variables. The dataset contains policies sold through different distribution channels, namely tied-agents, brokers and direct platforms, cf. first line of Table 4. Obviously, the way we sell insurance products plays a major role in the customer decision to renew or to lapse. The coverage types (Full Comprehensive, Partial Comprehensive and Third-Part Liability) have a lesser influence on the lapse according to the first table.

Coverage	FC	PC	TPL			Channel	Agent	Broker	Direct	
prop. size	36.16	37.61	26.23			prop. size	65.1	20.1	6.1	
lapse rate	14.26	12.64	12.79			lapse rate	7.4	10.3	12.1	
Claim nb.	0	1	2	3	(3 - 13]	Bonus evol.	down	stable	up	
prop. size	70.59	25.29	3.60	0.44	0.09	prop. size	33.32	62.92	3.76	
lapse rate	13.75	13.37	16.03	12.82	35.16	lapse rate	16.69	11.53	12.02	
Policy age	(0,1]	(1,2]	(2,7]	(7,34]		Vehicle age	(0,6]	(6,10]	(10,13]	(13,18]
prop. size	24.97	16.79	34.38	23.86		prop. size	26.06	31.01	21.85	21.08
lapse rate	17.43	15.27	11.26	8.78		lapse rate	15.50	13.56	12.72	10.67

Table 4: Impact on lapse rates (%)

The dataset also contains some information on claim history, e.g. the bonus/malus or the claim number. In Table 4, we present a dummy variable for the bonus evolution

(compared to last year). From this table, we observe that a non-stable bonus seems to increase the customer propency to lapse. This could be explained by the fact that decreasing or increasing bonus implies the biggest premium difference compared to last year premium, raising the customer attention. At this stage, the claim number does not seem to influence the lapse. The policy age has the same impact as in the previous dataset (cf. Table 2). The older is the policy the lower the customer lapses. However, the opposite effect is observed for the vehicle age compared to previous dataset.

3.2.2 GLM analysis

Now, we go to the GLM analysis. We apply a backward selection to select statistically significant variables. The regression summary is put in Appendix 8.4. The signs of coefficient β_{+p} are positive for the two categories of last year premium level[1], thus this is business consistent. The most significant variables[2] are the region code, the distribution channel and the dummy variable indicating the relative difference between the technical premium and the proposed premium and the dummy variable checking whether the policyholder is also the car driver.

In terms of prediction, the results presented in Table 5 are similar to the results of the previous subsection. As for the "poor" dataset, we use the average lapse function $\hat{\pi}_n(p)$ and delta lapse rate $\Delta_{1+}(\delta)$ defined in Equations (1) and (2), respectively. The overall central lapse rate is low compared to the previous set but the customers on that market seems more price sensitive, with bigger deltas for a 5% decrease or increase. Taken into account the distribution channel, the differences are huge: around 8.7% vs. 11.6% for agent and direct, respectively. Despite observing higher deltas, we think these estimates still underestimate the true price sensitivity.

[1] See lastpremgroup2(0,500] and lastpremgroup2(500, 5e+3].
[2] See diff2tech, region2, channel, diffdriverPH7.

	$\Delta_{1-}(5\%)$	$\hat{\pi}_n(1)$	$\Delta_{1+}(5\%)$
All	-0.833	8.966	1.187
Channel agent	-0.759	7.732	0.75
Channel broker	-1.255	9.422	1.299
Channel direct	-1.18	11.597	1.268
Coverage Full Comp.	-0.622	7.723	0.97
Coverage Part. Comp.	-0.714	9.244	1.063
Coverage TPL	-0.899	10.179	1.178

Table 5: Central lapse rates (%) and deltas (pts)

Looking at the bottom part, the impact of cover type on central lapse rates is considerably lower. Central rates are between 8% and 10%, regardless of the product purchased. Delta lapse rates Δ_{1+} are again surprisingly low around 1 pt. In Appendix 8.4, we also compare the observed lapse rate by channel and coverage type against the fitted lapse rate, see Table 16. The results are unsatisfactory.

3.3 Issues

The price-sensitivity assessment appears to be difficult. Getting outputs is easy but having reliable estimates is harder. We are not confident on the lapse prediction as well as the additional lapse rates Δ_{1+}. A first answer is shown in Table 17 of Appendix 8.4, where we present the predicted results when the dataset is split according to the distribution channel or the coverage type. This split provides more realistic lapse rates, each fit better catches the specificity of the distribution channel. Thus, we choose to fit nine regressions in the following in order to catch the full characteristics of the distribution channel and the coverage type.

However, this section reveals major issues of a quick application of GLMs with few or weakly relevant explanatory variables. We miss something as it does not really make sense that a 5% premium increase on the whole portfolio leads to a lapse rate increase less than 1pt. In such situation, the insurer has all reasons to increase premium by 5% and to get a higher gross written premium. The market competition level drives the level of customer price-sensitivity that we can estimate. Therefore, caution is needed when using GLMs predictions with few variables.

4. INCORPORATING NEW VARIABLES IN THE REGRESSION

This section focuses on identifying new key variables needed in the GLM regression in order to get more reliable results. Attentive readers have probably noticed that some

variables have been forgotten in this first analysis. As we will see, they have a major impact on the GLM outputs. Furthermore, taking into account previous conclusions on the large dataset of Subsection 3.2, all results presented in this section are obtained by nine different regressions, one for each channel and each coverage type.

4.1 Rebate levels

Firstly, we add to all regressions the rebate level variable, specifying the amount of rebate granted by the agent, the broker or the client manager to the customer. As reported in Table 6, the number of customers having rebates is considerably high. The broker channel grants a rebate to a majority of customers. Then comes the tied-agent channel and finally the direct channel.

	Full Comp.	Part. Comp.	TPL
Agent	56.62	36.84	22.26
Broker	62.25	52.5	36.24
Direct	23.05	22.89	10.37

Table 6: Proportion of granted rebates (%)

It seems logical that the direct channel does not grant rebates since the premium is generally lower through the direct channel than with other distribution channels. The influence of the coverage type is also substantial: it is harder to get a rebate for a third-part liability (TPL) product than a full comprehensive coverage product.

In order to catch the most meaningful features of the rebate on the lapse decision, the rebate variable has been categorized. Despite the dataset is subdivided into 9 parts, this variable is always statistically significant. For example in the TPL broker subgroup, the estimated coefficients $\hat{\beta}$ for the rebate variable are $\hat{\beta}_{10-20} = -0.368879$, $\hat{\beta}_{25+} = -0.789049$. In that case, the variable has three categories (0, 10-20 and 25+), thus two coefficients for two categories plus the baseline integrated in the intercept. The negative sign means that the rebate level has a negative impact on the lapse, i.e. a rebate of 15 decreases the linear predictor (hence the predicted lapse rate). This is perfectly natural.

Furthermore, when predicting lapse rate with the average lapse function $\hat{\pi}_n$, we force the rebate level to zero. That is to say, in the equation

$$\hat{\pi}_n(p) = \frac{1}{n}\sum_{i=1}^{n} g^{-1}\left(\hat{\beta}_0 + x_i(p)^T \hat{\beta}_{-p} + z_i(p)^T \hat{\beta}_{+p} \times p\right),$$

the explanatory variables $x_i(p), z_i(p)$ are updated depending on the price ratio p. The

rebate variable appearing in the vector $(x_i(p), z_i(p))$ is set to zero when predictions are carried out. So that a 5% increase really means such premium increase, and not 5% minus the rebate that the customer got last year.

	$\hat{\pi}_n(1)$	$\Delta_{1+}(5\%)$	$\hat{\pi}_n(1)$	$\Delta_{1+}(5\%)$	$\hat{\pi}_n(1)$	$\Delta_{1+}(5\%)$
Agent	7.278	0.482	8.486	0.896	8.549	0.918
Broker	10.987	2.888	9.754	2.776	10.972	3.437
Direct	12.922	1.154	11.303	1.263	11.893	1.490
	Full Comp.		Part. Comp.		TPL	

Table 7: Central lapse rates (%) and deltas (pts)

Table 7 presents GLM predictions for the nine subgroups. We can observe the major differences compared to the situation where the rebate level was not taken into account, cf. Table 5. Notably for the broker channel, the delta lapse rates are high and represent the broker's work for the customer to find the cheapest premium. The central lapse rates also slightly increase in most cases compared to the previous fit. This subsection shows how important the rebate variable is when studying customer price-sensitivity.

4.2 Market proxy

In this subsection, we add another variable to regressions, a market premium proxy by policy. The proxy is computed as the tenth lowest premium among competitor premiums of a standard third-part liabibility coverage product for which there is no deductible. Such computation is carried out on a market premium database which is filled by all insurers of the market. However, we don't have the choice of the market proxy. It would have been a good study to see the influence of the market proxy choice, e.g., the fifth, the first lowest or the mean premium, in the GLM fit.

Unfortunately, the market proxy information is only available on two subsets of the database, namely TPL agent and TPL direct subsets. As for the technical premium, we choose to insert that variable in the regression via the relative difference compared to the proposed premium. We consider

$$m_i = \frac{\text{market}_i - \text{premium}_i}{\text{premium}_i},$$

where market_i and premium_i denote the market premium and the proposed premium for the ith policy, respectively. In Table 8, we give a basic cross-table of lapse and relative

market premium variables. Among the lapsed policies, 65% of them have a higher premium than the market proxy, whereas for renewed policies it drops to 57%.

m	(-0.75,-0.5]	(-0.5,-0.25]	(-0.25,0]	(0,0.25]	(0.25,0.5]	(0.5,0.75]	(0.75,1]
Renew	0.69	18.484	33.248	28.254	9.735	0.571	0.066
Lapse	0.079	1.326	4.158	2.637	0.327	0.032	0.006

Table 8: Percentage of policies (%)

However, we cannot conclude that lapses result from a higher premium compared to the market, just based on this table. In fact, the market proxy is just a proxy for the third-part liability coverage, computed as the tenth lowest premium. Moreover, the market proxy is a theoretical premium based on the risk characteristics. If a client goes to another company, it may have a lower or a higher premium depending if he get a *rebate* or choose an *add-on* cover.

Now, that we have described the new explanatory variable, we turn our attention to the GLM regression. The residual deviance and Akaike Information Criterion (AIC) have slightly decreased with the addition of the market proxy (8866 to 8728 and 8873 to 8735, respectively). Regression summary for the GLM with market variable is available on request to the author.

The most instructive results are the average lapse prediction. Comparing the Table 9 with Table 7 reveals that the addition of the market proxy has a major impact on the delta lapse rate $\Delta_{1+}(5\%)$, cf. bolded figures. For the TPL agent subset, it goes from 0.918 to 1.652 pts, while for the TPL direct subset, from 1.490 to 2.738. Central lapse rates before and after the market proxy inclusion are consistent. The predicted results are plotted on Figure 1, where the x-axis represents central lapse rates ($\hat{\pi}_n(1)$), the y-axis delta lapse rates for a 5% premium increase ($\Delta_{1+}(5\%)$). The bubble radius are determined by the proportion of the subset in the whole dataset. The text order in the legends is the decreasing order of bubble radius.

	$\hat{\pi}_n(1)$	$\Delta_{1+}(5\%)$	$\hat{\pi}_n(1)$	$\Delta_{1+}(5\%)$	$\hat{\pi}_n(1)$	$\Delta_{1+}(5\%)$
Agent	7.278	0.482	8.486	0.896	**8.548**	**1.652**
Broker	10.987	2.888	9.754	2.776	10.972	3.437
Direct	12.922	1.154	11.303	1.263	**11.958**	**2.738**
	Full Comp.		Part. Comp.		TPL	

Table 9: Central lapse rates (%) and deltas (pts)

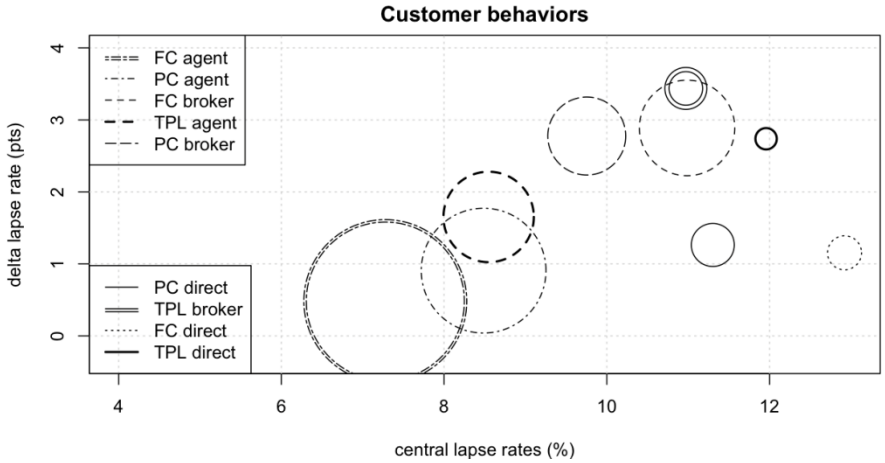


Figure 1: Comparison of distribution channels and cover types

On Figure 1, we clearly observe the difference between those two channels both in terms of central lapse rates and delta lapse rates. These two differences can be explained again by the fact the brokers are paid to find the cheapest premium. The direct channel shows higher central lapse rates $\hat{\pi}_n(1)$, but the *estimated* delta lapse rates are lower than those for Broker channel. Direct channel are designed for customers shopping around on internet, so it seems logical that their propency to lapse should be higher. We would have expected the same to hold for delta lapse rates $\Delta_{1+}(5\%)$. The estimated delta rate of the direct channel might still be underestimated. In addition to the absence of market proxies in the TPL direct database, the direct channel is also small in size. Hence, higher uncertainty on those estimates might explain low delta lapse rates for FC/PC direct subsets.

4.3 Backtesting

In this subsection, we present backtesting results for the fitted GLMs. We start by looking only at an aggregate level: channel and coverage. The results are given in Table 10, reporting observed and fitted lapse rates. The observed lapse rate r_j for the j th group is computed as the average lapse rate variable over the j th group, whereas fitted lapse rate is the average of the fitted probabilities $\hat{\pi}_i$ over the j th group given the observed explanatory variables for each individual $\frac{1}{n_j}\sum_{i=1}^{n_j}\hat{\pi}_i$.

The fitted results are good, since for each subgroup, the deviation is below one percentage point. Compared to the previous backfit table, the improvements with rebate

level, market proxy and datasplit are amazing. The two subgroups for which we use market proxy, the results are even better (deviation < 0.1 pt), see TPL agent and direct. However, we must recognize that observed price ratio are relatively small: for 85% of the portfolio, the difference is below 5%. Hence, the model appropriately catches the lapse phenomenon when the variation in premium remains *reasonable*.

	Observed	Fitted	Observed	Fitted	Observed	Fitted
Full Comp.	7.361	7.124	10.167	10.468	12.958	12.881
Part. Comp.	8.123	8.084	9.971	10.09	11.258	11.193
TPL	8.579	8.569	10.867	11.042	11.153	11.171
	Agent		Broker		Direct	

Table 10: Central lapse rates (%) and deltas (pts)

To further assess the predictive power of our GLM fits, we focus on the TPL coverage product. We consider three subpopulations representing three different behaviors: (i) old drivers with at least two contracts in the household, (ii) working class with a decreasing bonus-malus and an old vehicle, (iii) young drivers. We expect the population 3 to be the most price-sensitive.

	Prop.	Obs.	Fit.	Std.Err.	Prop.	Obs.	Fit.	Std.Err.	Prop.	Obs.	Fit.	Std.Err.
Pop. 1	13	4.98	5.16	0.22	5	7.99	8.24	0.49	3	6.98	8.187	0.65
Pop. 2	13	8.45	8.65	0.32	16	11.59	12.36	0.50	17	12.44	13.02	0.61
Pop. 3	10	10.01	9.91	0.42	14	13.25	12.45	0.62	13	14.91	14.184	0.74
	Agent				Broker				Direct			

Table 11: Lapse rates and proportions (%)

In Table 11, we report the backfit results for the three selected populations separating each distribution channel. Each block presents the proportion of population i in the total subset, the observed lapse rate for population i, the mean of fitted lapse rates and standard deviations. As expected the difference between the three populations is high whatever the channel. Population 1 can be tagged as a sluggish behavior, Population 2 a kind of medium behavior, while Population 3 represents highly sensitive customers.

4.4 Market scenarios

Having a market variable in the database allows us to perform market scenarios. In this subsection, we briefly present this topic particularly interesting for business line managers. We perform two basic scenarios: a 5% increase of market premium and a 5% decrease of market premium.

	Insurer			Insurer		
	-5%	0%	+5%	-5%	0%	+5%
Market -5%	8.007	8.763	10.481	12.538	14.143	17.589
Market 0%	7.801	8.548	10.152	9.604	11.958	14.696
Market +5%	7.645	8.359	9.916	8.638	10.943	13.589
	Agent			Direct		

Table 12: Market scenarios (%)

The results are summarized in Table 12. It is surprising to see how the tied-agent customers react very slowly when premium fluctuates. In particular when market decrease of 5% and the proposed premium increases by 5%, then the lapse rate goes only from 8.548% to 10.481%. While for the direct channel, the lapse rate rockets from 11.958% to 17.589%. Actually for any difference in premium, the lapse rate fluctuates largely for the direct channel.

4.5 Conclusion

The two previous sections demonstrate that GLMs are easy to implement, but care on the variable selection and appropriate data are needed to ensure reliable outputs. In this section, we show how incorporating new key variables in the GLM regression substantially improves the lapse rate predictions in the different premium scenarios. The rebate level partially reveals the agent or the broker actions on the customer decisions, while the use of market proxies illustrates how decisive the competition level is when studying customer price-sensitivity.

In conclusion, the GLM methodology, when used on appropriate data, fulfills the initial objective to derive average lapse rate prediction taking into account individual features. Furthermore, using the predicted lapse rate values of GLMs, it has been easy to identify customer segments, which react differently to premium changes. The back-fit of the GLMs on the identified populations is correct. At a customer segment level, GLMs provide a fair estimate of lapse rate and price sensitivity for reasonable premium changes. But at a policy level, we think lapse predictions should be treated carefully.

5. TESTING FOR ASYMMETRY OF INFORMATION

Asymmetry of information occurs when two agents (say a buyer and a seller of insurance policies) do not have access to the same amount of information. In such situations, one of the two agents might take advantage of his additional information in the

deal. Typically, two problems can result from this asymmetry of information : adverse selection and moral hazard. In insurance context, moral hazard can be observed when individuals behave in risker ways, when they are insured. Insurers cannot control the policyholder's actions to prevent risk.

Adverse selection depicts 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 himself. Generally, the buyer will choose a deductible in his favor based on its own risk assessment. Hence, high-risk individuals will have the tendency to choose lower deductibles. Adverse selection is caused by hidden information, whereas moral hazard is caused by hidden actions.

Joseph Stiglitz was awarded the Nobel price in economics in 2001 for his pioneer work in asymmetric information modelling. In insurance context, Rothschild & Stiglitz (1976) models the insurance market where individuals choose a “menu” (a couple of premium and deductible) from the insurer offer set. Within this model, they show that high-risk individuals choose contracts with more comprehensive coverage, whereas low-risk individuals will choose higher deductibles.

5.1 Testing for evidence of adverse selection

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

In this paper, we follow the framework of Dionne et al. (2001), which uses GLMs to test for the evidence of adverse selection[1]. Let X be an exogenenous variable vector, N an endogeneous variable and Z a decision variable. The absence of adverse selection is equivalent to the prediction of Z based on the joint distribution of X and N coincides with prediction with X alone. This indirect characterization leads to

$$l(Z \mid X, N) = l(Z \mid X), \tag{3}$$

where $l(.|.,.)$ denotes the conditional probability density function. A simple approach is to perform conditional dependence in the GLM parametric framework, such that the model is

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

contrained as $l(Z \mid X, N) = l(Z; aX + bN)$. Testing for the conditionnal independence of Z with respect to N given X is carried out by regressing the variable Z on X and N in order to check whether the coefficient for N is significant.

However, this approach may lead to spurious conclusions due to nonlinear effects between X and N. Dionne et al. (2001) recommend to use the following econometric model

$$l(Z \mid X, N) = l\Big(Z \mid aX + bN + c\widehat{E}(N \mid X)\Big), \tag{4}$$

where $\widehat{E}(N \mid X)$, denoting the conditionnal expectation of N given the variable X, will be estimated by a regression model initially. The introduction of the estimated expectation $\widehat{E}(N \mid X)$ allows to take into account nonlinear effects between X and N, yet not nonlinear effects with Z. We refer interested readers to Su & White (2003) and Huang (2009) for recent procedures of conditional independence testing in a nonparametric framework.

Summarizing the testing procedure, we have first a regression N on X to get $\widehat{E}(N \mid X)$. Secondly, we regress the decision variable Z on X, N, and $\widehat{E}(N \mid X)$. If the coefficient for N is significant in the second regression, then risk adverse selection is detected. The relevant choice for Z is the insured deductible choice, with X rating factors and N the observed number of claims. $\widehat{E}(N \mid X)$ will be estimated with a Poisson or more sophisticated models, see below.

5.2 A deductible model

The deductible choice takes values in the discrete set $\{d_0, d_1, \ldots, d_K\}$. The more general model is a multinomial model $\mathcal{M}(1, p_0, \ldots, p_K)$, where each probability parameter p_j depends on covariates through a link function. If we assume that variables Z_i are independent and identically distributed random variables from a multinomial distribution $\mathcal{M}(1, p_0, \ldots, p_K)$ and we use a logit link function, then the multinomial regression is defined by

$$P(Z_i = d_j) = \frac{e^{x_i^T \beta_j}}{1 + \sum_{l=1}^{K} e^{x_i^T \beta_l}},$$

for $j = 1, \ldots, K$ where 0 is the baseline category and x_i covariate for i th individual, see, e.g., McFadden (1981), Faraway (2006) for a comprehensive study of discrete choice modelling.

When reponses ($d_0 < d_1 < \ldots < d_K$) are ordered (as it is for deductibles), one can also use ordered logistic models for which

$$P(Z_i = d_j) = \frac{e^{\theta_j - x_i^T \beta}}{1+e^{\theta_j - x_i^T \beta}} - \frac{e^{\theta_{j-1} - x_i^T \beta}}{1+e^{\theta_{j-1} - x_i^T \beta}}.$$

Note that the number of parameters substantially decreases since the linear predictor for multinomial logit regression, we have $\eta_{ij} = x_i^T \beta_j$, whereas for the ordered logit, $\eta_{ij} = \theta_j - x_i^T \beta$.

The parameters θ, called thresholds, have a special interpretation since they link the response variable Z with a latent variable U by the equation $Z = d_k \Leftrightarrow \theta_{k-1} < U \leq \theta_k$. Hence, the trick to go from a Bernoulli model to a polytomous model is to have different ordered intercept coefficients θ_k's for the different categorical values.

As in Dionne et al. (2001), our choice goes to the ordered logit model for its simplicity. So Z is modelled by the following equation

$$P(Z_i \leq j \mid X_i, N_i) = g^{-1}\left(\theta_j + X_i^T \beta + N_i \gamma + \hat{E}(N \mid X_i)\delta\right),$$

for individual i and deductible j, with g^{-1} the logistic distribution function[1] and X_i exogeneous explanatory variables as opposed to endogeneous variables N_i. The parameters of this model equation are the regression coefficients β and γ and the threshold parameter θ_k's.

5.3 Application on the large dataset of Subsection 3.2

We want to test for evidence of adverse selection on the full comprehensive (FC) coverage product. So, we study in this subsection only the three datasets relative to that coverage. First, we model the claim number, and then we test for the asymmetry of information.

5.3.1 Modelling the claim number

Modelling count data in the generalized linear model framework can be done by choosing an appropriate distribution: the Poisson and overdispersed Poisson distribution, where the canonical link function is the logarithm. Since for a Poisson distribution $\mathcal{P}(\lambda)$, $P(N = 0) = e^{-\lambda}$, the GLM Poisson consists in assuming

[1] Note that in this form, it is easy to see that g^{-1} can be any distribution functions (e.g. normal or extreme value distributions).

$$E(N \mid x_i) = e^{x_i^T \beta} \Leftrightarrow -\log P(N = 0 \mid x_i) = x_i^T \beta.$$

where x_i denotes the covariates. In practice, this models suffers a subparametrization of the Poisson distribution, one single parameter.

One could think that the Negative binomial in an extended GLM[1] framework will tackle this issue, but in practice the mass in zero is so high, that both Poisson and negative binomial distributions are inappropriate. As presented in Table 13, the high number of zero-claim will compromise the good fit of regular discrete distributions.

Claim number	0	1	2	3	4	5	5 <
Frequency	43687	5308	667	94	17	2	38

Table 13: Claim number for Full Comp. agent subset

As presented in Zeileis et al. (2008) and the references therein, the issue is solved by using a zero-inflated distribution, e.g., a zero-inflated Poisson distribution. The mass probability function is given by

$$P(N = k) = \begin{cases} \pi & \text{if } k = 0, \\ (1-\pi)\dfrac{\lambda^k}{k!}e^{-\lambda} & \text{otherwise.} \end{cases}$$

Note that N is a mixture of a Bernoulli distribution $\mathcal{B}(\pi)$ with a Poisson distribution $\mathcal{P}(\lambda)$. The mean of the zero-inflated Poisson distribution is $(1-\pi)\lambda$. Using the GLM framework and the canonical link functions, a zero-inflated GLM Poisson model is defined as

$$E(N \mid x_i) = \frac{1}{1+e^{x_i^{1T}\gamma}} e^{x_i^{2T}\beta},$$

where the covariate vectors x_i^1, x_i^2 are parts of the vector x_i. Now there are two (vector) coefficients to estimate β and γ. The GLM is implemented in R base by the glm function. For the zero-inflated model, we need to use the **pscl** package, cf. Jackman (2011).

Still studying the FC agent dataset, we fit three distributions on the claim number: Poisson, zero-inflated Poisson and Negative binomial distributions. As shown in Table 18 in Appendix 8.6, the three models are similar in terms of log-likelihood or AIC. But, differences appear at the predictions.

[1] The negative binomial distribution does not belong to the exponential family, except if the shape parameter is known. So, the trick is to use a maximum likelihood procedure for that shape parameter at outer iteration whereas each inner iteration use a GLM fit given the current value of the shape parameter.

Despite being equivalent for first probabilities $P(X = 0,1,2)$, cf. Table 14, classic and zero-inflated Poisson distributions decrease too sharply compared to the observed number of claims. The negative Binomial distribution (fourth line) is far better. In Appendix 8.6, we give the regression summary for zero-inflated negative binomial distribution on the FC agent subset. We obtain the same conclusion for other FC subsets.

Claim number	0	1	2	3	4	5	6
Observed	43687	5308	667	94	17	2	2
Poisson	43337.9	5896.0	500.9	39.8	3.7	0.417	0.054
zeroinfl. Poisson	43677.6	5267.7	745.0	80.2	7.5	0.665	0.058
zeroinfl. NB	43704.6	5252.6	704.7	98.8	14.9	2.457	0.442

Table 14: Claim number prediction for Full Comp. agent subset

5.3.2 Testing for adverse selection

Now that we have modelled the claim frequency, we turn to the modelling of the deductible choice as described in the previous section: an ordered logistic model. We test for evidence of adverse selection on three datasets: agent, broker and direct with Full. Comp. products. Let us note that we cannot test adverse selection on TPL covers, since there is no deductible for this cover. As reported in Subsection 5.1, adverse selection testing is done by a fit of a GLM to explain the deductible choice Z_i. In addition to the exogeneous variables X_i for ith individual, the regression will use the observed claim number N_i (endogeneous) and its expected value coming from the zero-inflated negative binomial regression $\widehat{E}(N \mid X_i)$ (exogeneous).

The numerical illustrations reveal that it is more relevant to cluster some deductible values which are too few in the dataset. Actually, 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 choose to regroup deductible values greater than 500 together. In Table 15, we report the proportion of customers by deductible value for the first two datasets. Small 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 15: Frequency table for Full Comp. deductibles values

As shown in Appendix 8.7 for FC agent subset, the endogeneous variable N_i is not statistically significant despite being negative, i.e. the higher the loss number, the lower the deductible. But the expected value $\widehat{E}(N \mid X_i)$ is significant. For the two other FC datasets, both coefficients for N_i and $\widehat{E}(N \mid X_i)$ are not significant, but these datasets are also smaller in size. We conclude that there is no adverse selection for FC datasets.

After removing insignificant variables in the deductible regression, we integrate the deductible choice predicted probabilities to the lapse regression (Y). Let Z_i denote the deductible for the ith individual, we incorporate fitted probabilities $\hat{P}(Z_i = 0)$, $\hat{P}(Z_i = 150)$ and $\hat{P}(Z_i = 500+)$. We choose to consider 300 euros as the baseline category, as 300-euro deductible is the standard ``unchosen" deductible. For the FC agent dataset, the three probabilities, $\hat{P}(Z_i = 0)$, $\hat{P}(Z_i = 150)$ and $\hat{P}(Z_i = 500+)$, are significant, see Appendix 8.7, whereas for the two other FC datasets some probabilities are not significant. We perform the usual predictions for the lapse rate (-5%, 0% and +5% for the proposed premium). But we do not present here the lapse rate predictions since predictions are almost unchanged[1].

5.4 Conclusion

This section shows how to use GLM modelling to test for evidence of adverse selection. In our dataset, no adverse selection is detected. The inclusion of deductible choice probability neither improves the lapse predictions nor helps in understanding the lapse decision at aggregate level. But we believe that the deductible choice (especially non standard ones) by a customer plays a major role in the propensy of lapse when renewing its policy. Low-risk invididuals, i.e. with high deductibles, are likely to be the most sensitive customers, unlike to high-risk individuals.

6. OTHER REGRESSION MODELS

This section presents other regression models. There are mainly two (static) extensions to GLMs in two directions: (i) additive models where the linear predictor is composed of smooth terms and (ii) mixed models where we add a random term (as opposed to fixed term, i.e. deterministic). These two extensions are available for the exponential family distribution, leading to generalized additive models and generalized linear mixed models, respectively. In this paper, we discard mixed models as they are inefficient in our

[1] difference less than 0.1% pt.

context. The first subsection introduces generalized additive models, and then the second subsection is devoted to an application. The last subsection details other regression models than generalized additive models.

6.1 Model presentation

The Generalized Additive Models (GAM) were introduced by Hastie & Tibshirani (1990) by unifying generalized linear models and additive models. So, GAMs combine 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.

6.1.1 Theoretical presentation

In this subsection, we present Generalized Additive Models (GAM) in two steps: from linear to additive models and then from additive to generalized additive models. Fitting algorithms are then briefly presented, whereas smoothing techniques are detailed in Hastie & Tibshirani (1990) and Venables & Ripley (2002). Finally, we apply GAMs on the large dataset of Subsection 3.2.

Assuming observations X_i and response variables Y_i are identically and independently distributed random variables having the same distribution of generic random variables X and Y, respectively. Linear models assume by definition a linear relationship between X and Y motivated by mathematical tractability rather than empirical evidence. One candidate to extend linear models is the additive model for which the relation between X and Y is modelled by a smooth function.

Thus, 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 provides a smooth 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 invertible, 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), f_j's smooth functions and S a set of possible values of the expectation $E(Y_i)$. Note that linear models (and GLMs) are special

cases of additive models (and GAMs) with $f_j(x) = \beta_j x$.

We present here only the main idea of fitting algorithms and do not go into details, see Hastie & Tibshirani (1990) or Venables & Ripley (2002) for details. All smoothers have a smoothing parameter λ (the polynom degree, the bandwidth or the span). A first concern is how to choose a criterion on which to optimize λ (hence to have an automatic selection). Then, a second concern is to find a reliable estimate of the parameters α and smooths coefficients given a smoothing value λ.

We present the procedure in the reverse way. Assuming a value of λ, we present an algorithm to fit the model. Hastie & Tibshirani (1990) propose a local averaging generalized Fisher scoring method. However, Wood (2008) proposes a recent and reliable method: the Penalized Iteratively Reweighted Least Square method (PIRLS). The PIRLS is (unsurprisingly) an iterative method aiming to minimize the penalized deviance

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

where the second term penalizes the wiggly behavior of smooth functions.

Given a set of basis functions $(b_{jk})_{jk}$, we can express the smooth function f_j as $f_j(x) = \sum_{k=1}^{K_j} \beta_{jk} b_{jk}(x)$. So, in the end, 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 linear parameter α and coefficients β_{jk}'s. Thus, the first term is fully determined. Hence, the penalized deviance is given by

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

where $S_j = (b_{jk}(x_i))_{ik}$ contains known coefficients and $D(\beta)$ the GLM version of the deviance for the fixed-basis GAM model. See Wood (2008) for details on how the PIRLS algorithm solves the problem $\min \widetilde{D}(\beta)$.

The PIRLS algorithm gives for any λ the corresponding fitted coefficient $\hat{\beta}(\lambda)$, i.e. smooth functions $\hat{f}_j$. Now, we must find a criterion to select the appropriate vector λ. We cannot choose the smoothing parameter λ as the parameter minimizing the deviance, because the model will overfit the data. In the literature, there are many criteria to select the smoothing parameter: likelihood measures such as Restricted Maximum Likelihood (REML), Maximum Likelihood (ML) and cross validation measures such as Generalized Cross Validation (GCV), Generalized Approximate Cross Validation (GACV). These

methods differ whether 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-likelihood criterion (ML and REML) 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, see Wood (2010) for details.

An alternative approach seeks in minimizing the prediction error. The predictive error may seem difficult to assess, but the trick is to use a leave-one-out procedure. It 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 approximation is used to compute the GCV or GACV. Then again, a nested optimization procedure using the PIRLS scheme is used.

We test GCV and REML criteria with different polynomial bases on simple examples and conclude that the criterion and the choice of polynomial basis have few impact on the final model. Thus, in the following, we use the REML criterion to determine the appropriate λ and thin plate basis regression. The thin plate regression uses a basis of thin plate (also known as polyharmonic functions) functions $\phi_{md}(r) = \alpha_{md} r^{2m-d} \log(r)$ if d is even and $\alpha_{md} r^{2m-d}$ if d is odd. The smooth function is defined as $s(x) = \sum_{i=1}^{n} \delta_i \phi_{md}(\| x - x_i \|_2)$. A low-rank approximation of this smooth function is then used to decrease the computational burden. This method avoids the knot placement problems of traditional regression spline models.

6.1.2 Binary regression and model selection

As for GLMs, the binary regression means we assume that Y_i follows a Bernoulli distribution $\mathcal{B}(\pi_i)$, π_i being linked to explanatory variables. So, 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:

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

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 and are longer to compute (than linear terms). For instance, if a covariate represents the date or the time of events, it is "often" better to consider the effect 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 distribution of the estimators. The variable selection for GAMs is similar to those of GLMs. The true improvement is a 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 a term is dropped only if no smooth function and no linear function with this term is relevant. That is to say, a poor significance of a variable modelled by a smooth function might be significant when modelled by a single linear term.

We use the following acceptance rules of Wood (2001) to drop an explanatory variable:

- Is the estimated degrees of freedom for the term close to 1?
- Does the plotted confidence interval band for the term include zero everywhere?
- 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 GAMs, we use the recommended R package **mgcv** written by S. Wood.

6.2 Application to the large dataset

In Section 3.2, the GLM analysis of this large 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 what we could expect. Moreover, the price sensitivity gap measured by the lapse deltas is also high. Let us see this it still holds with GAM results.

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 8.8, we provide the regression summary for one of the nine subsets.

6.2.1 Comments on regression summary

In this subsection, we briefly comment on the nine regression summaries. Let us start with the Third-Part Liability cover. For the agent subset, for which we have a market proxy, we keep four non linear terms (premium difference variables and car class) all modelled jointly with the price ratio. We try to model these terms independently of price ratio, but this was worse in terms of REML scores. On the broker subset, we keep two non linear terms (difference to technical premium and vehicle age). Only the first term is modelled jointly with the price ratio, because the second term has a linear effect with the price ratio. Due to a small size, the direct subset was hard to handle with a GAM. We restrict the price ratio to be a smooth term of small order. This dataset also shows some strange results with a negative elasticity for small premium increase.

Studying Partial Comprehensive coverage is also challenging. For the agent subset, 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 sensitivity is explained by linear terms. As for the TPL covers, the same variables are modelled non linearly for the broker subset, jointly with the price ratio. The high estimated degrees of freedoms emphasize this non linearity. Turning to the direct channel, only the difference to technical premium variable is modelled through a smooth function, jointly with the price ratio.

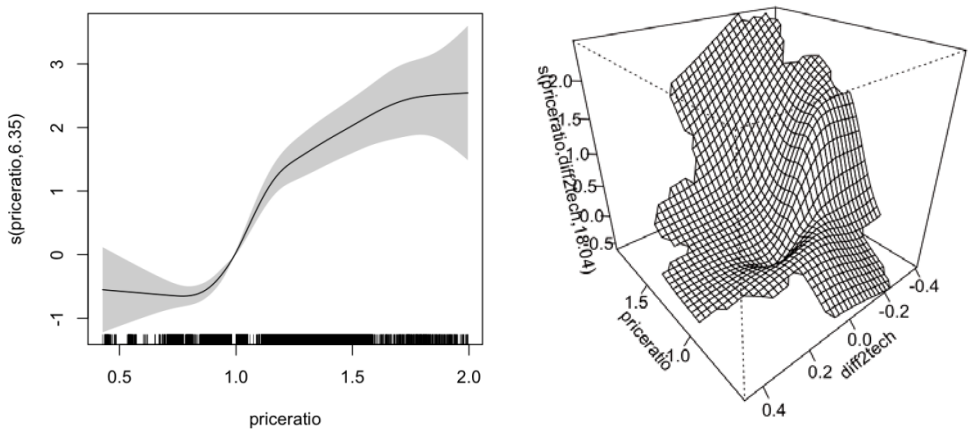
Finally, we study the Full Comprehensive coverage product. As always, the agent subset has many nonlinear terms. Three terms (driver age, difference to technical premium and car class) are smoothed together with the price ratio. Again, the estimated degrees of freedom are high, especially for the difference to technical premium variable. Regarding the broker subset, four terms (driver age, vehicle age, difference to technical premium and car class) are modelled non linearly. We retrieve 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. Ending with the direct subset, only two terms are

modelled non linearly (the driver age, difference to technical premium): 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.

6.2.2 Examples of fitted smooth functions

In the preceding analysis, we observe some trends between channel distributions. Notably, the broker channel results are more sensitive to the difference with technical premium and the vehicle age variables than the other two channels. There is also a data size effect, since the data sets gradually increase in size from TPL and PC to FC covers. Of course, the more we have data, the more the regression is reliable.

On Figure 2, we plot two fitted smooth functions from two different GAM regressions[1]. Figure 1 represents the smooth function for the price ratio variable of the PC-agent regression. We observe that the smooth function is highly non linear, i.e. a high degree of freedom of 6.35. The smooth function features a very sharp increase of the price ratio around 1: such steep increase is not possible with a linear predictor.



a) PC agent - price ratio smooth function b) FC broker - bivariate smooth function

Figure 2: GAM smooth functions

Figure 1 is the plot of the bivariate smooth function of the price ratio and the difference to technical premium variable for FC broker dataset. There is a small hollow in the curve around the point (1, 0), a price ratio of 1 and a zero difference with technical

[1] The grey area represents the standard error bandwidth around the smooth function. It is standard to use an area rather than two simples curves for the confidence interval: this suggests smooth functions lies in such area.

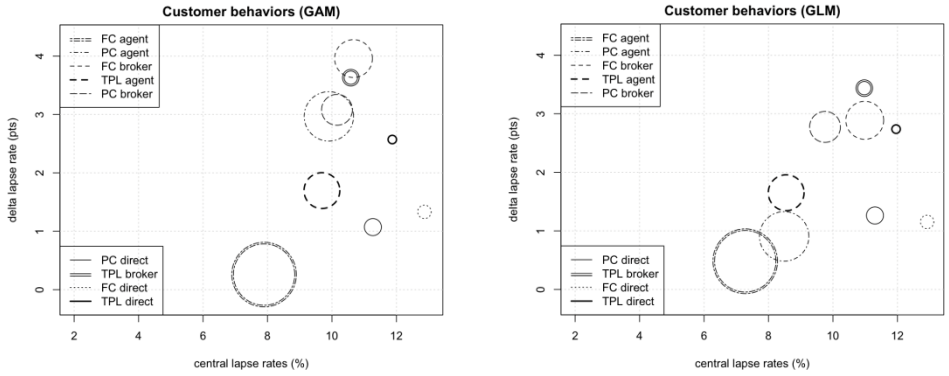
premium. Locally, the price elasticity of the lapse decision is negative. Fortunately, this business inconsistency is small and located. If we had market variables for this dataset, it could be of interest to check whether this anomaly vanishes.

6.2.3 Discussion on predictions

As for the GLM analysis, we turn to the analysis of the distribution channel and the coverage type by looking at the lapse rate predictions. We also consider an average lapse rate function defined as

$$\hat{\pi}_n(p) = \frac{1}{n}\sum_{i=1}^{n} g^{-1}\left(\hat{\mu} + x_i(p)^T \hat{\beta}_{-p} + z_i(p)^T \hat{\beta}_{+p} \times p + \sum_{j=1}^{p} \hat{f}_j(\tilde{z}_i(p), p)\right), \quad (5)$$

where $(\hat{\mu}, \hat{\beta}_{-p}, \hat{\beta}_{+p})$ are the fitted parameters, $\hat{f}_j$ are the fitted smooth functions, $(x_i, z_i, \tilde{z}_i)$ are parts of explanatory variables of the ith individual and g is the logit link function. What differentiates Equation (5) with Equation (1) is the inclusion of additive terms in the predictor.



a) GAM b) GLM

Figure 3: GAM vs. GLM - comparison of distribution channels and cover types

On Figure 3, we plot the usual bubble plot to compare GAMs and GLMs. We observe that GAM delta lapse rate 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. Only two channel-covers have a lower delta lapse rate $\Delta_{1+}(5\%)$ with GAMs: the FC 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.

In terms of central lapse rates, most of predictions $\hat{\pi}_n(1)$ are higher, i.e. shift to the right on Figure 3. It means that the customers in the portfolio are more price-sensitive even if we propose exactly the same premium as last year. On a private motor insurance, most people expect a better bonus-malus from year to another, hence a premium decrease.

Now, we stop the GAM analysis and conclude on the pros and cons of GAMs. GAMs are less known tools than GLMs in actuarial science. But since their introduction in the 90's, GAMs are well studied and use state-of-the-art fitting procedures. There are two ways to perform model selections: prediction errors vs. likelihoods. In this paper, we follow the Wood's rule to select variables based on the restricted maximum likelihood. We tested other statistical quantities, but the impact remains limited.

As for GLMs, GAMs allow us to assess an overall estimated price elasticity (via $\hat{\pi}_n(1)$ and $\Delta_{1+}(5\%)$) taking into account the individual features of each policy. The additional complexity coming with additive modelling compared to GLMs permit to really fit the data. Especially for broker lines, we get a more cautious view of customer price sensitivity. For small datasets, GAM predictions may lead to irrelevant results. Furthermore, as already noticed for GLMs, GAMs predictions are reliable for with a small range of price change: extrapolating outside observed price ratio range leads to doubtful results.

Finally, GAMs need a longer time to fit than GLMs and require a better computing power. This is a limitation for GAMs to be used easily by everyone. In addition, some user judgement is needed to select, to linearize or to reject explanatory variables in order to get the final model for GAMs. Even with Wood's rules, newcomers may find it hard to choose between two GAM models with the same ``score'', i.e. with the same likelihood or prediction errors.

6.3 Other regression models

GLMs and GAMs are static models. One option to take into account dynamics could have been to use time serie models on regression coefficients of GLMs. But this was impossible with our datasets due to a limited number of years and it is rather a trick than an appropriate solution. 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 to deal with heterogeneity across time.

Among many others, Frees (2004) presents GLMMs in the context of longitudinal and panel data. Since a panel data model cannot deal with right-censoring (that occurs when a policy is terminated), they are not appropriate to our policy termination problem, i.e. lapse. Despite discarding GLMMs for dynamic lapse modelling, we try to use the GLMMs on one period in order to model endogeneous effects such as dropping coverage with a random term. Unfortunately, the application of GLMMs to our lapse problem reveals inefficient on our datasets.

The Survival Regression Model of Cox (1972) allow to remove the inherent limits of the static regression models previously presented. By nature, they take into account the dynamic aspects of the response variable considering it as a lifetime variable. In our context, the lapse decision model can be expressed as the lifetime of an insurance policy, i.e. lapse means the termination of the policy for a given policy age. Thus, the dataset must contain the policy age if one wants to use the basic Cox model. Furthermore, if one allows explanatory variables to vary over time (i.e. extended Cox model), we need to observe multiple times the customer choices and the explanatory variables. Typically the dataset will look like below

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
			⋮			

where information for individual #2 is surrounded. As GLMs and GAMs demonstrate, renewing a policy for the first time is not motivated by the same factors as renewing one for the tenth time. This will remain true for survival regression models, see Brockett et al. (2008) and (Dutang, 2011, Chap. 4) for an application of such models.

The full power of survival models is not only to model one lapse reason. Other policy termination factors can be integrated so as to model the complete life cycle of a policy. With a full picture integrating other cash flows such as claims, and premiums, insurance risk could also be better assessed. Further advanced models than the Cox model regression exists, such as state-space models, e.g., Fahrmeir (1994) or stochastic counting

processes, see, e.g., Andersen et al. (1995), Aalen et al. (2008). Some attempts have been done to use Fahrmeir (1994)'s state space model, but the fitting process was too heavy to be quickly used.

7. CONCLUSION

Fitting price-sensitivity is a complex topic. Being dependent on the market's environment, price elasticity forecasts require rigorous attention to details to prevent the risk of erroneous conclusions. Not surprisingly, a data cleaning process is essential prior to any regression fitting. In short, some supplied explanatory variables substantially affect the results. 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. In Section 3, the small dataset only provides the driver age: this example leads to inconclusive results. On the large dataset, the coverage type, and the cross-selling indicators were added to the regression fit. This enabled us to refine our analysis. Having or not having a household policy with the same insurer was thus proven to be a driving factor in renewing or allowing a contract to lapse.

However, fully reliable predictions are only achieved when the rebate level and market premium proxies are used. In Section 4, the price sensitivity fit was considerably enhanced, along with our ability to fine tune the results, thanks to the inclusion of distribution channels, a market proxy, and a rebate level. With the gradual addition of explanatory variables, we have seen an increased accuracy of the lapse rate predictions. 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. If those market proxies are no longer available, we are likely to get back to less meaningful results.

Adverse selection resulting from an asymmetry of information is a widely known risk in insurance. Section 5 investigates for empirical evidence of adverse selection and studies its relationship to the lapse decision of customers. On our large dataset, no adverse selection is detected. At aggregate level, adverse selection does not have a big influence. Nevertheless, at individual level, choosing a non-standard deductible when underwriting a new policy will certainly have consequences on the termination of this policy.

Generalized Linear Models are widely known and respected methods in non-life

insurance. However, they have some inherent constraints with GLMs. Thus, in Section 6, we test Generalized Additive Models, which allow for non linear terms in the predictor. Like GLMs, the quality of the findings attained is directly related to the data provided. Using limited variables will produce approximate results, whereas, dealing with an extensive set of variables lead to proven results.

Applying GAMs, despite their additional complexity, can be justified in cases where GLMs fail to provide realistic lapse predictions and we have substantial datasets. Note 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.

In this paper, we have explored the price elasticity topic from various viewpoints. Once again, our research has further demonstrated that the quality of data used in actuarial studies unequivocally affects the findings reached. In addition, the key role of the market proxies in estimating price sensitivity has been established. Market competition modelling, see, e.g., Demgne (2010), Dutang et al. (2012), is therefore relevant.

The conclusions drawn from customer price sensitivity 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 sensitivity 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.

ACKNOWLEDGEMENTS

The author is indebted to Stéphane Loisel, Véronique Maume-Deschamps and Ragnar Norberg, who suggested numerous corrections to earlier drafts of this manuscript. The remaining errors, of course, should be attributed to the author alone. This work is partially funded by the Swiss National Science Foundation Project 200021-124635/1.

REFERENCES

AALEN, O., BORGAN, O. & GJESSING, H. (2008), *Survival and Event History Analysis*, Springer.

ANDERSEN, P., BORGAN, O., GILL, R. & KEIDING, N. (1995), *Statistical Models Based on Counting Processes*, Springer, Corrected Edition.

ATKINS, D. C. & GALLOP, R. J. (2007), 'Re-thinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models', *Journal of Family Psychology* **21**(4), 726–735.

BELLA, M. & BARONE, G. (2004), 'Price-elasticity based on customer segmentation in the italian auto insurance market', *Journal of Targeting, Measurement and Analysis for Marketing* **13**(1), 21–31.

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'.

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.

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

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

COX, D. R. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society*: *Series B* **34**(2), 187–200.

CUMMINS, J. D. & VENARD, B. (2007), *Handbook of international insurance*, Springer.

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

DEMGNE, E. J. (2010), Etude des cycles de réassurance, Master's thesis, ENSAE.

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.

DUTANG, C. (2011), Regression models of price elasticity in non-life insurance, Master's thesis, ISFA.

DUTANG, C., ALBRECHER, H. & LOISEL, S. (2012), A game to model non-life insurance market cycles. Working paper, ISFA.

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

FARAWAY, J. J. (2006), *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Parametric Regression Models*, CRC Taylor& Francis.

FREES, E. W. (2004), *Longitudinal and Panel Data*, Cambridge University Press.

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.

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.

HASTIE, T. J. & TIBSHIRANI, R. J. (1990), *Generalized Additive Models*, Chapman and Hall.

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

HUANG, M. (2009), Essays on testing conditional independence, PhD thesis, UC Sand Diego.

JACKMAN, S. (2011), pscl: *Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University*, Department of Political Science, Stanford University. R package version 1.04.1.

KAGRAOKA, Y. (2005), Modeling insurance surrenders by the negative binomial model. Working Paper 2005.

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'.

KIM, C. (2005), 'Modeling surrender and lapse rates with economic variables', *North American Actuarial Journal* **9**(4), 56–70.

LOISEL, S. & MILHAUD, X. (2011), 'From deterministic to stochastic surrender risk models: Impact of correlation crises on economic capital', *European Journal of Operational Research* **214**(2).

MCCULLAGH, P. & NELDER, J. A. (1989), *Generalized Linear Models*, 2nd edn, Chapman and Hall.

MCFADDEN, D. (1981), *Econometric Models of Probabilistic Choice*, The MIT Press, chapter 5.

MILHAUD, X., MAUME-DESCHAMPS, V. & LOISEL, S. (2011), 'Surrender triggers in Life Insurance: What main features affect the surrender behavior in a classical economic context?', *Bulletin Français d'Actuariat* **22**(11).

NELDER, J. A. & WEDDERBURN, R. W. M. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society* **135**(3), 370–384.

OHLSSON, E. & JOHANSSON, B. (2010), *Non-Life Insurance Pricing with Generalized Linear Models*, Springer.

R Core Team (2012), R: *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org

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.

SHAPIRO, A. F. & JAIN, L. C. (2003), *Intelligent and Other Computational Techniques in Insurance*, World Scientific Publishing.

SU, L. & WHITE, H. (2003), Testing conditional independence via empirical likelihood. UCSD Department of Economics Discussion Paper.

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

VENABLES, W. N. & RIPLEY, B. D. (2002), *Modern Applied Statistics with S*, 4th edn, Springer.

WOOD, S. N. (2001), 'mgcv: GAMs and Generalized Ridge Regression for R', R News 1, 20–25.

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).

WOOD, S. N. (2010), 'Fast stable reml and ml estimation of semiparametric glms', *Journal of the Royal Statistical Society: Series B* **73**(1), 3–36.

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.

ZEILEIS, A., KLEIBER, C. & JACKMAN, S. (2008), 'Regression models for count data in r', *Journal of Statistical Software* **27**(8).

8. APPENDIX

8.1 R outputs

8.1.1 Bronchitis dataset

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, see Appendix 8.1. Let us first regress the bronchitis indicator on all variables

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

with a logit link function. The regression summary is given below

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

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

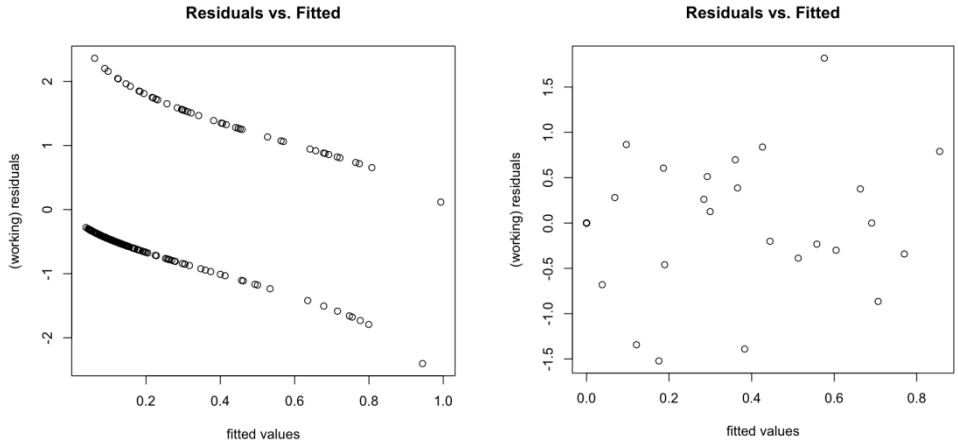
Null deviance: 221.78 on 211 degrees of freedom

Residual deviance: 174.21 on 209 degrees of freedom - AIC: 180.21

So the GLM fit seems good because all variables (including intercept) are significant with a very low p-value. However the plot of residuals[1] (see Figure 3) against fitted values[1]

[1] Working residuals are $\hat{\varepsilon}_i = Y_i - \hat{\pi}_i$. Note that using other residual types, Pearson, Studentized, do not change this behavior.

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 4: 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 Figure 3. So for the same data, with the same (significant) variables, the two analyses of residuals lead to different conclusions. Hence, conclusions of residual analysis must be taken with great care.

8.1.2 GLM outputs of Section 3.1

See below the summary table with coefficients values, standard errors, z-statistics and p-value. For confidentiality reason, all the deviance and AIC statistics shown in this paper have been scaled by the same positive coefficient. Here follows the regression summary when variables are categorical.

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

[1] Fitted values are $\hat{\pi}_i$.

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	***
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	1.018771	0.120903	8.426	< 2e-16	***
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
Null deviance: 53978 on 56026 degrees of freedom
Residual deviance: 53258 on 56012 degrees of freedom - AIC: 53261

8.1.3 Variable list for Subsection 3.2

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

- Policy: a dummy variable indicating the lapse, the policy age, the cover type (TPL, PC or FC) and the product, the bonus 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,
- Cross-selling: the number of insurer contracts 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.

8.1.4 GLM outputs for Subsection 3.2

The regression summary is given below

Call: glm(formula = lapse ~ lastprem_group2 + diff2tech + directdebit + product + nbclaim0708percust + vehiclage + householdNbPol + polholderage + maritalstatus2 + jobgroup2 + gender + polage + bonusevol2 + cover + priceratio:(lastprem_group2 + diff2tech + paymentfreq + glasscover + region2 + nbclaim08percust + householdNbPol + diffdriverPH7 + channel + typeclassTPL + bonusevol2), family = binomial("logit"), data = idata)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.1241	-0.4366	-0.3427	-0.2402	3.3497

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.6456876	0.1822517	-14.517	< 2e-16	***
lastprem_group2(500,5e+03]	0.2008839	0.0952157	2.110	0.034878	*
diff2tech	6.9600797	0.7949370	8.756	< 2e-16	***
directdebit	-0.0422104	0.0097823	-4.315	1.60e-05	***
productT1	-0.1060909	0.0185019	-5.734	9.80e-09	***
productT2	-1.0107703	0.0336376	-30.049	< 2e-16	***
productT3	-0.3869057	0.0193135	-20.033	< 2e-16	***
nbclaim0708percust	0.0802148	0.0061759	12.988	< 2e-16	***
vehiclage	-0.0172387	0.0010180	-16.934	< 2e-16	***
householdNbPol	-0.1638354	0.0156899	-10.442	< 2e-16	***
polholderage	-0.0106258	0.0003000	-35.417	< 2e-16	***
maritalstatus2b	-0.1455813	0.0266586	-5.461	4.74e-08	***
maritalstatus2d	-0.1088016	0.0119736	-9.087	< 2e-16	***
jobgroup2public	-0.1529926	0.0079183	-19.321	< 2e-16	***
gender	-0.0739520	0.0077666	-9.522	< 2e-16	***
polage	-0.0245842	0.0006806	-36.123	< 2e-16	***
bonusevol2up-down	1.9010618	0.1746998	10.882	< 2e-16	***
coverpartial compr.	0.0244814	0.0099107	2.470	0.013504	*
coverTPL	-0.0349025	0.0131839	-2.647	0.008112	**
priceratio:lastprem_group2(0,500]	1.0418939	0.1840274	5.662	1.50e-08	***
priceratio:lastprem_group2(500,5e+03]	1.0246974	0.2000580	5.122	3.02e-07	***
priceratio:diff2tech	-8.7933934	0.7867136	-11.177	< 2e-16	***
priceratio:paymentfreq	-0.0136538	0.0010577	-12.909	< 2e-16	***
priceratio:glasscover	-0.0865708	0.0139001	-6.228	4.72e-10	***
priceratio:region2_02-04-05-11	0.3608514	0.0207136	17.421	< 2e-16	***
priceratio:region2_03-09-10	0.1368317	0.0109978	12.442	< 2e-16	***
priceratio:region2_04-05-06-07	0.0935641	0.0103280	9.059	< 2e-16	***
priceratio:region2_12-13	0.3938396	0.0166819	23.609	< 2e-16	***
priceratio:region2_14-15-16	0.4424354	0.0160587	27.551	< 2e-16	***
priceratio:region2_17_	0.4812002	0.0243385	19.771	< 2e-16	***
priceratio:nbclaim08percust	-0.0374916	0.0102707	-3.650	0.000262	***
priceratio:householdNbPol	0.0794544	0.0157004	5.061	4.18e-07	***
priceratio:diffdriverPH7learner 17	0.2768748	0.0578518	4.786	1.70e-06	***
priceratio:diffdriverPH7only partner	0.0976821	0.0077879	12.543	< 2e-16	***
priceratio:diffdriverPH7young drivers	0.1684370	0.0148135	11.371	< 2e-16	***
priceratio:channelbroker	0.3954067	0.0089064	44.396	< 2e-16	***
priceratio:channeldirect	0.3715832	0.0132034	28.143	< 2e-16	***

priceratio:typeclassTPL	0.0108773	0.0016963	6.412	1.43e-10	***
bonusevol2up-down:priceratio	-1.8295464	0.1740807	-10.510	< 2e-16	***

--- Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
Null deviance: 62279 on 121813 degrees of freedom
Residual deviance: 58739 on 121809 degrees of freedom - AIC: 58747

Group j	Observed r_j	Fitted $\frac{1}{n_j}\sum_{i=1}^{n_j}\hat{\pi}_i(p_i)$	Group j	Observed r_j	Fitted $\frac{1}{n_j}\sum_{i=1}^{n_j}\hat{\pi}_i(p_i)$
Agent	8.840	7.714	FC	8.962	7.492
Broker	9.245	8.896	PC	9.464	8.846
Direct	11.837	9.005	TPL	10.222	12.522

Table 16: Lapse rates (%)

	$\Delta_{1-}(5\%)$	$\hat{\pi}_n(1)$	$\Delta_{1+}(5\%)$	$\Delta_{1-}(5\%)$	$\hat{\pi}_n(1)$	$\Delta_{1+}(5\%)$
Channel agent	-0.983	8.652	1.23	-0.759	8.732	0.75
Channel broker	-1.344	9.123	1.841	-1.255	9.422	1.299
Channel direct	-1.246	12.341	1.143	-1.18	11.597	1.268
Channel	One fit by channel			One fit for all channels		
	$\Delta_{1-}(5\%)$	$\hat{\pi}_n(1)$	$\Delta_{1+}(5\%)$	$\Delta_{1-}(5\%)$	$\hat{\pi}_n(1)$	$\Delta_{1+}(5\%)$
Coverage FC	-0.926	8.297	1.01	-0.622	8.723	0.97
Coverage PC	-0.635	9.347	1.195	-0.714	9.244	1.063
Coverage TPL	-0.973	12.011	1.876	-0.899	10.179	1.178
Coverage	One fit by coverage			One fit for all coverages		

Table 17: Predicted lapse rates by channel and coverage

8.1.5 GLM outputs for Subsection 4.2

The regression summary without using the market proxy is given below.
Call: glm(formula = lapse ~ diff2tech + product2 + region2 + cumulrebate3 + nbclaim0608percust + isinsuredinhealth + isinsuredinlife + vehiclage + householdNbPol + polholderage + maritalstatus2 + jobgroup2 + gender + typeclassTPL + bonusevol2 + priceratio:(diff2tech + paymentfreq + nbclaim08percust + nbclaim0608percust + nbclaim0708percust + isinsuredinaccident + householdNbPol + gender + typeclassTPL + bonusevol2), family = binomial("logit"), data = idata)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.2613	-0.4104	-0.3482	-0.2792	3.1127

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.3513224	0.1034727	-13.060	< 2e-16	***
diff2tech	7.8972018	1.4461272	5.461	4.74e-08	***
product2T1	-0.1275087	0.0321359	-3.968	7.25e-05	***

product2T2	-0.2762145	0.0348857	-7.918	2.42e-15	***
region2_02-04-11	0.2886433	0.0427885	6.746	1.52e-11	***
region2_05	0.1878357	0.0277600	6.766	1.32e-11	***
region2_08-09	0.0661201	0.0259573	2.547	0.010857	*
region2_10	0.4506006	0.0906820	4.969	6.73e-07	***
region2_12-13	0.3729663	0.0404406	9.223	< 2e-16	***
region2_14-15-16	0.4591227	0.0406760	11.287	< 2e-16	***
region2_17	0.4469127	0.0609890	7.328	2.34e-13	***
cumulrebate3	0.0131512	0.0220328	0.597	0.550581	
nbclaim0608percust	0.2538161	0.0861386	2.947	0.003213	**
isinsuredinhealth	-0.2117021	0.0737189	-2.872	0.004082	**
isinsuredinlife	-0.0904838	0.0403864	-2.240	0.025061	*
vehiclage	-0.0418472	0.0024594	-17.015	< 2e-16	***
householdNbPol	-0.1608386	0.0347312	-4.631	3.64e-06	***
polholderage	-0.0142367	0.0007987	-17.824	< 2e-16	***
maritalstatus2b	-0.2473493	0.0756033	-3.272	0.001069	**
maritalstatus2d	-0.1026557	0.0339761	-3.021	0.002516	**
jobgroup2public	-0.1564253	0.0212887	-7.348	2.01e-13	***
gender	-0.8573031	0.1748974	-4.902	9.50e-07	***
typeclassTPL	-0.1127455	0.0320514	-3.518	0.000435	***
bonusevol2up-down	3.5129944	0.6064173	5.793	6.91e-09	***
priceratio:diff2tech	-8.7833478	1.4474939	-6.068	1.30e-09	***
priceratio:paymentfreq	-0.0314041	0.0025894	-12.128	< 2e-16	***
priceratio:nbclaim08percust	-0.1047064	0.0383473	-2.730	0.006324	**
priceratio:nbclaim0608percust	-0.2269052	0.0913726	-2.483	0.013017	*
priceratio:nbclaim0708percust	0.1429228	0.0365854	3.907	9.36e-05	***
priceratio:isinsuredinaccident	-0.1395317	0.0505194	-2.762	0.005746	**
priceratio:householdNbPol	0.0817417	0.0347087	2.355	0.018519	*
priceratio:gender	0.7813407	0.1758044	4.444	8.81e-06	***
priceratio:typeclassTPL	0.1300911	0.0320887	4.054	5.03e-05	***
priceratio:bonusevol2up-down	-3.3300573	0.6048578	-5.506	3.68e-08	***

Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
Null deviance: 9151 on 18893 degrees of freedom
Residual deviance: 8866 on 18860 degrees of freedom - AIC: 8873

8.1.6 GLM outputs for Subsection 5.3.1

	Poisson	zeroinfl. Poisson	zeroinfl. NB
$\log \mathcal{L}$	-27571	-28372	-28105
AIC	45197	46797	46258
Deg. of free.	27	26	26

Table 18: Model adequacy for claim frequency of FC agent

Here follows the regression summary for zero-inflated NB distribution fit.
Call: zeroinfl(formula = nbclaim08FC ~ bonuspercentnew + bonusevol2 + lastprem_group2 + isinsuredinhealth + isinsuredinlife + isinsuredinaccident + polage + vehiclage + polholderage + typeclassFC + diffdriverPH2 + gender | lastprem_group2 + diff2tech + isinsuredinaccident + polage + polholderage, data = subdata, dist = "negbin")

Pearson residuals:

Min	1Q	Median	3Q	Max
-0.6907	-0.3701	-0.3263	-0.2836	27.6615

Count model coefficients (negbin with log link):

	Estimate	Std. Error	zvalue	Pr(>|z|)	
(Intercept)	-2.5053555	0.0463173	-54.091	< 2e-16	***
bonuspercentnew	-0.0045481	0.0004473	-10.168	< 2e-16	***
bonusevol2up-down	0.2814031	0.0108215	26.004	< 2e-16	***
lastprem_group2(500,5e+03]	0.2867385	0.0125864	22.782	< 2e-16	***
isinsuredinhealth	0.2536512	0.0129962	19.517	< 2e-16	***
isinsuredinlife	0.1500995	0.0101994	14.716	< 2e-16	***
isinsuredinaccident	0.1545091	0.0132603	11.652	< 2e-16	***
polage	-0.0045662	0.0008071	-5.657	1.54e-08	***
vehiclage	-0.0116381	0.0012641	-9.207	< 2e-16	***
polholderage	0.0052154	0.0006398	8.152	3.59e-16	***
typeclassFC	0.0259947	0.0012908	20.139	< 2e-16	***
diffdriverPH2all drivers > 24	0.1603390	0.0110572	14.501	< 2e-16	***
diffdriverPH2commercial	0.5143316	0.0338102	15.212	< 2e-16	***
diffdriverPH2learner 17	0.2501158	0.0642750	3.891	9.97e-05	***
diffdriverPH2same	-0.1661160	0.0111876	-14.848	< 2e-16	***
diffdriverPH2young drivers	0.2524112	0.0158128	15.962	< 2e-16	***
gender	-0.0593577	0.0088454	-6.711	1.94e-11	***
Log(theta)	0.2848294	0.0330418	8.620	< 2e-16	***

Zero-inflation model coefficients (binomial with logit link):

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-7.299505	0.367536	-19.861	< 2e-16	***
lastprem_group2(500,5e+03]	-0.484487	0.081025	-5.979	2.24e-09	***
diff2tech	-7.214606	0.562964	-12.815	< 2e-16	***
isinsuredinaccident	-0.256634	0.098848	-2.596	0.00942	**
polage	-0.011704	0.004260	-2.747	0.00601	**
polholderage	0.094674	0.004658	20.326	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Theta = 1.3295
Number of iterations in BFGS optimization: 77
Log-likelihood: -2.81e+04 on 24 Df

8.1.7 GLM outputs for Subsection 5.3.2

We give the regression summary for ordered logistic regression for FC agent subset. The GLM regression summary for lapse on the FC agent subset including deductible choice probabilities is available on request to the author.

Call: polr(formula = deductibleFC3 ~ nbclaim08FC + ClaimNBhat + bonuspercentnew + lastprem_group2 + diff2tech + isinsuredinaccident + polage + vehiclage + polholderage + typeclassFC, data = subdata, Hess = TRUE, method = "logistic")

Coefficients:

	Value	Std. Error	t value	pvalue
nbclaim08FC	-2.900e-02	8.425e-03	-3.442e+00	0.180
ClaimNBhat	1.656e+00	9.401e-02	1.762e+01	0.036
bonuspercentnew	1.391e-02	3.357e-04	4.143e+01	0.015
lastprem_group2(500,5e+03]	-3.026e-01	1.129e-02	-2.679e+01	0.024
diff2tech	-1.720e+00	6.900e-02	-2.493e+01	0.026
isinsuredinaccident	-2.964e-01	9.988e-03	-2.968e+01	0.021
polage	-2.789e-02	3.594e-04	-7.759e+01	0.008
vehiclage	4.625e-02	1.056e-03	4.381e+01	0.015
polholderage	-9.538e-03	2.921e-04	-3.266e+01	0.019
typeclassFC	1.169e-01	1.154e-03	1.013e+02	0.006

Intercepts:

	Value	Std. Error	t value
0|150	-2.3565	0.0354	-66.5322
150|300	-0.4060	0.0334	-12.1655
300|500	4.1764	0.0341	122.4217

Residual Deviance: 664289.21
AIC: 664315.21

8.1.8 GAM outputs for Subsection 6.2

Below we give the regression summary for the TPL agent dataset. Other summaries are available on request to the author.

Formula: lapse ~ product2 + region2 + cumulrebate3 + nbclaim0608percust + isinsuredinhealth + isinsuredinlife + vehiclage + householdNbPol + polholderage + maritalstatus2 + jobgroup2 + gender + bonusevol2 + priceratio:(paymentfreq + nbclaim08percust + nbclaim0608percust + nbclaim0708percust + isinsuredinaccident + bonusevol2) + s(priceratio, diff2tech) + s(priceratio, diff2top10agent) + s(priceratio, diff2top10direct) + s(priceratio, typeclassTPL)

Parametric coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)			Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.9881832	0.0744176	-13.279	< 2e-16	***						
product2T1	-0.2957239	0.0365839	-8.083	6.30e-16	***	product2T2	-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	***	householdNbPol	-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	*
priceratio:nbclaim0608percust	-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	**	priceratio:bonusevol2up-down	-7.2677291	1.0573222	-6.874	6.26e-12	***

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,diff2top10agent)	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	***

R-sq.(adj) = 0.0176 Deviance explained = 3.46 – REML score = 44028 Scale est. = 1 n = 187733