32nd International Summer School (2019)
of the Swiss Association of Actuaries

Insurance Data Science: Use and Value of Unusual Data

Part I: Insurance with Vehicle Telematic Data

Jean-Philippe Boucher
Professor
Chaire Co-operators en analyse des risques actuariels
Département de mathématiques
Université du Québec à Montréal (UQAM)

August 2019

Table des matières

Chapitre 1

Introduction to Insurance with Telematic Data

> ⭐ **Telematic**
>
> Telematics is a translation of the French word *télématique* which was first coined by Simon Nora and Alain Minc in a 1978 report to the French government on the computerization of society.
>
> It refers to the transfer of information over telecommunications and is a merge of the French words *télécommunications* (telecommunications) and *informatique* (computing science).
>
> Source : Wikipedia.

Generally, the expression *telematic* refers to the technology of sending, receiving and storing information using telecommunication devices to control remote objects. In this short course, it will refer to the use of such systems within road vehicles, mainly for insurance purpose. This can also be called **vehicle telematics**. It is interesting to note that, in the future, smart houses, intelligent homes with internet of things, may also be of interest for insurers and actuaries.

1.1 Insurance and Vehicle Telematics

Insurance companies have long been interested in the driving habits of their policyholders, but it is only recently that technology has been at an interesting level for insurers to equip their policyholders with devices that measure various driving statistics.

Regarding the data we will use for the course, it is important to understand that insureds have access to two types of devices :

	In-car Device	Mobile App.
Technology	Old	New
Cost	High	Low
Precision	High	Low
Informations	High : Fuel Consumption, RPM, etc.	Low
Practical Problems	Connexion problems	Low
Basis for insurance	Car	Insured (driver)

The car-device, installed on the car, collects information about the vehicle. That means that several different drivers can drive the same car !

On the other hand, the mobile app only collects driving behavior of a single insured. The app can record driving activities even if the insured does not drive his insured vehicle. Indeed, the app could collect driving statistics when the insured is in a bus, in a taxi, as a passenger, etc. To correct this situation, the app gives the opportunity to the insured to indicate that he did not drive certain trips. An insured will obviously be tempted to indicate that he was not driving a trip at high speed, with several many hard braking for example. However, by analyzing the driving style, will it be possible for actuaries to identify that the insured is lying ?

Even if the insurance industry now favors mobile app for telematic insurance, in the end, the choice between the in-car device and the app refers to a more fundamental question : what is the basis of insurance, a vehicle or an insured ?

1.1.1   Challenges

For insurers, vehicle telematic involves some challenges, that can be separated into different components :

1. **Data managing**

2. **Analytics**

3. **Others**

1.1.2 Objectives

In the next pages, we will explain and develop models that use these new vehicle telematics data. However, it may not be easy to use this new data directly in ratemaking, for example. In fact, the legislation in place, as in Canada for example, makes it impossible to use telematics data to rate an active contract. Indeed, the premium must be calculated when the contract is issued, and the information collected during the insurance period can not be used to change the premium for the current insurance period. However, the collection and use of telematic data is still a useful exercise. Despite restrictions on the use of these data, it is possible to mention some important uses of telematic data :

1. **Discounts or surcharges upon renewal**
2. **New segmentation variables**
3. **Pay-as-you-drive, Pay-how-you-drive or other variants**
4. **Identify risky behavior patterns**
5. **Autonomous driving**
6. **Claims management**
7. **Others**

1.2 Literature Review

Academic research, or even private research (but published in journals) on telematic automobile insurance is quite new. Here is short list of papers published :

1. Litman, T. (2007). Distance-based vehicle insurance feasibility, costs and benefits. Victoria, 11.

2. Ferreira Jr, J., & Minikel, E. (2012). Measuring per mile risk for pay-as-you-drive automobile insurance. Transportation Research Record, 2297(1), 97-103.

3. Boucher, J. P., Pérez-Marín, A. M., & Santolino, M. (2013). Pay-as-you-drive insurance : the effect of the kilometers on the risk of accident. In Anales del Instituto de Actuarios Españoles (Vol. 19, No. 3, pp. 135-154). Instituto de Actuarios Españoles.

4. Paefgen, J., Staake, T., & Fleisch, E. (2014). Multivariate exposure modeling of accident risk : Insights from Pay-as-you-drive insurance data. Transportation Research Part A : Policy and Practice, 61, 27-40.

5. Tselentis, D. I., Yannis, G., & Vlahogianni, E. I. (2016). Innovative insurance schemes : pay as/how you drive. Transportation Research Procedia, 14, 362-371.

6. Lemaire, J., Park, S. C., & Wang, K. C. (2016). The use of annual mileage as a rating variable. ASTIN Bulletin : The Journal of the IAA, 46(1), 39-69.

7. Weidner, W., Transchel, F. W., & Weidner, R. (2016). Classification of scale-sensitive telematic observables for riskindividual pricing. European Actuarial Journal, 6(1), 3-24.

8. Weidner, W., Transchel, F. W., & Weidner, R. (2017). Telematic driving profile classification in car insurance pricing. Annals of Actuarial Science, 11(2), 213-236.

9. Verbelen, R., Antonio, K., & Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. Journal of the Royal Statistical Society : Series C (Applied Statistics), 67(5), 1275-1304.

10. Ma, Y. L., Zhu, X., Hu, X., & Chiu, Y. C. (2018). The use of context-sensitive insurance telematics data in auto insurance rate making. Transportation Research Part A : Policy and Practice, 113, 243-258.

11. Ayuso, M., Guillén, M., & Pérez-Marín, A. M. (2014). Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. Accident Analysis & Prevention, 73, 125-131.

12. Ayuso, M., Guillen, M., & Nielsen, J. P. (2019). Improving automobile insurance ratemaking using telematics : incorporating mileage and driver behaviour data. Transportation, 46(3), 735-752.

13. Ayuso, M., Guillen, M., & Marín, A. M. P. (2016). Using GPS data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. Transportation research part C : emerging technologies, 68, 160-167.

14. Ayuso, M., Guillen, M., & Pérez-Marín, A. (2016). Telematics and gender discrimination : some usage-based evidence on whether men's risk of accidents differs from women's. Risks, 4(2), 10.

15. Boucher, J. P., Côté, S., & Guillen, M. (2017). Exposure as duration and distance in telematics motor insurance using generalized additive models. Risks, 5(4), 54.

16. Denuit, M., Guillen, M., & Trufin, J. (2019). Multivariate credibility modelling for usage-based motor insurance pricing with behavioural data. Annals of Actuarial Science, 1-22.

17. Wüthrich, M. V. (2017). Covariate selection from telematics car driving data. European Actuarial Journal, 7(1), 89-108.

18. Gao, G., & Wüthrich, M. V. (2018). Feature extraction from telematics car driving heatmaps. European Actuarial Journal, 8(2), 383-406.

19. Gao, G., & Wüthrich, M. V. (2019). Convolutional neural network classification of telematics car driving data. Risks, 7(1), 6.

The paper of Litman, who is not an actuary, worths reading.

Note, however, that maybe new papers have been published recently, in 2019. But, nevertheless, this list is short. And we can note that many papers are published in Transportation journals, and not in Actuarial journals. We can also see that the team of Barcelona (with Guillén and Ayuso) is one of the leading team in the area. In the PhD seminar on telematic I gave this winter, we analyzed and summarized almost all the papers. So, questions on the papers can be asked.

1.3 Available Databases

We mainly have a databases to work with when we want to use telematic information. The dataset used are based on true database, but are **simulated datasets**. We will analyse the database, but for now a quick overview of the available data can be done :
The Telematic Database (*CanadaData.csv*) mimics the form and content of the classic actuarial databases used in ratemaking. Telematic information is only seen as additional covariates. As a result, the database is constructed to have only one observation per vehicle/year, in which we find :

— The characteristics of the driver (sex, age, civil status, etc.) ;
— The characteristics of the car (make, model, year, etc.)
— **A very large amount of telematic data** ;
— Information on the loss experience of the vehicle/year (time covered, number of claims, details about claims, severity of each claims, etc.).

Such a database makes it possible to estimate and model the claim frequency, the severity or the pure premium, for example. The same traditional actuarial tools in ratemaking can be used to verify if covariates are statistically significant for pricing.

Typically, companies that provide a telematic information collection service will offer to build such databases for insurers. Often, these same companies will go as far as summing up all telematic information and convert it into arbitrary scores :

— A braking score summarizing all the braking activities of the year for a specified insured/vehicle ;
— An acceleration score ;
— A turn score ;
— etc.

We think, however, that it is more interesting for actuaries to work with source data instead of using these scores.

For the database, the names of the fields and variables contained in the database can be found in the file Dictionnary.xlsx, where we can find a variety of interesting variables summarizing driver behavior. It should be noted that for this short-course, this shared file is not an original database : it is a simulated one, based on a real database owned by a private Canadian insurance company.

We took a sample of the database *CanadaData.csv*, to have the possibility to observe some vehicles over time. For this dataset, only few covariates have been chosen to illustrate how we can work with this kind of data. A variable named *Falsevin* have been added to the database to identify each observation of a single vehicle.

1.3.1 Type of drivers in the databases

1. Policyholders who love this type of gadget, and want detailed information about their driving habits (summary data is continously available to policyholders via a mobile app) ;

2. Young drivers and bad drivers. To motive policyholders to take the telematics option, insurance companies offer a first discount of 5%, and the renewal discounts range from 0% to 25% (remember that it is not possible for an Ontario insurance company to increase the insurance premium based on the telematics information collected). Because auto insurance in Ontario is very expensive and often unaffordable, all discounts are welcome for policyholders with high insurance premiums. As a result, an unusually high proportion of risky insured uses telematics devices or telematic app.

1.4 Automobile Insurance in Ontario, Canada

To properly analyze the claims experience of available databases, it is important to understand the way car insurance works in Canada. Insurance is under provincial jurisdiction in Canada, and each province has its own rules and regulations. Since we will be analyzing the data from the province of Ontario, we only need to analyze this province.

Ontario is Canada's most populous province, with nearly 14 million people, and the largest city is Toronto.

Ontario has a completely private plan for auto insurance. Thus, the entire auto insurance industry is controlled and shared by private insurance companies. As a result, and by a large margin, Ontario's auto insurance market is the largest in Canada [1].

1.4.1 Mandatory Coverages

Ontario's law requires four mandatory protections :

1. **Third party liability** (or TPL)

2. **Direct Compensation - Property Damage Coverage** (or DCPD)

3. *Accident Benefits Coverage* (or AB)

---

1. The province of Quebec, the second most populous province (main city : Montreal), has a mixed insurance system where a public company is responsible for compensation for bodily injury in car accidents. The province of British Columbia, the third-most populous province (main city : Vancouver), has a public insurance plan covering all compensation related to car accidents.

4. *Uninsured Automobile Coverage* (or UA)

Ontario's auto insurance compensation system is partially fault-free, as we have seen with the *Accident Benefits Coverage* coverage. On the other hand, we also saw that it was possible to sue the responsible driver for an accident. This is allowed only if the injuries caused by the accident exceed a *threshold of severity*.

In Ontario, the threshold of severity is a *verbal threshold*. This kind of threshold explicitly describes the type of injury, or the type of pain, that gives the victim the opportunity to sue the driver responsible for an accident. In Ontario, the verbal threshold mentions death, serious and permanent damage to the body, or severe and permanent impairment of a significant bodily, mental or psychological function.

> ☢ **Responsability**
>
> When there is an accident, the responsibility of each driver is useful to determine which coverages apply. For bonus-malus rating, the responsability is then important for determining the insurance premium of the next years. Being involved in a responsible accident causes large premium increases over several years.
>
> In Ontario, the responsibility of every driver involved in an accident is determined by a simple method : the use of diagrams. In the Motor Insurance Act, a series of drawings illustrating all possible car accidents is indicated. For each drawing, the percentage of responsibility of each driver can be found.
>
> (See www.ontario.ca/laws/regulation/900668).

1.4.2 Optionnal Coverages

In addition to mandatory coverage, Ontario drivers may also take optional coverages. Among the most important are :

— (B2) Collision Coverage :

  Collision usually includes a vehicle hitting another vehicle. But the term collision can also includes an accident where a vehicle hits a tree, a guardrail or any other object. This warranty covers damage to the insured's vehicle when he is responsible for a collision with one or more vehicles or when the accident does not involve any other vehicle. When purchasing insurance coverage, the insured will have to choose a deductible for this chapter. Typically, deductibles of $ 250, $ 500 or $ 1000 are offered.

— (B3) Comprehensive :

The Chapter B3 covers damages to the vehicle caused by non-collision damage. Thus, it can be understood that the guarantee covers in particular :

— Damage caused by projectiles (breakage of glass) ;

— Fire ;

— Total or partial theft of the vehicle ;

— Animals ;

— Explosion ;

— Earthquake ;

— Storm ;

— Hail ;

— Flooding ;

— Riots ;

— etc.

Unlike Chapter B2, it is possible to buy the B3 guarantee without buying the B2 guarantee. As for chapter B2, when purchasing the insurance cover, the insured will have to choose a deductible for this chapter.

It exists many other chapters or garanties in Ontario, but we will restrict ourselves to those listed. Generally speaking, it means that we have three kinds of drivers in Ontario :

1. Drivers with Full Coverage : the insured will have the 4 mandatory covarages, with B2 (Collision) and B3 (Comprehensive) Coverages.

2. Drivers with Partial Coverage : the insured will have the 4 mandatory covarages, with B3 (Comprehensive) Coverage.

3. Drivers with Minimum Coverage : the insured will have the 4 mandatory covarages, without B2 (Collision) or B3 (Comprehensive) Coverages.

Chapitre 2

Summary of the Database TelematicDB.csv

Before working directly with the database, we need to explain the way we will analyze the data. For ratemaking, we usually model the claim frequency and the claim severity separately, and for more complex modeling, we can add a dependence structure between them. if we suppose that the cost associated with the contract of insured i is noted S_i, we can then suppose the following mathematical structure :

$$S_i = \sum_{j=1}^{N_i} X_{i,j}$$

where N_i is the number of claims of contract i, and $X_{i,j}$, i.i.d. r.v., are the cost of each claim $j = 1,\ldots,N_i$ of contract i. By convention, if there is no claim during contract i, the total cost is zero (if $N_i = 0$ then $S_i = 0$). We can also model directly the total amount of claims S_i by a a composed Tweedie distribution for example. Without consideration to premium principles, we can say that the expected value of S_i represents the annual premium. If N_i and the $X_{i,j}$ are supposed independent, we have

$$E[S_i] = E[N_i] \times E[X_{i,j}]$$

For this short-course, we will focus on the claim frequency N_i. In the examples, we will then suppose that the premium is $E[N_i]$, meaning that the severity of the claims are not considered, or that each claim costs 1\$. Neverthelsess, we think that telematics informations can be used to improve our modeling of the severity. Indeed, speed of the crash, for example, could be used to better estimate the final cost of a claim. However, to introduce the use of telematics in insurance, we think that claims frequency analysis is a better starting point.

We are also interested in modeling the number of claims related to car accidents, which mean that we excluded everything related to Comprehensive coverage. This modeling choice is made to simplify our analyzes for this introductory course on telematics pricing. Indeed, as mentioned by several authors, some elements of Comprehensive coverage are related to driving and can even be qualified as a car accident. For example broken glass or collision with animals can clearly be linked with driving

behavior. Even partial or total car theft could be related to the use of the vehicle : an insured who does not use his car, lea-
ving his car it in his driveway or garage, is much less likely to be stolen than an insured who parks his car in a lot of different places.

In any case, in our examples, we will focus on the number of car accidents. This variable is available in our database and is called
RA_ACCIDENT_IND [1]. As we saw in the previous chapter, there are different coverages in a car insurance contract and insurers
want those coverages to be priced. Because there is a clear dependence between each coverage, it is not possible to model each
coverage independently. That means that we need a two-step model :

1. A model to model the number of accidents ;
2. For each car accident, a model that indicates which coverage is affected for each accident.

Exemple 2.0.1. Frees and Valdez (2008) proposed to model the number of claims by a standard count distributions, and proposed
to use a multivariate vector indicating if coverage is affected by an accident. For illustration, if we suppose a vector that consists
of { TPL/DCPD/AB/Collision }, and a binary variable {0 , 1 } indicating whether the coverage is affected or not by a specific
accident, a series of vectors of similar to {1,1,0,0}, {0,1,1,0}, {0,0,0,1}, etc. would be observed for each accident.

To rate each coverage (and find the dependence structure between coverages), we only need to model correctly the $(2^4 - 1)$
accident possibilities, where minus one comes from the fact that the case {0,0,0,0} does not exist (there is no claim if no coverage
is implied).

---

1. Note that the insurer names its variable "Accident", but it would be more correct to name it "Claims" since some accidents are not reported to
insurers.

2.1 Structure and basic statistics of the Database

The simulated database, based on a database from *The Co-operators General Insurance Company*, consists of 93,682 auto insurance policies. For each insurance policy, corresponding to a vehicle/driver, this database contains the standard pricing variables, including several variables relevant for ratemaking :

1. Age of the principal driver ;
2. Sex of the principal driver ;
3. Age of the vehicle ;
4. Marital status of the principal driver ;
5. Vehicle use (commute or not) ;
6. etc.

Other pricing variables are available. What is interesting about the database is the presence of several telematics variables giving information on driving habits. So, we have the total distance traveled by each driver, the proportion of night driving, or the number of abrupt acceleration, to name a few. We will return to the use of some of those variables later in this course.

We begin with basic work on the available data. The dataset used is in .csv format. The dataset of 93,682 auto insurance policies has been separated into a train dataset (or in-sample data, with 65,489 observations), and a validation/test dataset (or out-of-sample data, with 28,193 observations). Each dataset contains 114 potential variables (covariates).

2.1.1 Basic Statistics

Basic summary statistics of the train dataset can be computed to understand the subset of the data used.

Table 2.1.1 shows the results, where we can observe the average distance driven (in km), the average exposure time (in year), the average number of claims, as well as the distribution of the insured's age and the vehicle's age.

Variable	Mean	Variance	Minimum	Maximum	Centile k		
					$k = 25$	$k = 50$	$k = 75$
Distance Driven (km)	7 994	53 126 697	0.1000	76 271	2 616	5 903	11 255
Exposure time (year)	0.5012	0.0960	0.0027	1.0795	0.2438	0.4959	0.7479
Age of the insured	48.7227	295.3452	16.0000	103.0000	34.0000	49.0000	62.0000
Vehicle Age	5.7562	20.0174	-2.0000	20.0000	2.0000	5.0000	9.0000
Nb. Claims	0.0705	0.0737	0.0000	3.0000	0.0000	0.0000	0.0000

Table 2.1.1 − Summary Statistics

We can also take a look on some other covariates, such as the sex of the driver, the civil status and the use of the car.
Table 2.1.2 shows the results. The dataset contains approximately the same number of male drivers and female drivers, and half
of the drivers use their car to go to work.

Statistics	RA_GENDER			RA_MARITALSTATUS		RA_VEH_USE		
	M	F	Other	Maried	Other	Commute	Pleasure	Other
Percentage (%)	45.7%	52.9%	1.3%	66.0%	33.9%	53.7%	41.9%	4.3%

Table 2.1.2 – Summary Statistics

The absolute and relative frequencies of the number of accidents are presented in Table 2.1.3. We see that a large proportion of
the insured (93.3%) did not claim on the covers studied.

Number of Accidents	Number of Contracts	Pourcentage (%)
0	61122	93.3
1	4128	6.3
2	225	0.34
3	14	0.02
Total	65 489	100

Table 2.1.3 – Number of accidents in the database

Even if the information is not available (we mut be careful before sharing private informations), other characteristics of the dataset
could be mentioned. On average in the porfolio, an insured is 49 years' old. This dataset is then really different from the one
used by the team of Barcelona (see papers of Montserrat Guillén and her team), which a portfolio of young drivers. Insured are
mostly from urban agglomeration. Figure (2.1.1) shows how sex and marital status are linked.

Figure (2.1.2) exposes the zip codes of our insureds. With zip codes, we can usually find social data on the Canada's government websites (mean salary, population, etc.).
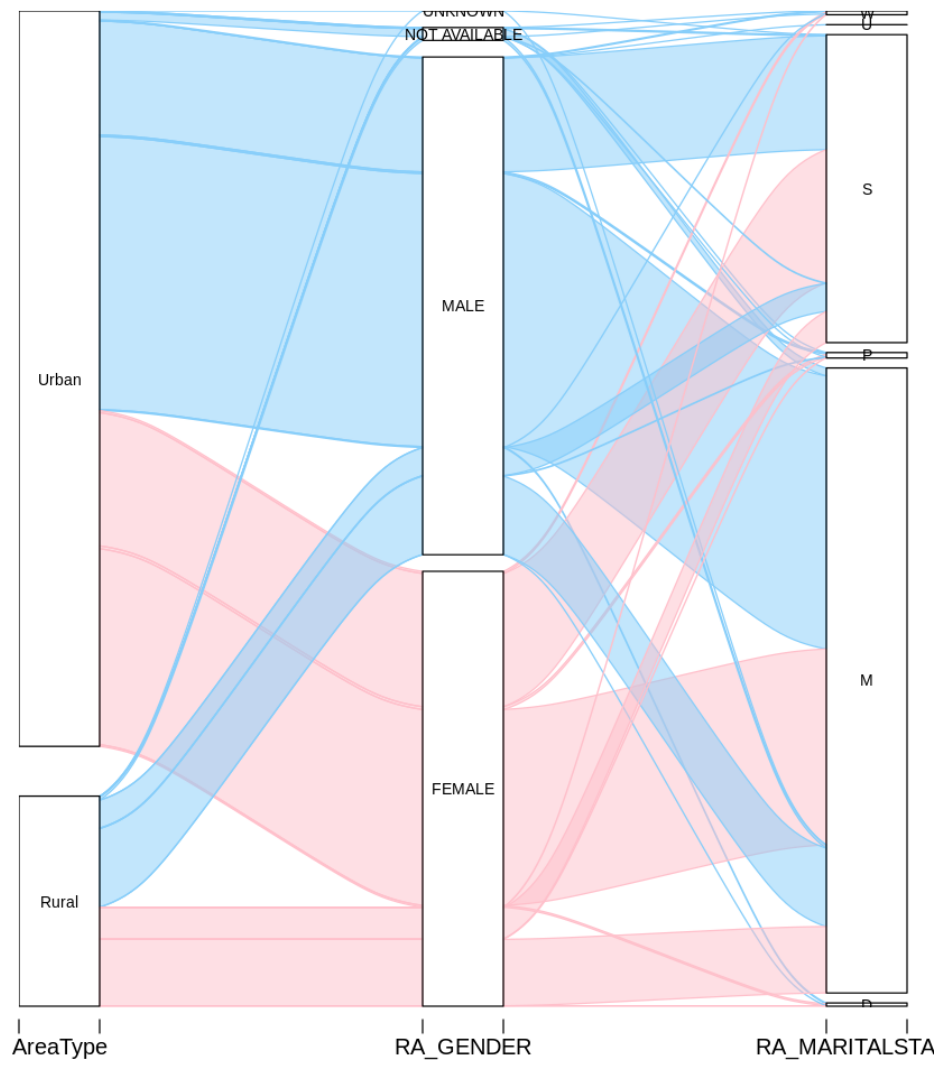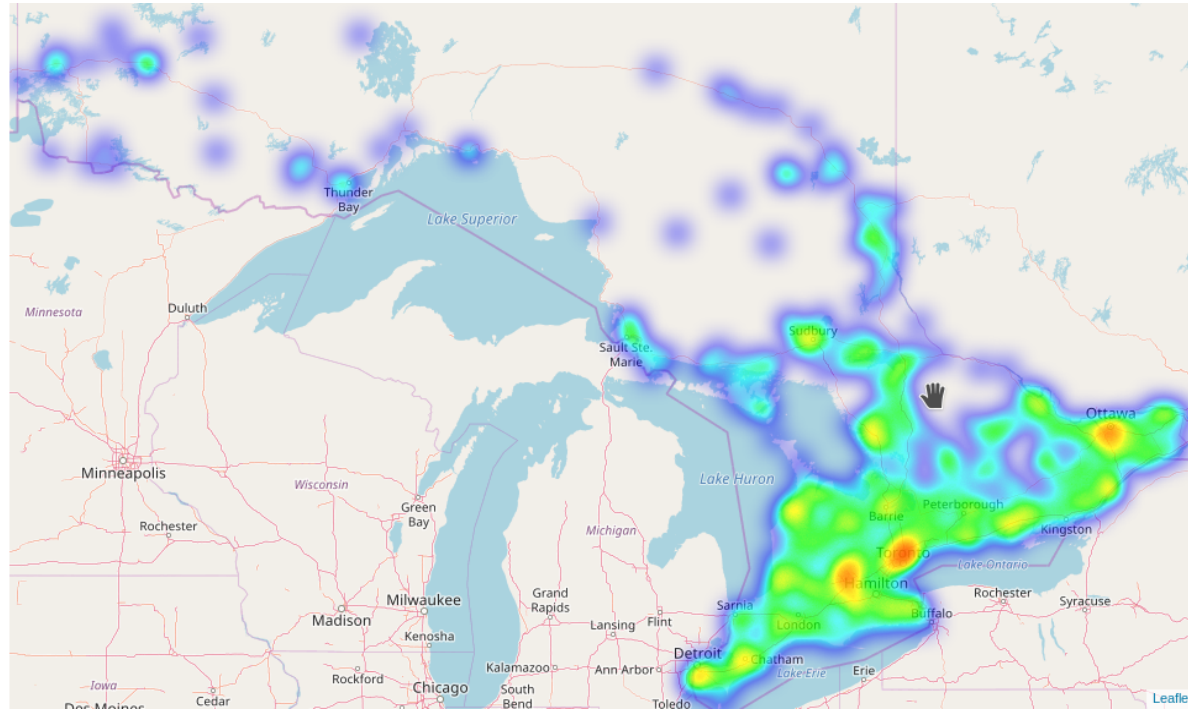
Figure 2.1.1 – Gender and Marital Status

Figure 2.1.2 – Zip Codes

Figure 2.1.3 – Distribution des assurÃ©s par code postal

Figure (2.1.3) shows the distribution of insured in Ontario, where we can see again that most on the insureds live near big cities in Ontario (Toronto, Mississauga, Ottawa, Windsor).

The vehicle identification number (or VIN) was originally available for our analysis. Using the library *vin_decoder*, we were able to link many vehicle informations on our database. For example, figure (2.1.4) shows the distribution of mark and model of each vehicle in the database. Most of them are (*passenger car*), with Toyota, Honda and Ford begin the most popular.

In Figures (2.1.5), (2.1.6) and (2.1.7)We also draws the number of claims on the Ontario map to illustrate where are the insureds who report one, two or three accidents in the database.

2.1.2 Telematic Statistics

The interest in this short-course in telematic insurance is to look at telematics information. In Figure (2.1.8), we analyse the following three variables :

— Percentage of week-end/week driving ;

— Percentage of night/day driving ;

— Percentage of rush/non-rush driving.

We observe that most of the insured drive during the week days, on the day, during rush hours. On the same figure, we have some doubt about the way rush hours is defined since we saw some observations of rush hours drive during the week-end, and on the night. We learned that it is because rush hours are defined as specific hours, even during the weekend. Similarly, nigh and day are defined at specific hours, without consideration for the seasons.

In Figure (2.1.9), we see that the percetnage of use of the car during the day is similar : around 15% for week days, and 11% for Sunday. On the same figure, the standard error of the driving time on week-end is higher than during week days, meaning that, maybe, insureds drive longer distance on the week-end.
We also analyse the distance driven by zipcodes. In Figure(2.1.11), we seen that insureds in Toronto drive less kilometers than insureds who live in Ottawa or Mississauga. This can be explained by the population density, or the availibility of public transport to go to work.
We have a lot of telematics information regarding the way insureds are driving. It is worth looking the file **Dictionnary.xlsx** to understand each variable.

To analyse the correlation between variables, we selected some of the most popular variables from classic actuarial ratemaking and some telematic information. A simple Pearson correlation statistics between those variables are shown at Figure (2.1.13). We see a strong correlation between some variables.
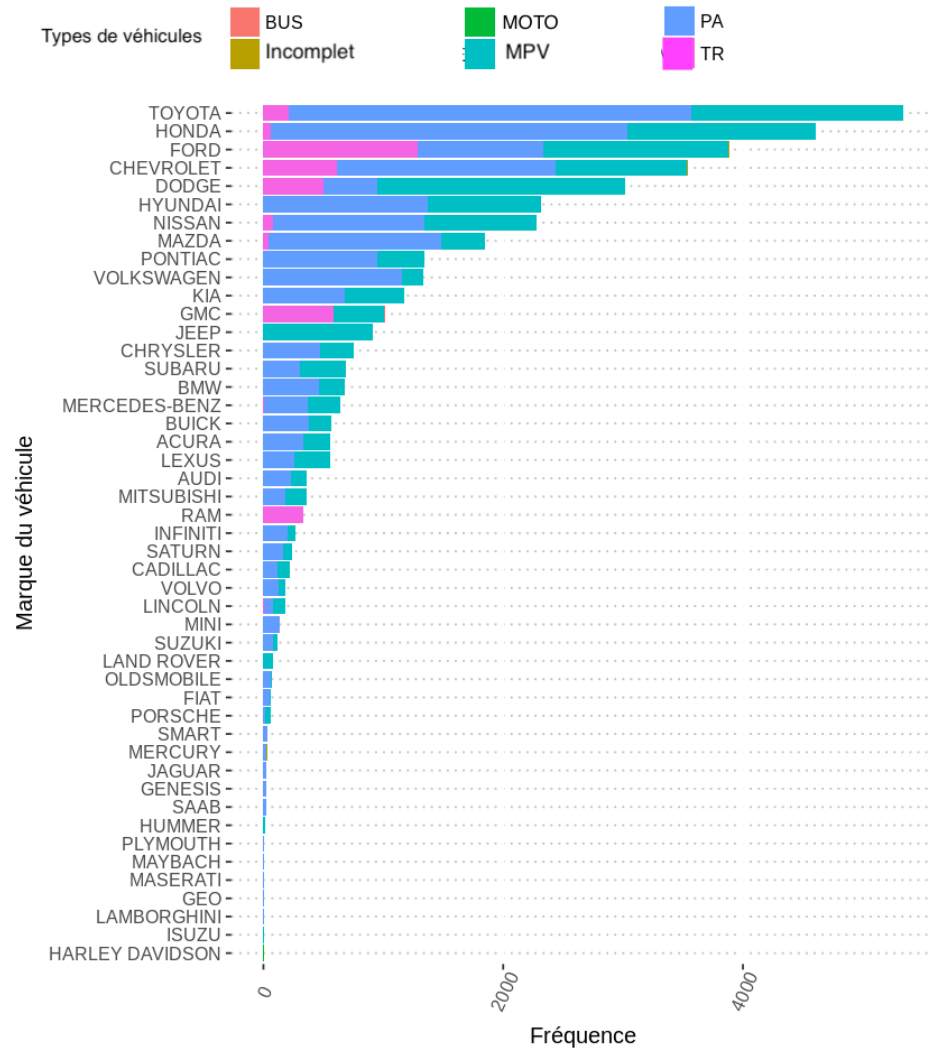
Figure 2.1.4 – Distribution of mark and model of insured cars
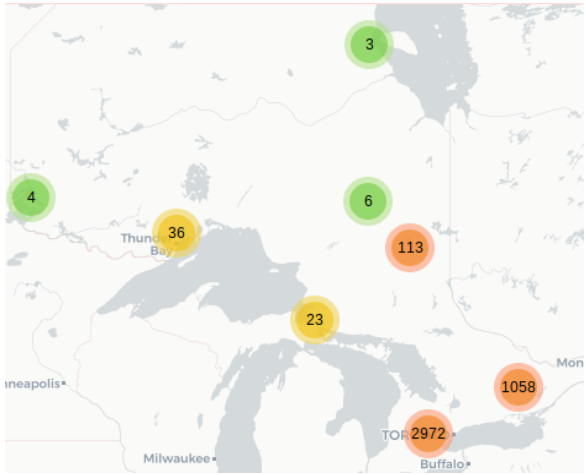
Figure 2.1.5 – Insureds with at least one claim

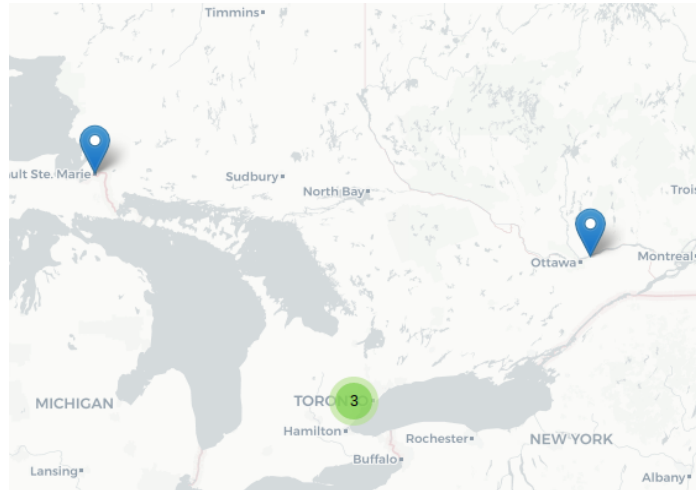Figure 2.1.6 – Insureds with at least two claims

Figure 2.1.7 – Insureds with at least three claims

It can be very difficult to analyse all those available statistics about the way the insureds drive. On of my master student, Noureddine Meraihi, performs a Gradient-Boosting Poisson model for the number of claims. Gradient-Boosting algorithms can be classified in statistical/machine learning techniques. It allows us to include almost all variables in a predictive modeling, and verify their importance for prediction. We will not cover this kind of models in this short-course. However, the interesting conclusions of his analysis on the same dataset (the original one, not this simulated) is to highlight what are the more important variables to explain the number of claims. The image, in Figure 2.1.14, exposes one of the major conclusion of his study.

The two main variables to explain the number of claims are the distance driven (measured by the telematic device), and the exposure day (not measured by the telematic device, it corresponds only the number of days the insured has been covered). On the best covariates to explain the number of claims, many of them are not related to telematics (credit score, vehicle age, martial status, etc.). Also, quite surprisingly, variable related to driving behavior or what we call "how you drive" (brakings, accelerations, turning, etc.) are not so predictive [2]. The variables that seem to better explain the risk are related to what we call risk exposure.

---

2. Note that in Canada, we are driving a the right side of the road, meaning that turning left is more dangerous than turning right
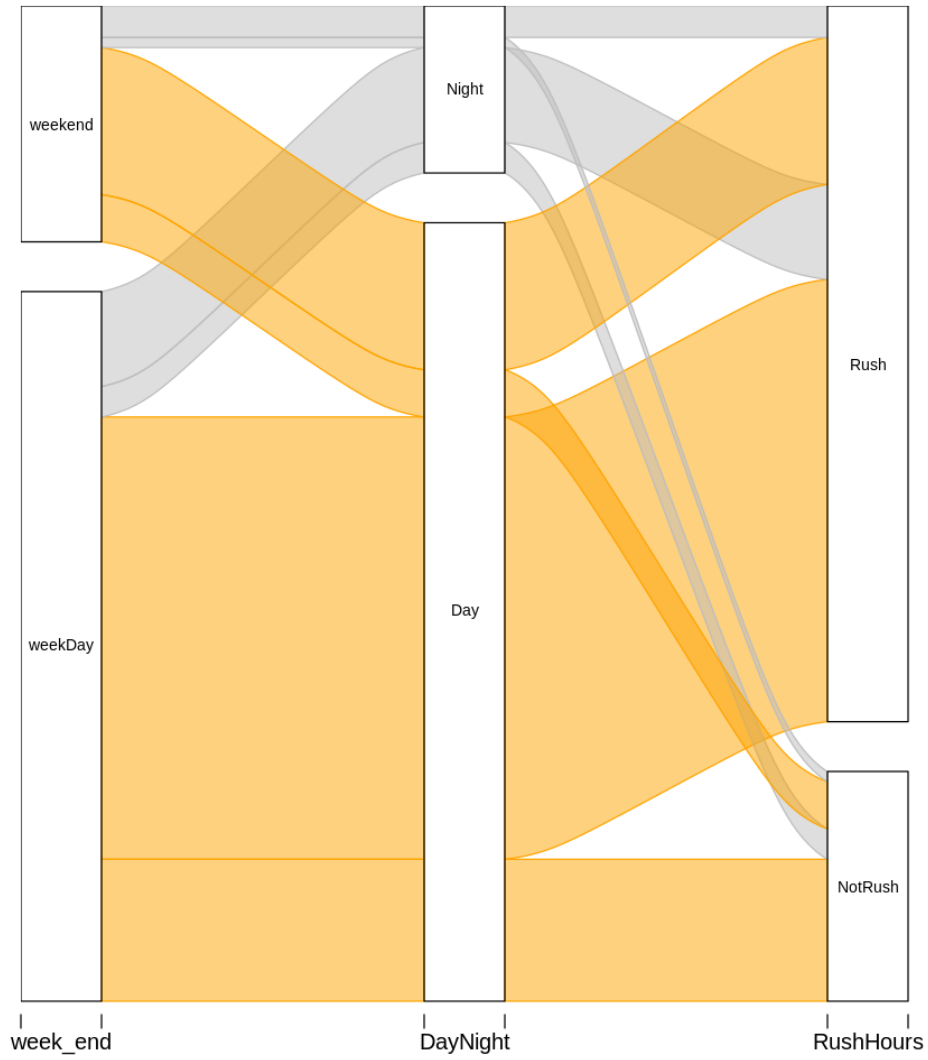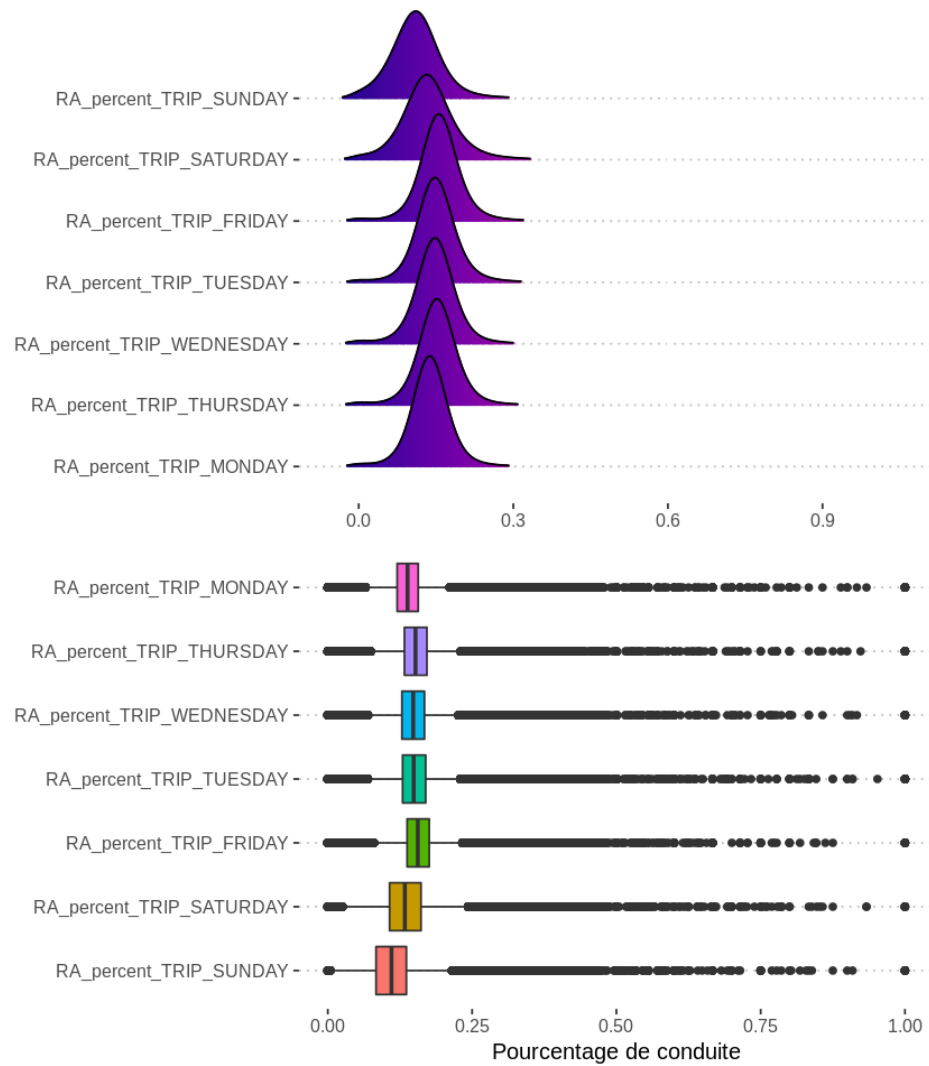
Figure 2.1.8 – Use of the car

Figure 2.1.9 – Use of the car
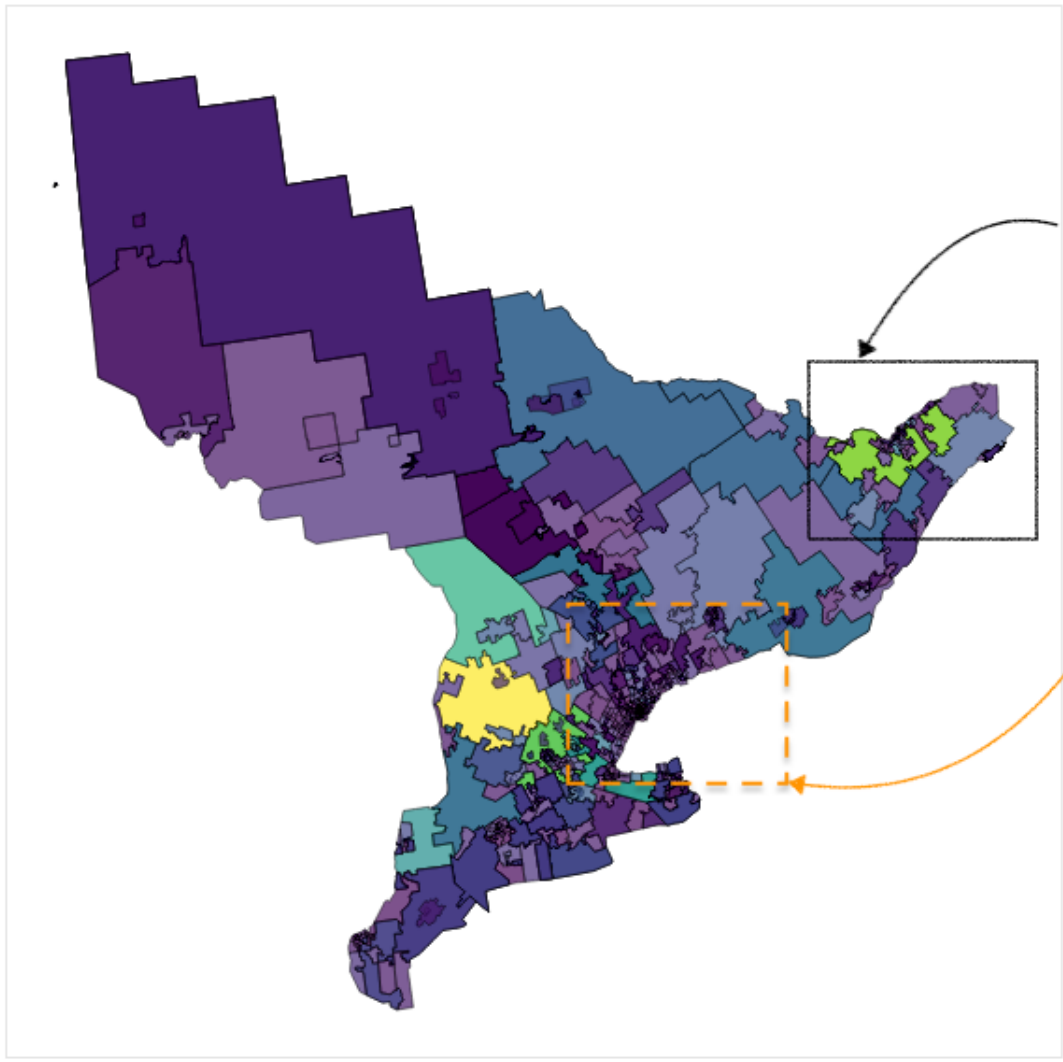
RA_DISTANCE_DRIVEN

Figure 2.1.10 – Ontario

RA_DISTANCE_DRIVEN

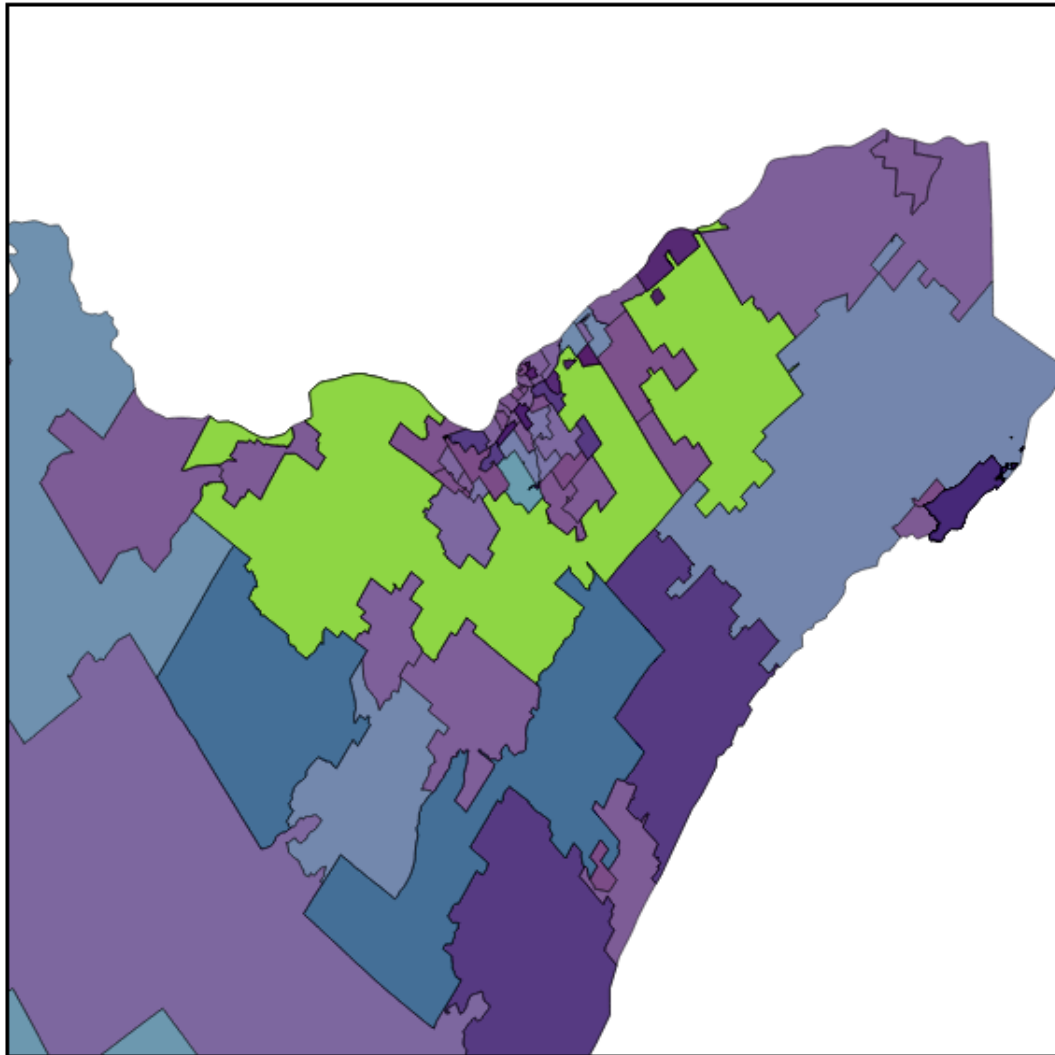Figure 2.1.11 – Toronto Region

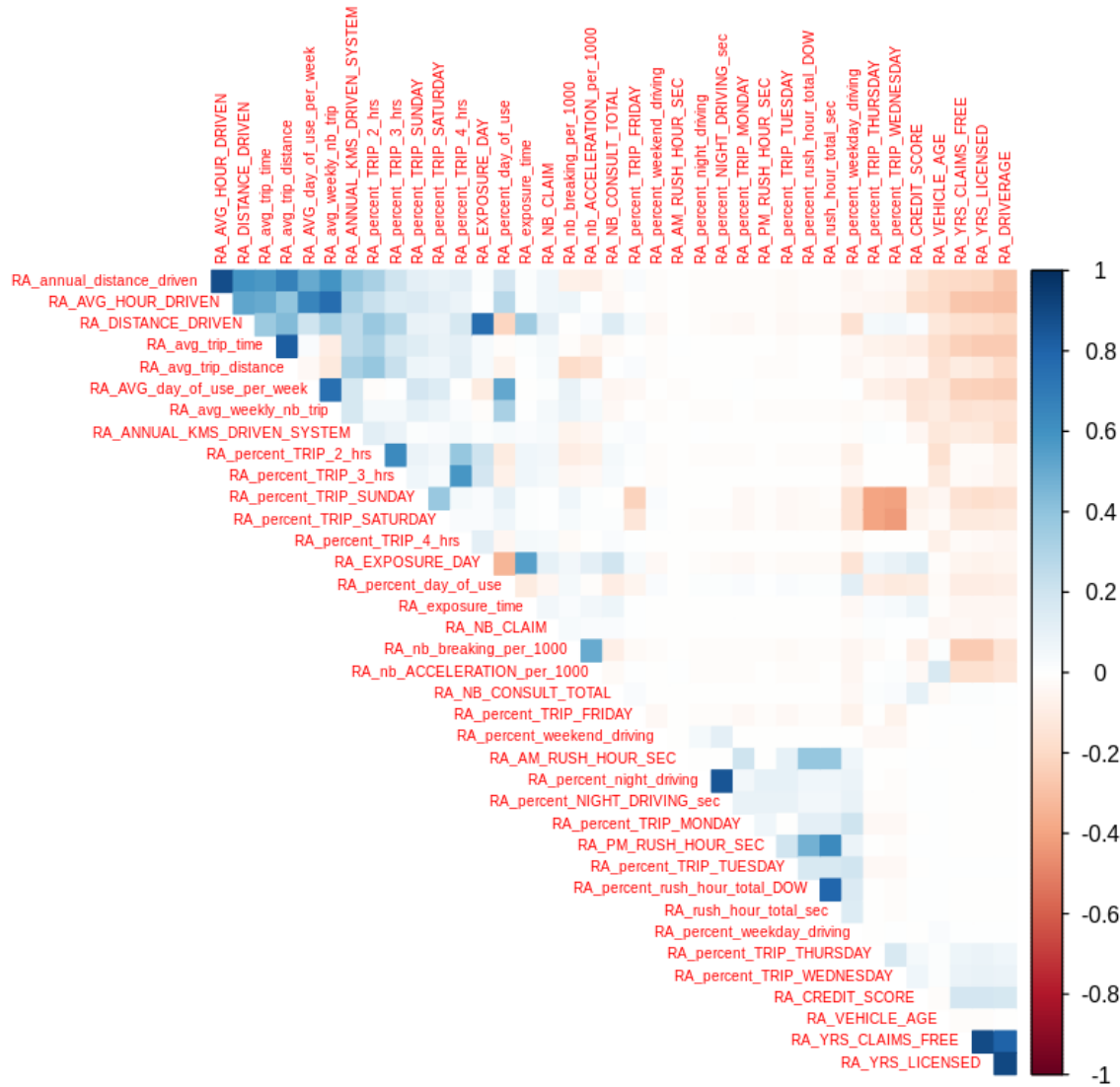**RA_DISTANCE_DRIVEN**

Figure 2.1.12 – Ottawa Region

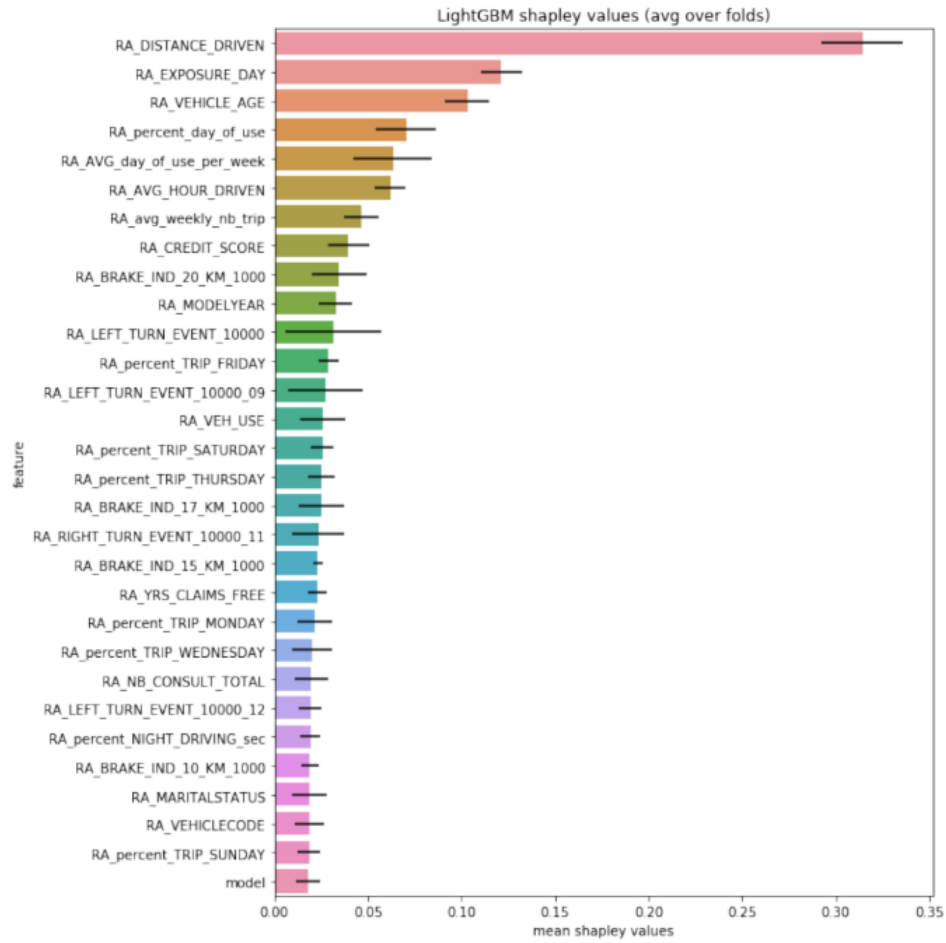Figure 2.1.13 – Correlation between covariates

Figure 2.1.14 – Importance of the best 30 variables (for a Gradient-Boosting Poisson model for the number of claims)

2.2   Risk Exposure

Special attention should be paid to two variables we have just encountered : the distance driven and duration of the contract (exposure time which is the exposure day divided by 365). Classically in insurance, the premium of an insurance contract is proportional to its duration : a one-year contract is twice as expensive as a 6-month contract, all things being equal. Such proportionality is not always present when analyzing risk. We can think of some specific examples.

Consequently, it might be interesting to ask ourselves whether the exposure of risk in automobile insurance should be calculated only in terms of the duration of the contract. It is obvious that the use of the vehicle is an important factor in the risk of accident. Indeed, it is expected that a person who drives a lot will have a high propensity to have an automobile accident compared to a person who uses his car occasionally.

One way to quantify this use of the vehicle is simply to consider the number of kilometers traveled annually (distance driven). Several studies have, in the past, shown that there is a significant link between the number of kilometers traveled and the risk of being involved in an automobile accident. As we see with our database, this telematics information is available. But before studying the link between risk exposure and automobile use, it is worth summarizing what has been proposed before in the literature :

1. Vickrey (1968)

2. Lourens et al. (1999)

3. Limtan (2005)

4. Bordoff and Notel (2008)

Until recently, most insurers did not have accurate data on the mileage traveled annually by their policyholders. Rather, insurers use an estimate provided directly by the insured before the insurance period. In most cases, this information is inaccurate. For example, as indicated by Butler et al. (1988), a US insurance company has already reported that 60% to 70% of the vehicles it insures were categorized as vehicles traveling less than 12,000 km annually, while the actual average distance traveled per vehicle was order of 20,000 kilometers.

Generally, it seems to be advantageous for the insured to lie about his mileage, as no periodic check of the odometer is made. It is therefore difficult for insurers to prove the dishonesty of their policyholders. Some form of control measure exists, however. In the event that a claim is submitted to an insurer, an odometer check is performed and if an inconsistency or a lie is detected, certain financial penalties may be applied. In this sense, there is a certain deterrent effect on not revealing the distance traveled correctly, but it seems that this effect is not sufficient for the data to be completely reliable. It is therefore difficult to construct a pricing model based on the use of the vehicle that is fair.

Vickrey (1968) had some interesting proposals to reform the automobile insurance product so that it would be priced according to the use of the vehicle. This type of insurance is known as Pay-As-You-Drive Insurance (PAYD). Among the suggested suggestions :

— the *insured gasoline*, where the cost of insurance is included in the cost of gasoline, via a surcharge at the gaz pump ;

— the *insured tires*, where the purchase of tires from a dealer associated with an insurer provides coverage in the event of an accident.

Some recent articles by Litman (2005, 2011) offer some other solutions, which seem well suited to telematics devices that are added to cars :

— A possible structure for PAYD is to consider mileage as a pricing criterion in the calculation of insurance premiums. This is called **Mileage Rate Factor** (MRF).

— A rate structure **Per-Mile Premiums** (PMP). This changes the traditional exposure unit, which is the duration of the contract (usually one year), for a unit of distance. So, the insured pays a cost for each unit of distance traveled and not a cost per day of coverage. The author proposes a cost of insurance which decreases with the number of kilometers, in connection with his observations on the insured.

With the advent of new technologies, it is now possible for actuaries to develop new forms of pricing. Typically, one could classify forms into two categories :

1. Pay-As-You-Drive Approach (PAYD)

2. Pay-How-You-Drive Approach (PHYD)

2.2.1 Our Database

Let's go back to our database and look at the available statistics. We obviously have, for each policy, the observed duration of each insurance contract.

The empirical distribution for distance driven, and exposure time are shown in Figures 2.2.2 and 2.2.1.

To link the accident risk to the various exposure measures we have grouped insured by exposure duration of their contract by considering intervals of time of 0.01 year. For each group, the average of the $RA_A CCIDENT_I ND$ variable (i.e. the average number of claims) was plotted. We did the same exercise with distance driven, by 500 km jump.
Figure 2.2.3 shows the claim frequency per exposure time for our database. Color of the dots indicates the number of insureds for each group. By looking at the figure, we can see that the relation between the exposure time and the claim frequency is linear with a positive slope. We also observe this kind of relationship in other research papers, such as papers using Spanish data (see Boucher et al (2012) and Boucher et al (2017)).
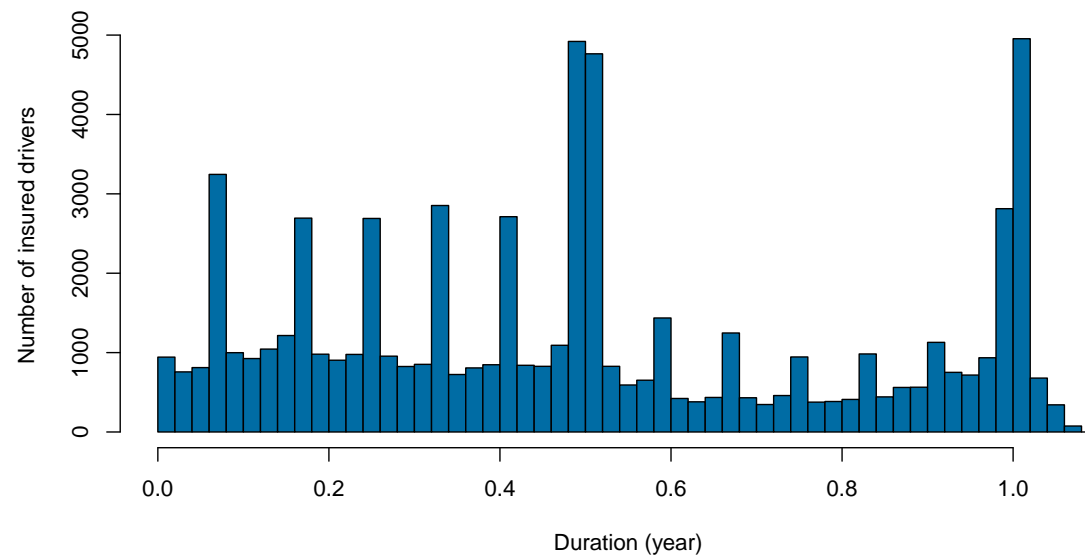
Figure 2.2.1 – Distribution of the risk exposure (in years). Each band has a lenght of 0.02 year.
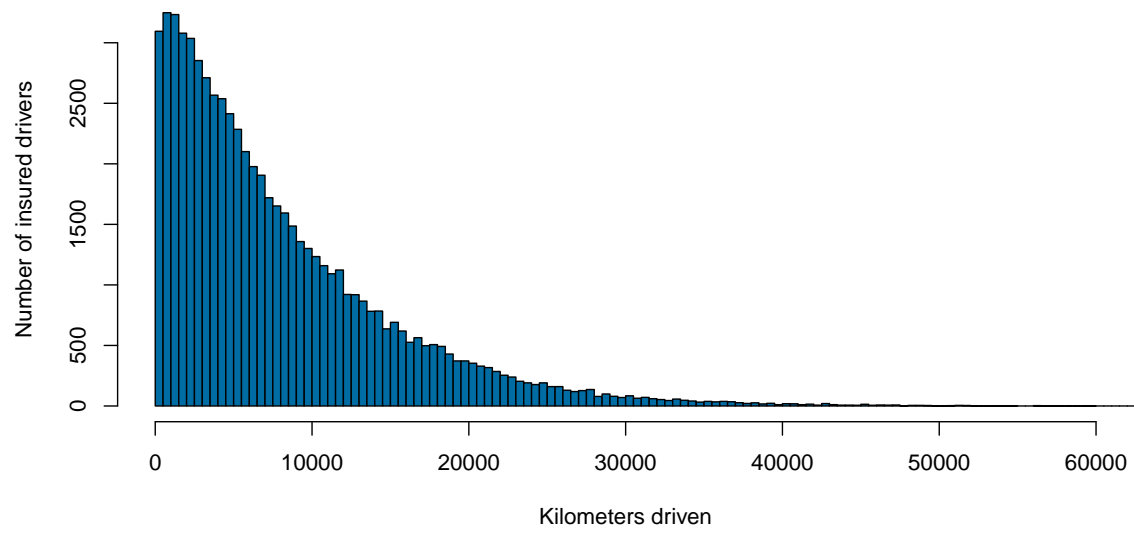
Figure 2.2.2 – Distribution of the distance driven(en km). Each band has a lenght of 500km.
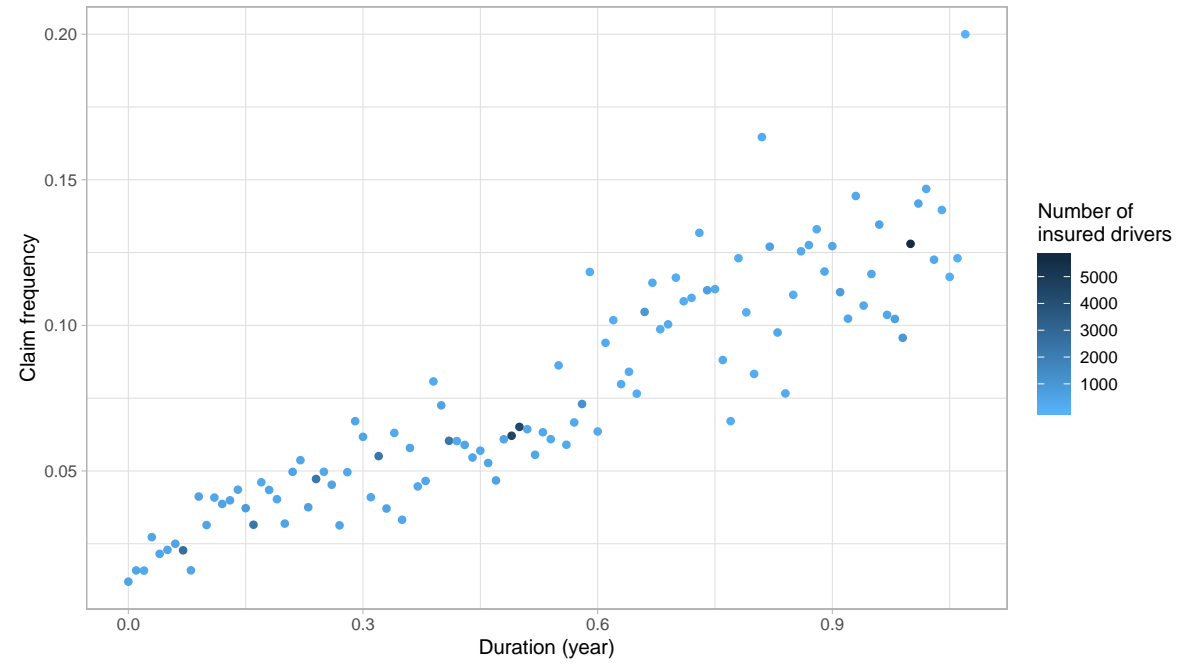
Figure 2.2.3 – Claim Frequency vs Exposure Time

Figure 2.2.4 does the same exercise as the previous figure, but groups the insured according to the number of kilometers traveled, by 500 km jump. The graph also shows a trend similar to what was observed with the Spanish data of Boucher et al. (2012) and Boucher et al. (2017). There is a strong linear and positive relationship in the first kilometers traveled. Then, the effect of distance traveled on propensity to claim seems to fade, but the relationship is still positive.
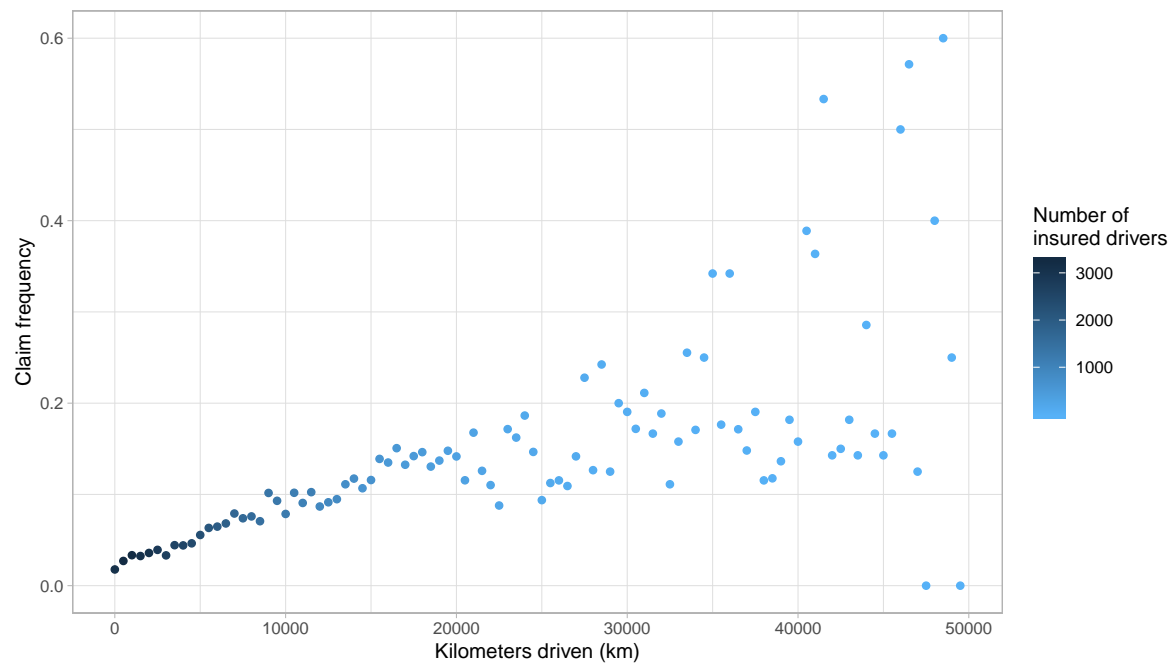
Figure 2.2.4 – Claim frequency vs Distance Driven

Boucher et al (2017) indicates that this trend could be explained by a learning effect, where policyholders take more and more driving experience and thus improve their driving quality. This explains why the marginal effect of an extra kilometer on the frequency has less effect for a high distance value. Although I was the author of this article, I now disagree with this explanation : we can imagine a learning effect for very young drivers (maybe like the Spanish Database ?), but for the general population, the learning effect on a few extra kilometers do not exist for drivers who already have a lot of driving experience.

Chapitre 3

Duration Models

Different risk exposures allow us to revisit classic dependence definitions, more precisely the occurrence dependence and the duration dependence.

1. **Occurrence dependence**

2. **Duration dependence**

3.1 "Time" between Accidents

Because we have now access to various risk exposition, we will focus on duration dependence to model the *time* between two claims.

Among many properties of the distributions, such as the mean (denoted μ), the variance (denoted σ^2), the coefficient of variation (denoted $\kappa = \sigma/\mu$) or other higher moments, the hazard function is an interesting function that can be used to compare waiting time distributions. The hazard function described the underlying dependence of the process and is defined as :

$$\gamma(t) = \frac{f(t)}{1 - F(t)}$$

where $f(t)$ and $F(t)$ are the density and the cumulative density of the waiting time between two events. For now, the variable t will refer to the calendar time between two claims, i.e. the number of days/weeks/years between two claims. As we want to redefine the risk exposure, we want to know if other measures, such as i) the distance driven, ii) the number of trips, or iii) the hours driven are better to express the duration dependence.

If the hazard function is monotonic, we know that :

$$\frac{d\gamma(t)}{dt} \begin{cases} > 0 \Rightarrow \kappa < 1 \\ = 0 \Rightarrow \kappa = 1 \\ < 0 \Rightarrow \kappa > 1. \end{cases}$$

The distribution displays negative duration dependence for $\frac{d\gamma(t)}{dt} < 0$ and positive duration dependence for $\frac{d\gamma(t)}{dt} > 0$. The count distribution implied by the waiting time process exhibits overdispersion i.i.f. $\kappa > 1$, while it exhibits underdispersion i.i.f. $\kappa < 1$. Equidispersion, overdispersion and underdispersion are key concepts for count data. We will return to this.

When the hazard function does not depend on how much time has been spent, the model does not have duration dependence.

3.2   Count Model Construction

As for the modeling of the hazard rate function, there are many ways to model a count distribution based on a specific form of a hazard rate function. We can use full parametric models where the waiting time between two events is modeled by a specific distribution.

The idea of the models using waiting times distribution is to model the number of reported events of a time horizon of t. Let τ_i be the waiting time between the the $i - 1^{th}$ event and the ith event. Consequently, it follows that the kth event occurs at time defined as :

From this construction, we can state that the relationship between the arrival time $\tau(i)$ and the count process $N(t)$ is :

From this last relationship, we can compute the probability function of a count process, using waiting time process :

Note that even if we suppose different "time" process for the arrival of a claim, we still work with count data and we want to model the number of claims N_i of insured $i = 1, \ldots, n$. To derive count models, many processes of arrival time can be used. The

only restriction is on distributions having domain in the positive values since it models the waiting time between two successive events.

3.2.1 Exponential Waiting Time

If is assumed that the waiting time between two events is following an exponential distribution with mean parameter λ :

$$f(\tau; \lambda) = \lambda e^{-\lambda \tau}.$$

Tt is well known that the underlying count process is a Poisson distribution of mean $t\lambda$. Indeed, the density of τ_k corresponds to a sum of k independent exponential distributions. This represents a gamma distribution of parameters λ and k, and then :

$$F_k(t) = \frac{1}{\Gamma(k)} \int_0^t \lambda^k u^{k-1} e^{-\lambda u} du,$$

This last integral can be expressed as :

which can be expressed as a Poisson probability function :

$$P(N(t) = k) =$$

We can show that $E[N] = Var[N] = \lambda$. Because of this equality, we say that the Poisson has an equidispersion property. In other words, when we know the mean of the distribution, the variance is set.

The hazard function of the waiting time between two events is equal to λ. The hazard function gives other interesting properties of the Poisson distribution. Indeed, since the hazard function does not depend on t (memoryless property of the exponential distribution), the model does not imply duration dependence. Finally, the hazard function confirms what we have seen in the construction of the Poisson count distribution : the mean parameter of the model is proportional to the observed time length.

As we do with classic GLM models, covariates can also be added in the mean parameter of the Poisson distribution. If we suppose p covariates $X_{i,1}, \ldots, X_{i,p}$, we can use the following structure :

$$\begin{aligned}
\lambda_i &= \exp(\beta_0 + X_{i,1}\beta_1 + \ldots + X_{i,p}\beta_p) \\
&= \exp(X_i'\beta).
\end{aligned}$$

where X_i is a $(p+1)$ vector of risk characteristics, and β a $(p+1)$ vector of parameters to be estimated. Classically, we use maximum likelihood to estimate the parameters.

In empirical situations, the Poisson distribution should always be the starting point when we want to model count data. The Poisson is a member of the linear exponential family, which means that the GLM theory applies. And without going into details, the Poisson distribution has other some very useful statistical properties. One of the most important is the consistency of the MLE β parameters, even when the distribution is not Poisson. In this situation, the standard error of the estimated parameters are not necessarily consistent, but it exists ways to correct it.

However, the equidispersion property of the Poisson causes problem since the mean of a dataset is almost never equal to the variance. If we take a look on the expected value of the number of claims, in chapter 2, Table 2.1.1, we see that $Var[N] = 0.0737 > E[N] = 0.0705$. In practice, the Poisson distribution is then often rejected in favor of other count distributions.

Alternatives to the Poisson distribution

The first alternative of the Poisson distribution for count regression is the negative binomial. It can be expressed in at-least two forms :

$$\Pr(X = k) = \frac{\Gamma(\lambda/\theta + k)}{\Gamma(\lambda/\theta)\Gamma(k+1)} \left(\frac{1}{1+\theta}\right)^{\lambda/\theta} \left(\frac{\theta}{1+\theta}\right)^k \quad (NB1(\lambda,\theta))$$

$$\Pr(X = k) = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)\Gamma(k+1)} \left(\frac{\alpha}{\alpha+\lambda}\right)^\alpha \left(\frac{\lambda}{\alpha+\lambda}\right)^k \quad (NB2(\lambda,\alpha))$$

For both distributions, $E[X] = \lambda$, but $Var[X] = \lambda(1 + \theta)$ for the NB1, and $Var[X] = \lambda + \lambda^2\alpha^{-1}$ for the NB2 [1]. For each distributions, overdispersion can be handled.

In actuarial sciences, we justify the use of the NB1/NB2 by saying that the mean of the count process should have an heterogeneity

---

1. Be careful with the name of the distribution : some books or R packages did not chose the same name for the distribution (NB1 instead of NB2, and NB2 instead of NB1)

term, caused by the omission of some important classification variables (swiftness of reflexes, aggressiveness behind the wheel, consumption of drugs, etc.).

Indeed, if we suppose $S|\theta \sim Poisson(\lambda\theta)$, with $\theta \sim Gamma(\alpha, \alpha)$, i.e. :we have :

$$
\begin{aligned}
\Pr[S = s_t|\theta] &= \frac{(\lambda\theta)^s e^{-\lambda\theta}}{s!} \\
f(\theta) &= \frac{\alpha^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\alpha\theta}
\end{aligned}
$$

we have :

which is the NB2 distribution. If we instead suppose $S|\theta \sim Poisson(\lambda\theta)$, with $\theta \sim Gamma(\alpha^{-1}\lambda, \alpha^{-1}\lambda)$. It can be shown that we obtained the NB1 distribution. Fun fact : the sum of the NB1 distribution is a NB1 distribution, but the sum of the NB2 distribution is not a NB2 distribution.

A classic analysis of the distribution of claim counts clearly shows an excess of zero (insureds without claim) compared to the fit of a Poisson distribution[2]. for that reason, another count distribution to use is the zero-inflated Poisson distribution. This distribuiton has the following probability function :

$$
\Pr(N_i = k) = \begin{cases} \phi + (1-\phi)\exp(-\lambda_i) & \text{for } k = 0 \\ (1-\phi)\frac{\lambda_i^k \exp(-\lambda_i)}{k!} & \text{for } k = 1,2,... \end{cases}
$$

We can show that $E[N_i] = (1-\phi)\lambda_i$ and $Var[N_i] = E[N_i] + E[N_i](\lambda_i - E[N_i])$. If we inflate the zero-part of the distribution, $\phi > 0$, and the model can deal with dispersion. Another interesting part of the model is that it is possible to add covariates in the phi_i parameter, meaning that we can suppose :

$$
\phi_i = \frac{\exp(x_i'\gamma)}{1 + \exp(x_i'\gamma)},
$$

with γ a vector of parameters.

---

2. Some authors claimed that this is caused by the fact that some accidents are not claimed

3.2.2   Gamma Waiting Time

We come back to "time" between claims. After the exponential distribution, it is also possible to use an other distribution with a nonconstant hazard function, such as the Gamma distribution. The Gamma distribution has the following density :

$$f(\tau; \alpha, \lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\lambda \tau}$$

The gamma distribution has mean equal α/λ and variance equal to α/λ^2. The distribution nests the exponential distribution for $\alpha = 1$. The construction of the count model can be done as for the exponential distribution. Indeed, since we have to consider the arrival time of the kth event, such as $\nu(k) = \sum_{i=1}^{k} \tau_i$ The reproductive property of the gamma distribution, for distributions sharing the same λ parameter implies that $\nu(k)$ is also gamma distributed with density :

$$f(\nu; \alpha, \lambda) = \frac{\lambda^{k\alpha}}{\Gamma(k\alpha)} \nu^{k\alpha-1} e^{-\lambda \nu}$$

Consequently, the cumulative density function of the waiting time process can be expressed as :

$$
\begin{aligned}
F_k(t) &= \frac{1}{\Gamma(k\alpha)} \int_0^t \lambda^{k\alpha} \nu^{k\alpha-1} e^{-\lambda \nu} d\nu \\
&= G(t; \text{scale} = \alpha k, \text{shape} = \lambda)
\end{aligned}
$$

where the integral part is known an as the incomplete gamma function. The Gamma count distribution (GCD), using $\lambda_i = \exp(x_i \beta)$, can be expressed as :

$$P(N_i(t_i) = n_i) = G(t_i; \alpha n_i, \lambda_i) - G(t_i; \alpha n_i + \alpha, \lambda_i)$$

with $G(t_i; 0, \lambda_i) = 1$. This probability can be calculated using preprogrammed functions such as the *pgamma* function in R. Additionally, the incomplete gamma function can be evaluated using integrations approximations or asymptotic expansions, such as :

$$\int_0^z e^{-t} t^{a-1} dt = \sum_{n=0}^{\infty} \frac{(-1)^n z^{a+n}}{(a+n)n!}$$

For $\alpha = 1$, the waiting times is exponentially distributed and the count distribution is the Poisson distribution as seen earlier.

For other integer values of α (2,3,4,...), the gamma distribution has a Erlangian form and the related count distribution is :

For non-integer value of α, which is what is usually observed for claims number, the Gamma Waiting Time process does not produce a closed-form hazard function, since it obeys to the following equation :

$$\frac{1}{\gamma(t)} = \int_0^\infty e^{-\lambda u} \left(1 + \frac{u}{\tau}\right)^{\alpha - 1} du$$

Nevertheless, we can see that the hazard function is increasing for $\alpha > 1$ and is decreasing for $\alpha < 1$ (and shows constant hazard of $\alpha = 1$). Thus, for $\alpha \neq 1$, the model exhibits duration contagion since the probability to have an event depends on the time of the last event. The expected value of the gamma distributed waiting times do not have a closed-form, but can be computed given by :

$$
\begin{aligned}
E[N_i(t_i)] &= \sum_{j=0}^\infty P(N_i(t_i) = j) \\
&= \sum_{j=0}^\infty \left(G(t_i; \alpha j, \lambda_i) - G(t_i; \alpha j + \alpha, \lambda_i)\right)
\end{aligned}
$$

Weibull Waiting Time

An other common model used in the duration analysis is the Weibull distribution. As for the Gamma distribution, the Weibull has the attractive property of nesting the Exponential distribution, meaning the we can test directly the Weibull count distribution against the Poisson. The density of this distribution is the following :

$$f(\tau; c, \lambda) = \lambda c \tau^{c-1} \exp(-\lambda \tau^c)$$

for $\lambda > 0$ and $c > 0$.

As opposed to the Gamma distribution, however, the Weibull distribution does not have the reproductive property. Indeed, the sum of Weibull distributions is not a Weibull distribution. Nevertheless, to find the time arrival of the kth event, **?** use the k-fold convolution of the interarrival time distribution, which has the form $\int_0^t F(t - s) f(s) ds$.

This gives the following expression for the Weibull count distribution (WCD) :

$$P(N(t) = k) = \sum_{j=k}^\infty \frac{(-1)^{j+k} (\lambda t^c)^j \alpha_j^p}{\Gamma(cj + 1)}$$

where

$$
\begin{aligned}
\alpha_j^0 &= \frac{\Gamma(cj+1)}{\Gamma(j+1)} \\
\alpha_j^{p+1} &= \sum_{m=p}^{j-1} \alpha_m^p \frac{\Gamma(cj - cm + 1)}{\Gamma(j - m + 1)}, \quad p = 0,1,2,\dots \ \ j = p+1, p+2, \dots
\end{aligned}
$$

Under the parameter restriction $c = 1$, the distribution is exponentially distributed.

Expected value and variance of the WCD are as follow :

$$
\begin{aligned}
E[N(t)] &= \sum_{p=1}^{\infty} \sum_{j=p}^{\infty} \frac{p(-1)^{j+p}(\lambda t^c)^j \alpha_j^p}{\Gamma(cj+1)} \\
Var[N(t)] &= \sum_{p=2}^{\infty} \sum_{j=p}^{\infty} \frac{p^2(-1)^{j+p}(\lambda t^c)^j \alpha_j^p}{\Gamma(cj+1)} - \left(\sum_{p=1}^{\infty} \sum_{j=p}^{\infty} \frac{p(-1)^{j+p}(\lambda t^c)^j \alpha_j^p}{\Gamma(cj+1)} \right)^2
\end{aligned}
$$

Modified Counts Distributions

Analyzing the WCD allows us to see that the probability distribution and the mean of the distribution use λt^c. Seeing that the mean of the Poisson distribution is equal to λt, we can construct another form of Poisson distribution having the following probability distribution :

$$
P(N(t) = n) = \frac{e^{-\lambda t^c}(\lambda t^c)^n}{n!}.
$$

where c is now a parameter to estimate (by maximum likelihood, for example).

Based on real insurance data, Boucher and Denuit (2004) showed that the GCD and WCD are way better than the Poisson distribution to model the number of claims. They also showed that the estimated value of the c parameter of the WCD is really close to $\hat{c}$ of the modified Poisson distribution. Moreover, they calculated that the estimated premiums based on different value of t, which showed that the modified Poisson distribution seems to offer a good approximation to the WCD, as we can see in table 3.2.1.

In consequences, the modified Poisson, simpler to use than the WCD (in addition to its interesting theoretical properties), could be favored in modeling. The estimated value of the c parameter could claim to be a proxy of the c parameter of the WCD. We then suppose that the number of claims N_i of insured i follows a Poisson(λ_i) distribution, with :

Time	Poisson		GCD		WCD		Mod. Poisson	
1.0	0.1975	(0.2749)	0.1713	(0.1960)	0.1718	(0.1967)	0.1645	(0.2416)
0.9	0.1777	(0.2404)	0.1643	(0.1871)	0.1650	(0.1883)	0.1638	(0.2312)
0.8	0.1580	(0.2075)	0.1571	(0.1780)	0.1577	(0.1791)	0.1577	(0.2202)
0.7	0.1382	(0.1762)	0.1491	(0.1679)	0.1491	(0.1680)	0.1511	(0.2084)
0.6	0.1185	(0.1463)	0.1404	(0.1571)	0.1407	(0.1576)	0.1438	(0.1957)
0.5	0.0987	(0.1181)	0.1309	(0.1454)	0.1311	(0.1455)	0.1356	(0.1817)
0.4	0.0790	(0.0914)	0.1205	(0.1327)	0.1205	(0.1329)	0.1262	(0.1662)
0.3	0.0592	(0.0662)	0.1079	(0.1178)	0.1080	(0.1179)	0.1150	(0.1482)
0.2	0.0395	(0.0426)	0.0927	(0.1001)	0.0922	(0.0995)	0.1010	(0.1265)
0.1	0.0197	(0.0205)	0.0719	(0.0764)	0.0713	(0.0757)	0.0808	(0.0971)

Table 3.2.1 – Expected values for the Poisson distribution, the GCD, the WCD and the modified Poisson distribution for different exposure time

If we can directly correct the Poisson distribution with the addition of a c parameter, we can obviously use other count distributions to model the number of claims. For more informations, we refer the reader to the book of Denuit et al.(2007), or Boucher et al.(2007).

3.2.3 Applications

We can use the Canada dataset with the proposed models. Note, that normally, covariates selection should be done, but for illustration, only gender, marital status and vehicle use will be used in the X vector. However, you can test other covariates in the models if you want. We will also focus our analyses on two definition of the exposure : the calendar time (variable RA_EXPOSURE_TIME) and the distance driven (RA_DISTANCE_DRIVEN). In exercices, other definitions of exposure could be used (number of trip, hours driven, etc.)

Validation Set

An easy way to analyse the dataset is to model the number of claims with a Poisson distribution, using the GLM package of R.

Fitting results are shown in Table 3.2.2. By looking at the AIC statistic (all models have the same number of parameters), we clearly see how the addition of an offset variable improves the fitting : 34,093 versus 33,396 or 33,288. Comparing the model with distance as an offset, with the more traditional model that instead uses exposure time shows us that the distance traveled seems to be a better way to define the risk exposure in automobile insurance.

Differences seen in the values of $\widehat{\beta_0}$ for all models come from the offset variables, where exposure time is in [0,1], while distance driven can goes up to 60,000. For illustration, we also added three covariates in the modeling to verify the impact of the models on the risk segmentation. We can see that the importance of the regressors is not the same for each model. We can also see that some covariates do not seem statistically significant. More interestingly, although the standard deviation of $\widehat{\beta}$ is about the same for each model, $\widehat{\beta}$ seem much less important for the model using distance as an offset variable. We can even draw a different interpretation of the impact of the variable RA_VEH_USE on the claim frequency for the last model : having a positive impact of the first two models, the use of the vehicle to commute seems to decrease the expected frequency when the distance driven is used as an offset variable.

Parameter	Without offset		Exp.Time		Distance	
	Estimate	Std. Err.	Estimate	Std. Err.	Estimate	Std. Err.
β_0	-2.72750	0.03486	-2.02841	0.03476	-11.53777	0.03494
β_1 (GENDER=1)	-0.02971	0.02956	0.02194	0.02953	-0.02502	0.02950
β_2 (MSTATUS=1)	-0.08082	0.03099	-0.08385	0.03092	-0.06829	0.03091
β_3 (VEHUSE=1)	0.25715	0.03040	0.19597	0.03036	-0.06628	0.03038
AIC	34,093		33,396		33,288	

Table 3.2.2 – Summary Statistics of Poisson regression

Exemple 3.2.1. Compute the expected frequency (that we can refer to the frequency part of the insurance premium) for the three models, for a single male driver, who drove exactly 12,000 kilometers for 8 months, who use his car to commute.

Solution

Note that instead of using the GLM function, we can maximise the loglikelihood with exposure as an exponential distribution. This way to express the model suppose an exponential distribution "time" between claims. We already showed that this corresponds to a classic Poisson distribution, with an offset variable. The estimated parameters obtained by this way of estimating the model will give the exact same results as the previous GLM example. The interest of this other algorithm is to let α be estimated (instead of setting $\alpha = 1$ for the exponential) : we then obtain a more general gamma distribution for the "time" between claims, meaning that we are deriving the GCD distribution.

Fitting results for the two duration models are shown in Table 3.2.3.

Once again, the best model is the one that considers the distance driven instead of the exposure time as the exposure to risk

Parameter	Exp. Time		Distance	
	Est.	Std.Err.	Est.	Std.Err.
β_0	-4.011	0.187	-13.833	0.167
β_1 (GENDER)	0.012	0.047	-0.032	0.051
β_2 (M. STATUS)	-0.136	0.050	-0.122	0.054
β_3 (VEH.USE)	0.351	0.051	0.096	0.054
α	0.602	0.022	0.558	0.017
Log-likelihood	-16,559.7		-16,357.18	
AIC	33,129.4		32,724.36	

Table 3.2.3 – Summary Statistics for Duration Models

(33,129.4 versus 32,724.36 for the AIC). We can also compare the fit of the GCD with the Poisson distributions. However, instead of comparing the AIC values, standard statistical tests can be done. Indeed, we saw that a GCD with $\alpha = 1$ is equivalent to a Poisson distribution. We can then do a Wald test on the estimated value of $\widehat{\alpha}$ to see is the estimated parameter is statistically different from the null hypothesis $\alpha = 1$. A log-likelihood ratio test can also be done. In our cases, with $\widehat{\alpha}(std.err) = 0.602(0.022)$ for the exposure time, and $\widehat{\alpha}(std.err) = 0.558(0.017)$, we can clearly see that the Poisson distribution is rejected against the GCD. In other words, we can say that the exponential distribution for the "time" between claims is rejected against the Gamma distribution for the exposure time, and for the distance driven.

Exemple 3.2.2. Compute the expected frequency (that we can refer to the frequency part of the insurance premium) for the the GCD, for a single male driver who use his car to commute if :

1. 12,000 kilometers for 8 months ;

2. 6,000 kilometers for 4 months ;

Solution

We can see that the premiums generated by the GCD model are not the same proportion as the exposure time, or the driving distance. This is an interesting aspect of the GCD model : it seems well suited for PAYD insurance.

With the value of $\widehat{\alpha}$, we can deduce the time between claims. This is interesting because **we did not observe the time between**

claims, but only observed the number of claims, from which we infer the time between claims.

Note that the estimated value of α is less than one. That means that the model suppose positive duration dependence : the report of an accident decreases the expected time to report an other claim. This is usually observed with insurance data. However, it is not clear if it is caused by duration dependence directly, or because of the heterogeneity of the insurance portfolio. In insurance, situation with $\alpha > 1$ are observed less often. Earthquakes might be an example of negative dependence, where the occurence of an earthquakes increases the expected time of another earthquake (related to the tension forces of tectonic plates).

To conclude this approach, note that an important criticism of this GCD model is the fact that we do not observe the starting time of the count process. Indeed, we consider the beginning of the contract as the starting point to estimate the time "between" claims. We can still use this model for ratemaking, but we should be careful about what the model means.

We see that we used exposure time or the distance driven for the definition of what might be exposure to risk. Ideally, it is conceivable that adequate risk exposure could be defined as a combination of these two variables, or several other variables. For example, in the database, we also have access to the number of trips and the total driving time. We could imagine that in certain circumstances, the driving time might better represents the risk (in traffic or rush hour), while in other situations, the distance traveled might be a better choice (on the highway). By analogy, we can even take the example of taxi pricing.

⭐ **Taxi in Montreal**

— Fees at the start 3,50 $

— Price by km : 1,75 $ / km

— Price for each waiting minute 0,65 $ / minute (39,00 $ / hour)

Note : When the taxi is traveling less than 22,286 km/h, the taximeter is in timer mode and it totals 0.65$ per minute(or fraction of a minute). As soon as the speed of the taxi is equal to or greater than 22,286 km/h, it is the distance traveled of which the taximeter takes into account 1.75 $ per kilometer (or fraction of a kilometer).

Source : Commission des transports, Gouvernement du Québec

We will skip the WCD (Weibull count distribution), and instead expose the results of the modified Poisson distribution. This distribution is interesting because it can handle both the distance and the exposure time as covariates. Resultas are shown in table .

If we only focus on some fitting statistics, for example the AIC, the modified Poisson with both the logarithm of exposure time and the logarithm of the distance driven put as covariates, generates the best AIC among all distributions studied so far. As a simple Poisson distribution, properties of the GLM still hold, and the model is easy to use for ratemaking. Other modified count

Parameter	Exp. Time		Distance		Exp. Time and Distance	
	Estimate	Std. Err.	Estimate	Std. Err.	Estimate	Std. Err.
β_0	-2.24134	0.03762	-7.67383	0.15603	-7.08170	0.24775
β_1 (GENDER=1)	0.00700	0.02956	-0.01830	0.02952	-0.01354	0.02956
β_2 (MSTATUS=1)	-0.08616	0.03094	-0.07860	0.03095	-0.07963	0.03095
β_3 (VEHUSE=1)	0.22303	0.03041	0.05689	0.03086	0.06922	0.03115
β_4 (Exp. Time)	0.61992	0.02276	.	.	0.10325	0.03395
β_5 (Distance)	.	.	0.57337	0.01708	0.51422	0.02574
AIC	33150		32,728		32,721	

Table 3.2.4 − Summary Statistics of the Modified Poisson regression

distributions can also be used, such as the NB2, the NB1 or the ZIP. For the ZIP, we can even imagine putting exposure base variables in the extra-parameter that inflates the probability of zero.

Count regression is often complex because, for now, it does not seem to exist a goodness-of-fit test to verify if the estimated model fits correctly the data. In insurance, most of the insureds do not report at all (>85-90%), at the estimated mean for each insured is always around (0.05-0.20). Straight residuals ($n_i - \lambda_i$) of count regression cannot be used directly. This mean that it is difficult to verify if the model predict correctly the count distribution. One way to compare model is to check on out-of-sample data.

Test Set (Out-of-Sample)

For now, we mainly focus on the estimation part of the dataset. It can also be interesting to analyse how each model perform on out-of-sample data. For count data, as mentioned for the fitting diagnostics, it is difficult to see if the model accurately estimate the mean parameter for a specified insured. That is the reason why we instead use scores to compare the performances of models. As mentioned in Verbelen at al. (2017), these scores measures the quality of probabilistic forecasts, for the predictive distribution P and the observed count n. A lower scores indicate a better quality of the forecast. Table 3.2.5 shows some of the score that we can use, where pk means $P(N = k)$ and Pk means $P(N \leq k)$. The mean is represented by μ_p, and standard deviation by σ_p. Finally, $\|p\| = \sum_{k=0}^{\infty} p_k^2$.

Results of the out-of-sample analysis, for the logarithm score and the squared error score are shown in Table 4.3.3. The results with out-of-sample dataset are similar to what we saw previously, where the modified Poisson (with exp. time and the distance) is the best model, while the GCD model with distance is second.

To conclude those parametric models, and before introducing the next chapter, note that if $\log(t_i)$ and $\log(d_i)$ (for t as exposure base and d as distance) are used as a covariate in a count regression model, it is obvious that one might wonder why not use another function of time t or d in the Poisson mean parameter. For example :

Score	Formula
Logarithmic	$\mathrm{logs}(P,n) = -\log(p_n)$
Quadratic	$\mathrm{qs}(P,n) = -2p_n + \|p\|$
Spherical	$\mathrm{sphs}(P,n) = -p_n/\|p\|$
Ranked probability	$\mathrm{rps}(P,n) = \sum_{k=0}^{\infty}\{P_k - \mathbf{1}(n \leq k)\}^2$
Dawid-Sebastiani	$\mathrm{dss}(P,n) = \left(\frac{n-\mu_p}{\sigma_p}\right)^2 + 2\log(\sigma_p)$
Squared error	$\mathrm{ses}(P,n) = (n - \mu_p)^2$

Table 3.2.5 − Scores for out-of-sample count data

Model	Logarithmic	Squared Error
Modified Poisson (Exp.Time)	1974.909	7009.240
Modified Poisson (Distance)	1964.170	6939.918
Modified Poisson (Exp.Time + Distance)	1963.541	6938.488
GCD (Exp.Time)	1974.868	7007.247
GCD (Distance)	1964.295	6941.219

Table 3.2.6 − Scores for out-of-sample

$$\lambda_i \;=\; \exp(\boldsymbol{X}_i^T\boldsymbol{\beta} + \gamma f(t_i)),$$

for some parametric function $f(.)$. A polynomial function of t_i should then be something to consider :

$$\lambda_i \;=\; \exp(\boldsymbol{X}_i^T\boldsymbol{\beta} + \gamma_1 t_i + \gamma_2 t_i^2 + ... + \gamma_s t_i^s).$$

where $\gamma_1, \ldots, \gamma_s$ would be new parameters to estimate. Such an approach is obviously not recommendable in statistics, and it is appropriate to use a much more efficient approaches like semi-parametric methods.

If we use a parametric approach for a single criterion defining risk exposure, like the exposure time t_i, it would be easy to generalize the model into a multi-dimensional approach where, for example, the number of claims N_i of insured i would follow a Poisson distribution (λ_i) with :

$$\lambda_i \;=\; \exp(\boldsymbol{X}_i^T\boldsymbol{\beta} + \gamma_1 f_1(t_i) + \gamma_2 f_2(km_i)),$$

or more generally :

$$\lambda_i \;\; = \;\; \exp(\boldsymbol{X}_i^T \boldsymbol{\beta} + \gamma_3 f_3(t_i, km_i)),$$

where $f_j(t_i)$ is a function of the time exposure, $f_2(km_i)$ would be a function of the distance driven and $f_3(t_i, km_i)$ a joint function of time exposure and distance.

Before addressing this kind of approaches, it should be mentioned that from the moment we dropped the GCD and the WCD, and use the modified Poisson distribution (or NB2, NB1, etc.), we kind of lose the interpretation of the model in terms of risk exposure definitions. While the λ, α and c parameters of the GCD and the WCD allowed us to know the time distribution between two claims, it will be much more difficult to explain what the parameters of the modified count distribution models mean. That does not mean that we cannot use them in practice, or for ratemaking since we saw that the fitting of the distribution is nice, as they have nice predictive capacities (scores on out-of-sample). Simply said : there is thus a tradeoff between adjustment and understanding of the model.

Chapitre 4

Flexible Models : GAM Models

Parametric approaches, like the duration models seen in the previous chapter, are often too rigid and can hide some properties of the data studied. To go beyond those models and the conclusions based on empirical observations and simple graphs, the number of claims will now be modeled using semi-parametric approaches, namely generalized additive models (which we will later call GAM). Such an approach will provide a better understanding of the impact of various definitions of risk exposure on the number of claims. For illustration, as done with the duration models, we will focus on the distance traveled (or driven) and the duration of the contract. Other definitions of risk exposure could be used for exercises.

4.1 Independent Cubic Splines

First, we model the number of claims N_i of insured i using a generalized additive model where we will study in details the impact of the distance traveled (km_i) and the duration of the insurance contract (d_i) of insured i. We will first suppose that the number of claims is following a Poisson distribution of mean λ_i, with :

$$log(\lambda_i) = \beta_0 + f_1(km_i) + f_2(d_i),$$

where β_0 is the intercept. GAM are quite flexible, but for illustration in this short-course, functions f_1 and f_2 are cubic splines, defined as a univariate smoothing function and having the following linear form :

$$f(x) = \sum_{k=1}^{q} b_k(x)\beta_k$$

where $\beta_k, k = 1 \ldots, q$ are parameters to be estimated by the GAM, and $b_k(x), k = 1,...,q$ are functions from cubic spline basis of dimensions q. Details can be found on functions $b_k(x), k = 1,...,q$ in Wood (2006, section 4.1.2). By doing this, similar techniques used to estimate a GLM can be used to estimate a GAM.
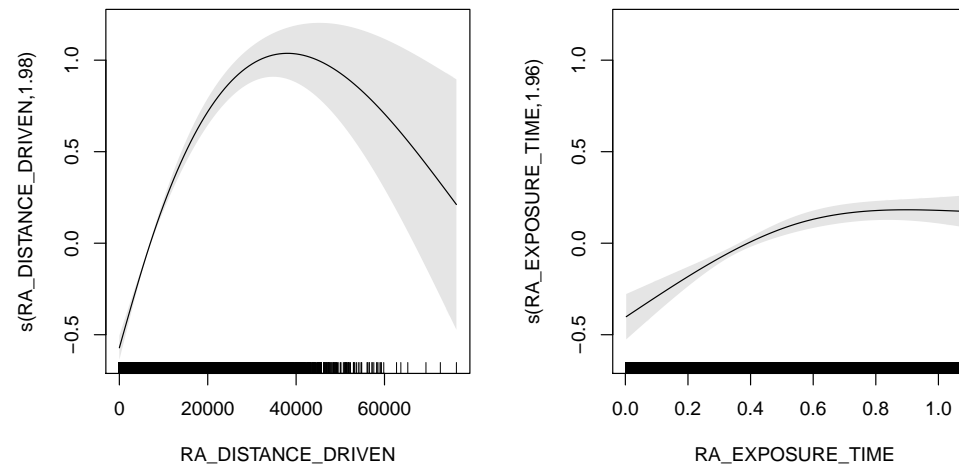
Figure 4.1.1 − Smoothing functions for $k = 2,5$ knots

For now, no β_p regressors, capturing the potential differences between insureds via risk characteristics such as age of the drive, driver's sex or marital status will be used. This will allow us to make a graphical and visual analysis more interesting for $f_1(km_i)$ and $f_2(d_i)$ since all insureds has the same basic frequence $(\exp \beta_0)$. Of course, it is possible that some insured show different behaviors than those illustrated here. GAM are flexible enough to easily add covariates in the mean parameter of the Poisson.

For this model, we have to chose the number of knots for f_1 and the number of knots for f_2. The choice of the number of knots is a manual process that depends on the desired degree of flexibility. Nevertheless, this is an important step because too few knots will produce an adjustment that may not capture significant trends in the data, while too many knots may lead to over-adjustment. To our knowledge, there does not seem to be a consensus among the scientific community about the determining the optimal number of knots. The choice of the number of knots therefore remains an important part of the modeling process and may depend on the application domain.

For illustration, we finally use $k = 6$ knots for the distance driven, and $k = 3$ for the exposure time, which results on smoothing function shown in figure 4.3.1. The black curves found in the two illustrations correspond to the predicted values for each function. The gray boxes correspond to 95 % confidence intervals on the predictions. At the bottom of the graphs, the density of the observations used in the modeling is displayed. Tables 4.1.1 and 4.1.2 shows the value of the parameters for the model.

The concept of degrees of freedom which is generally used in statistics is adapted in the context of GAM and is known as the
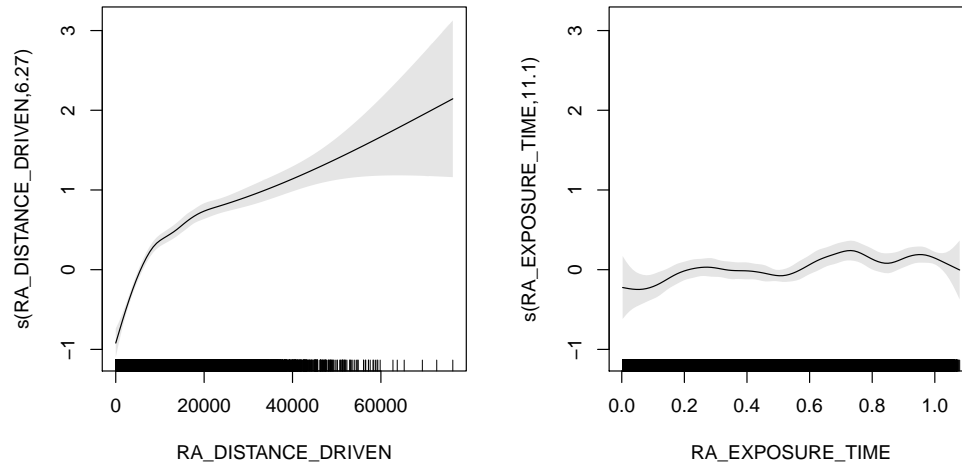
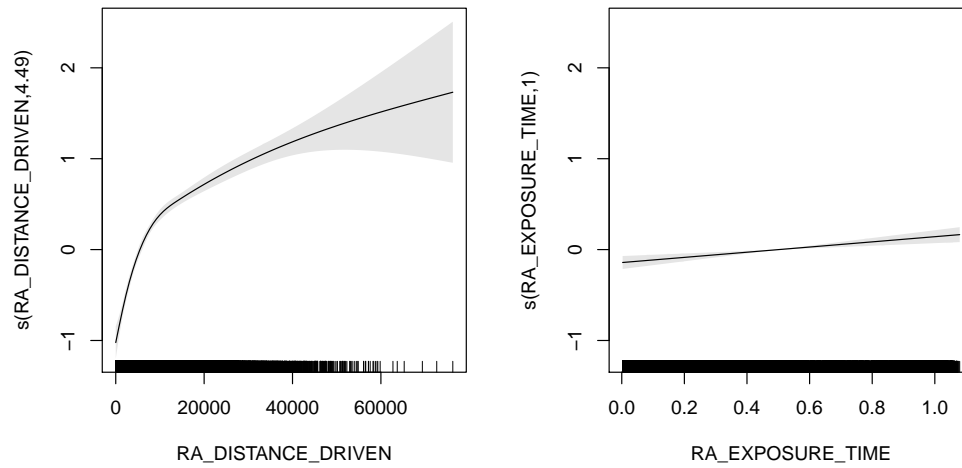Figure 4.1.2 – Smoothing functions for $k = 20$ knots

Figure 4.1.3 – Smoothing functions for the chosen GAM

	Estimate	Standard Error	t value	p-value
$\hat{\beta}_0$	-2.82252	0.01766	-159.8	2×10^{-16}

Table 4.1.1 − Results for the parametric parameter of model 4.3

	EDF	F Value	p-value
$\hat{f}_1(km)$	4.494	107.80	$< 2 \times 10^{-16}$
$\hat{f}_2(d)$	1.001	1.001	$< 2 \times 10^{-16}$
GCV	0.36282		

Table 4.1.2 − Results for the non-parametric parameters of model 4.3

effective degrees of freedom (EDF). If the smoothing parameters are zero, then the number of degrees of freedom for a smoothing function is simply the number of parameters to estimate minus one (due to the constraint that the function should add up to 1 for any given observation). If the smoothing parameters are not zero, the number of degrees of freedom is necessarily reduced and then the concept of effective degrees of freedom is considered in order to quantify the flexibility of a smoothing function or a general model. The F value shown in tables is the result of a statistical test similar to the Wald test to verify the significance of nonparametric terms. Finally, the generalized cross validation (GCV) is a method associated to the minimization of the smoothing parameter, where small score values are preferred. For more details we refer to Wood (2006), sections 3.2.3 and 4.8.5.

In our case, the low p-values show that both the parametric and the nonparametric effects in the model are significant. The GCV score is equal to 0.36282, but does not let us conclude anything at the moment, since it must be compared to another model. The GCV score is a statistical measure that only makes sense when several models are compared. We can see how the distance driven influences the probability of having a claim, as well as the exposure time to a lesser extent. For comparison, figure 4.1.4) exposes the same analysis with a different dataset (from Spain).

Among others, cubic splines, Generalized Cross Validation (GCV), Effective Degrees of Freedom (EDF), concepts that we have tackled quickly, are key elements to understand correctly generalized additive models. For more details on the theory, we refer to Wood (2006). We do not have enough time in this short-course to cover all this theory. The idea here is only to mention that semi-parametric modeling can be a useful tool to analyse telematic data.

4.2   Dependant Splines (with tensor product)

For the first model, the distance traveled and exposure duration were introduced as explanatory variables by using cubic smoothing splines. Firstly, they were parameterized completely independently, and later they were estimated. Now, we are going to show how the addition of an interaction term between the distance traveled and the exposure time changes the results obtained from the previous model. We will use the same modeling procedure, i.e., GAM with cubic splines. The difference is that instead of
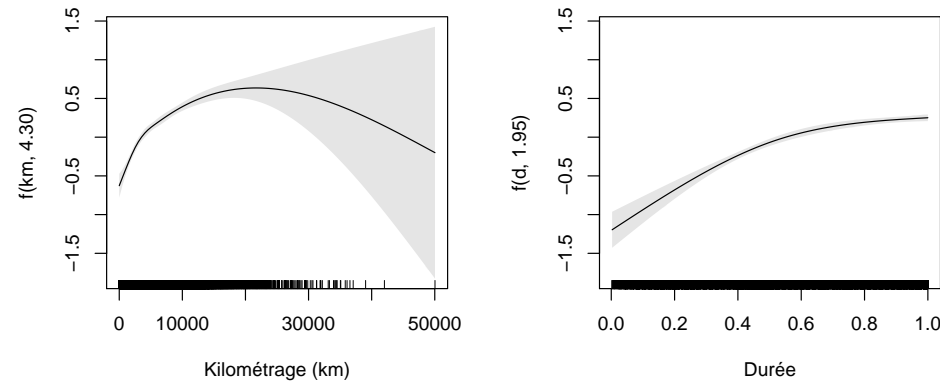
Figure 4.1.4 − Smoothing functions for spanish insurance data (from Boucher et al.(2016))

using two separate cubic splines to include the distance travelled and exposure time in the model, we will use a smoothing tensor product base. Broadly speaking, the idea is to introduce a smoothing function for the interaction.

Similarly to model with independent splines, only the distance traveled and exposure duration are used, as these are the variables explaining the number of claims of each insured. We will still denote the number of claims by N_i and assume that it follows a Poisson distribution. A multiplicative link function is used to relate the expectation with the linear predictor. As mentioned above, the tension product smoothing base is used to define a two variable function which is introduced in the GAM. Specifically, the model is formulated by the following equation

$$log(\mu_i) = \beta_0 + f(km_i, d_i),$$

where β_0 corresponds to the independent term of the model and $f(km_i, d_i)$ is a smoothing function which depends on the distance driven (km) and the duration (d). The tensor product smoothing base is implemented both in R and SAS softwares. However, it is also possible to apply the technique by using the procedure described by Wood (2006, section 4.1.8). Similarly to the model with independent cubic splines, the function is also parameterized by using 6 nodes for the distribution of the distance travelled and 3 nodes for the exposure duration. The model defined by 4.2 will be called model 2, while the first one presented in the previous section will be called model 1. Tables 4.2.1 and 4.2.2 show the estimation results.

The low p-values found in tables show that the nonparametric part of the model 2 is important to explain the number of claims of the policyholders. The value of the GCV score, which is equal to 0.36284, suggests no improvement of model 2 with respect to model 1, which provided a GCV score equal to 0.36282. In other words, no added flexibility is provided by the tensor product smoothing base. Since function $f(km_i, d_i)$ in model 2 is not expressed in terms of km or d separately, it is not possible to analyze

	Estimate	Standard Error	t value	p-value
$\hat{\beta}_0$	-2.82234	0.01766	-159.8	$< 2 \times 10^{-16}$

Table 4.2.1 – Results for the parametric parameter of model 4.2

	EDF	Valeur F	valeur-p
$\hat{f}(km,d)$	11.58	103.6	$< 2 \times 10^{-16}$
GCV	0.36283		

Table 4.2.2 – Results for the non-parametric parameters of model 4.2

the impact of the distance travelled and exposure time independently as we did in model 1. However, while Figure 4.2.1 shows the surface derived from the predictions produced by the estimation of model 1, Figure 4.2.2 shows the same surface for model 2 for every possible pair (km, d). Even if the two figures seem to have differences, those differences appear to be at places where almost no insured has been observed (more than 50,000 km driven for example).

4.3   Pricing Application

The purpose of this section is to compare the conventional methodology used in practice by insurance companies to model the frequency of claims reported by policyholders and the models presented in the previous section. The starting point considered by actuaries in modeling the frequency of automobile insurance claims is a GLM that assumes that the number of reported claims follows a Poisson distribution.

As in model 1 and 2, we will only consider the distance travelled (km) and exposure duration (d) to explain the number of accidents reported to the company by the insured. In order to generate a pricing system based on the results obtained by the GAM, the distance travelled will be categorized into several classes, including a reference category. Therefore, new regressors related to the distance travelled will be included in the model. To find the classes of distance, it is useful to look again at the smoothing functions chosen for the GAM.

First, for simplicity, we will completely exclude the exposure time in the model, as we see in the GAM models that his effect seems to be almost flat over [0,1]. For the distance driven, there is a sharp increase of claim frequency for the interval [0; 10,000], and the slope is lower (but constant) for distance greater than 10,000 km. As shown in Table 4.3.1, we can then proposed some binary variables to approximate the smoothing function of the distance driven. The reference category corresponds to a distance travelled ranging from 25,000 km and 40,000 km.

The goal here is to see if it is possible to easily replicate the previous results obtained with the GAM, but using a Poisson
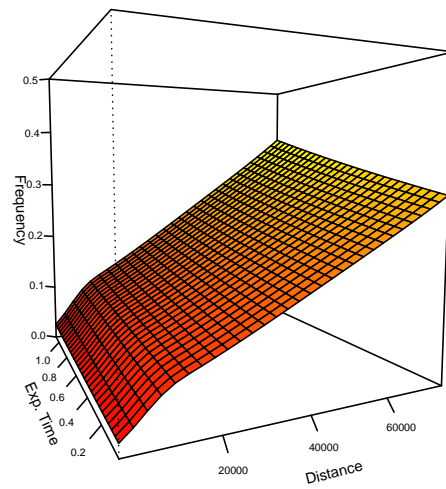
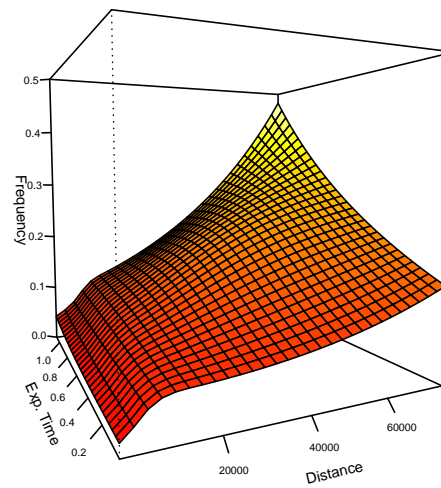Figure 4.2.1 – Smoothing functions for independent cubic splines

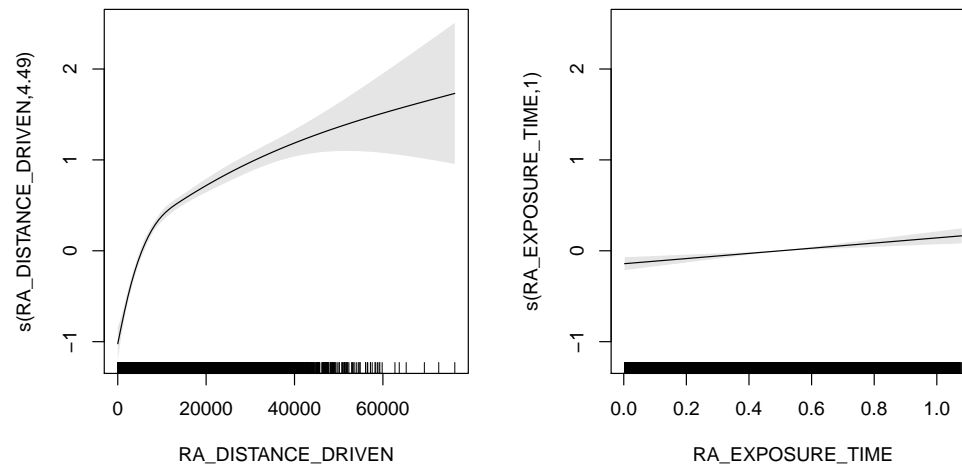Figure 4.2.2 − Smoothing functions for dependent cubic splines(tensor product)

Figure 4.3.1 – Smoothing functions for the chosen GAM

Variable	Description
x_1	Takes value 1 if km $\leq$ 3000
x_2	Takes value 1 if 3000 $\leq$ km $\leq$ 7000
x_3	Takes value 1 if 7000 $\leq$ km $\leq$ 12000
x_4	Takes value 1 if 12000 $\leq$ km $\leq$ 25000
x_5	Takes value 1 if km > 40000

Table 4.3.1 – Binary variables used for the segmentation of the distance traveled

GLM model, which is nowadays the standard model used in insurance companies. Let us assume that the number of reported claims follows a Poisson distribution. A multiplicative link is used to specify the relationship between the linear predictor and the expectation of the response variable. Specifically, the model is represented by the following equation :

$$log(\lambda_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_1 x_{3,i} + \beta_4 x_{4,i} + \beta_5 x_{5,i} + log(t_i)$$

where β_0 is the constant term in the model. Table 4.3.2 shows the results of the model. In particular, we observe that all parameters in the model are significant (knowing that it does not mean that all parameters are statistically different). Figures 4.3.2 and 4.3.3 show the surface of the predictions generated by this model. Note that this parametric model hardly approximates to the surface illustrated in the previous figures.

Variable	Estimate	Std. Err.	z value	Pr($>|z|$)
β_0	-1.689	0.054	-31.008	<0.001
β_1	-0.330	0.070	-4.750	<0.001
β_2	-0.440	0.063	-6.968	<0.001
β_3	-0.280	0.062	-4.538	<0.001
β_4	-0.188	0.060	-3.123	0.002
β_5	0.435	0.138	3.148	0.002

Table 4.3.2 – Results of the GLM with discretized distances

4.3.1   Comparison between GAM and the ratemaking by GLM structure

If the generalized additive models estimated at the beginning of this chapter seemed to be more flexible, and seem to offer a better fit to the auto insurance data than the generalized linear model with steps for the distances, a great way to assess the actual performance of a model (and thus avoid the risk of over-adjustment) is to put it to the test on out-of-sample data. We saw in the previous chapter that some scores can be used for count data. Once again, we compare each models.

Model	Logarithmic	Squared Error
GAM (ind. splines)	1963.363	6930.271
GAM (tensor prod.)	1963.311	6939.918
GLM	1964.053	6943.799

Table 4.3.3 – Scores for out-of-sample

On Figures 4.3.4, 4.3.5 et 4.3.6, another tool of studying out-of-sample performance for count data are proposed. Boxplots show the predicted residuals for each model, depending on the number of claims. The analysis of this kind of figures is not so easy. We expect the residual to be as close as zero as possible. On Figures 4.3.7, 4.3.8 et 4.3.9, a slightly different statistic is computed for each insured. Based on the logarithmic score, we computed the probability to observe n_i claims, for each insured i, depending on the model. In this case for a good model, we expect the boxplots to show that a lot on insured are close to a probability of one.
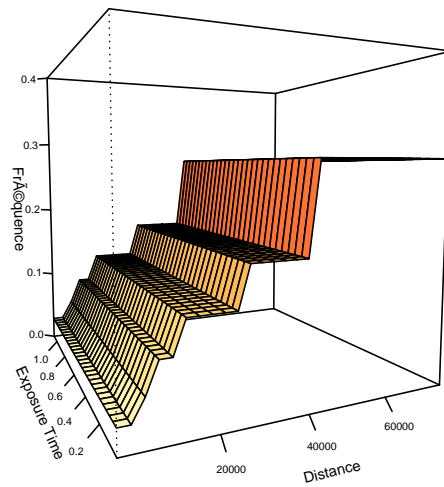
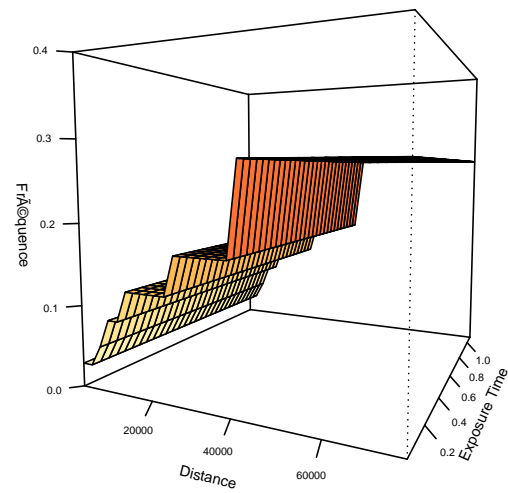Figure 4.3.2 − Smoothing functions for dependent cubic splines(tensor product)

Figure 4.3.3 – Smoothing functions for dependent cubic splines(tensor product)
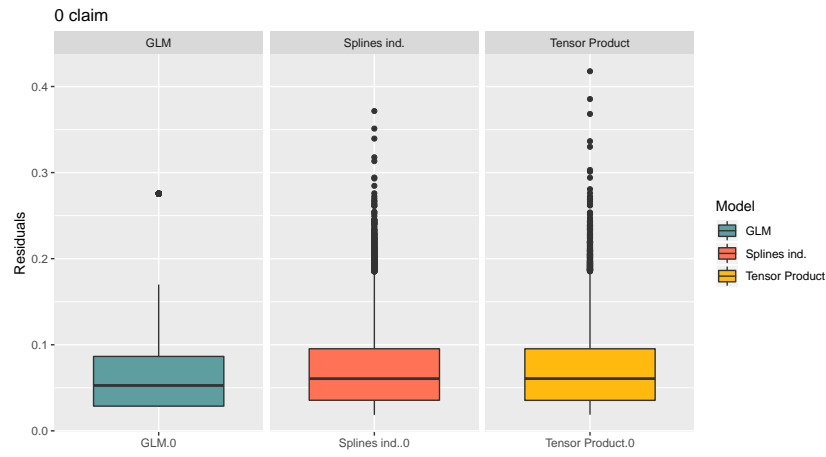
Figure 4.3.4 – Prediction residuals for each model, for insured without claims
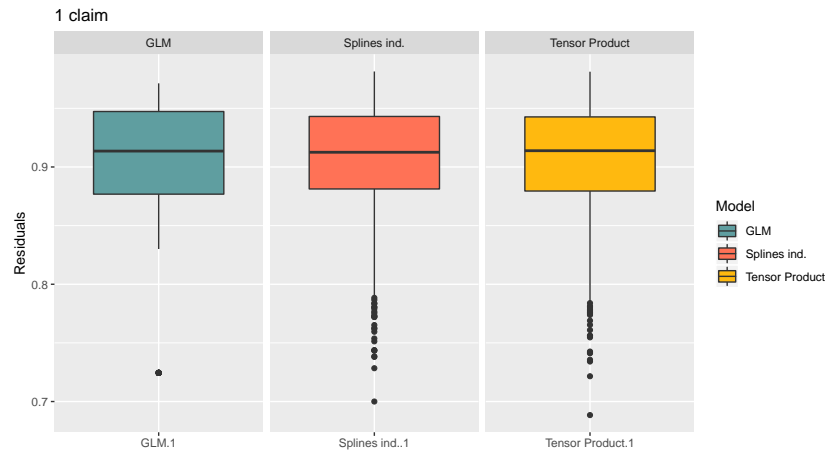

Figure 4.3.5 – Prediction residuals for each model, for insured with one claim
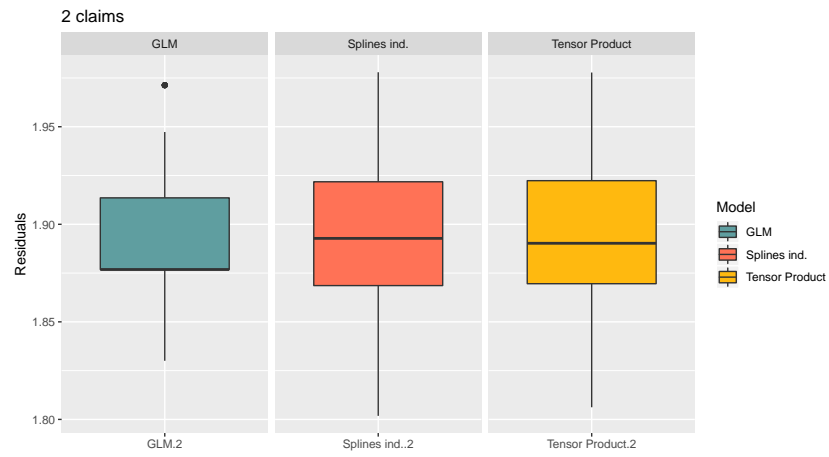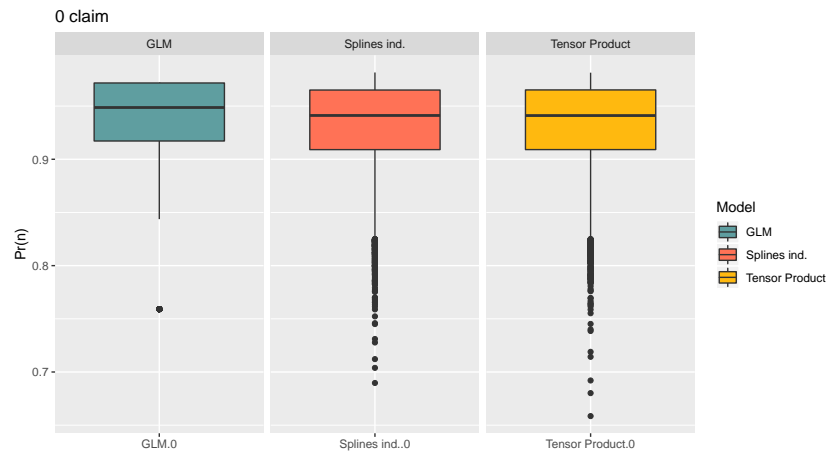
Figure 4.3.6 – Prediction residuals for each model, for insured with two claims

Figure 4.3.7 – Probability of observing n_i, for each insured i, for each model, for insured without claims
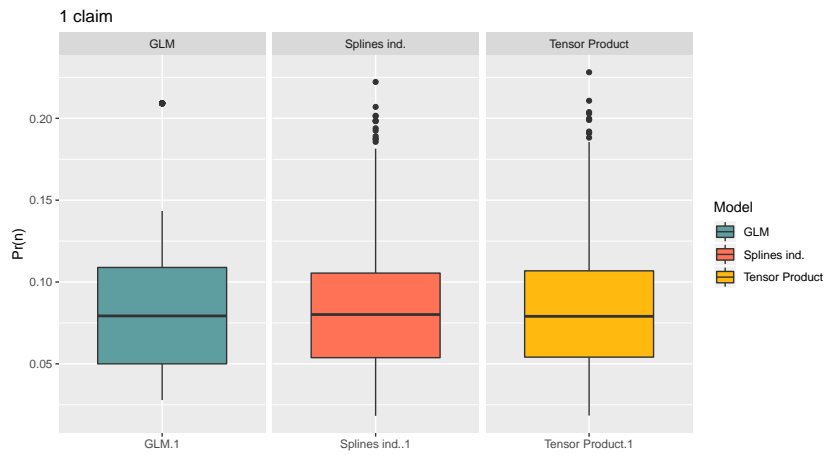
Figure 4.3.8 – Probability of observing n_i, for each insured i, for each model, for insured with one claim
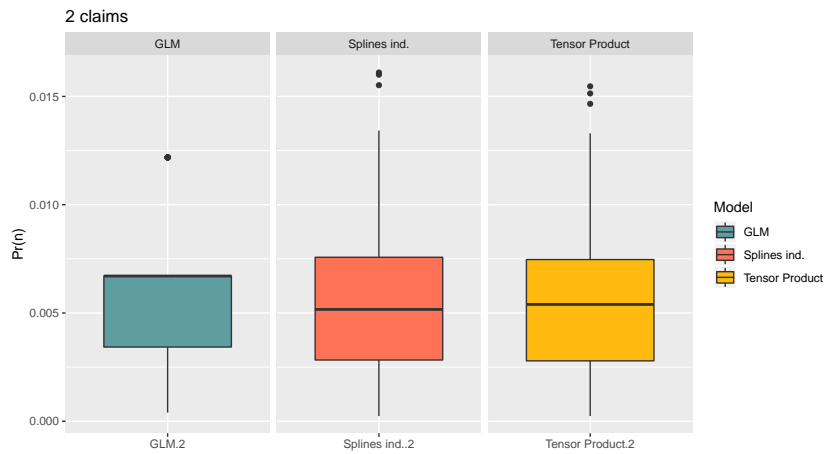

Figure 4.3.9 – Probability of observing n_i, for each insured i, for each model, for insured with two claims

The GAMs perform better than our simple GLM constructed to approximate the distance driven structure given by both GAM, but surprisingly, the difference is quite small. We then see that the GAMs can overfit the data, but by looking at the smoothing functions, it can be simple for an actuary to determine which distance intervals to choose in the ratemaking. An important element to remember is the fact that we totally excluded the exposure time in our ratemaking process : only the distance has been used. That means that an insured with 20,000 km should pay the same even if he drove his vehicle for 1 months, 2 months or one year. Note that we can even compare the performance of the GAM to the duration models presented in the last chapter. Even if no covariates have been added to the GAM models so far, the GAMs show the best score on the out-of-sample study.

4.4   GAMLSS

The *Generalized additive models for location, scale and shape* or GAMLSS are a generelizations of the GAM. Proposed by **?**, the GAMLSS remove the assumption that the modeled random variable should be a member of the exponential linear family. The Poisson distribution that we used for the modeling of the smoothing functions of the exposure time, or the distance driven, is a member of the exponential linear family. However, as overdispersion is often present when we modeled claim counts with in insurance data, negative binomial distributions (NB1 or NB2) could be helpful. When the extra parameter α of the NB1 or the NB2 is unknown, negative binomials are not member of the exponential linear family. Use the GAMLSS theory is then needed.

The GAMLSS give us an additional advantage : it allows to include smoothing function $g(.)$ up to four distribution parameters. Generally, we will add smoothing function to the variance parameter (σ), but also for any other parameters of a distribution (for example shape parameters, as skewness(ν) and kurtosis (τ) parameters. Formally :

$$g_1(\boldsymbol{\mu}) \;=\; \boldsymbol{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \boldsymbol{Z}_{j1}\boldsymbol{\gamma}_{j1},$$

$$g_2(\boldsymbol{\sigma}) \;=\; \boldsymbol{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \boldsymbol{Z}_{j2}\boldsymbol{\gamma}_{j2},$$

$$g_3(\boldsymbol{\nu}) \;=\; \boldsymbol{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \boldsymbol{Z}_{j3}\boldsymbol{\gamma}_{j3},$$

$$g_4(\boldsymbol{\tau}) \;=\; \boldsymbol{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \boldsymbol{Z}_{j4}\boldsymbol{\gamma}_{j4},$$

where parameters μ, σ, ν and τ are represented in a vectorial form ($n \times 1$) for n observations. As before, $\boldsymbol{X}\boldsymbol{\beta}$ represents the parametric form of the the model, and the termes $\boldsymbol{Z}\boldsymbol{\gamma}$ is the non-parametric part. For the parametric form, $\boldsymbol{X}_k$, for $k = 1,\dots,4$, is a design matrix $n \times J'_k$, where J'_k is the number of covariates used in $\boldsymbol{X}_k$, and $\boldsymbol{\beta}_k$ is a coefficient vector of dimension J'_k. For the non-parametric part, $\boldsymbol{Z}_{jk}$ is a design matrix with known dimension $n \times q_{jk}$ and $\boldsymbol{\gamma}_{jk}$ ia vector of dimension q_{jk}. The model allows us to add as many additive terms as we want, J_k, for each function $g_k(.)$.

We will skip the theoretical details abouth GAMLSS, as we did for the GAM, but it is worth mentioning that we will use B-splines, instead of cubic splines, for the smoothing functions. The number of knots of each smoothing function is still important, but as long as the number of knots is enough, the flexibility of the model will be control by a smoothing parameter λ that control the effective degrees of freedom (EDF).

4.4.1   Examples

We apply the GAMLSS for the two negative binomial distributions, that can be expressed as [1] :

$$\Pr(X = k) = \frac{\Gamma(\lambda/\theta + k)}{\Gamma(\lambda/\theta)\Gamma(k+1)} \left(\frac{1}{1+\theta}\right)^{\lambda/\theta} \left(\frac{\theta}{1+\theta}\right)^{k} \quad (NB1(\lambda, \theta))$$

$$\Pr(X = k) = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)\Gamma(k+1)} \left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha} \left(\frac{\lambda}{\alpha+\lambda}\right)^{k} \quad (NB2(\lambda, \alpha))$$

As done with the GAM, the smoothing functions can be used within the mean function (i.e. the parameter called "mu" in the R package), but another possibility is to include smoothing functions within the α parameter (i.e "sigma" in the R package).

We can also use the zero-inflated Poisson distribution with the GAMLSS package. Remember that we introduced this distribution is the last chapter, where the probability function can be expressed as :

$$\Pr(N_i = k) \quad = \quad \begin{cases} \phi + (1-\phi)\exp(-\lambda_i) & \text{for} \quad k = 0 \\ (1-\phi)\frac{\lambda_i^k \exp(-\lambda_i)}{k!} & \text{for} \quad k = 1,2,... \end{cases}$$

We saw that $E[N_i] = (1-\phi)\lambda_i$ and $Var[N_i] = E[N_i] + E[N_i](\lambda_i - E[N_i])$. This distribution can be used with *family = "ZIP1"*. The GAMLSS offers another parametrization of the zero-inflated Poisson distribution, the *family = "ZIP2"*. The probability function of this count distribution is expressed as :

$$\Pr(N_i = k) \quad = \quad \begin{cases} \phi + (1-\phi)\exp(-\frac{\lambda_i}{1-\phi}) & \text{for} \quad k = 0 \\ (1-\phi)\frac{\lambda_i^k \exp(-\frac{\lambda_i}{1-\phi})}{k!\,1-\phi^k} & \text{for} \quad k = 1,2,... \end{cases}$$

In this case, $E[N_i] = \lambda_i$ and $Var[N_i] = \lambda_i + \lambda_i\frac{\phi}{1-\phi}$. Both the ZIP1 and the ZIP2 converge to a Poisson distribution for $\alpha \to 0$.

The ZIP2 distribution seems more natural for 2 reasons :

1. More easy to use with smoothing functions, as the ZIP has two parameters that affects the mean. The α parameter (i.e "sigma"), for both ZIP and ZIP2 is used to model the extra-weight given to the probability of not claim (i.e $n = 0$).

2. It is more natural to use ZIP2 models with exposure. For example, using the traditional offset variable, we have $E[t_i N_i] = t_i \lambda_i$.

---

1. Note that, strangely, the GAMLSS packages inverses the name of the distributions. So, by the use of *family = "NBI"*, we have the NB2 distribution), and with *family = "NBII"*, we have the NB1 distribution.

We first wanted to apply the NB1, the NB2 and the ZIP1/ZIP2 distribution to our simulated data. However, even if the data were correctly simulated, they remained simulated data based on Poisson distribution. Consequently, we cannot find any interesting smoothing functions for the α parameter of the NB1 or the NB2 distributions, nor for the ϕ parameter of the ZIP1/ZIP2 distributions. For illustration, however, we show here what kind of smoothing function we can obtained with real data.

The first three figures, 4.4.1 shows the prediction surface for the NB1, the NB2 and the ZIP distribution. As expected, surfaces are almost the same, and would be similar to what we could obtained with a Poisson GAM.

For sure, it is interesting to use NB or ZIP distribution with smoothing functions on the mean parameter. However, the interest lies in the modeling of the α parameter for the NB distributions, or for the ϕ parameter for the ZIP distribution. First, figure show how the exposure time impacts the α parameter of the NB1 distribution. As the exposure (in calendar time) increases, we see that the variance of the number of claims decreases, meaning that insureds with longer contracts seem more stable. We must be careful with the conclusion : we already saw that insureds with higher distance driven or higher exposure time have higher claim frequency. We know that count data are overdispersed, but with the GAMLSS, we see that that insureds with higher exposure time are simply less overdispersed. Note, to conclude, that other covariates, and not only smoothing functions, ca be added into the α parameter of the NB1 or the NB2 distributions.

Finally, let's take a look on the ZIP distribution. Figure 4.4.3 also shows how exposure time impacts the ϕ parameter of the ZIP distribution. As already noted with the NB distribution, even for the ZIP, we still know insureds with higher distance driven or higher exposure time have higher claim frequency. Here, by the modeling of the ϕ parameter of the ZIP distribution, we want to understand how the extra weight put at the probability of having no claim is impacted by some exposure bases. As the exposure time increases, the ϕ parameter decreases, which is coherent with what we obtained with figure 4.4.2.

We see, interestingly, that if the use of car increases, the ϕ parameter of the ZIP distribution decreases. We can link this result with a well-known property of the Poisson distribution, namely the fact that the Poisson distribution is the limit of a binomial distribution when $n \to \infty$ et $p = \lambda/n$.
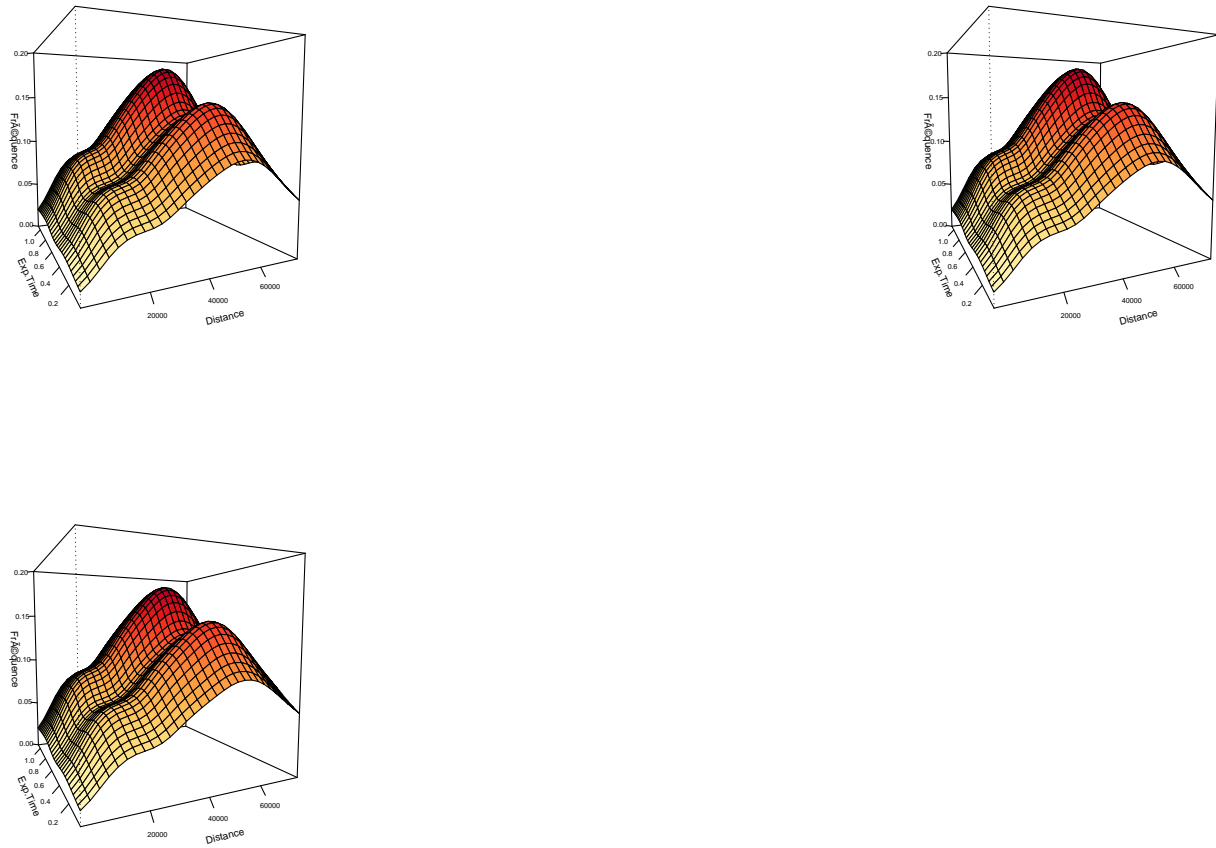
Figure 4.4.1 – Smoothing functions for the mu (λ) parameter of the NB1, the NB2 and the ZIP distributions

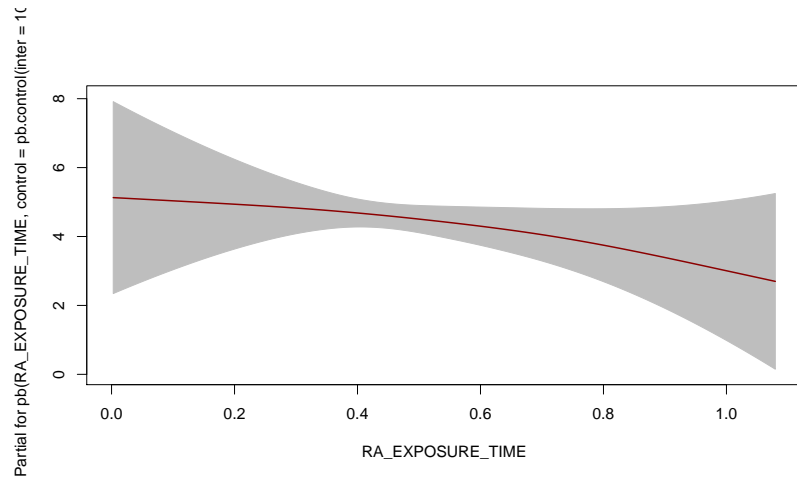Figure 4.4.2 – Smoothing functions for exposure time on the sigma (α) parameter of the NB1 distribution
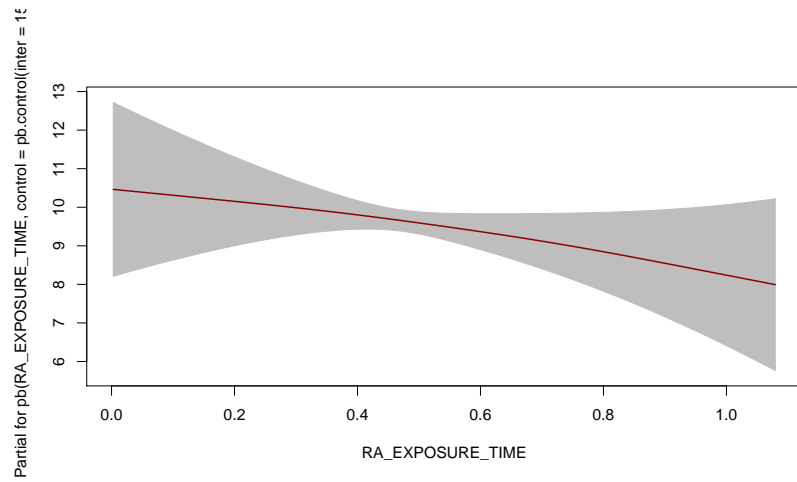

Figure 4.4.3 – Smoothing functions for exposure time on the sigma (ϕ) parameter of the ZIP distribution

For insurance, when modeling claim counts, this can be interpreted as if the number of trials corresponds to the number of uses, where the probability of having an accident for a specific trip, a specific day or a specific kilometer is almost zero. But, if the insured drives a lot, it becomes possible to have an accident. Maybe the approximation by the Poisson is not perfect as the number of trials maybe to small. The ZIP, by the extra parameter α can be seen as the approximation error between the Poisson and what we really observe. As the graph shows, if the number of trips increases, α goes to zero and the Poisson distribution seems better suited to model the claim counts.

4.5   Conclusion

Until now, in a ratemaking context, it was assumed that there could be insurance contracts indicating the premium that the insured will have to pay according to the duration of his contract (this kind of clarification is in fact included in the contracts) but also depending on the distance the insured will drive, the number of trips he will have, or his number of driving hours. In many jurisdictions, it is not possible to include this type of particularities in the insurance contracts. The premium is set in advance, and the insurer knows the total distance driven by an insured at the same time he knows if his insured claimed or not. So, the models that we have seen so far can be rather limited in their use. On the other hand, all is not completely lost because the models seen can still allow to better understand the risk, to better segment the portfolio, etc. Still, it would probably be interesting to predict these risk exposure variables as well.

Chapitre 5

Ratemaking with Panel Data

In what we call cross-section data models, we supposed that each contract was independent. With panel data models, or lon-
gitudinal data models, we will suppose a form of dependence between the contracts of the same insured. In this structure, an
insurance contract i can be observed T_i years, with $T_i > 1$. So instead of modeling univariate random variables N_i or Y_i, we are
interested in the vector of random variables $\boldsymbol{N}_i = \{N_{i,1}, \ldots, N_{i,T_i}\}$ (or $\boldsymbol{Y}_i$).

We can then express such database as :

| obs. | pol. | Sex | Terr. | Year 2016 | | | | Year 2017 | | | Year 2018 | |
				...	Nb.sin.	Expo.	...	Nb.sin.	Expo.	...	Nb.sin.	Expo.
1	1831	F	Rural	...	0	1,000	...	1	1,000	...	0	1,000
2	1932	F	Rural	...	1	1,000	...	2	1,000	...	1	0,633
3	2033	H	Rural	...	0	0,912	...	0	1,000	...	.	.
...	...	...	...	...	...	...	...	...	...	...	...	...
4654	9511	H	Urban	...	3	1,000	...	0	1,000	...	0	1,000
4655	9522	F	Rural	...	0	0,144	...	1	0,791	...	2	1,000

Such structure, where all insureds are observed exactly T periods is a balanced dataset. In practice, we mainly work with unbalanced
panel dataset, as some insureds only stay one year with an insurance company and other says longer. This can be expressed as :

| obs. | pol. | Sex | Terr. | Year 2016 | | | | Year 2017 | | | Year 2018 | |
				...	Nb.sin.	Expo.	...	Nb.sin.	Expo.	...	Nb.sin.	Expo.
1	1831	F	Rural	...	0	1,000	...	1	1,000	...	0	1,000
2	1932	F	Rural	...	1	1,000	...	2	1,000	...	1	0,633
3	2033	H	Rural	...	0	0,912	...	.	.	...	.	.
...	...	...	...	...	...	...	...	...	...	...	...	...
4654	.	.	.	...	.	.	...	0	1,000	...	0	1,000
4655	.	.	.	...	.	.	...	1	0.977	...	.	.

In the models that we will use for this chapter, it will be easy to generalize the balanced approach to the unbalanced. To simplify
the notation, vwe will suppose $T_i = T$, for all $i = 1,\ldots,n$. Covariates and risk characteristics are also time dependent. For each

individual policy i and period t, $t = 1, \ldots, T$, some covariate information exists, because the insurer knowns a vector of observable characteristics ($\mathbf{x}_{i,t}$ related to the individual. We are now looking to derive the following joint count distribution for each insured i, that will we suppose observed t times :

$$\Pr(N_1 = n_1, N_2 = n_2, ..., N_t = n_t) =$$

Panel data, compared to cross-section data seen earlier, offer many new possibilities :

1. Ratemaking based on past claim experience

2. Ratemaking based on past telematic statistics

3. PAYD ratemaking

5.1   Modeling

To create a dependence between the number of claims of a single driver, the starting point is again the Poisson distribution :

$$N_{it} \sim Poisson(\mu_{i,t} = \alpha_i \lambda_{i,t}), \quad \lambda_{i,t} = \exp(x'_{i,t}\beta), \ i = 1,...,n. t = 1,...,T$$

where α is an individual parameter. We then have two different situations :

1. Fixed effects model : α_i, $i = ,...,n$ are **unknown parameters**, that need to be estimed (which means n parameters).

2. Random effects model : α_i are **i.i.d. random variables**, that come from a prior distribution.

In both cases, as a single α_i parameter affects all contracts of the same insured, we end up with a panel data model that gives us the flexibility to create a form time dependence.

Like it was mentioned earlier when we introduce the heterogeneity of the Poisson distribution (to obtain the NB2 or the NB1 distribution), it is well known that many important factors cannot be taken into account in the *a priori* risk classification. Consequently, tariff cells are still quite heterogeneous despite the use of many classification variables. Intuitively, it seems reasonable to believe that the hidden characteristics are partly revealed by the number of claims at fault reported by the policyholders. Several empirical studies have shown that, if insurers were allowed to use only one rating variable, it should be some form of merit rating : the best predictor of the number of claims incurred by a driver in the future is not age or vehicle type but past claims history. Hence the adjustment of the premium from the individual claims experience in order to restore fairness among policyholders. In that respect, the allowance of past claims in a rating model derives from an exogenous explanation of serial correlation for longitudinal data. In this case, correlation is only apparent and results from the revelation of hidden features in the risk characteristics.

5.2   Random Effects

In random effects models, we do not suppose that the α_i are simple parameters to estimate, but are random variables, with density $g(\cdot)$. Conditionally on the random effects α_i^{RE}, all number of claims $N_{i,1}$, $N_{i,2}$, . . . ,$N_{i,T}$ from insured i are independent. The joint distribution of $N_{i,1},...,N_{i,T}$ can be derived as :

If α_i^{RE} follows a gamma distribution of mean 1 and variance $\frac{1}{\nu}$, the joint distribution can be expressed as :

$$\Pr[N_{i,1} = n_{i,1},...,N_{i,T} = n_{i,T}] \ = \ \left(\prod_{t=1}^{T} \frac{(\lambda_{i,t}^{RE})^{n_{i,t}}}{n_{i,t}!}\right) \frac{\Gamma(n_{i,\bullet} + \nu)}{\Gamma(\nu)} \left(\frac{\nu}{\sum_{i=1}^{T} \lambda_{i,t}^{RE} + \nu}\right)^{\nu} \left(\sum_{i=1}^{T} \lambda_{i,t}^{RE} + \nu\right)^{-n_{i,\bullet}},$$

where $n_{i,\bullet} = \sum_{t=1}^{T} n_{i,t}$.

This is the multivariate negative binomial, or MVNB. This is the basic distribution for panel count data modeling. In this case :

$$\mathbb{E}[N_{i,t}] = \lambda_{i,t}^{RE} < \mathbb{V}[N_{i,t}] = \lambda_{i,t}^{RE} + (\lambda_{i,t}^{RE})^2 / \nu.$$

It can be shown that the first-order condition to obtain the ML estimate of the β for the MVNB is :

$$\sum_{i=1}^{n}\sum_{t=1}^{T} x_{i,t} \left(n_{i,t} - \lambda_{i,t} \frac{\sum_t n_{i,t} + \nu}{\sum_t \lambda_{i,t} + \nu} \right) = 0$$

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{n_{i,\bullet}-1} \frac{1}{j+\nu} \right) - \log \left(1 + \frac{\sum_t \lambda_{i,t}}{\nu} \right) + \sum_t \frac{\lambda_{i,t} + n_{i,t}}{\sum_t \lambda_{i,t} + \nu} = 0.$$

5.2.1   Others Count Distributions

Instead of supposing that the random effects are gamma distributed, we can choose other distributions for α. Two popular choices are the Inverse-Gaussian (of mean 1 and variance τ), having the following density :

$$g(\alpha) = \frac{\alpha^{-3/2}}{\sqrt{2\pi\tau}} \exp \left(-\frac{(\alpha-1)^2}{2\tau\alpha} \right)$$

For cross-section data, this count distribution is known to be Poisson-Inverse Gaussian 2, or PIG2. PIG1 distribution, having a similare variance form as the NB1 distribution also exists. For panel data, we have the MPIG distribution.

Another mixing distribution with the Poisson distribution is the lognormal. The advantage of this distribution comes from the fact that the lognormal distribution is a transformation of the gaussian distribution, where the correlation matrix allows us to put different form of dependence between claim counts. In this case, if α_i is a LogNormal of parameters $\mu = -\sigma^2/2$ and σ^2, we have

$$f(\alpha) = \frac{1}{\alpha\sigma\sqrt{2\pi}} \exp \left(\frac{-(\log(x)-\mu)^2}{2\sigma^2} \right),$$

we have the Poisson LogNormal ($PLN2$), expressed as :

$$\Pr(N_i = k) = \int_{-\infty}^{\infty} \frac{\exp(-\lambda_i\alpha)(\lambda_i\alpha)^k}{k!} \frac{1}{\alpha\sigma\sqrt{2\pi}} \exp \left(\frac{-(\log(x)-\mu)^2}{2\sigma^2} \right) d\alpha,$$

with $\lambda_i = \exp(X_i'\beta)$. We can also use this form, with $\gamma_i = \exp(X_i'\beta + \epsilon)$, where ϵ is gaussian :

$$\Pr(N_i = k) = \int_{-\infty}^{\infty} \frac{\exp(-\gamma_i)\gamma_i^k}{k!} \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{1}{2} \left(\frac{\epsilon}{\sigma} \right)^2 \right) d\epsilon$$

We can not express the PLN2 distribution in a closed way. There are several techniques for estimating this model and many R packages can be used. Although we will not discuss these models in detail for this course, note that the Poisson-Lognormal is now widely used to model hierarchical dependence structure.

5.2.2 Predictive Ratemaking

One of the interest of using panel data framework in actuarial sciences is to produce predictive ratemaking, i.e. a premium based on the past claim experience of the insured. Indeed, by using the Bayes theorem, we can fin the posteriori distribution of the random effect $(g(\alpha_i^{RE}|n_{i,1}, \ldots, n_{i,T-1})$ to compute the insured's premium at time T.

For the MVNB, where we removed the subscript i for convenience, and use $\alpha = \alpha_i^{RE}$, we have :

$$N_t|\alpha \sim Poisson(\lambda_t\alpha), \quad \text{with} \quad \alpha \sim gamma(\nu, \nu).$$

We can show that :

1. The posterior distribution of the random effects, $\alpha|n_1, \ldots, n_{T-1}$, is :

2. The predictive distribution of the number of claims, $N_T|n_1, \ldots, n_{T-1}$, is :

The mean of the predictive distribution is expressed as :

$$E[N_T|n_1, \ldots, n_{T-1} = \lambda_t \frac{\nu + \sum_{t=1}^{T-1} n_t}{\nu + \sum_{t=1}^{T-1} \lambda_t}$$

which can be seen as the predictive premium for an insured with experience. It is intereting to analyse quickly this predictive premium :

1. A new driver, without claim experience, will have a premium equal to $E[N_1] = \lambda_1$. As $\lambda_1 = \exp(X_1\beta)$, only his risk characteristics are used in the ratemaking (what about the telematics informations ?) ;

2. An insured will see his premium increases if $\sum_{t=1}^{T-1} n_t > \sum_{t=1}^{T-1} \lambda_t$, in other words, if he claims more accidents than what was expected of him.

5.2.3 Telematic covariates

As mentioned earlier, it is possible to use past information on telematic to predict the future driving behaviour. Counting the number of claims is often difficult (discrter distribution with between 85%-95% do not claim at all). However, distance driven or exposure time are continuous variables. Consequently, basic time series models, such as ARMA(p) model might be used to estimate future telematic variables.

However, as done by Denuit et al. (2019), it might be better to use panel data models with random effects. Let's take a look at one possible approach.

We note $Y_{i,t}$ the telematic statistic that we want to model for insured i for his $t^t h$ contract. We can suppose that $Y_{i,t}$ is the distance driven, or the number of trips, for example. As a strictly positive value, we can suppose that $Y_{i,t}$ can follow a gamma distribution. Formally, we have (we remove substript i) :

$$Y_t|\theta \sim Gamma(\phi, \frac{\mu_t}{\theta}), \quad \text{with} \quad \theta \sim Gamma(\delta, \tau).$$

More precisely, we have :

$$f(y_t|\theta) = \frac{y_t^{\phi-1}}{\Gamma(\phi)\left(\frac{\mu_t}{\theta}\right)^\phi} \exp\left(\frac{-y_t}{\frac{\mu_t}{\theta}}\right)$$

and

$$g(\theta) = \frac{\theta^{\delta-1}}{\Gamma(\delta)\tau^\delta} \exp\left(\frac{-\theta}{\tau}\right)$$

We then have :

$$E[Y_t|\Theta] = \frac{\phi\mu_t}{\Theta}$$

meaning that

$$E[Y_t] =$$

In this case, we can also show that the joint distribution of $Y_1, \dots, Y_T$ can be expressed as a form of multivariate gamma distribution :

$$
\begin{aligned}
f(y_1, \dots, y_T) &= \int_0^\infty \prod_{t=1}^T f(y_t|\theta)g(\theta)d\theta \\
&= \left(\prod_{t=1}^T \frac{y_t^{\phi-1}}{\Gamma(\phi)\mu_t^\phi}\right) \frac{\Gamma(t\phi+\delta)}{\Gamma(\delta)\tau^\delta} \left(\sum_{t=1}^T \frac{y_k}{\mu_k} + \frac{1}{\tau}\right)^{-(t\phi+\delta)}
\end{aligned}
$$

The posterior distribution of θ, giving $y_1, \dots, y_T$ is still a gamma distribution with updated parameters

$$\delta^* = \delta + t\phi \text{ and } \tau^* = \frac{1}{\tau} + \sum_{t=1}^{T} \frac{y_k}{\mu_k}$$

This allows us to estimate the future expected value of Y_{T+1}, given $y_1, \ldots, y_T$, as :

$$E[Y_{T+1}|y_1, \ldots, y_T] =$$

The general approach is classic. Lognormal distribution with gaussian random effects is also a possibility. The main interest is to develop a potential generalization where the random effects of the count distribution might be correlated with the random effects used with the telematic statistics. One can use a copula between α and θ (or use a bivariate gaussian distribution) for this dependence. A interesting candidate might be the bivariate Sarmanov distribution that can be expressed as :

$$u^S(\alpha, \theta) = g(\alpha) g(\theta) (1 + \omega \phi_1(\alpha) \phi_2(\theta)),$$

where ϕ_ℓ, $\ell = 1, 2$ are two bounded non-constant functions such that $\int_{-\infty}^{\infty} \phi_\ell(t) u_\ell(t) dt = 0$ and ω is a real number that satisfies the condition :

$$1 + \omega \phi_1(\alpha) \phi_2(\theta) \geq 0$$

In this case, it can be shown that an analytical form for the joint distribution of $N_1, \ldots, N_T, Y_1, \ldots, Y_T$ can be expressed. Expected values and predictive expected values based on past claims experience and past telematic statistics.

5.3   Fixed Effects

We covered a random effects model for panel data. Instead of working with the last approaches, I want conclude this short-course by another possible approach. Indeed, since the α_i, $i = ,...,n$ parameters are unknown, an interesting approach would be to estimate all of them by maximum likelihood. That means a panel data model with $n + (p + 1)$ parameters... which can be high.

Beyond the gigantic amount of parameters to estimate, the problem with this ML estimation of the α's is that it does not necessarily generate convergent estimates in the classical case in insurance, where T is fixed and quite small, with $n \to \infty$. Indeed, the large number of parameters in the model causes what is called *incidental problem*, which means that an incorrect estimation of the fixed effects α generates incorrect estimates of β. In the case of a logistic regression, for example, it was shown that the β MLEs were indeed biased. We will have to be careful.

In any case, let us find the MLE in the Poisson case. For a specified insured i, we can easily find the MLE of α_i :

So the idea is to substitute each α_i (for all insureds) by his estimator $\hat{\alpha}_i$, in the Poisson model :

$$\prod_{i=1}^{n} \Pr\left(n_{i,1},...,n_{i,T}|\beta\right) = \prod_{i=1}^{n} \exp(-\lambda_{it} \frac{\sum_{t=1}^{T} n_{i,t}}{\sum_{t=1}^{T} \lambda_{i,t}})(\frac{\sum_{t=1}^{T} n_{i,t}}{\sum_{t=1}^{T} \lambda_{i,t}})^{n_{i,t}}(\lambda_{i,t})^{n_{i,t}}/n_{i,t}!$$

$$\propto \prod_{i=1}^{n}[\prod_{t=1}^{T} \left(\frac{\lambda_{i,t}}{\sum_s \lambda_{i,s}}\right)^{n_{it}}]$$

In the case where the individual estimate of α_i (i.e. fixed effects) would be considered, this would be the equation to maximize to find the MLE of β.

Like mentioned earlier, the possibility of the incidental problem is still present. Hopefully, there are ways to get around this problem of potential incorrect β estimates when fixed effects are present. Conditional maximum likelihood is a solution. The goal of this

method is to condition on the exhaustive statistics of α_i, $i = 1, ..., n$ for distributions that are members of the linear exponential family. In our case, the Poisson is a member of this family, and, as we have just seen, the exhaustive statistic for α_i is $\sum_{t=1}^{T} n_{i,t}$.

This conditional MLE of β can then be derived as follow :

$$
\begin{aligned}
\Pr\left(n_{i,1}, ..., n_{i,T} \Big| \sum_{t=1}^{T} n_{i,t}\right) &= \Pr\left(n_{i,1}, ..., n_{i,T}, \sum_{t=1}^{T} n_{i,t}\right) \Big/ \Pr\left(\sum_{t=1}^{T} n_{i,t}\right) \\
&= \Pr\left(n_{i,1}, ..., n_{i,T} \Big| \sum_{t=1}^{T} n_{i,t}\right) \Big/ \Pr\left(\sum_{t=1}^{T} n_{i,t}\right) \\
&= \frac{\prod_{t=1}^{T} \exp(-\mu_{i,t})\mu_{i,t}^{n_{i,t}}/n_{i,t}!}{\exp(-\sum_{t=1}^{T}\mu_{i,t})(\sum_{t=1}^{T}\mu_{i,t})^{\sum_{t=1}^{T} n_{i,t}}/(\sum_{t=1}^{T} n_{i,t})!} \\
&= \frac{(\sum_{t=1}^{T} n_{i,t})!}{\prod_{t=1}^{T} n_{i,t}!} \times \prod_{t=1}^{T}\left(\frac{\mu_{i,t}}{\sum_{s=1}^{T}\mu_{i,s}}\right)^{n_{i,t}}
\end{aligned}
$$

where $\mu_{i,t} = \alpha_i\lambda_{i,t}$. We can even note the following ratio :

$$
\mu_{i,t} \Big/ \sum_{s=1}^{T}\mu_{is} = \lambda_{i,t} \Big/ \sum_{s=1}^{T}\lambda_{is}
$$

where all the α_i can be removed. Finally, we obtain the following first order condition for the β parameters :

$$
\prod_{i=1}^{n}\Pr\left(n_{i,1}, ..., n_{i,T}|\beta\right) \quad \propto \quad \prod_{i=1}^{n}[\prod_{t=1}^{T}\left(\frac{\lambda_{i,t}}{\sum_{s}\lambda_{i,s}}\right)^{n_{it}}]
$$

which is exactly the same equation to maximize for the (un-conditional) MLE !

Since the conditional MLE does not generate an incidental problem, it means that the classical MLE of a Poisson model with fixed effects do not generate this incidental problem. It is therefore "easy" to find the parameters of a fixed effects model : we only have to add (many !) regressors in the model to estimate in a classic Poisson regression.

More precisely, with $\lambda_{i,t}^{FE} = \exp(X_{i,t}\beta^{FE})$, it can be shown that the first-order condition to obtain the ML estimate of the β^{FE} is :

$$
\sum_{i=1}^{n}\sum_{t=1}^{T} x_{i,t}\left(n_{i,t} - \lambda_{i,t}^{FE}\frac{\sum_t n_{i,t}}{\sum_t \lambda_{i,t}^{FE}}\right) = 0.
$$

5.3.1 Fixed Effects or Random Effects

In practice, fixed effects are often used in econometrics, and almost never in insurance or in actuarial sciences. Random effects, on the other hand, are one of the bases of actuarial science. Indeed, we can conceive credibility theory as applied random effects models, and bonus-malus systems are also constructed using random effects.

We must take some time to understand the differences between each model. First, let's compare the first order conditions for the fixed effects (FE) and the random effects (RE) model (the MVNB for the example) :

$$\sum_{i=1}^{n}\sum_{t=1}^{T} x_{i,t} \left(n_{i,t} - \lambda_{i,t}^{FE} \frac{\sum_t n_{i,t}}{\sum_t \lambda_{i,t}^{FE}} \right) = \mathbf{0}.$$

$$\sum_{i=1}^{n}\sum_{t=1}^{T} x_{i,t} \left(n_{i,t} - \lambda_{i,t}^{RE} \frac{\sum_t n_{i,t} + \nu}{\sum_t \lambda_{i,t}^{RE} + \nu} \right) = 0$$

Some remarks can be made :

1. For fixed effects model, an insured without claim in all his T contracts do not participate in the estimation of the β parameters. Indeed, we have $n_{i,t} = 0, t = 1,...,T$ and $n_{i,|\bullet} = 0$;

2. Again for fixed effects model, an insured who does not change covariates during his T contracts do not participate in the estimation of the β parameters. In this case, $\lambda_{i,t}^{FE} = \lambda_i^{FE}$, et $\lambda_{i,\bullet}^{FE} = T\lambda_i^{FE}$.

3. The fixed effects models cannot have intercept, nor covariates that cannot change over time (sex, year of birth, etc.). All those effects are captured by the fixed effect α_i

4. There is no difference between random effects and fixed effects when T is large. However, for small values of T, estimated parameters might be different for each models.

The reasons explaining the differences between RE and FE models come from the construction of the random effects model. This difference comes from the third line of this developement :

$$\Pr[N_{i,1} = n_{i,1},...,N_{i,T} = n_{i,T}]$$

$$= \int_0^{\infty} \Pr[N_{i,1} = n_{i,1},...,N_{i,T} = n_{i,T} | x_{i,1},...,x_{i,T}, \alpha_i^{RE}] g(\alpha_i^{RE} || x_{i,1},...,x_{i,T}) d\alpha_i^{RE}$$

$$= \int_0^{\infty} \left(\prod_{t=1}^{T} \Pr[N_{i,t} = n_{i,t} | x_{i,1},...,x_{i,T}, \alpha_i^{RE}] \right) g(\alpha_i^{RE}) d\alpha_i^{RE}$$

$$= \int_0^{\infty} \left(\prod_{t=1}^{T} \exp(-\alpha_i^{RE}\lambda_{i,t}^{RE}) \frac{(\alpha_i^{RE}\lambda_{i,t}^{RE})^{n_{i,t}}}{n_{i,t}!} \right) g(\alpha_i^{RE}) d\alpha_i^{RE}.$$

We first use $g(\alpha_i^{RE}||x_{i,1},...,x_{i,T})$, and then $g(\alpha_i^{RE})$, meaning that the random effects are i.i.d (Independent and identically distributed random variables - the problem is "identically distributed"...) and are not influence by the covariates $x_{i,1},...,x_{i,T}$.

The fixed effects model is more general and needs less assumptions than the random effects model. Consequently, a statistical test comparing β^{FE} and β^{RE} can be done to verify if the assumptions of the random effects are respected. Sadly, studies and analyses done with real insurance data show that is clear that the α_i are not i.i.d. For example, it has been shown that young drivers exhibit larger heterogeneity, as well as drivers from the city. In consequence, theoretically, random effects model should not be used.

However, insurers still use those model, and scientific papers also uses those models. How to legitimize the use of random effects then ? Further analysis has shown that it is possible to use random effects in insurance, even though there is clearly a dependence between regressors and heterogeneity. In such a case, however, one must pay attention to the interpretation of the $\hat{\beta}^{RE}$ parameters. Indeed, the parameters obtained in the models with random effects indicate only the apparent effect of the regressors (such as the premium to charge), and not a causal effect, or what might be called the *real* impact.

As we will see in the application section, however, the fixed effects model still has an importance for telematic pricing.

5.4 Applications

5.4.1 Summary of the Database CanadaPanel.csv

We took the original database used at the beginning of the course to illustrate panel data models, and their use for telematics. More precisely, we only took vehicle (or insured) seen at least twice in the dataset. So, as done previously, we can model our claim count model to better understand the insurance risk. Table 5.4.1 sows a short summary of the dataset.

	Average	Var	Min	Max	25th pct	50th pct	75th pct
Exp. Time	0.652	0.061	0.277	1.079	0.485	0.521	0.940
Dist.Driven	10 143	54 522 659	1.1	76 271	4 805	8 277	13 544
# Trips	1 074	390 413	15	3 283	607	937	1 415
Hours Driven	14 344	86 107 621	176	89 233	7 502	12 064	19 205
#claims	0.035	0.036	0.000	3	0	0	0

In Figures 5.4.1 to 5.4.4, histograms of various definition of risk exposure are shown. In Figures 5.4.7 to 5.4.8, the link between claims frequency and risk exposures are exposed.
The first model we want to apply for our panel dataset with classic covariates : gender, marital status and vehicle use. To compare, we also adjust the Poisson distribution. Some remarks must be addressed before analyzing the results :

1. The MVNB has already been programmed in the package **multinbmod** ;

2. In the package, we must include *id=VIN* to specify that all observations from the same VIN have a dependence implied by a common random effects α_i (i.e. $i =$VIN) ;
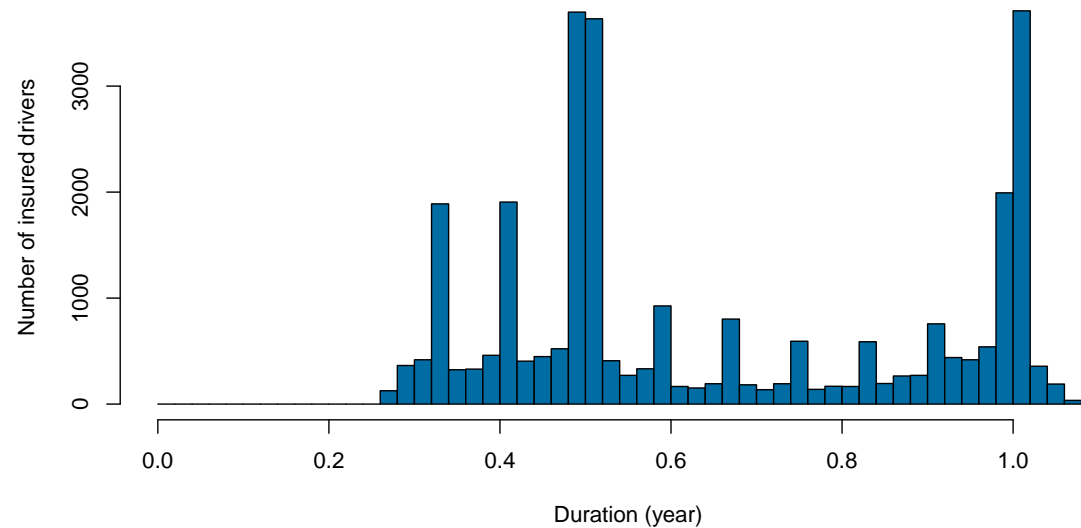
Figure 5.4.1 – Distribution of the risk exposure (in years). Each band has a lenght of 0.02 year.

3. The offset variable RA_EXPOSURE_TIME is not similar in the package multinbmod than in the glm package. Indeed, we have to use the logarithm of RA_EXPOSURE_TIME for the Poisson GLM, while the multinbmod package only needs RA_EXPOSURE_TIME.

A first analysis shows that the gender (RA_GENDER) is not statistically significant. We exclude it from our models, and run again the estimation. Results of the parameters estimations are shown in Table 5.4.2.

The ν parameter is the key parameter to include time dependence between the contracts of the same insured. In our case, we see that the value of ν is 0.4456, with a standard error of 0.1837. When $\nu \to 0$, the MVNB converges toward the Poisson distribution. So it is possible to test the null hypothesis $H0$ of the Poisson distribution against the MVNB using classic statistical test. However, as α must be always positive, it is bounded on the left. In this case, we must be careful we statistical test and use a one-sided test instead of the classic two-sided. In other word, for a probability $\delta = 5\%$, we must test at the value $1 - \delta$ instead of $1 - \delta/2$. In our case, we can use the loglikelihood ratio test ($-4334.1 - -4338,0 = 3.9$) or the Wald test ($0.4456/0.1837 = 2.58$) against a χ_1 square distribution (3.841). The Poisson is then rejected against the MVNB using the loglikelihood ratio test, but not using the Wald test (asymptotically, both tests are equivalent). Note that we are working with a small dataset. As studied many times with a lot of insurance dataset, the Poisson is almost always rejected against the MVNB. To illustrate panel data models with telematic, we will keep the MVNB distribution.
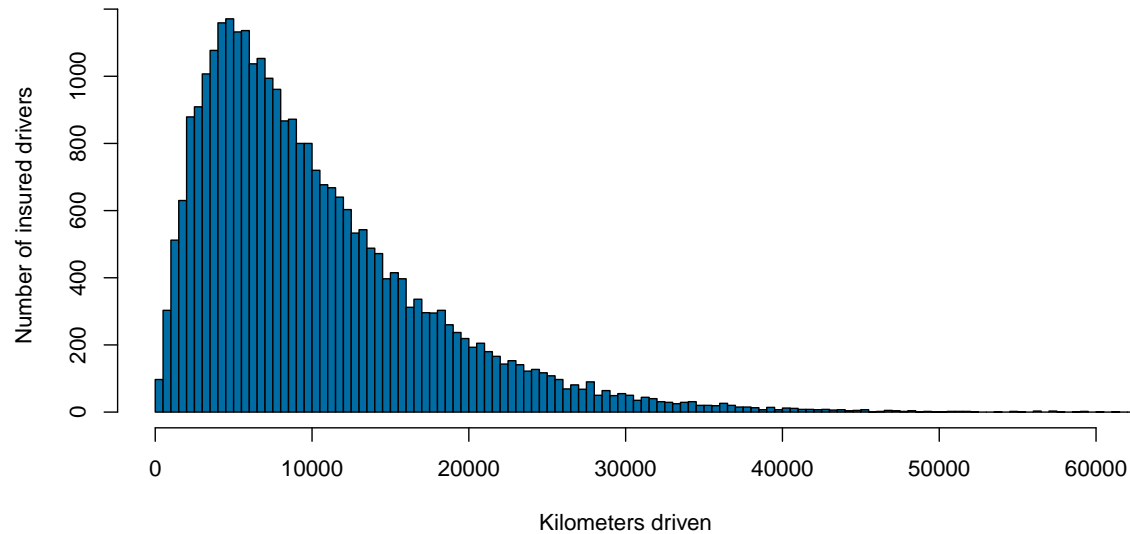
Figure 5.4.2 − Distribution of the distance driven(in km). Each band has a lenght of 500km.

	Poisson		MVNB	
Parameter	Estimate	Std. Err.	Estimate	Std. Err.
β_0	-3.1503	0.0554	-3.1493	0.0563
β_1 (RA_MARITALSTATUS)	0.2007	0.0655	0.2005	0.0669
β_2 (RA_VEH_USE)	0.2462	0.0659	0.2451	0.0671
ν	.	.	0.4456	0.1837
Log-likelihood	-4,338.0		-4,334.1	
AIC	8,681.9		8676.2	

Table 5.4.1 − Summary Statistics of Poisson regression and the MVNB

Beside, we see that the addition of a random effects, through the estimated parameter ν does not change a lot the values of the β parameters. As we will see, it only means that the *a priori* premium for the Poisson and the MVNB will be approximately the same. But the time dependence implied by the MVNB have great impact on the predictive premium.
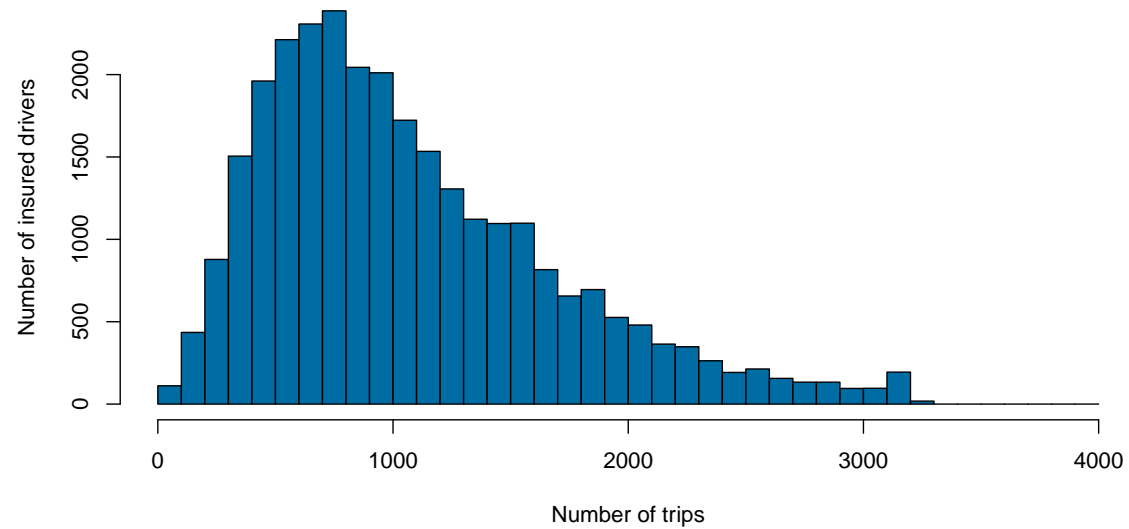
Figure 5.4.3 – Distribution of the number of trips. Each band has a lenght of 100 trips.

Exemple 5.4.1. Compute the premium (that we can refer to the frequency part of the insurance premium) for the the Poisson and the MVNB, for a new driver : a single male driver who use his car to commute. Compute the predictive premium of the same insured, depending on the past number of accidents he claimed to his insurer.
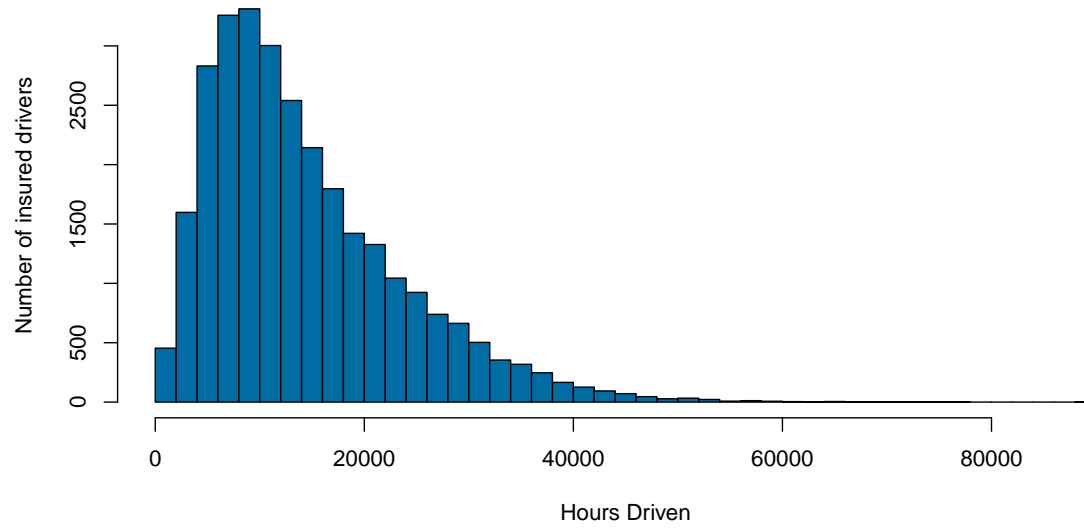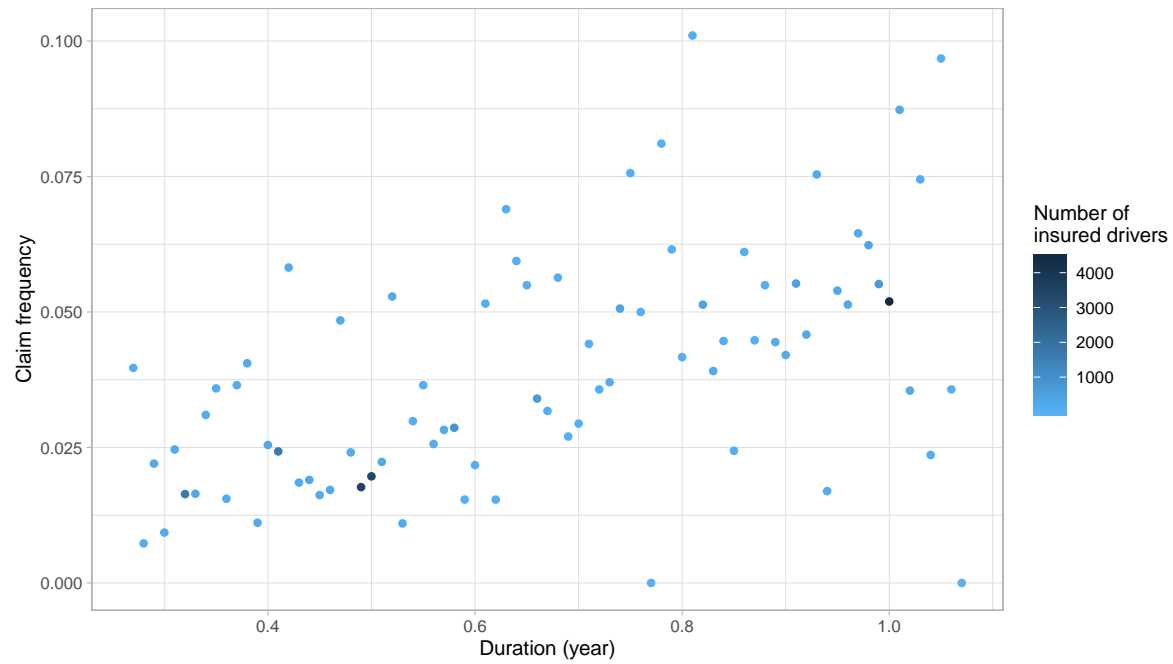
Figure 5.4.4 – Distribution of the hours driven. Each band has a lenght of 2000 hours.

Figure 5.4.5 – Claim Frequency vs Exposure Time
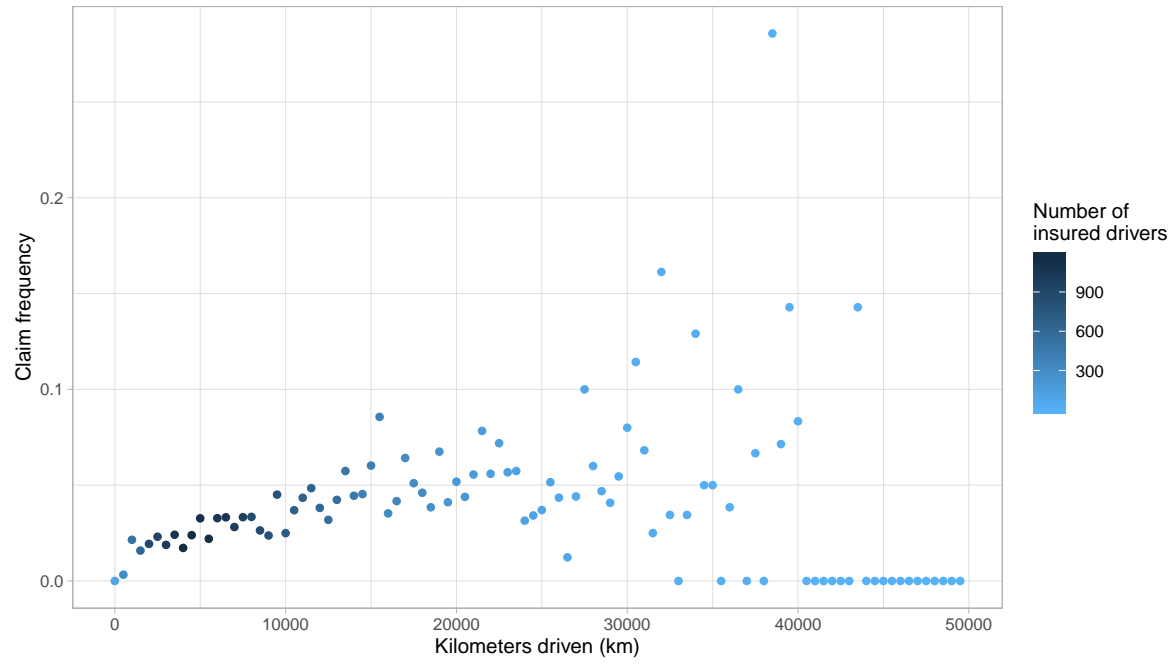
Figure 5.4.6 – Claim frequency vs Distance Driven
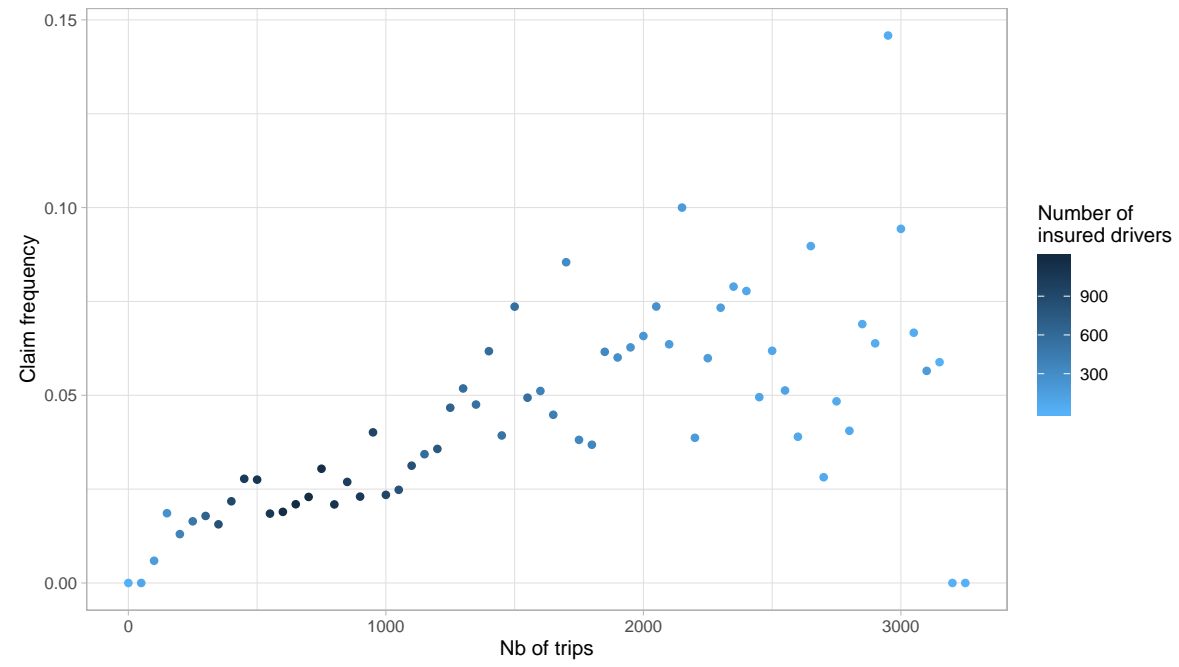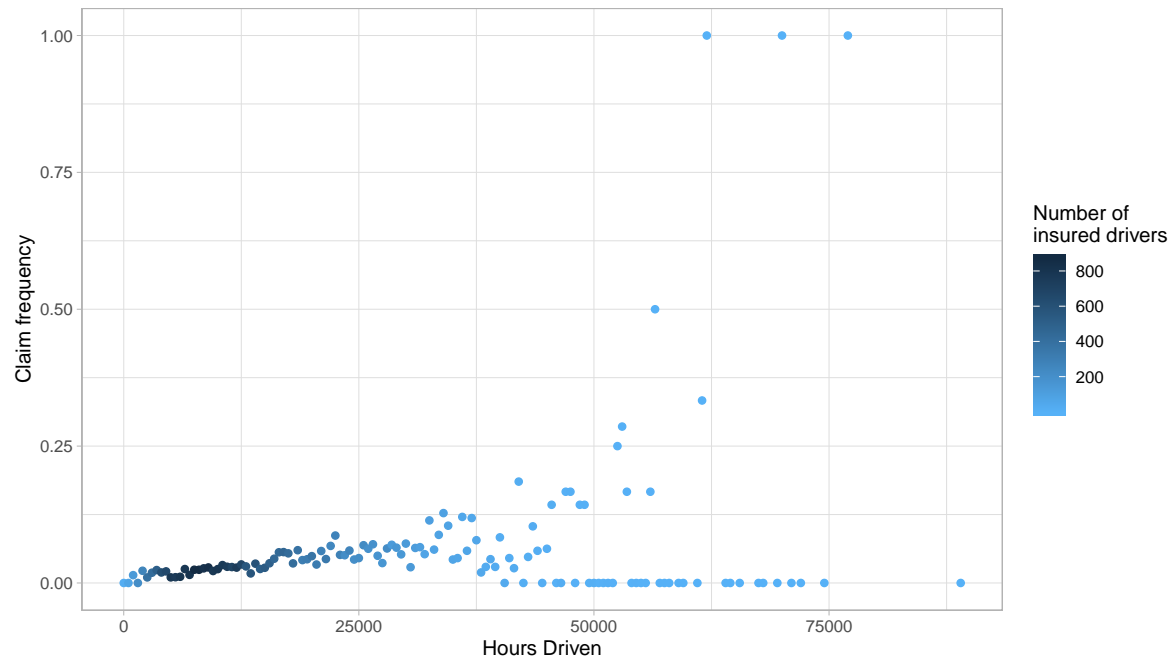
Figure 5.4.7 – Claim Frequency vs Number of trips

Figure 5.4.8 – Claim frequency vs Hours Driven

Solution

Because the Poisson distribution does not imply time dependence, the insured in the example (by supposing that he will stay single and will use his car to commute) will always have a premium of 0.0548, even if he does not claim, or even if he claims.

The MVNB supposes, instead, that his premium will depend on his claim experience. Table 5.4.2 shows some results of this computation, where the driving experience covers 1,2,5 and 10 years, and the total number of claims is between 0 and 4. Note that other random effects distributions can be used with the Poisson, but it won't be covered here.

Out-of-sample analysis should also be done to compare the model, but we want to focus on other properties of the panel data models for now. Indeed, the interest of this short-course is to analyse telematics data. So, has done previoulsy, we can take a look on the smoothing function of the distance driven. Ideally, as we work with panel data, we should use a more general way

Total number of claims	Years of coverage			
	1	2	5	10
0	0.0488	0.0440	0.0339	0.0246
1	0.1583	0.1427	0.1101	0.0797
2	0.2678	0.2414	0.1862	0.1349
3	0.3773	0.3401	0.2624	0.1900
4	0.4868	0.4388	0.3385	0.2452

Table 5.4.2 – Summary Statistics of Poisson regression and the MVNB

to analyse the distance driven. We can use a GAM based on a Poisson distribution, as shown in Figures 5.4.9 and 5.4.10. We should instead use a smoothing function applied to a MVNB distribution. This is theoretically possible, and the theory that we should use is the GAMLSS, already seen. That was done by one of my master student, Gabriel Alepin. We will not cover this theory in this short-course, however. But, you can note that the smoothing function obtained are almost the same as the ones shown in the previous figures. This is not surprising, as we saw that the β parameters, for the *a priori* risk segmentation, are similar between the Poisson and the MVNB.

It is surprising to see the curve of the distance driven passed 10,000km. As mentionned in the introduction of this course, this has been observed for a long time in the scientific community. If we want to use this result in ratemaking, it then means that an insured who drove, for example, 20,000km, should not pay more insurance if he decides, one day, to drive 40,000km. This does not make sense. As pointed by Litman(2007), a specific driver that drives more becomes necessarily riskier. Insureds who drive 40,000km are different insureds than those who drive 20,000km !

5.4.2   Fixed Effects

Maybe insurers are interested in the *real* impact for some telematic variables, like the driving distance. Indeed, if an insurance product covers 100km, 200 km, or 500 km, an insurer might be interested in finding the real impact of the driving distance, not only the *average frequency of claims of those who drive this distance*.

We think that fixed effects models, seen at the beginning of thius chapter is the solution. Remember the first order condition :

$$\sum_{i=1}^{n} \sum_{t=1}^{T} x_{i,t} \left(n_{i,t} - \lambda_{i,t}^{FE} \frac{\sum_t n_{i,t}}{\sum_t \lambda_{i,t}^{FE}} \right) = 0.$$

This equation can give us an explanation on how parameters of the fixed effects model are estimated. If we are interested in estimating the impact of a specific covariate X_j that takes two values $\{0,1\}$, for example, the estimation procedure of the model *will only consider* policyholders who changed from 0 to 1, or from 1 to 0, during the contracts observed. In other words, to estimate the impact of X_j, the same insureds must be observed when they are 0 and when they are 1. We think that these kind
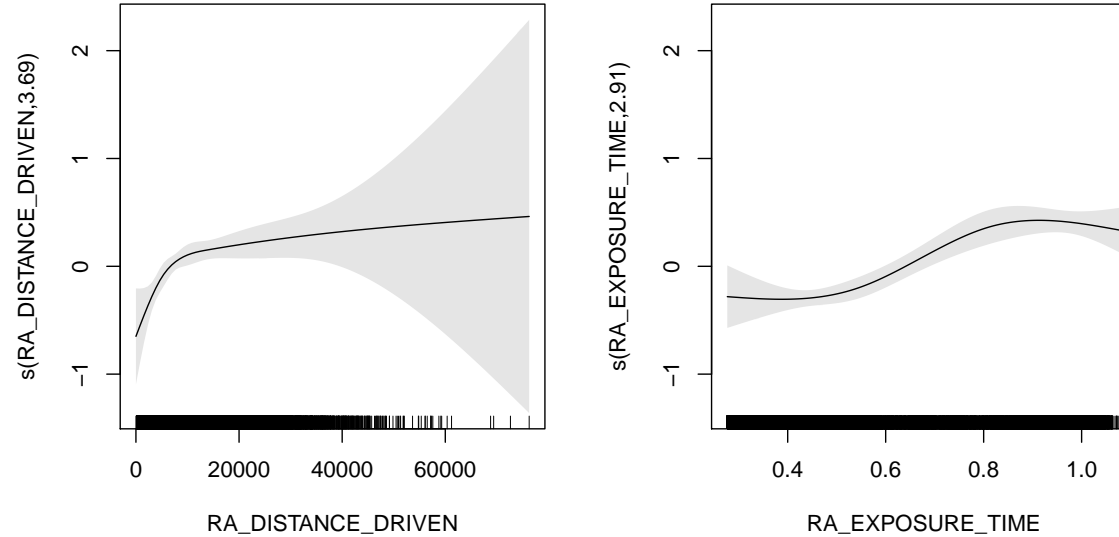
Figure 5.4.9 – GAM Smoothing Functions

of model should be used for PAYD insurance.

More precisely, we should use smoothing functions to see the real impact of distance driven on the claim frequency. GAM ou GAMLSS with fixed effects seems difficult, but are in fact quite direct and simple. Indeed, we saw that a direct MLE of α_i^{FE} can be done. That means that a classic GAM regression with each VIN as covariates can give us the wanted result!
However, some adjustments should be made :

1. An insured without claim in all his T contracts should be deleted from the dataset ;

2. The model should not have intercept, nor covariates that cannot change over time.

Using this Poisson GAM with fixed effects, the Figure 5.4.11 shows what the model gives us for the smoothing function for the distance driven. As we worked with a logarithm link, the exponential transformation of the smoothing function is showed in Figure 5.4.12. We also added a straight line, with intercept 0.15 and slope $\frac{1}{21800}$, to compare.
The expected value of a specific insured i, who drive ν kilometers is then :

In other words, for a specific insured i, according to this model, each kilometer driven increases the expected claims frequency by
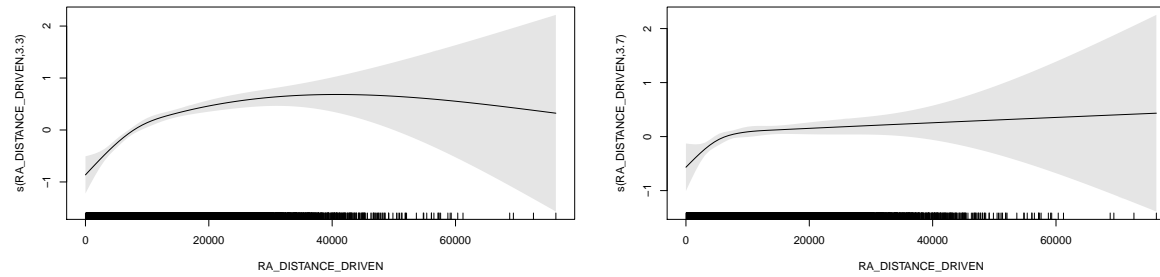
Figure 5.4.10 – GAM Smoothing Functions, without exposure time and with exposure time as offset variable

$\alpha_i \times 1/21800$. This result seems more logical, as additional kilometer driven increases lineary the claim frequency.

Using each α_i, $i = 1, \ldots, n$, and the kilometer driven for each contract, we can even understand how the curve from the Poisson GAM Smoothing Functions (without fixed effect) is generated.
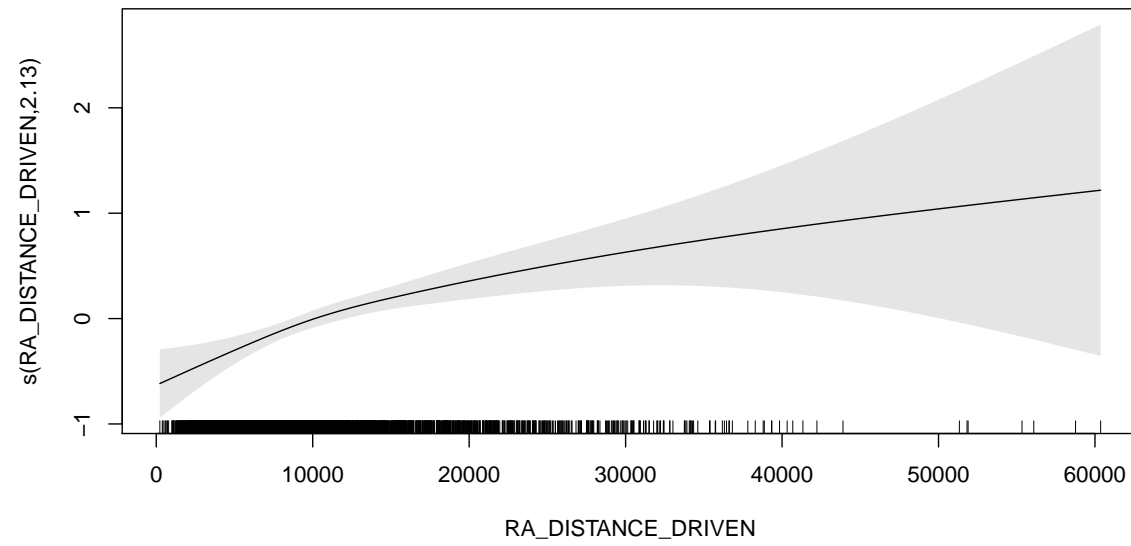
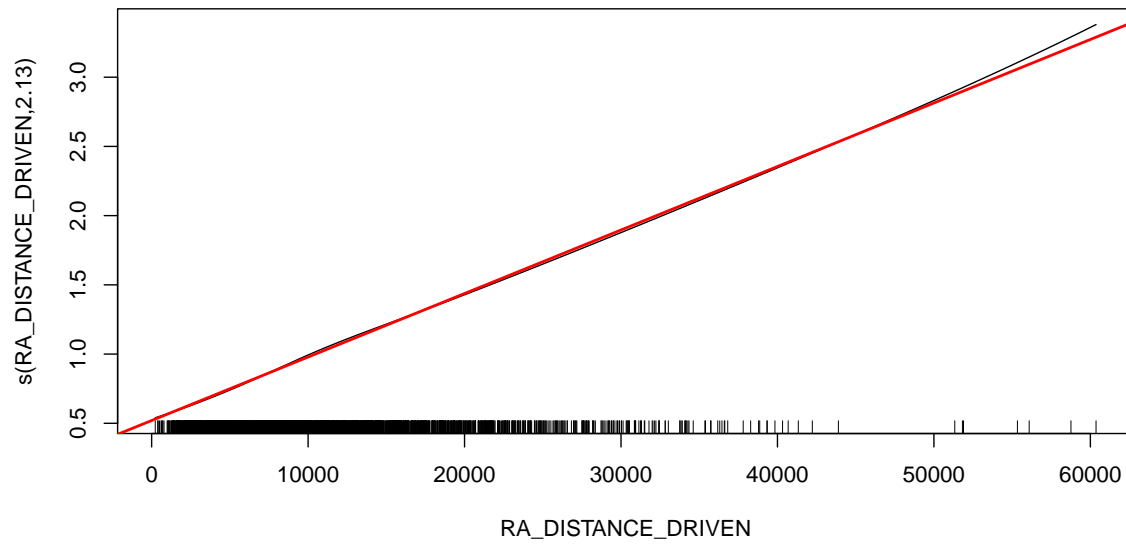Figure 5.4.11 − GAM Smoothing Functions with Fixed Effects

Figure 5.4.12 − Exponential trasform of the GAM Smoothing Functions with Fixed Effects