Analyse longitudinale de l'impact de la distance parcourue sur la probabilité d'un accident automobile

Séminaire d'été d'actuariat et de statistique de l'UQAM

Roxane Turcotte

14 juillet 2021

Université du Québec à Montréal

Chaire Co-operators en analyse des risques actuariels

Introduction

GPS-collected data

- Offer **new ways** to approach car insurance pricing.
- **Reliable information**.
- **Distance driven** is directly related to the risk insured.

Covariates such as territory, gender and age only describe the **general behavior** of insured in those groups.

Relevance

Ex : Use of gender in ratemaking

- Ayuso et al. (2016b) shows that the **differences** observed in claims frequency between men and women are largely attributable to **vehicle use** ;
- Verbelen et al. (2018) reached a similar conclusion

Calculating premiums on **more objective information** is of interest.

Introduction

Overview

Using telematics data, we study the relationship between **claim frequency** and **distance driven** through different models by observing **smooth functions**.

Search for a "marginal" effect

1. The objective is not to compute a premium.
2. The objective is mainly to understand **how** the **distance impacts** the claim frequency when **all individual characteristics** of policyholders have been **considered**.
3. Understanding that relationship provides clues on how to use it in ratemaking.

Introduction

Model frameworks considered :

1. Generalized Additive Models (**GAM**),
2. Generalized Additive Models for Location, Scale, and Shape (**GAMLSS**) that we generalize for panel count data (random effects).

A First Model

Starting Point

GAM Poisson model.

- GAMs : introduced by Hastie and Tibshirani (1986).
- Extension of the generalized linear models (GLM) theory : relax the hypothesis of linearity, and smoothing functions s of the covariates could be included in the predictor.
- Example : the mean for an individual i could be given by $g(\mu_i) = s_0 + s_1(x_{1,i}) + s_2(x_{2,i}) + s_3(x_{3,i})$.

Boucher et al. (2017)

Boucher et al. (2017) analyzed the influence of **duration** and **distance driven** on the number of claims with **independent cubic splines**.

A First Model

Notation

$N_i \sim Poisson(\mu_i)$, where $\log(\mu_i) = \beta_0 + s_1(km_i) + s_2(d_i)$.

$$\begin{aligned} \mu_{i,t} &= \exp(\mathbf{X_{i,t}}\beta + s_1(km) + s_2(d)) \\ &= \exp(s_1(km))\exp(s_2(d))\exp(\mathbf{X_{i,t}}\beta), \end{aligned} \tag{1}$$

A linear trend is **not imposed** by the model structure

Question

What the relation between $\exp(s_1(km))$ and claim frequency would look like **when a linear trend** is **not imposed** by the model structure ?

A First Model

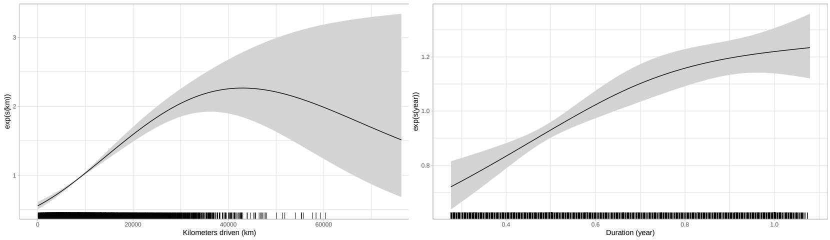

Figure 1 – $\exp(\hat{s}_1(km))$ and $\exp(\hat{s}_2(year))$ from the Poisson GAM

Case Study

1. All models are illustrated using data from a **major Canadian insurance company**.
2. The model yields **similar results** to those obtained by Boucher et al. (2017) **(Spanish data)**.
3. In the study by Boucher et al. (2017), a **learning effect** is advanced to justify the look of $\exp(\hat{s}_1(km))$.

A First Model

Consistency problem

As distance increase, the **risk** is **greater** :

- The **slope** could change, but it should always be **strictly positive**.
- The smoothing function should always be increasing.

Results Analysis

Distance driven is **correlated** with **other driving habits** (Ferreira and Minikel (2010)).

- The lower quantiles of the distribution come from **different** (type of) drivers than the higher quantiles.

Resulting relationship do **not** give an **appropriate representation** of how the claim frequency could change **when** insureds **change their driving habits**.

Problem

- This first model **supposes independence** between all contracts of the same insured.
- Insureds are observed over **many** contracts.

Construct Multivariate Count Models

- Instead of modeling the marginal distribution of each $N_{i,t}$ for $t = 1, \dots, T$, we are now looking for the **joint distribution** :

$$\begin{aligned}\Pr(N_1 = n_1, N_2 = n_2, ..., N_T = n_T) \quad &= \quad \Pr(N_1 = n_1) \times \Pr(N_2 = n_2 | N_1 = n_1) \times \\ &\quad \dots \times \Pr(N_T = n_T | N_1 = n_1, ..., N_{T-1} = n_{T-1}),\end{aligned}$$

- One popular way, is to **include an individual parameter** α in the mean parameter of the count distribution of each contract t :

$$N_{i,t} | \alpha_i \sim \text{Poisson}(\mu_{i,t} = \alpha_i \lambda_{i,t}), \tag{2}$$

Random vs Fixed effects

1. Random effects model
   - All α_i, $i = 1, \ldots, n$ are i.i.d. **random variables** that come from a selected **prior distribution**
2. Fixed effects model
   - All α_i, $i = 1, \ldots, n$ are **unknown parameters** that need to be **estimated**.

Random Effects

Random Effects Model

Model Specification

- $\alpha_i^{RE} \sim Gamma(\nu,\nu),\ i=1,\ldots,n.$
- **Conditionally** on the random effects α_i^{RE}, all numbers of claims $N_{i,1},N_{i,2},\ldots,N_{i,T}$ from insured i are **independent**.

$$\begin{aligned}
\Pr[N_{i,1}=n_{i,1},...,N_{i,T}=n_{i,T}] &= \int_0^\infty \left(\prod_{t=1}^{T} \exp(-\alpha_i^{RE}\lambda_{i,t}^{RE})\frac{(\alpha_i^{RE}\lambda_{i,t}^{RE})^{n_{i,t}}}{n_{i,t}!}\right) f(\alpha_i^{RE})d\alpha_i^{RE} \\
&= \left(\prod_{t=1}^{T}\frac{(\lambda_{i,t}^{RE})^{n_{i,t}}}{n_{i,t}!}\right)\frac{\Gamma(n_{i,\bullet}+\nu)}{\Gamma(\nu)}\left(\frac{\nu}{\lambda_{i,\bullet}^{RE}+\nu}\right)^{\nu}\left(\lambda_{i,\bullet}^{RE}+\nu\right)^{-n_{i,\bullet}}
\end{aligned}$$

(where $n_{i,\bullet}=\sum_{t=1}^{T} n_{i,t}$ and $\lambda_{i,\bullet}^{RE}=\sum_{t=1}^{T}\lambda_{i,t}^{RE}$)

MVNB

- This distribution is a **generalization** of the **negative binomial distribution**.

Random Effects Model

MVNB

1. It is a **basic** distribution for panel count data modeling with **overdispersion**.
2. It is **not** a member of the **linear exponential family**.
3. We use GAMLSS theory to include smooth functions into the mean parameter.

Generalized Additive Models for Location, Scale and Shape

- Any distribution.
- **More flexible** : can model a location parameter $\boldsymbol{\mu_i}$, a variance parameter $\boldsymbol{\sigma_i}$ (scale), a skewness parameter $\boldsymbol{\nu_i}$ and a kurtosis parameter $\boldsymbol{\tau_i}$ as additive functions of the covariates.

$$g_k(\boldsymbol{\theta_k}) = \mathbf{X_k}\beta_\mathbf{k} + \sum_{j=1}^{J_k} s_{j,k}(x_{j,k}), \tag{3}$$

where $s_{j,k}$ is a smooth non-parametric function.
- $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}\}$. $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}$ and $\boldsymbol{\tau}$ are vectors with n elements
- Can specify **only the location parameter** : $\boldsymbol{\theta} = \{\boldsymbol{\mu}\}$.

Random Effects Model

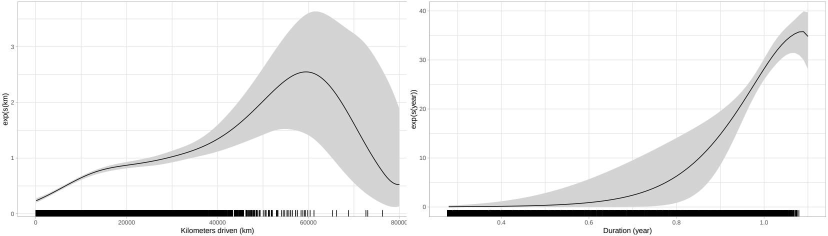

Figure 2 – $\exp(\hat{s}_1(km))$ and $\exp(\hat{s}_2(year))$ from the GAMLSS with random effects model

Model Specification

$N_{i,t}|\alpha_i \sim \text{Poisson}(\mu_{i,t} = \alpha_i \lambda_{i,t})$, where

- $\alpha_i^{RE} \sim Gamma(\nu, \nu),\ i = 1, \ldots, n.$
- $\lambda_{i,t} = \exp(\mathbf{X_{i,t}}\beta + s_1(km) + s_2(d))$
- $\boldsymbol{\nu}$ is kept **constant** for all individuals.

Fixed Effects

A Fixed Effects Approach

The model

$N_{i,t}|\alpha_i \sim \text{Poisson}(\mu_{i,t} = \alpha_i \lambda_{i,t})$, where

- α_i^{FE}, $i \in \{1, \ldots, n\}$ are unknown parameters.
- $\lambda_{i,t} = \exp(\mathbf{X_{i,t}}\beta + s_1(km) + s_2(d))$.
- GAM theory.

Parameters estimation

1. At least $n + p + 1$ parameters should be estimated.
2. The large number of parameters in the model causes what is called **incidental problem** : an incorrect estimation of the fixed effects α generates **incorrect estimates** of β associated with covariates in the mean.
3. It has been shown that a fixed effects model based on a Poisson distribution **does not have this problem** (see (Cameron and Trivedi, 2013)) for a detailed explanation).

- The splines are estimated correctly.

A Fixed Effects Approach

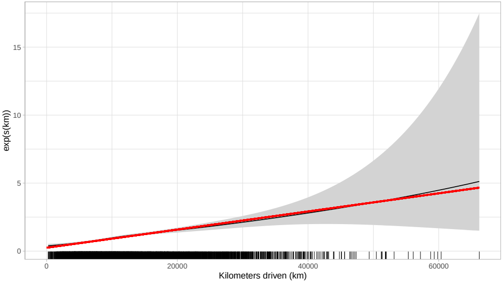

Figure 3 – GAM with fixed effects estimated with Canadian data

Results Analysis

1. The relationship is **always increasing**, and is even **almost linear**.
2. What has been called the "learning effect" has **disappeared**.
3. We observe a **much more logical** and **coherent** relationship than in the previous models.

A Fixed Effects Approach

Marginal impact of each additional kilometer

1. We approximated **exp(s(km))** by $0.25 + \frac{1}{15000}\,km_{i,t}$ (the red line), we then have

$$
\begin{aligned}
N_{it} &\sim Poisson\left(\exp(\alpha_i)\exp(s(km))\right) \\
&\sim Poisson\left(\exp(\alpha_i)(a + b\,km_{i,t})\right) \\
&\sim Poisson\left(0.25\exp(\alpha_i) + \frac{1}{15000}\exp(\alpha_i)\,km_{i,t}\right).
\end{aligned}
$$

2. The **slope** is **not the same** for each insured because it **depends on $\boldsymbol{\alpha_i}$**.

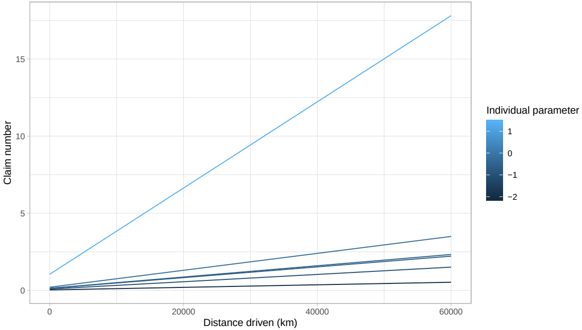

Figure 4 – Exposure measure for different individual parameters.

Results Analysis II

- With this model, we then **reconcile** the intuition that **each kilometer** should increase the risk for an individual, but that this increase could be **different for each driver**.

A Fixed Effects Approach

"Learning effect"

Instead of referring to the "learning effect", we **should understand** instead that

1. Typical insureds who drive more than 60,000 km per year are **better risks per kilometer** than insureds who drive approximately 40,000 km per year.
2. The difference between insureds related to their risk *per kilometer* can be explained by many factors : more frequent use of the highway, higher proportion of driving outside rush hours, etc.
3. For each driver, independently of their driving risk *per kilometer*, the risk of an accident will always **increase for each additional kilometer driven** (by approximately $\frac{1}{15{,}000}$).

Comparative Analysis

Which Effect Should Be Used in Practice ?

The **fixed effects model** is **more general** than the random effects model

- In case of contradictory results, fixed effects should always be **preferred**.

Random effects model

$$
\begin{aligned}
\Pr[N_{i,1} &= n_{i,1},...,N_{i,T} = n_{i,T}] \\
&= \int_0^\infty \Pr[N_{i,1} = n_{i,1},...,N_{i,T} = n_{i,T} | \mathbf{x}_{i,1},...,\mathbf{x}_{i,T}, \alpha_i^{RE}] f(\alpha_i^{RE} | \mathbf{x}_{i,1},...,\mathbf{x}_{i,T}) d\alpha_i^{RE} \\
&= \int_0^\infty \left(\prod_{t=1}^{T} \Pr[N_{i,t} = n_{i,t} | \mathbf{x}_{i,1},...,\mathbf{x}_{i,T}, \alpha_i^{RE}] \right) f(\alpha_i^{RE}) d\alpha_i^{RE}
\end{aligned}
$$

- **Additional assumption** : $f(\alpha_i^{RE} | \mathbf{x}_{i,1},...,\mathbf{x}_{i,T})$ becomes $f(\alpha_i^{RE})$.
- The interpretation of random effects results are tricky.

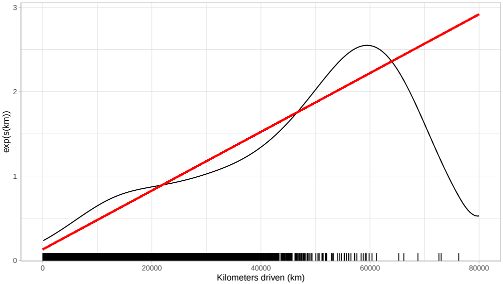

Figure 5 – Comparison between the random effect approach and the fixed-effect approach for the median value of the individual parameter

To Conclude

Take-home points

1. **Fixed effects** should be used to understand the **"true" relationship** between covariates and claims experience.
2. **Instead** of referring to the **"learning effect"**, we should understand instead that typical insureds who drive more than 60,000 km per year are **better risks per kilometer** than insureds who drive approximately 40,000 km per year.
3. For ratemaking, fixed effects should be used to compute the **premium surcharge** for each additional kilometer the insureds drive.
4. **Fixed effects** can be used to construct PAYD insurance solely based on kilometers driven for **self-service vehicles**, where drivers' profile cannot be directly used for ratemaking.

References

[1] Boucher, Jean-Philippe, and Roxane Turcotte. 2020. A Longitudinal Analysis of the Impact of Distance Driven on the Probability of Car Accidents *Risks* 8 : 91.

[2] Ayuso, Mercedes, Montserrat Guillen, and Ana María Pérez-Marín. 2016b. Telematics and gender discrimination : Some usage-based evidence on whether men's risk of accidents differs from women's. *Risks* 4 : 10.

[3] Boucher, Jean-Philippe, Steven Côté, and Montserrat Guillen. 2017. Exposure as duration and distance in telematics motor insurance using generalized additive models. *Risks* 5 : 54.

[4] Cameron, A. Colin, and Pravin K. Trivedi. 2013. *Regression Analysis of Count Data*. Cambridge : Cambridge University Press, vol. 53.

[5] Ferreira, Joseph, and Eric Minikel. 2010. *Pay-as-You-Drive Auto Insurance in Massachusetts : A Risk Assessment and Report on Consumer, Industry and Environmental Benefits*. Boston : Conservation Law Foundation.

[6] Hastie, Trevor, and Robert Tibshirani. 1986. Generalized additive models. *Statistical Science* 1 : 297–310.

[7] Lemaire, Jean, Sojung Carol Park, and Kili C. Wang. 2016. The use of annual mileage as a rating variable. *ASTIN Bulletin* 46 : 39–69.

[8] Verbelen, Roel, Katrien Antonio, and Gerda Claeskens. 2018. Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society : Series C (Applied Statistics)* 67 : 1275–304.