Chaire Co-operators en
analyse des risques actuariels

Ratemaking with Telematics Data

Casualty Actuaries of Greater New York - 2021 Virtual Spring Meeting Webinar

Roxane Turcotte

May 12th 2021

Université du Québec à Montréal

Longitudinal analysis of distance traveled

Longitudinal analysis of distance traveled

Context

- New technologies such as **GPS-collected data** have emerged, which offer new ways to approach car insurance pricing.
- Processing these data provides **reliable information** about drivers' behavior.

One piece of GPS-collected information that is directly related to the risk insured is distance driven.

Relevance

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

- 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

In a social-political context where the use of gender in ratemaking is restricted or criticize, calculating premiums on **more objective information** is of interest.

Overview

Objective

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

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

Why GPS-collected data ?

- As shown by many authors, such as Lemaire et al. (2016), the **self-reported** approximation of the distance driven is **not reliable** and is often very different from the exact distance driven.
- There are **important differences between driving uses and driving habits**, which justifies consideration of other measures than exposure time in the modeling.

A First Model

Starting Point

Boucher et al. (2017), by using a **GAM Poisson model**, analyzed the influence of duration and distance driven on the number of claims with **independent cubic splines** : $\log(\mu_i) = \beta_0 + s_1(km_i) + s_2(d_i)$.

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

GAM

- 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 $g(\mu_i) = s_0 + s_1(x_{1,i}) + s_2(x_{2,i}) + s_3(x_{3,i})$.

A First Model

What do you think ?

We model $N_{it} \sim Pois(\boldsymbol{\mu_{it}})$, where $\boldsymbol{\mu_{i,t}} = \exp(\mathbf{s_1}(\mathbf{km}))\exp(\mathbf{s_2}(\mathbf{d}))\boldsymbol{\lambda_{i,t}}$ with real canadian insurance data.

Questions :

1. 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 ?
2. And $\exp(s_2(d))$?

To help you :

- Would it be nonetheless nearly linear ?
- Would it stop increasing at some point ?
- Would it start to decline at some point ? Would it go up again ?
- Any other intuition ?

A First Model

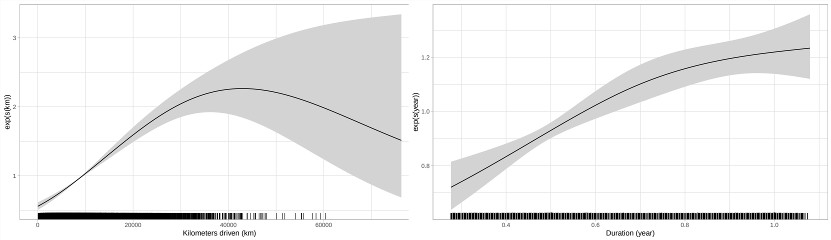

Figure 1: $\exp(\hat{s}_1(km))$ and $\exp(\hat{s}_2(year))$ from the Poisson GAM

Case Study

1. All models are illustrated using data from a **major Canadian insurance company**.
2. The model $\log(\mu_i) = \beta_0 + s_1(km_i) + s_2(d_i)$ 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

The **slope** could change as distance increases, but it should always be **strictly positive** since the **risk** is **greater**, meaning that the smoothing function should always be increasing.

- **One explanation** comes from the fact that GAM **supposes independence** between all contracts of the same insured.

Results Analysis

One can argue that distance driven is **correlated** with **other driving habits** resulting from driving experience, (Ferreira and Minikel (2010)).

1. The model **does not** take this correlation **into account**.
2. The resulting relationship between claim frequency and the distance driven do **not** give an **appropriate representation** of how the claim frequency could change **when** insureds **change their driving habits**.

We think that the shape of the smoothing function comes from the **driver profiles** : the lower quantiles of the distribution of the distance driven does not come from the same (type of) drivers as the higher quantiles.

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**.

Panel Data Modeling

In non-life insurance, however, we can observe the same insured over **many** contracts.

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

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

Construct Multivariate Count Models

- 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

We can consider two different situations regarding this parameter :

1. All α_i, $i = 1, \ldots, n$ are i.i.d. **random variables** that come from a selected **prior distribution** (we call this the random effects model) ;
2. All α_i, $i = 1, \ldots, n$ are **unknown parameters** that need to be **estimated** (we call this the fixed effects model).

Random Effects Model

Model Specification

In random effects models, we suppose that α_i, $i = 1, \ldots, n$, are **random variables**, with prior density $f(\cdot)$.

- 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.

$$\Pr[N_{i,1} = n_{i,1}, \ldots, N_{i,T} = n_{i,T}] = \int_0^\infty \left(\prod_{t=1}^{T} \exp(-\alpha_i^{RE} \lambda_{i,t}^{RE}) \frac{(\alpha_i^{RE} \lambda_{i,t}^{RE})^{n_{i,t}}}{n_{i,t}!} \right) f(\alpha_i^{RE}) d\alpha_i^{RE}. \tag{3}$$

- Many distributions can be used for α_i^{RE}, such as the **gamma** or the inverse Gaussian.

Gamma Distribution

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

$$\Pr[N_{i,1} = n_{i,1}, \ldots, N_{i,T} = n_{i,T}] = \left(\prod_{t=1}^{T} \frac{(\lambda_{i,t}^{RE})^{n_{i,t}}}{n_{i,t}!} \right) \frac{\Gamma(n_{i,\bullet} + \nu)}{\Gamma(\nu)} \left(\frac{\nu}{\lambda_{i,\bullet}^{RE} + \nu} \right)^{\nu} \left(\lambda_{i,\bullet}^{RE} + \nu \right)^{-n_{i,\bullet}} \tag{4}$$

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

Random Effects Model

MVNB

This well-known distribution is the multivariate negative binomial distribution.

1. This distribution is a **generalization** of the **negative binomial distribution**.
2. It is a basic distribution for panel count data modeling with overdispersion ($\mathbb{E}[N_{i,t}] = \lambda_{i,t}^{RE} < \mathbb{V}[N_{i,t}] = \lambda_{i,t}^{RE} + (\lambda_{i,t}^{RE})^2/\nu$).
3. It is **not** a member of the **linear exponential family**.
4. GAM theory **cannot be used** to include smoothing functions.

It can be shown that the first-order condition to obtain $\widehat{\beta}_{\textbf{MLE}}$ is :

$$\sum_{i=1}^{n} \sum_{t=1}^{T} \mathbf{x}_{i,t} \left(n_{i,t} - \lambda_{i,t}^{RE} \frac{n_{i,\bullet} + \nu}{\lambda_{i,\bullet}^{RE} + \nu} \right) = 0. \tag{5}$$

Random Effects Model

GAMLSS

Instead, we use Generalized Additive Models for Location, Scale and Shape theory, that can be used for other distributions than the members of the linear exponential family of distribution.

- **More flexible** : can model a location parameter μ_i, a variance parameter σ_i (scale), a skewness parameter ν_i and a kurtosis parameter τ_i as additive functions of the covariates.

$$g_k(\boldsymbol{\theta}_\mathbf{k}) = \mathbf{X}_\mathbf{k}\boldsymbol{\beta}_\mathbf{k} + \sum_{j=1}^{J_k} \mathbf{Z}_{\mathbf{j,k}}\boldsymbol{\gamma}_{\mathbf{j,k}} \tag{6}$$

- $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}\}$. $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}$ and $\boldsymbol{\tau}$ are vectors with n elements
- If a smooth function can be expressed in linear form, Equation (6) can be rewritten as

$$g_k(\boldsymbol{\theta}_\mathbf{k}) = \mathbf{X}_\mathbf{k}\boldsymbol{\beta}_\mathbf{k} + \sum_{j=1}^{J_k} h_{j,k}(x_{j,k}),$$

where $h_{j,k}$ is a smooth non-parametric function.

Random Effects Model

Model Specification

It is possible to use a GAMLSS that specify **only the location parameter**. In this case, $\boldsymbol{\theta}$ would simply become $\boldsymbol{\theta} = \{\boldsymbol{\mu}\}$.

1. We choose to model the parameter $\boldsymbol{\lambda}_{\mathbf{i,t}}$ with smoothing function ;
2. $\boldsymbol{\nu}$ is kept **constant** for all individuals.

R package

1. To use GAMLSS, many distributions are available in the R package *gamlss*.
2. Unfortunately, the MVNB distribution is not one of them.
3. The distribution is however implemented by itself in the package *multinbmod*).

Consequently, we have to **write our own code** for convenience.

Random Effects Model

What do you think ?

We model $\mathbf{N} \sim MVNB(\mu, \nu)$, where $\mu = \exp(\mathbf{s_1}(\mathbf{km}))\exp(\mathbf{s_2}(\mathbf{d}))\boldsymbol{\lambda}$ with real canadian insurance data.

Questions :

1. What the **relation** between $\exp(\mathbf{s_1}(\mathbf{km}))$ [$\exp(s_2(d))$] and **claim frequency** would look like ?
2. **How** would the results **differ** from the **previous model** ?

To help you :

- Would it be nonetheless nearly linear ?
- Would it stop increasing at some point ?
- Would it start to decline at some point ? Would it go up again ?
- Any other intuition ?

Random Effects Model

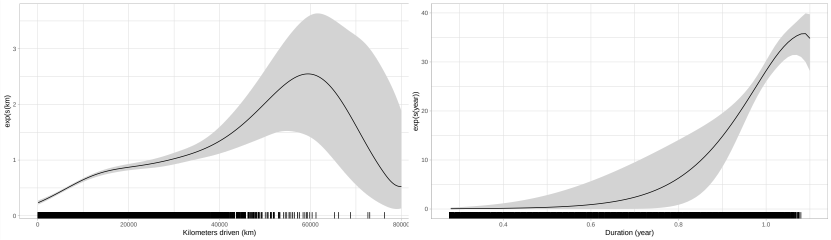

Figure 2: $\exp(\hat{s}_1(km))$ and $\exp(\hat{s}_2(year))$ from the GAMLSS with random effects model

Model Fitting

1. To fit the model, we maximize a **penalized log-likelihood** function l_p, integrating a quadratic penalty $\gamma^{\mathbf{T}}\mathbf{G}\gamma$.
2. **Penalty matrix G** : very often define as $\mathbf{\Lambda D_r^T D_r}$ (different formulations possible).
3. A hyper-parameter, noted here $\Lambda \in \mathbb{R}^+$, controls the **weight given to the penalty**. The **greater** its value, the **smoother** the resulting estimated function.
4. To select the penalty parameters in $G(\Lambda)$ associated with both p-splines, we **test out** multiple **combinations** of values of $\Lambda = \{\Lambda_1, \Lambda_2\}$.

A Fixed Effects Approach

The model

Poisson fixed effects model can be seen as a **basic Poisson** regression model without an intercept. Being part of the linear exponential family of distribution, **GAM theory** can then be used when smoothing functions are added to the mean parameter of the distribution.

In practice, as mentioned, it is **relatively easy** to implement the fixed effects model with R ; we simply used the *gam* function from the package *mgcv*.

1. To include fixed effects in the model the **intercept** of the model is **dropped**.
2. We include a **unique identifier** variable for each policyholder as a factor variable and we include the **distance driven** in the model using a **cubic spline s.**

A Fixed Effects Approach

Parameters estimation

In the fixed effects model, we consider each α_i, $i \in \{1, \ldots, n\}$ as an **unknown parameter**.

1. At least $n+p+1$ parameters should be estimated, which is quite a high number of parameters given that T_i is usually small for insurance datasets.
2. 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 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).

First-order condition equation

For the β parameters, the first condition by MLE can be shown to be equal to :

$$\sum_{i=1}^{n} \sum_{t=1}^{T_i} \mathbf{x}_{i,t} \left(n_{i,t} - \lambda_{i,t}^{FE} \frac{n_{i,\bullet}}{\lambda_{i,\bullet}^{FE}} \right) = 0. \tag{7}$$

When we compare the first-order condition equation of the random effects model and (7), we see that when T is large, or when $\boldsymbol{\nu \to 0}$, **random and fixed** effects models are **equivalent**.

A Fixed Effects Approach

What do you think ?

We model $N_{i,t} \sim Pois(\boldsymbol{\mu_{i,t}})$, where $\boldsymbol{\mu_{i,t}} = \mathbf{exp(a_i)}\exp(\mathbf{s(km)})$.

Questions :

1. What the **relation** between $\exp(s(km))$ and claim frequency would look like ?
2. Will the **"learning effect"** be there again ?

Rating structure based on distance driven

We decided to model the Poisson fixed effects by **not including** a smoothing function for the duration.

1. Our objective is to measure the **marginal effect** of the distance on the claim frequency. If we want to measure the risk of each additional kilometer the insured decides to drive, the **duration** of the contract is **not important**.
2. We want to construct a rating structure based solely on the distance driven as a risk measure.

A Fixed Effects Approach

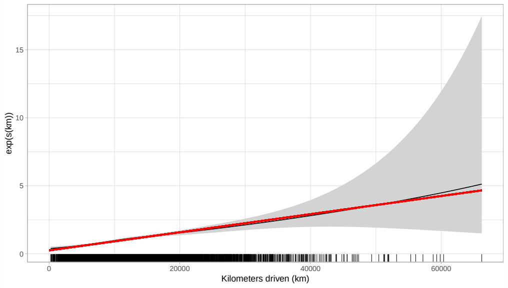

Figure 3: GAM with fixed effects estimated with Canadian data

Results Analysis

1. The relationship between distance traveled and claim frequency 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 between distance traveled and frequency than before.

A Fixed Effects Approach

Marginal impact of each additional kilometer

1. The relationship between claim frequency and the distance driven should be understood as the **marginal impact** of **each** additional kilometer driven or not-driven.
2. Explicitly, as we approximated $\exp(\mathbf{s}(\mathbf{km}))$ by $0.25 + \frac{1}{15000}\, km_{i,t}$ (the red line), we then have

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

3. We see that the **slope**, i.e., the marginal impact of each additional kilometer driven or not-driven, is **not the same** for each insured because it **depends on** $\boldsymbol{\alpha_i}$.

A Fixed Effects Approach

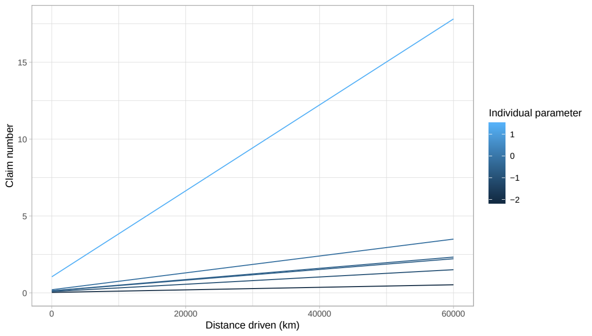

Figure 4: Exposure measure for different individual parameters.

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"

In summary, **instead** of referring to the "learning effect" to understand the left-hand graph of Cross-sectional data model, 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. However, 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 $\frac{1}{15{,}000}$).

Which Effect Should Be Used in Practice ?

The fixed effects model is **more general** than the random effects model, which means that in case of contradictory results, **fixed effects** should always be **preferred**.

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

We can see that we have to suppose an **additional assumption** : from the first to the second line of development, $f(\alpha_i^{RE} | \mathbf{x}_{i,1},\ldots,\mathbf{x}_{i,T})$ becomes $f(\alpha_i^{RE})$. **The interpretation of random effects results are tricky.**

Comparative Analysis

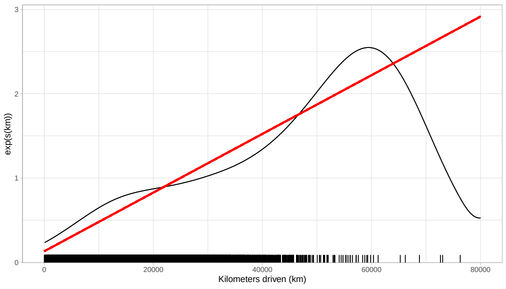

Figure 5: Comparison between the random effect approach and the fixed-effect approach for the median value of the individual parameter

Which Effect Should Be Used in Practice ?

1. Fixed effects modeling, even if theoretically better, is **not amenable** to ratemaking.
2. On the other hand, the **MVNB** can be used for **predictive rating**, where it can be shown that the predictive distribution of $N_{i,T}$ depends on past values of $\lambda_{i,t}$ and $n_{i,t}$, for $t = 1, \ldots, T-1$.

To Conclude

Take-home points

1. **Fixed effects** should be used to understand the **"true" relationship** between covariates and claims experience.
2. For ratemaking, fixed effects should be used to compute the **premium surcharge** for each additional kilometer the insureds drive.
3. In our case, it represents an increase of $\widehat{\alpha}_i \frac{1}{15{,}000}$ per km, for claim frequency.
    - Using this approach, insurers will **avoid** the situation where an insured could see a **premium reduction** if, for example, he decides to drive 50,000 km instead of 40,000 km, as we saw with a **basic GAM approach**.
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.
5. Research is required in this area.

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.