Reacfin

Know-How to Risk

Geographical ratemaking with R

Xavier Maréchal
Head of Innovation & Quality
xavier.marechal@reacfin.com

R in Insurance

14th July 2014

Non-Life pricing

- The Generalized Linear Models (GLM) are a classical statistical tool for non-life pricing
- GLM are a generalization of linear regressions
- In comparison with linear models, two main assumptions are relaxed
  - The response variable Y does not need to be a linear combination of explanatory variables $\Rightarrow$ Y is a function of a linear combination of the explanatory variables
  - Errors (and so, Y) do not need to be Gaussian
    - They have to be a member of the exponential family [Normal, Poisson, Gamma, Inverse Gaussian, Binomial]
    - Errors can then have non constant variance

Explanatory variables

Parameters

- The main components of GLM are
  - A score which is a linear predictor for response Y_i : $score_i = \boxed{\beta_0} + \sum_{j=1}^{p} \boxed{\beta_j} \boxed{x_{ij}}$
  - A link function that makes the link between the score and the mean μ_i of the responses : $g(\mu_i) = score_i \Leftrightarrow \mu_i = g^{-1}(score_i)$
- E.g.: in a Poisson regression for claims numbers we have $\boldsymbol{N_i} \sim \boldsymbol{Poi}\ [\boldsymbol{d_i}\ \boldsymbol{exp}(\boldsymbol{\beta}^T \boldsymbol{X}_i)]$

Continuous explanatory variables

- The linearity in the GLMs is not restrictive for categorical explanatory variables coded by means of binary variables, but well for continuous explanatory variables which may have a nonlinear effect on the score
- In practice, even if many explanatory variables are categorical (eg: gender, use of the car, occupation of the policyholder,…), some important rating variables are "continuous"
  - Ex: in motor ratemaking, age, power of the car or ZIP code are "continuous"
- Let x* be such a "continuous " explanatory variable.
  - Entering x* directly in the linear predictor boils down to assume a linear effect of the x* on the score scale: in log-linear models, this means that the mean is constrained to vary exponentially with x*
  - Such a monotone exponential behavior may not be supported by the data

Continuous explanatory variables

- Solutions
  - Treat x* as a categorical explanatory variable … but
    - This may introduce a large number of additional regression parameters
      - For example, for the age, it would result in around 70 coefficients for each integer age present in the portfolio
      - We could consider grouped age classes (e.g. 18-19,20-24,25-29,30-34,35-39,40-44,…) but we should determine how to build the groups
    - This may fail to recognize the possible smooth variation of the mean in x*
  - Entering different transformations of x* in the linear predictor (x*,$\sqrt{x*}$, sin x*, ln x*,…), but this obscures the model and any parametric specification may be erroneous
  - Semi-parametric approach: if we are not sure about the type of influence of x*, we would prefer fitting a model with an additive score of the form

$$linear\ part\ +\ f\ (x^{*})$$

where f is left unspecified and estimated from the data

Generalized additive model (GAM) for continuous variable

- As for the GLM, the response Y_i probability distribution has to be any member of the Exponential Dispersion family

- The mean μ_i of Y_i is linked to the nonlinear score via

$$g(\mu_i) = \ \beta_0 + \sum_{j=1}^{p_{cat}} \beta_j x_{ij} + \sum_{j=p_{cat}+1}^{p} f_j(x_{ij}) = score_i$$

for some smooth unspecified functions f_j, where g is the link function

Application in geographical ratemaking

- In motor insurance, most companies have adopted a risk classification according to the geographical zone where the policyholder lives (urban / non urban for instance, or a more accurate splitting of the country according to Zip codes).

- In order to predict the underlying risk in a geographical region, we use claims data which are near or relatively near to the region of interest.

- The geographical location is contained within the postcode for each policy

Application in geographical ratemaking

- For each of the policyholder i, we have
  - d_i which is the risk exposure
  - N_i^{obs} which is the observed number of claims
  - CM_i^{obs} which is the observed mean cost of the claims
- The aim is to introduce in the tariff a new explanatory variable based on the policyholder's district (ZIP code)
- Zones (group of different districts with similar risks) have to be constructed and their relativities have to be established
  - This new categorical variable has usually between 3 and 6 different modalities
- To perform this exercise, it is usually better to take into account both the frequency and the mean cost even if a lot of companies concentrate only on the frequency in order to establish their geographical ratemaking

Application in geographical ratemaking

- Modelling
  - First of all, we assume that

$$N_i^{obs} \sim Poi\ (d_i\ exp(\boldsymbol{\beta}^T\ \boldsymbol{x}_i))$$

    - Thus, the predicted number of claims can be obtained by:

$$N_i^{pred} = d_i \exp\left(\hat{\beta}_0^{freq} + \sum\nolimits_{j=1}^{p} \hat{\beta}_j^{freq}\ x_{ij}\right)$$

  - For the mean cost, we use the following model

$$CM_i^{obs} \sim Gamma\ (\mu_i = \ exp(\boldsymbol{\beta}^T\ \boldsymbol{x}_i))$$

    - And obtain, for the predicted mean cost:

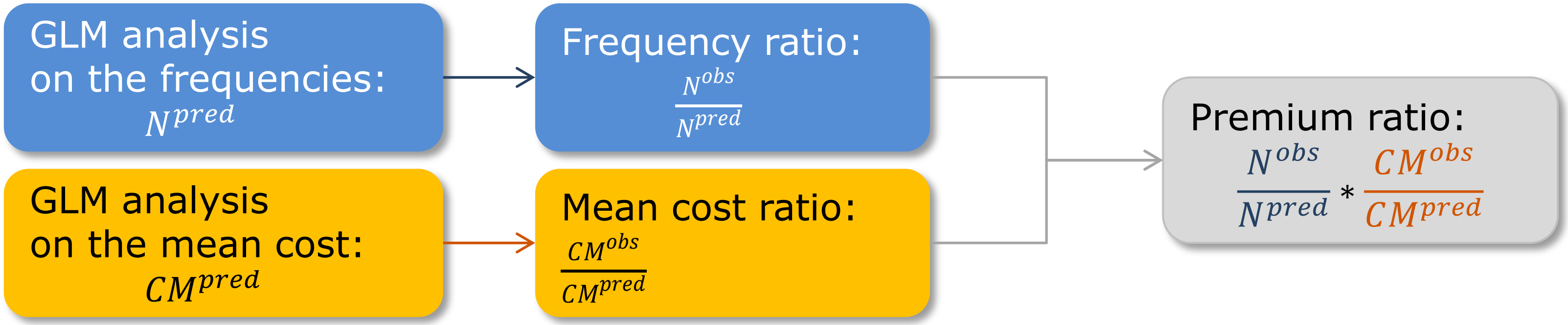
$$CM_i^{pred} = \exp\left(\hat{\beta}_0^{CM} + \sum\nolimits_{j=1}^{p} \hat{\beta}_j^{CM}\ x_{ij}\right)$$

  - To take into account both effects (frequency and mean cost), we compute the predicted pure premium

$$PP_i^{pred} = N_i^{pred} CM_i^{pred} = d_i \exp\left[\hat{\beta}_0^{freq} + \hat{\beta}_0^{CM} + \sum\nolimits_{j=1}^{p} \left(\hat{\beta}_j^{freq} + \hat{\beta}_j^{CM}\right) x_{ij}\right]$$

Application in geographical ratemaking

- Modelling
  - To determine the zones similar in terms of risk, we first aggregate the predicted and observed number of claims by district (ZIP code). The same is also performed for the mean cost.
  - We can then calculate the ratios between the observed pure premium and the predicted pure premium for each district



GLM analysis on the frequencies: N^{pred}

Frequency ratio: $\frac{N^{obs}}{N^{pred}}$

GLM analysis on the mean cost: CM^{pred}

Mean cost ratio: $\frac{CM^{obs}}{CM^{pred}}$

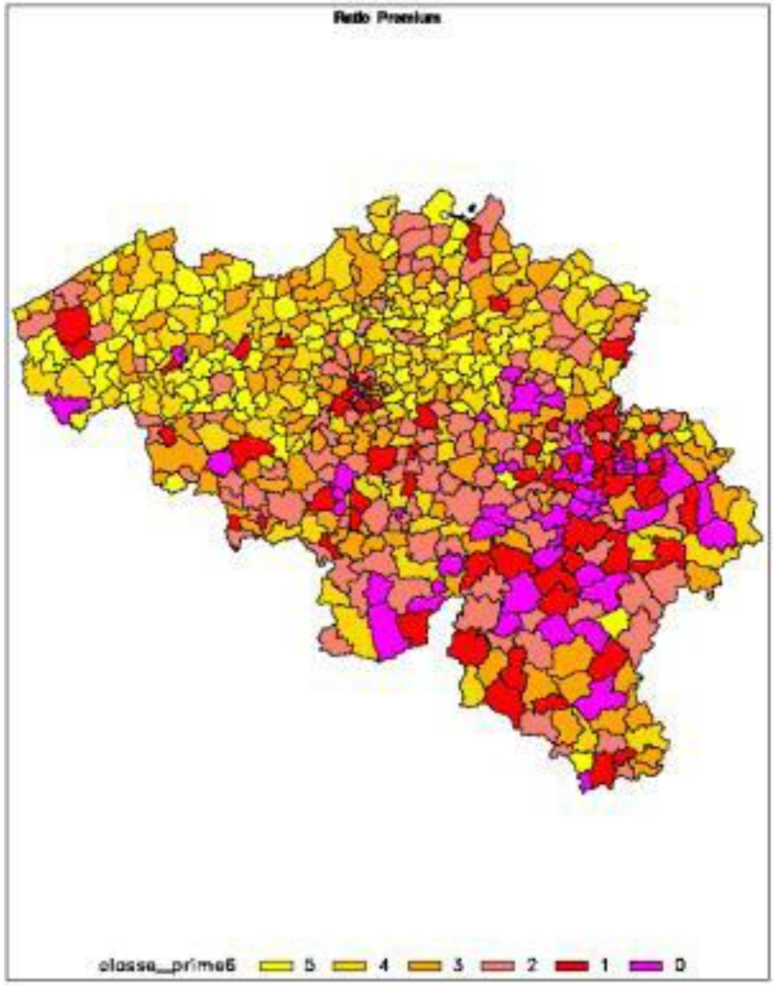
Premium ratio: $\frac{N^{obs}}{N^{pred}} * \frac{CM^{obs}}{CM^{pred}}$

  - These ratios can be interpreted as “residuals” by district
  - The idea is then to “structure” these residuals in order to define a categorical variable that will improve the risk prediction of the model

Application in geographical ratemaking

- A first idea would be to sort these premium ratios to create classes for the new variable

- Results
  - Using these “sorted” premium ratios, we obtain the following zones and for each of them, the corresponding relativities (thanks to an additional GLM analysis with the new categorical variable)
  - Unfortunately these results are not very smooth and could be difficult to explain from a commercial point of view



Application in geographical ratemaking

- To obtain smoother results, we can use another methodology and add in the model the coordinates of the district as a continuous explanatory variable
  - For each of the 589 Belgian districts, we know the coordinates (u_j, v_j) of the location of the centre of the district
- Thus, we can use the model

$$N_j^{obs} \sim Poi\left(N_j^{pred}\ exp\left(f\left(u_j, v_j\right)\right)\right)$$

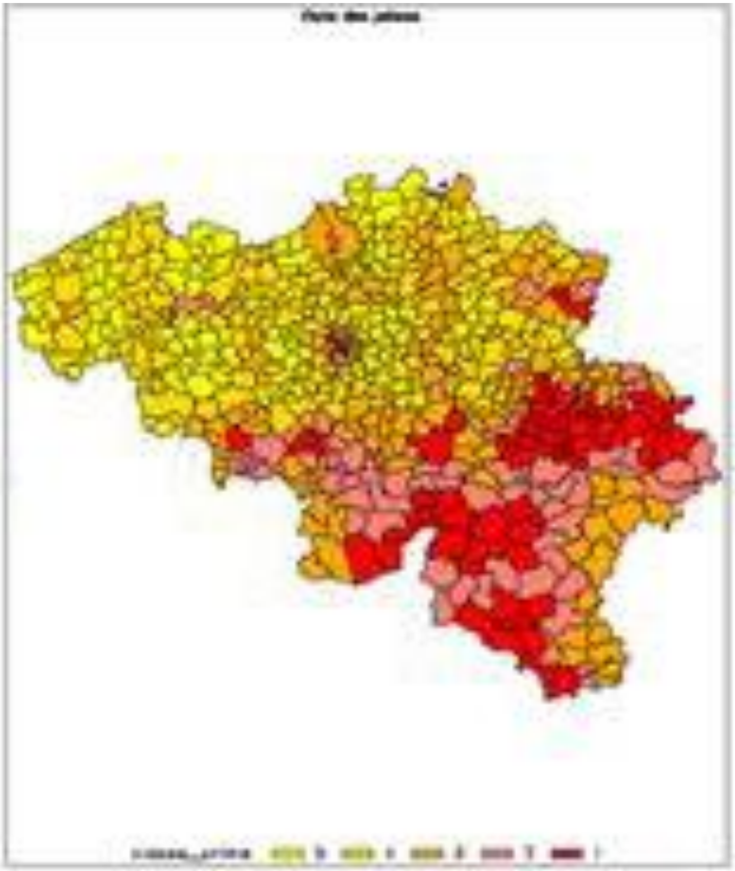
to estimate the function f assessing the geographic risk variations.

- The estimate $\exp\left(\hat{f}\left(u_j, v_j\right)\right)$ quantifies the relative risk of district j, everything else being equal
- For the mean cost, the following model is used

$$CM_j^{obs} \sim Gamma\left(\mu_j = CM_j^{pred} exp\left(g\left(u_j, v_j\right)\right)\right)$$

Application in geographical ratemaking

- To build zones on the basis of the pure premium, we use the ratio between the pure and the predicted premium
- Thus, we obtain:



$$\frac{\text{Pure Premium}}{\text{Predicted Premium}}$$

$$= \frac{N_j^{pred} \exp\left(\hat{f}(u_j, v_j)\right) * CM_j^{pred} \exp\left(\hat{g}(u_j, v_j)\right)}{N_j^{pred} CM_j^{pred}}$$

$$= \exp\left(\hat{f}(u_j, v_j)\right) \exp\left(\hat{g}(u_j, v_j)\right)$$

- As the $\hat{f}$ and $\hat{g}$ have been estimated with regressions, the results obtained are smooth with this methodology.

How is it performed in R?

- Example for claims frequency
- Gam function can be used to include coordinates of the district as a continuous explanatory variable

$$gam(\underbrace{Data[["NobsZIP"]]}_{\text{Observed number of claims}} \sim offset(LnPred)$$

$$+\, s\left(Data[["LAT"]], Data[["LONG"]]\underbrace{, bs = "sos"}_{\text{2 dimensional splines}}\right), family = poisson)$$

- Where LnPred is the logarithm of the predicted number of claims, resulting from the glm analysis (Poisson regression for claims counts) previously performed.
- The readShapeSpatial function is used to exploit and then plot the coordinates of the frontiers of each district

Alternative solutions (1/2)

- Instead of using the prediction Pred from the GLM model as an offset, the regression coefficients for the linear part of the score can be estimated directly in the GAM procedure taking into account the influence of (the non-linear effect of) geographical location

$$gam(Data[["\mathrm{Nbclaims}"]] \sim offset(\log(Data[["Exposure"]]))$$

$$+Data[["sexp"]] + Data[["fleetc"]] + Data[["usec"]] + \ \ldots$$

$$s(Data[["LAT"]], Data[["LONG"]], bs = "sos"), family = poisson)$$

- Unfortunately, if too many categorical variables are included, this methodology can lead to non-convergence of the fitting algorithm

Alternative solutions (2/2)

- An alternative to using the GAM would be the following
  - Compute the residuals by district: e.g. the deviance residuals
  - Smooth these residuals with local polynomial regression (Loess function) taking into account the latitude and longitude of the districts
  - The risk exposure in each district can also be used as weight to improve the smoothing

Disclaimer:

The recipient of this document should treat all information as strictly confidential and only in the context stated below. Information may not be disclosed to any third party without the prior join-consent of Reacfin.

Estimates given in this presentation are based on our current knowledge, they can be based upon our previous experience within the Undertaking, as well as taking into account similar projects in the same context as the Undertaking, either locally, within majority of the EU countries as well as overseas.

This presentation is only the supporting document of a verbal presentation. Hence, it is not intended to be exhaustive. Quoting or using this document on its own might be misleading. As a result, these materials may not be used by anybody except their authors nor should they be relied upon in any way for any purpose other than as contemplated by joint written agreement with Reacfin.

Reacfin

Place de l’Université 25
B-1348 Louvain-la-Neuve
www.reacfin.com