GLM vs XGBM – Part 2a GLM Modelling

Introduction

The goal of this presentation is to compare the results of modelling a Car insurance Pure Premium using a classic GLM vs a XGBM in R.

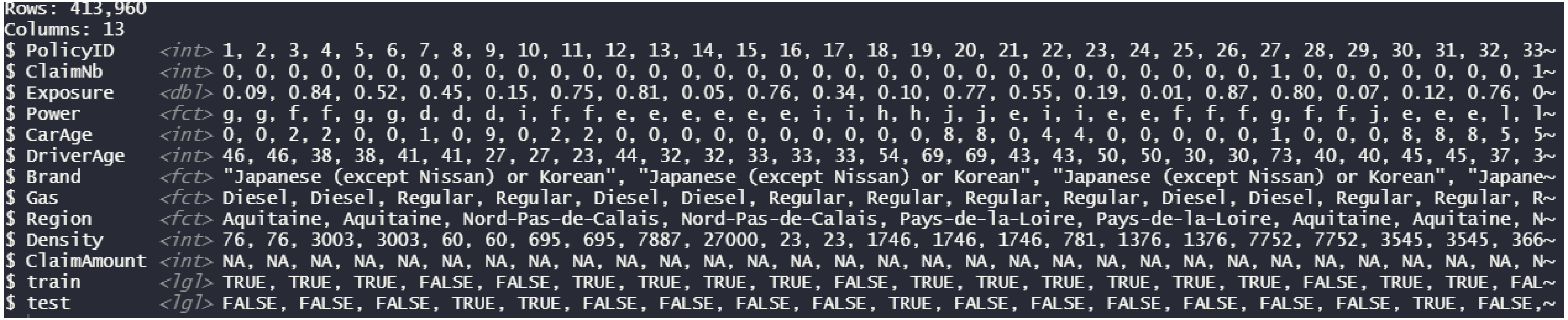
The thought process shown here is not exhaustive, we are just presenting the main visuals that can ease the overall understanding.

The data in use is a join of 2 database present in the R {CASdatasets} package (link: https://github.com/dutangc/CASdatasets)

- freMTPLfreq.rda
- freMTPLsev.rda

- The codes for these study is enclosed in the repo: https://github.com/william-tiritilli/GLM-vs-GBM---Part-2---Modelling-/tree/main

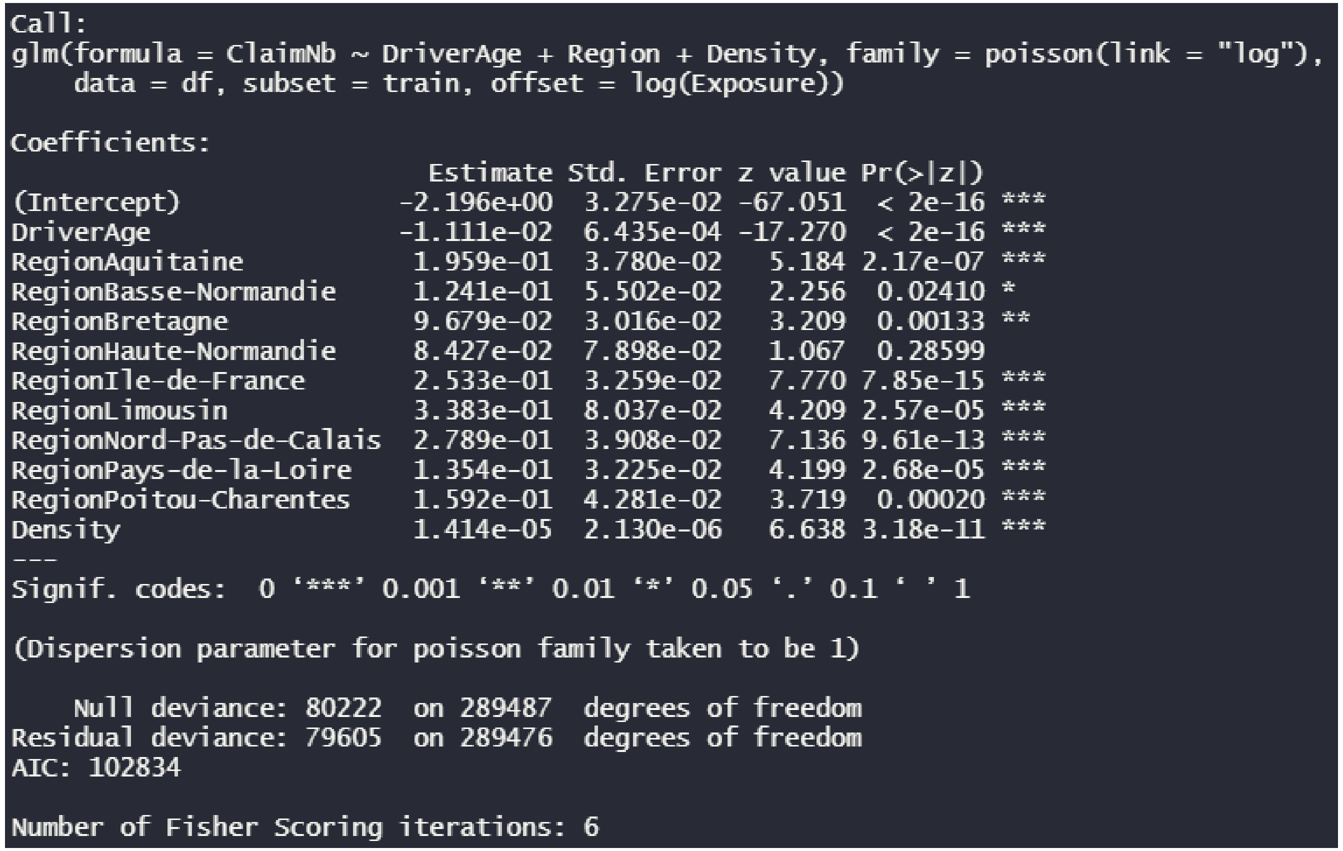
Reminder – Data set



```
Rows: 413,960
Columns: 13
$ PolicyID    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33~
$ ClaimNb     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1~
$ Exposure    <dbl> 0.09, 0.84, 0.52, 0.45, 0.15, 0.75, 0.81, 0.05, 0.76, 0.34, 0.10, 0.77, 0.55, 0.19, 0.01, 0.87, 0.80, 0.07, 0.12, 0.76, 0~
$ Power       <fct> g, g, f, f, g, g, d, d, d, i, f, f, e, e, e, e, e, e, i, i, h, h, j, j, e, i, i, e, e, f, f, f, g, f, f, j, e, e, e, l, l~
$ CarAge      <int> 0, 0, 2, 2, 0, 0, 1, 0, 9, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 8, 8, 0, 4, 4, 0, 0, 0, 0, 0, 1, 0, 0, 0, 8, 8, 8, 5, 5~
$ DriverAge   <int> 46, 46, 38, 38, 41, 41, 27, 27, 23, 44, 32, 32, 33, 33, 33, 54, 69, 69, 43, 43, 50, 50, 30, 30, 73, 40, 40, 45, 45, 37, 3~
$ Brand       <fct> "Japanese (except Nissan) or Korean", "Japanese (except Nissan) or Korean", "Japanese (except Nissan) or Korean", "Japane~
$ Gas         <fct> Diesel, Diesel, Regular, Regular, Diesel, Diesel, Regular, Regular, Regular, Regular, Diesel, Diesel, Regular, Regular, R~
$ Region      <fct> Aquitaine, Aquitaine, Nord-Pas-de-Calais, Nord-Pas-de-Calais, Pays-de-la-Loire, Pays-de-la-Loire, Aquitaine, Aquitaine, N~
$ Density     <int> 76, 76, 3003, 3003, 60, 60, 695, 695, 7887, 27000, 23, 23, 1746, 1746, 1746, 781, 1376, 1376, 7752, 7752, 3545, 3545, 366~
$ ClaimAmount <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ train       <lgl> TRUE, TRUE, TRUE, FALSE, FALSE, TRUE, TRUE, TRUE, TRUE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, TRUE, TRUE, FAL~
$ test        <lgl> FALSE, FALSE, FALSE, TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, FALSE,~
```

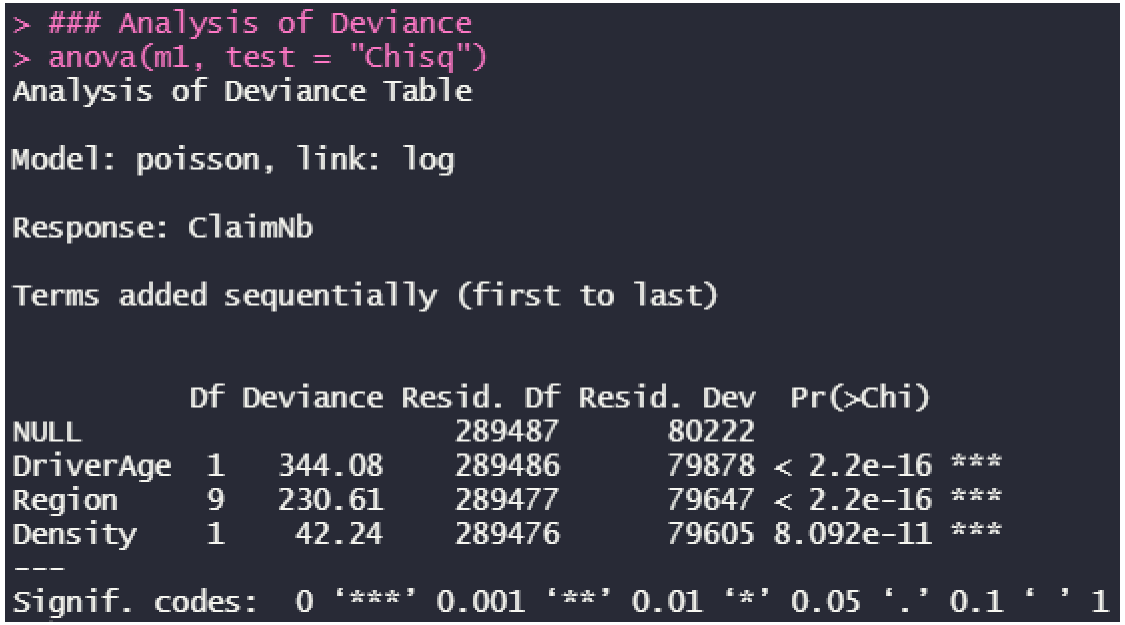
10 variables available.

First model



```
Call:
glm(formula = ClaimNb ~ DriverAge + Region + Density, family = poisson(link = "log"),
    data = df, subset = train, offset = log(Exposure))

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -2.196e+00  3.275e-02 -67.051  < 2e-16 ***
DriverAge                -1.111e-02  6.435e-04 -17.270  < 2e-16 ***
RegionAquitaine           1.959e-01  3.780e-02   5.184 2.17e-07 ***
RegionBasse-Normandie     1.241e-01  5.502e-02   2.256  0.02410 *  
RegionBretagne            9.679e-02  3.016e-02   3.209  0.00133 ** 
RegionHaute-Normandie     8.427e-02  7.898e-02   1.067  0.28599    
RegionIle-de-France       2.533e-01  3.259e-02   7.770 7.85e-15 ***
RegionLimousin            3.383e-01  8.037e-02   4.209 2.57e-05 ***
RegionNord-Pas-de-Calais  2.789e-01  3.908e-02   7.136 9.61e-13 ***
RegionPays-de-la-Loire    1.354e-01  3.225e-02   4.199 2.68e-05 ***
RegionPoitou-Charentes    1.592e-01  4.281e-02   3.719  0.00020 ***
Density                   1.414e-05  2.130e-06   6.638 3.18e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 80222  on 289487  degrees of freedom
Residual deviance: 79605  on 289476  degrees of freedom
AIC: 102834

Number of Fisher Scoring iterations: 6
```

Analysis of Deviance



```
> ### Analysis of Deviance
> anova(m1, test = "Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: ClaimNb

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     289487      80222
DriverAge  1   344.08    289486      79878 < 2.2e-16 ***
Region     9   230.61    289477      79647 < 2.2e-16 ***
Density    1    42.24    289476      79605 8.092e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

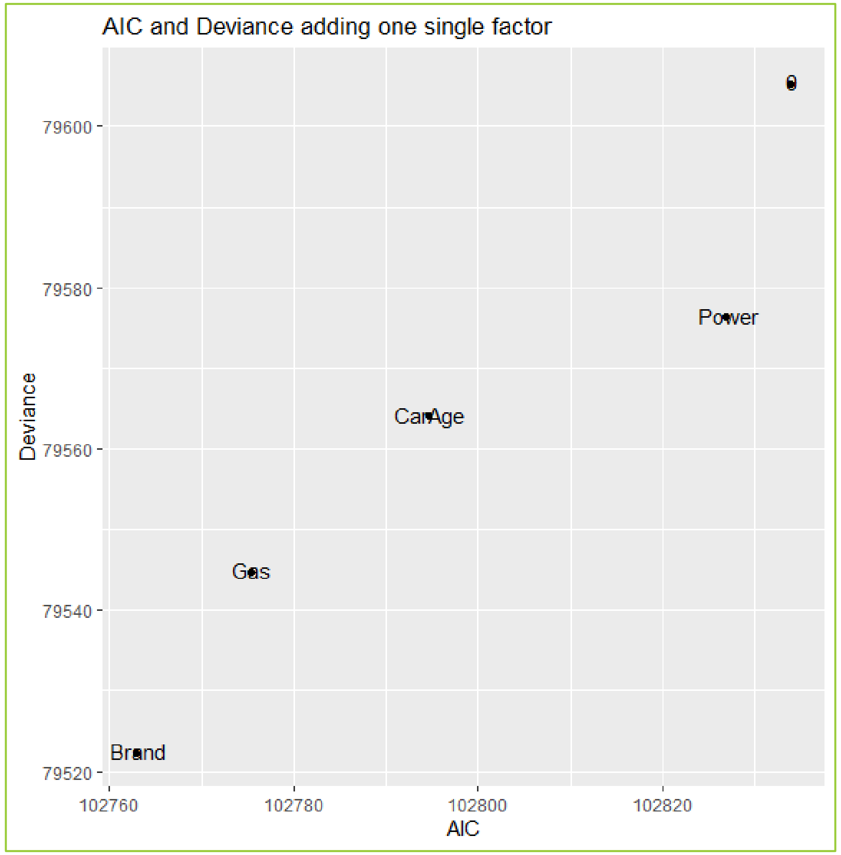
The table show how the Deviance changes as we add term to the model.

Second column: Degree of Freedom: the contribution t the number of parameters estimated by adding the variable.

Third column: the Deviance for the variable being added.

Here we had 3 variables. If we add the variable Region, we estimate 9 additional parameters, and the residual Deviance would drop by 230.61, which can be considered as a good investment. Moreover, the last column shows that it is significant at the 1% level.

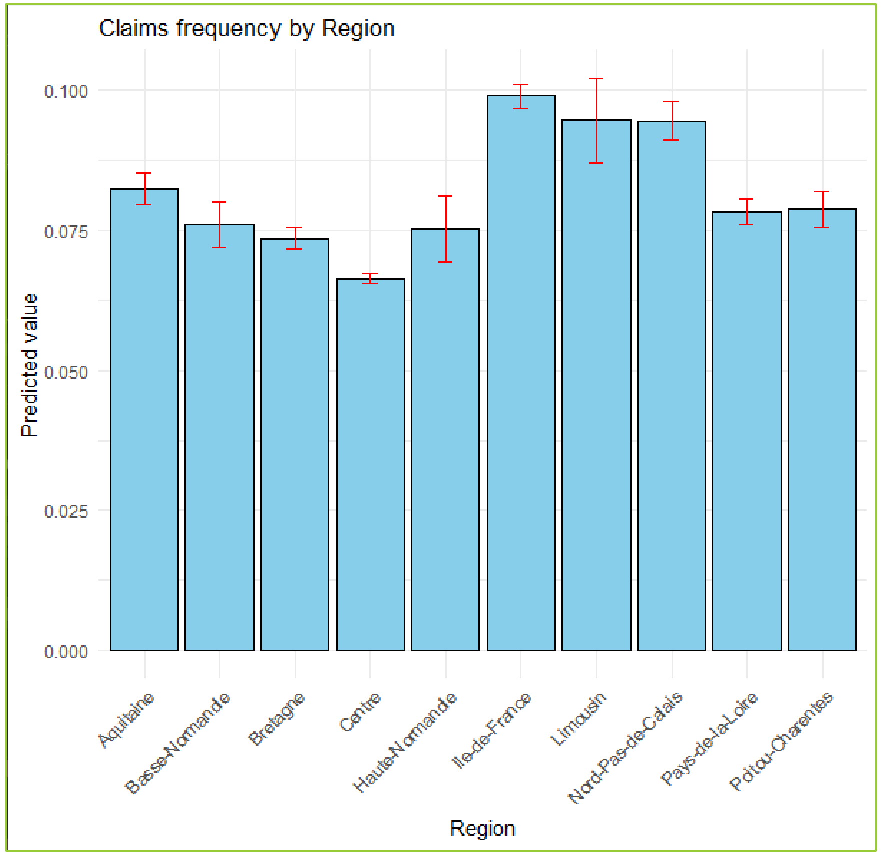
GLM - AIC and Deviance representation



After having set a simple model for Claims frequency with only 3 predictors, we want to understand the potential gain of adding the other predictors to this first model.
To do that, we successively add a factor to the equation and compute the AIC and Deviance for the new model with 4 variables.

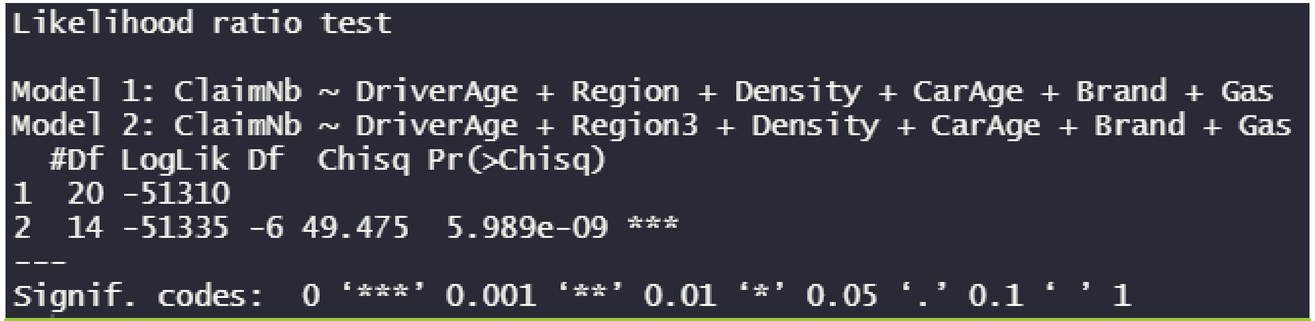
“Brand” has the biggest impact in decreasing both AIC and the Deviance. However, “Power” appears behind.

Binning categorical



As we saw in part 1 that some Region can be binned together.

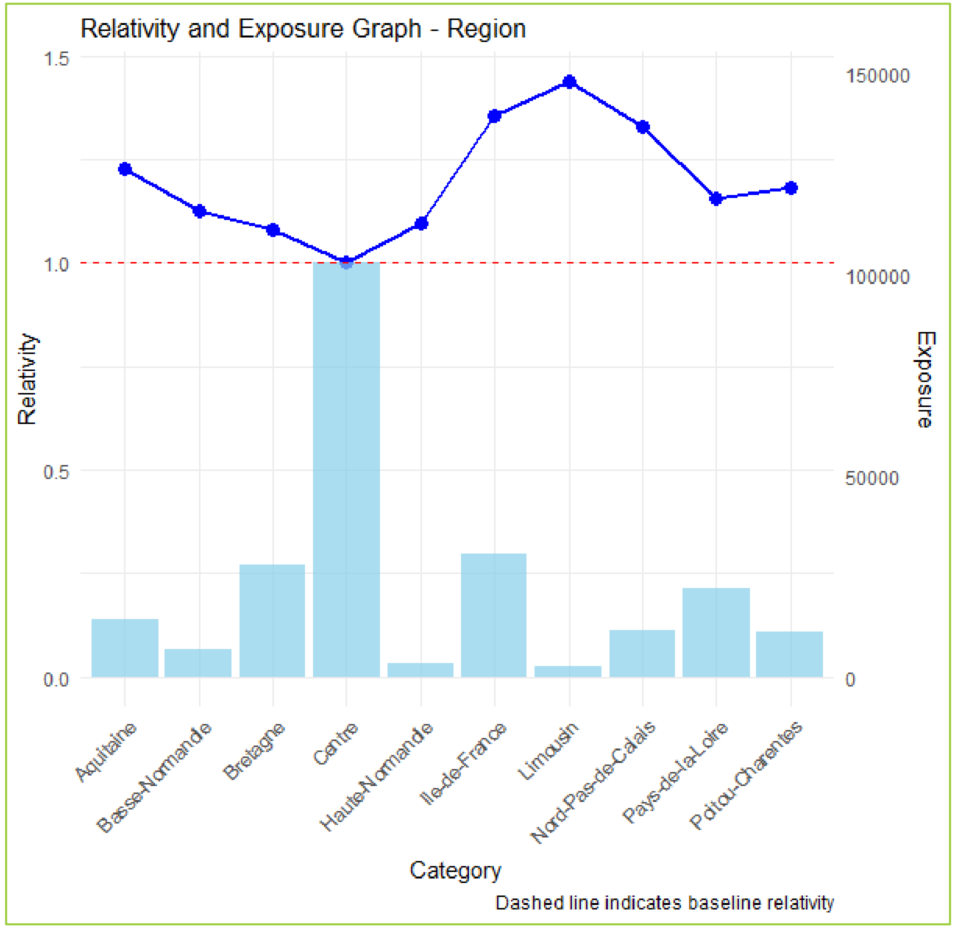
```
df <- df %>%
  mutate(Region3 = case_when(
    Region %in% c("Basse-Normandie", "Haute-Normandie", "Bretagne", "Centre","Aquitaine","Pays-de-la-Loire", "Poitou-Charentes") ~ "Group_Ouest",
    TRUE ~ Region   # Keep other levels unchanged
  ))
df %>% group_by(Region3) %>% summarise(count=n())

# New model with new Region
m3tris <- glm(ClaimNb ~ DriverAge + Region3 + Density + CarAge + Brand + Gas,
            data = df,
            subset = train,
            family = poisson(link = "log"),
            offset = log(Exposure))
summary(m3tris)
```



```
Likelihood ratio test

Model 1: ClaimNb ~ DriverAge + Region + Density + CarAge + Brand + Gas
Model 2: ClaimNb ~ DriverAge + Region3 + Density + CarAge + Brand + Gas
  #Df LogLik Df  Chisq Pr(>Chisq)
1  20 -51310
2  14 -51335 -6 49.475  5.989e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

But, according to the LRT test in R, keeping the original model makes sense. Here we reject H0. The first model provides a significant better fit than the 2nd.

Relativities – Example on Regions

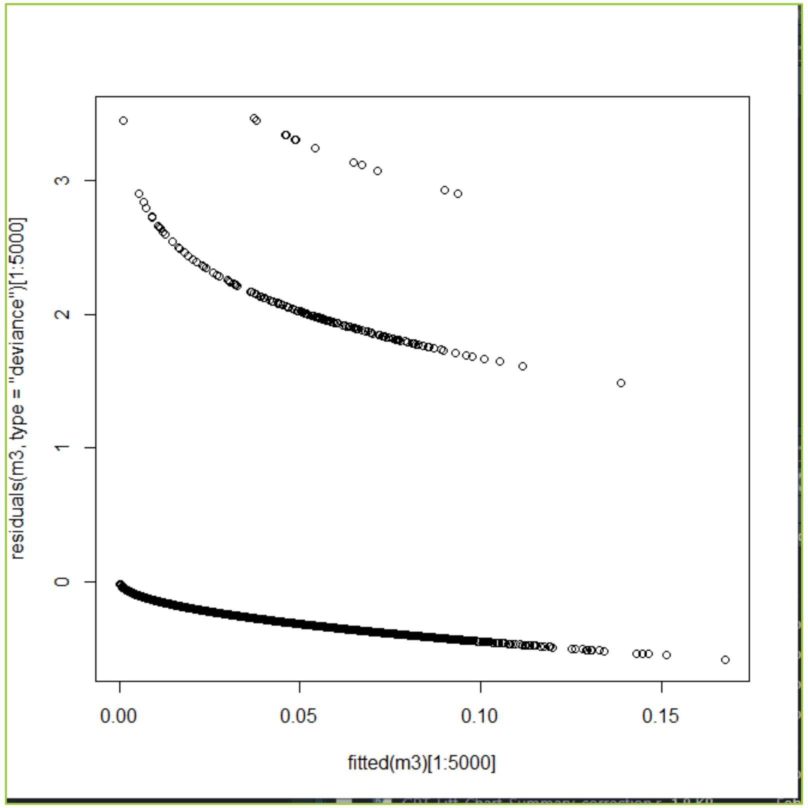


We can compare the different coefficients of the model, here we represent the variable Region where the “Centre” has been chosen as the base level.

The blue line shows the exponential of the fitted parameter estimates.

We observe that Aquitaine’s is 1.23: This means the model estimates that - all other factor being constant - the exposure from the region Aquitaine will have a relativity of 23% times that expected for exposures at the region Centre.

Deviance Residuals



Residuals are used to check the appropriateness of a chosen response distribution, and for outlying values. The residuals -ˆi = yi − yˆi, and their studentized version are central to model checking for the normal linear model. Let’s check the look of this graph:

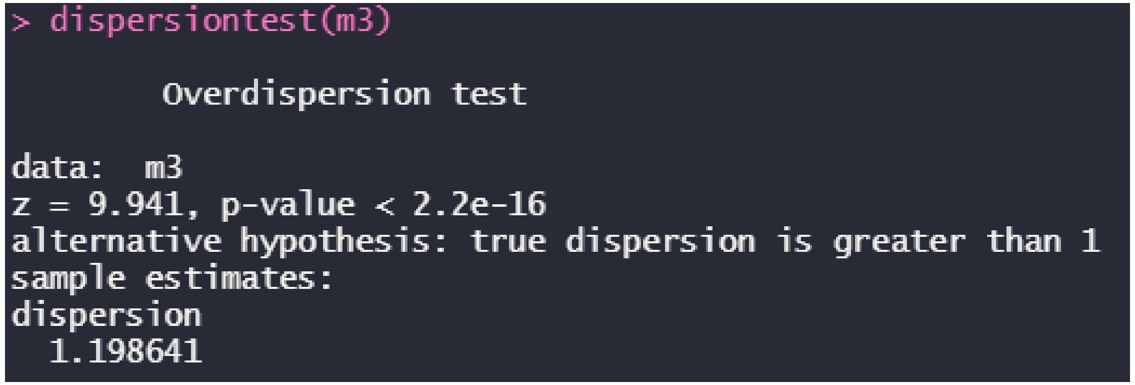
Centered Residuals: The crunched residuals seem to be mostly centered around 0, which is a good sign. It means the model is generally unbiased in its predictions.

Systematic Patterns: There doesn't appear to be a clear pattern in the residuals relative to the fitted values, which suggests that the model captures the main structure in the data.

Spread of Residuals: The spread of residuals appears to be relatively constant, which is important for assuming homoscedasticity (constant variance). However, there are a few negative outliers (points with residuals around -0.2) that might warrant further investigation. These could represent extreme cases or points where the model struggles to fit the data accurately.

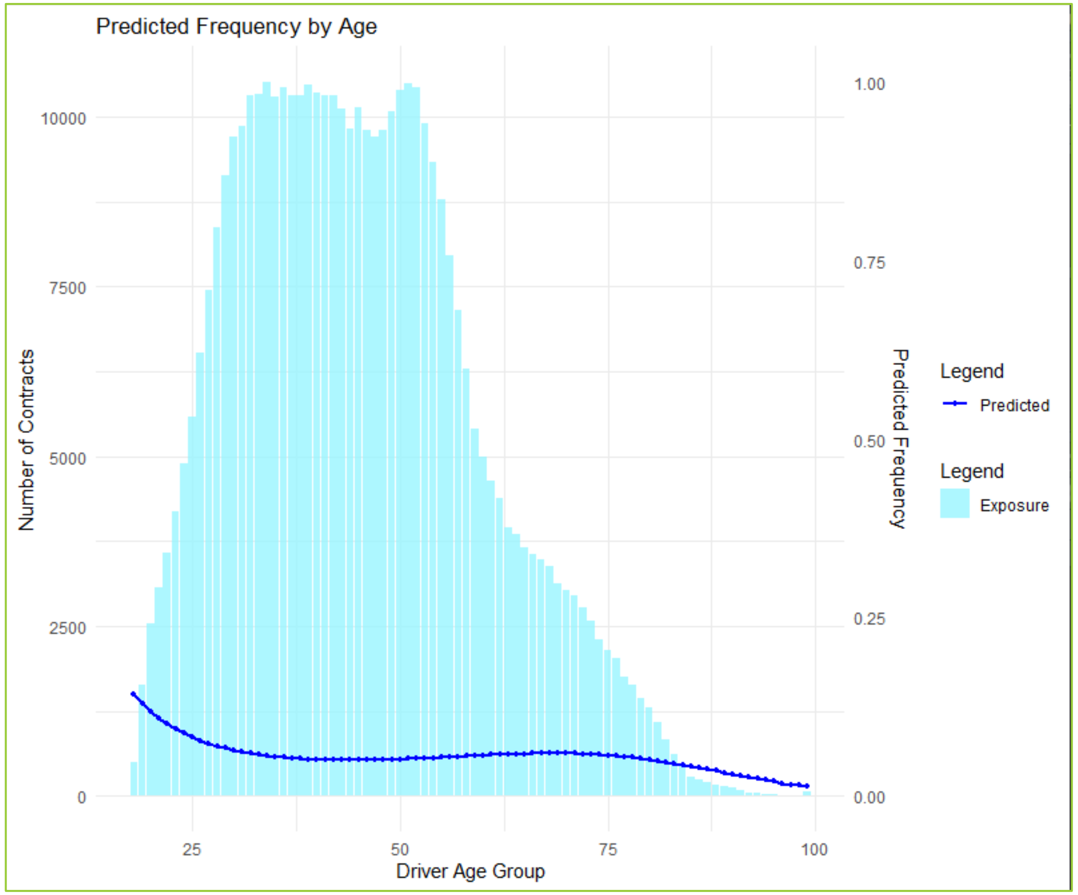
Horizontal Bands: The horizontal bands in the residuals reflect the discrete nature of the claims frequency data (integer counts). This is expected in a Poisson regression model.

Overdispersion?



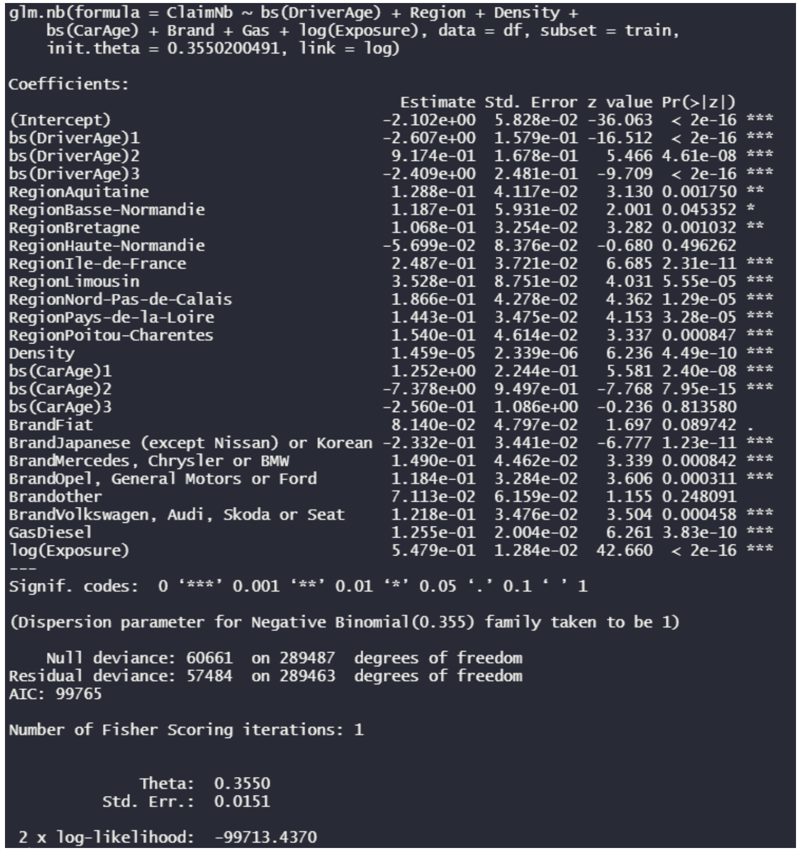
```
> dispersiontest(m3)

        Overdispersion test

data:  m3
z = 9.941, p-value < 2.2e-16
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  1.198641
```

Predicted claims frequency by Driver age

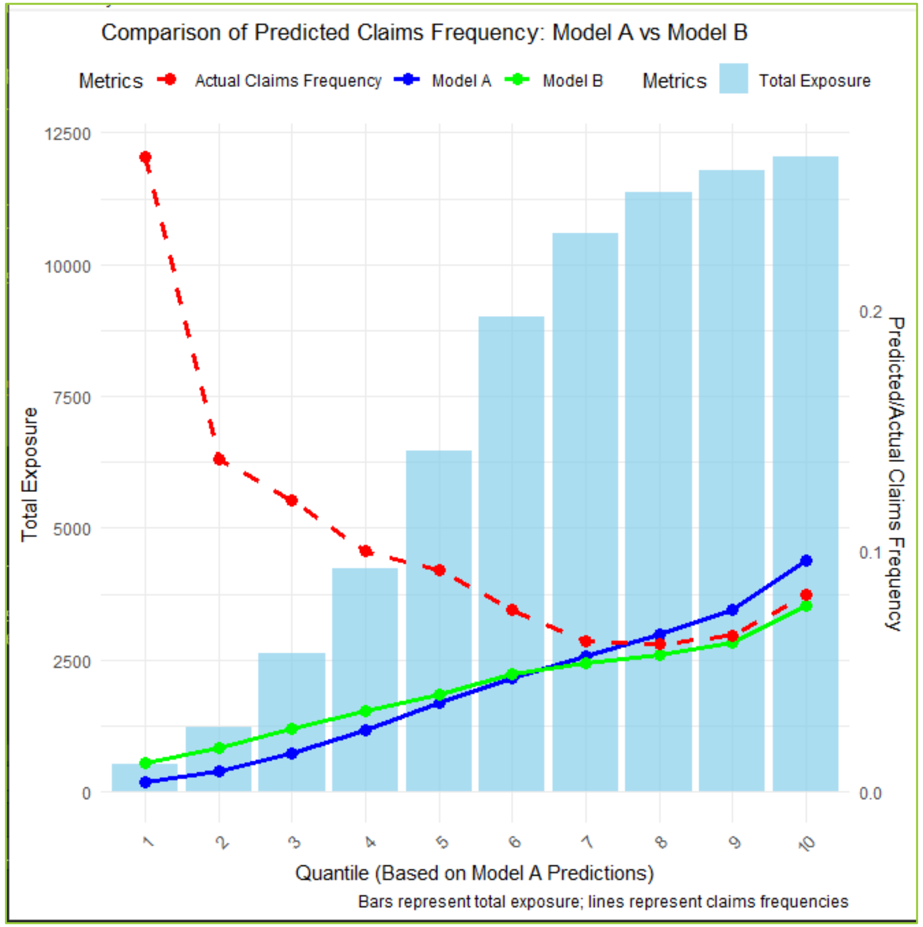


A representation of the claims frequency for each Driver age.

Lift Chart

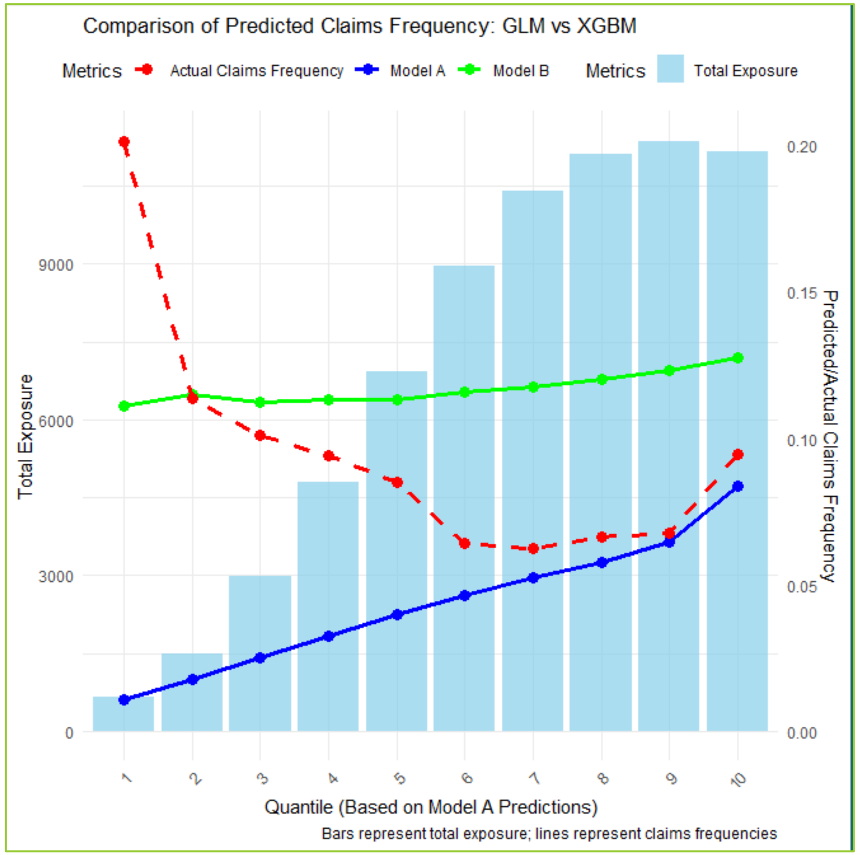


```
glm.nb(formula = ClaimNb ~ bs(DriverAge) + Region + Density +
    bs(CarAge) + Brand + Gas + log(Exposure), data = df, subset = train,
    init.theta = 0.3550200491, link = log)

Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -2.102e+00  5.828e-02 -36.063  < 2e-16 ***
bs(DriverAge)1                       -2.607e+00  1.579e-01 -16.512  < 2e-16 ***
bs(DriverAge)2                        9.174e-01  1.678e-01   5.466 4.61e-08 ***
bs(DriverAge)3                       -2.409e+00  2.481e-01  -9.709  < 2e-16 ***
RegionAquitaine                       1.288e-01  4.117e-02   3.130 0.001750 **
RegionBasse-Normandie                 1.187e-01  5.931e-02   2.001 0.045352 *
RegionBretagne                        1.068e-01  3.254e-02   3.282 0.001032 **
RegionHaute-Normandie                -5.699e-02  8.376e-02  -0.680 0.496262
RegionIle-de-France                   2.487e-01  3.721e-02   6.685 2.31e-11 ***
RegionLimousin                        3.528e-01  8.751e-02   4.031 5.55e-05 ***
RegionNord-Pas-de-Calais              1.866e-01  4.278e-02   4.362 1.29e-05 ***
RegionPays-de-la-Loire                1.443e-01  3.475e-02   4.153 3.28e-05 ***
RegionPoitou-Charentes                1.540e-01  4.614e-02   3.337 0.000847 ***
Density                               1.459e-05  2.339e-06   6.236 4.49e-10 ***
bs(CarAge)1                           1.252e+00  2.244e-01   5.581 2.40e-08 ***
bs(CarAge)2                          -7.378e+00  9.497e-01  -7.768 7.95e-15 ***
bs(CarAge)3                          -2.560e-01  1.086e+00  -0.236 0.813580
BrandFiat                             8.140e-02  4.797e-02   1.697 0.089742 .
BrandJapanese (except Nissan) or Korean -2.332e-01  3.441e-02  -6.777 1.23e-11 ***
BrandMercedes, Chrysler or BMW        1.490e-01  4.462e-02   3.339 0.000842 ***
BrandOpel, General Motors or Ford     1.184e-01  3.284e-02   3.606 0.000311 ***
Brandother                            7.113e-02  6.159e-02   1.155 0.248091
BrandVolkswagen, Audi, Skoda or Seat  1.218e-01  3.476e-02   3.504 0.000458 ***
GasDiesel                             1.255e-01  2.004e-02   6.261 3.83e-10 ***
log(Exposure)                         5.479e-01  1.284e-02  42.660  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.355) family taken to be 1)

    Null deviance: 60661  on 289487  degrees of freedom
Residual deviance: 57484  on 289463  degrees of freedom
AIC: 99765

Number of Fisher Scoring iterations: 1


              Theta:  0.3550
          Std. Err.:  0.0151

 2 x log-likelihood:  -99713.4370
```



We can compare two models by using a double Lift plot. Here, the model B (Negative Binomial) is closer to the observation than the model A (Poisson). We will keep the NB model for the exercise.

Lift Chart: GLM vs XGBM



Bibliography

Generalized Linear Models for Insurance Data (P. de Jong, G. Heller)

A Practitioner's Guide to Generalized Linear Models (D. Anderson and al.)

Predictive Modelling Application in Actuarial Science (E. Frees and Al.)