MODÉLISATION DE LA FRÉQUENCE DES SINISTRES EN ASSURANCE AUTOMOBILE

avec R

MAARAF Oumayma

Base de donnée

La base de données utilisée dans ce document, est "freMTPL2freq". Elle contient des informations sur les contrats et les clients d'une compagnie d'assurance française, reliées à un portefeuille d'assurance automobile. Elle est extraite de la libraire "CASdatasets" de R.

```
> install.packages("CASdatasets", repos = "http://cas.uqam.ca/pub/", type="source")
> library(CASdatasets)
> data("freMTPL2freq")
> Contrats<-freMTPL2freq
> str(Contrats)
'data.frame': 678013 obs. of 12 variables:
$ IDpol : num 1 3 5 10 11 13 15 17 18 21 ...
$ ClaimNb : 'table' num [1:678013(1d)] 1 1 1 1 1 1 1 1 1 1 ...
$ Exposure : num 0.1 0.77 0.75 0.09 0.84 0.52 0.45 0.27 0.71 0.15 ...
$ VehPower : int 5 5 6 7 7 6 6 7 7 7 ...
$ VehAge : int 0 0 2 0 0 2 2 0 0 0 ...
$ DrivAge : int 55 55 52 46 46 38 38 33 33 41 ...
$ BonusMalus: int 50 50 50 50 50 50 50 68 68 50 ...
$ VehBrand : Factor w/ 11 levels "B1","B10","B11",..: 4 4 4 4 4 4 4 4 4 4 ...
$ VehGas : chr "Regular" "Regular" "Diesel" "Diesel" ...
$ Area : Factor w/ 6 levels "A","B","C","D",..: 4 4 2 2 2 5 5 3 3 2 ...
$ Density : int 1217 1217 54 76 76 3003 3003 137 137 60 ...
$ Region : Factor w/ 21 levels "Alsace","Aquitaine",..: 21 21 18 2 2 16 16 13 13 17
..
```

On a 678013 observations et 12 variables. La significations de ces variables est la suivante :
- IDpol : numéro du contrat
- ClaimNb : nombre de sinistres déclarés durant la durée d'exposition
- Exposure : durée d'exposition en années
- VehPower : puissance de la voiture (ordre catégoriel)
- VehAge : age de la voiture en année
- DrivAge : age du conducteur en année (commence de 18 ans)
- BonusMalus
- VehBrand : marque de la voiture (par catégorie)
- VehGas : diesel ou régulier
- Area
- Density : nombre d'habitants pour un kilomètre carrée
- Region : région où habite la conducteur

La variable dépendante ClaimNb du nombre de sinistres est composée majoritairement par des zéros (94%)

```
> Contrats$ClaimNb<-as.integer(Contrats$ClaimNb)
> DataClaimNb<-data.frame(table(Contrats$ClaimNb))
> ZeroProportion<-DataClaimNb[1,2]/sum(DataClaimNb[1:11,2])
> ZeroProportion
  0.949765
```

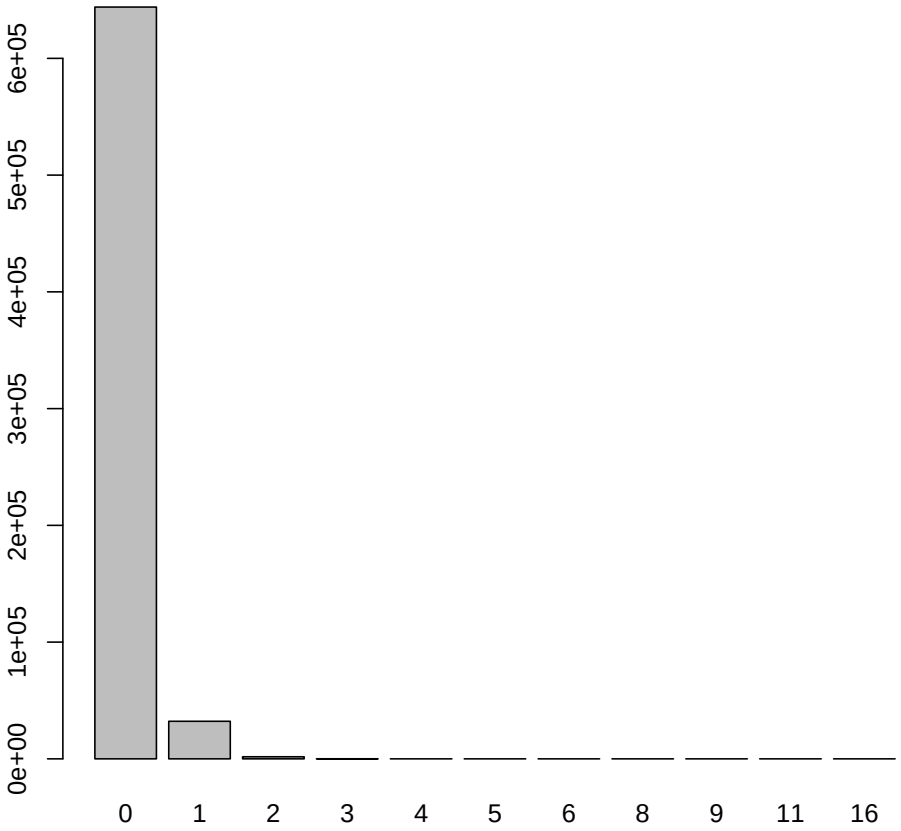

FIGURE 1 – Distribution de la variable dépendante "ClaimNb"

On va travailler avec des variables qualitatives,
- AgeF : age du conducteur, [18,31[, [31,44[, [44,57[, [57,70[, [70,83[, [83,96[, [96,+∞[.
- VehageF : age de la voiture, [0,4[, [4,8[, [8,12[, [12,16[, [16,+∞[.
- DensityF : densité, [0,40[, [40,200[, [200,500[, [500,4000[, [4000,+∞[.
- VehpowerF : puissance de la voiture,[4,7[, [7,9[, [9,15[.
- VehbrandF :marque de la voiture, contients 4 modalités, dont B1, B2, B12 et une nouvelle modalité B qui regroupe les modalités suivantes B10, B11, B13, B14, B4, B5, B6, B3.

```
> breakage=c(seq(18,100,13),Inf)
> Contrats$AgeF<-cut(Contrats$DrivAge,breaks=breakage,right=FALSE)
> breakagevoiture=c(seq(0,16,4),Inf)
> Contrats$VehageF<-cut(Contrats$VehAge,breaks=breakagevoiture,right=FALSE)
> breakdensity=c(0,40,200,500,4000,Inf)
> Contrats$DensityF<-cut(Contrats$Density, breaks=breakdensity,right=FALSE)
> breakpower=c(4,7,9,15)
```

```
> Contrats$VehpowerF<-cut(Contrats$VehPower,breaks=breakpower,right=FALSE)
> library(forcats)
> Contrats$VehbrandF<-fct_recode(Contrats$VehBrand,A="B10",A="B11",A="B13",A="B14",
A="B4",A="B5",A="B6",A="B3")
```

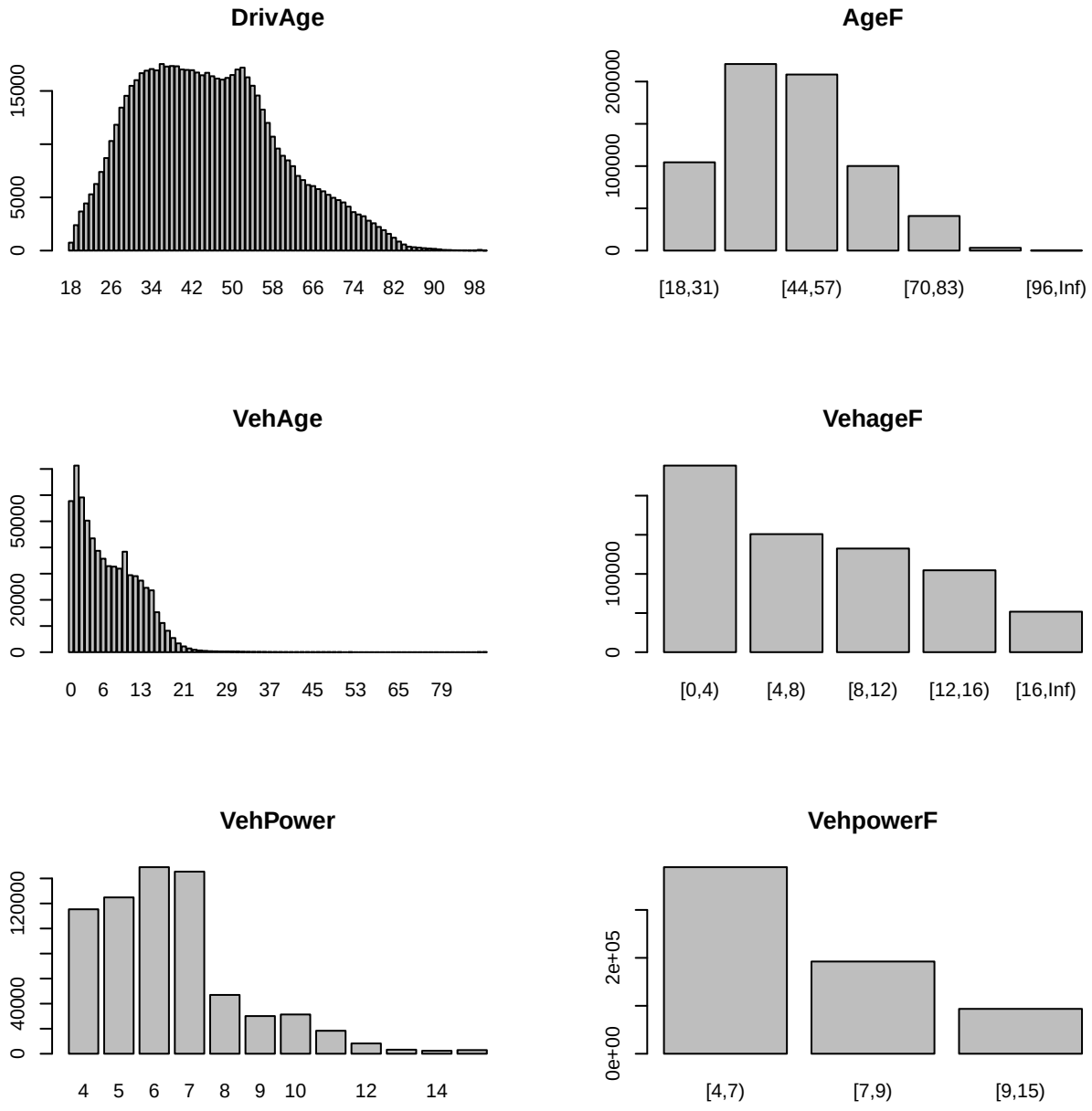

FIGURE 2 – Distribution de l'age du conducteur, de l'age de la voiture et de la puisance de la voiture avant (gauche) et après (droite) la catégorisation

Modèle Poisson

Soit N_i le nombre de sinistre **annuel** pour l'assuré i. On suppose que $N_i \sim P(\lambda_i)$. En général N_i est inobservable, seulement Y_i le nombre de sinistre pendant une durée E_i est observable. En fait, la supposition $N_i \sim P(\lambda_i)$ peut être écrite de manière équivalente comme $Y_i \sim P(E_i.\lambda_i)$. En utilisant, la fonction de lien logarithme, on aura $\lambda_i = e^{x_i'\beta}$, où x_i ensemble de variable mesurées pour chaque assuré i. Donc

$$Y_i \sim P(e^{x_i'\beta + log(E_i)})$$

Pour estimer les paramètres de ce modèle on utilise la méthode du maximum vraisemblance. Le log-vraisemblance est :

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} [y_i log(\lambda_i E_i) - (\lambda_i E_i) - log(y_i!)]$$

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} [y_i(x_i'\beta + log(E_i)) - exp(x_i'\beta + log(E_i))log(y_i!)]$$

Et

$$\frac{\partial \mathcal{L}(\beta)}{\partial \beta} = \sum_{i=1}^{n} (y_i - exp(x_i'\beta + log(E_i))x_i'$$

La maximisation de cette vraisemblance se fait numériquement, en utilisant l'algorithme de Newton-Raphson.
Pour estimer les paramètres de ce modèle avec R, on utilise :

```
> regp <- glm(ClaimNb ~ AgeF+VehageF+DensityF+VehpowerF+VehGas+VehbrandF+Area
+Region+offset(log(Exposure)),data=Contrats,family=poisson(link="log"))
> summary(regp)
```

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-2.1714	0.1013	-21.43	0.0000
AgeF[31,44)	-0.3614	0.0184	-19.61	0.0000
AgeF[44,57)	-0.2384	0.0179	-13.30	0.0000
AgeF[57,70)	-0.3745	0.0213	-17.59	0.0000
AgeF[70,83)	-0.3125	0.0265	-11.80	0.0000
AgeF[83,96)	-0.1421	0.0703	-2.02	0.0432
AgeF[96,Inf)	-0.0426	0.3539	-0.12	0.9042
VehageF[4,8)	-0.1597	0.0164	-9.74	0.0000
VehageF[8,12)	-0.1651	0.0177	-9.33	0.0000
VehageF[12,16)	-0.3272	0.0203	-16.13	0.0000
VehageF[16,Inf)	-0.5578	0.0286	-19.49	0.0000
DensityF[40,200)	0.0966	0.0368	2.63	0.0086
DensityF[200,500)	0.1543	0.0434	3.56	0.0004
DensityF[500,4e+03)	-7.2463	38.1692	-0.19	0.8494
DensityF[4e+03,Inf)	-7.2279	38.1692	-0.19	0.8498
VehpowerF[7,9)	-0.0738	0.0139	-5.31	0.0000
VehpowerF[9,15)	0.0630	0.0179	3.52	0.0004
VehGasRegular	0.0448	0.0122	3.66	0.0003

TABLE 1 – Estimation des paramètres du modèle regp

Plusieurs coefficients ne sont pas significatifs. La non significativité, peut renvoyer à ce que certains coefficients de modalités (de la même variable) peuvent est significativement égaux, ou à ce que certains variables n'expliquent pas la variable dépendante.
On commence donc par la sélection de variables, (Table 2)

```
> library(MASS)
> selection<-stepAIC(regp,direction="backward",k=log(nrow(Contrats)))
> summary(selection)
```

la sélection faite grâce à stepAIC, ne garde que les variables suivantes, AgeF, VehageF, DensityF, VehpowerF, VehGas et VehbrandF. On remarque que quelques coefficients ne sont pas significatifs, on fait le test d'égalité de coefficients de AgeF[83,96[et AgeF[96,+∞[, et de VehbradFA et VehbrandB2.

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-2.1109	0.0279	-75.60	0.0000
AgeF[31,44)	-0.3611	0.0184	-19.62	0.0000
AgeF[44,57)	-0.2367	0.0179	-13.22	0.0000
AgeF[57,70)	-0.3724	0.0212	-17.54	0.0000
AgeF[70,83)	-0.3133	0.0264	-11.87	0.0000
AgeF[83,96)	-0.1387	0.0703	-1.97	0.0483
AgeF[96,Inf)	-0.0325	0.3539	-0.09	0.9269
VehageF[4,8)	-0.1573	0.0164	-9.60	0.0000
VehageF[8,12)	-0.1636	0.0177	-9.26	0.0000
VehageF[12,16)	-0.3248	0.0202	-16.06	0.0000
VehageF[16,Inf)	-0.5561	0.0286	-19.47	0.0000
DensityF[40,200)	0.0926	0.0216	4.29	0.0000
DensityF[200,500)	0.1545	0.0234	6.61	0.0000
DensityF[500,4e+03)	0.2738	0.0209	13.12	0.0000
DensityF[4e+03,Inf)	0.3573	0.0251	14.25	0.0000
VehpowerF[7,9)	-0.0717	0.0139	-5.16	0.0000
VehpowerF[9,15)	0.0686	0.0179	3.84	0.0001
VehGasRegular	0.0481	0.0122	3.94	0.0001
VehbrandFA	0.0085	0.0166	0.51	0.6087
VehbrandFB12	0.2093	0.0190	11.02	0.0000
VehbrandFB2	-0.0051	0.0170	-0.30	0.7648

TABLE 2 – Estimation des paramètres du modèle qui contient les variables sélectionnées

```
> library(car)
> linearHypothesis(regression,'AgeF[83,96)=AgeF[96,Inf)')
> linearHypothesis(regression,'VehbrandFA=VehbrandFB2 ')
```

	Res.Df	Df	Chisq	Pr(>Chisq)
1	540060			
2	540059	1	0.0871	0.7679

TABLE 3 – Résultat du test d'égalité de coefficients de AgeF[83,96) et AgeF[96,Inf)

	Res.Df	Df	Chisq	Pr(>Chisq)
1	540060			
2	540059	1	0.6644	0.415

TABLE 4 – Résultat du test d'égalité de coefficients de VehbrandFA et VehbrandFB2

On recode les variables concernés de manière à ce que ces modalités soient groupées.

```
> breakage1=c(seq(18,83,13),Inf)
> Contrats$AgeF1<-cut(Contrats$DrivAge,breaks=breakage1,right=FALSE)
> Contrats$VehbrandF1<-fct_recode(Contrats$VehbrandF,B="A",B="B2")
```

Tous les coefficients sont significatifs - pour le modèle qui contient AgeF1 et VehbrandF1 à la place de AgeF et VehbrandF- sauf pour VehbrandF1B. Donc pour la variable de la marque de la voiture, il vaudrait mieux la coder en seulement deux modalités, celle de B1 (modalité de référence) et une autre Other (qui regroupe toutes les autres marques).

```
> Contrats$VehbrandF2<-fct_recode(Contrats$VehbrandF1,other="B",other="B12")
```

Les résultats de l'estimation des paramètres du nouveau modèle (regp2) sont les suivants

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-2.0747	0.0277	-74.95	0.0000
AgeF1[31,44)	-0.3547	0.0184	-19.28	0.0000
AgeF1[44,57)	-0.2298	0.0179	-12.85	0.0000
AgeF1[57,70)	-0.3636	0.0212	-17.15	0.0000
AgeF1[70,83)	-0.3130	0.0264	-11.87	0.0000
AgeF1[83,Inf)	-0.1493	0.0690	-2.16	0.0304
VehageF[4,8)	-0.2168	0.0156	-13.88	0.0000
VehageF[8,12)	-0.2419	0.0164	-14.73	0.0000
VehageF[12,16)	-0.4090	0.0189	-21.59	0.0000
VehageF[16,Inf)	-0.6416	0.0276	-23.22	0.0000
DensityF[40,200)	0.0923	0.0216	4.27	0.0000
DensityF[200,500)	0.1542	0.0234	6.60	0.0000
DensityF[500,4e+03)	0.2813	0.0208	13.51	0.0000
DensityF[4e+03,Inf)	0.3768	0.0250	15.08	0.0000
VehpowerF[7,9)	-0.0628	0.0139	-4.53	0.0000
VehpowerF[9,15)	0.0969	0.0176	5.51	0.0000
VehGasRegular	0.0607	0.0122	4.99	0.0000
VehbrandF2other	0.0474	0.0141	3.37	0.0008

TABLE 5 – Estimation des paramètres du modèle regp2

Pour s'assurer que ce modèle s'ajuste bien aux données, on utilise le test statistique basé sur la déviance. La déviance D est égale à $2(\mathcal{L}_{saturé} - \mathcal{L})$, et asymptotiquement suit une $\mathcal{X}^2_{n-p-1}$, où n le nombre d'observations et p le nombre de paramètres considérés.(on désigne par $\mathcal{L}$ le log-vraisemblance du modèle considéré, et $\mathcal{L}_{saturé}$ le log-vraisemblance du modèle saturé)

```
> pchisq(deviance(regp2),regp2$df.residual,lower.tail = FALSE)
1
```

Le modèle est donc adéquat. Pourtant, il ne faut pas oublier de tester l'équi-dispersion, car si il y a une dispersion, on n'est conduit à une sur ou sous-estimation des écarts-type des coefficients.
Le test de dispersion utilisé en R a pour hypothèse nulle, et hypothèse alternative (où $\mu = \mathsf{E}[y]$, et $\text{trafo}(x) = x^{trafo}$) :

$$\begin{cases} H_0 : \ \mathsf{VAR}[y] = \mu \\ \\ H_1 : \ \mathsf{VAR}[y] \quad = \quad \mu \ + \ \alpha \cdot \text{trafo}(\mu). \end{cases}$$

```
> library(AER)
> dispersiontest(regp2)
```

Le test indique qu'il n'y a pas d'équi-dispersion. Si on refait le test, avec la condition trafo = 1, l'hypothèse H_0 est rejetée, donc la variance peut s'écrire de la manière suivante $\mathsf{VAR}[y] = \mu + \alpha \cdot \mu$, qui correspond à la variance de NB1, ou bien d'un quasi-poisson. En tous cas, on est sûr qu'on a une sur-dispersion ($\alpha > 0$), car si on avait une sous dispersion, les modèles binomiales négatives ne peuvent pas être utilisés.

Modèle Binomiale Négative

On suppose que $N_i|_{x_i} \sim \mathcal{BN}(\mu_i, \alpha)$, et donc que

$$P(N_i = y_i|x_i) = \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i + 1)\Gamma(1/\alpha)} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(1 - \frac{1}{1+\alpha\mu_i}\right)^{y_i}$$

Le log-vraisemblance est donné par

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} y_i ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \frac{1}{\alpha} ln(1+\alpha\mu_i) + ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - ln\Gamma\left(y_i + 1\right) - ln\Gamma\left(\frac{1}{\alpha}\right)$$

En supposant que $\mu_i = e^{(x_i'\beta)}$ (fonction de lien logarithme), on a

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} y_i ln\left(\frac{\alpha exp(x_i'\beta)}{1+\alpha exp(x_i'\beta)}\right) - \frac{1}{\alpha} ln(1+\alpha exp(x_i'\beta)) + ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - ln\Gamma\left(y_i + 1\right) - ln\Gamma\left(\frac{1}{\alpha}\right)$$

Et on a

$$\frac{\partial\mathcal{L}(\beta)}{\partial\beta} = \sum_{i=1}^{n} \frac{x_i(y_i + \mu_i)}{1+\alpha\mu_i}$$

La résolution se fait numériquement, sachant qu'on travaille avec les Y_i et non pas les N_i(comme déjà dit dans la section du modèle de Poisson), c'est-à-dire $E(Y_i|_{x_i}) = E_i.\mu_i$. On suit alors la même démarche pour la sélection des variables, et on se retrouve avec les mêmes variables déjà sélectionnées pour le modèle final de Poisson.

```
> regnb<-glm.nb(ClaimNb ~ AgeF1+VehageF+DensityF+VehpowerF+VehGas+VehbrandF2+
offset(log(Exposure)),data=Contrats)
> summary(regnb)
```

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-2.0484	0.0288	-71.01	0.0000
AgeF1[31,44)	-0.3641	0.0192	-18.97	0.0000
AgeF1[44,57)	-0.2372	0.0187	-12.68	0.0000
AgeF1[57,70)	-0.3768	0.0221	-17.02	0.0000
AgeF1[70,83)	-0.3267	0.0276	-11.84	0.0000
AgeF1[83,Inf)	-0.1651	0.0730	-2.26	0.0236
VehageF[4,8)	-0.2321	0.0163	-14.20	0.0000
VehageF[8,12)	-0.2542	0.0172	-14.82	0.0000
VehageF[12,16)	-0.4228	0.0197	-21.47	0.0000
VehageF[16,Inf)	-0.6569	0.0285	-23.05	0.0000
DensityF[40,200)	0.0933	0.0224	4.16	0.0000
DensityF[200,500)	0.1529	0.0243	6.30	0.0000
DensityF[500,4e+03)	0.2850	0.0216	13.18	0.0000
DensityF[4e+03,Inf)	0.3790	0.0260	14.56	0.0000
VehpowerF[7,9)	-0.0667	0.0144	-4.62	0.0000
VehpowerF[9,15)	0.0951	0.0184	5.17	0.0000
VehGasRegular	0.0707	0.0127	5.57	0.0000
VehbrandF2other	0.0507	0.0147	3.46	0.0005

TABLE 6 – Estimation des paramètres du modèle regnb

Ici encore, on utilise le test de la déviance pour s'assurer que notre modèle s'ajuste bien aux données, ce qui est d'ailleurs confirmé.

```
> pchisq(deviance(regnb),regnb$df.residual,lower.tail = FALSE)
1
```

Le modèle Binomiale Négative **est meilleur** que celui de Poisson,en terme d'AIC. En effet, $\mathsf{AIC}_{NB} = 231635.7$ et $\mathsf{AIC}_P = 232651.3$.

Modèle de Poisson à inflation de zéros : ZIP

La distribution de la variable ClaimNb (nombre de sinistre), laisse à penser que le modèle peut être à inflation de zéros.
Ici, on s'intéresse au cas où le modèle de comptage est Poisson. On a donc :

$$\mathsf{P}(N_i = k|x_i) = \begin{cases} \pi_i + e^{-\lambda_i}(1-\pi_i) & \mathsf{Si\ k} = 0 \\ \\ e^{-\lambda_i}\frac{\lambda_i^k}{k!}(1-\pi_i) & \mathsf{Si\ k} = 1, 2, ... \end{cases}$$

On suppose dans un premier temps, que la probabilité π_i est la même pour tous les assurés, on l'estime donc par une constante.

```
> regzip=zeroinfl(ClaimNb AgeF1+VehageF+DensityF+VehpowerF+VehGas+VehbrandF2+
offset(log(Exposure))|1,data = Contrats,dist = "poisson",link = "logit")
> summary(regzip)
```

Count model coefficients	(poisson	with log	link)	
	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-1.30544	0.03721	-35.079	0.0000
AgeF1[31,44)	-0.36352	0.01928	-18.858	0.0000
AgeF1[44,57)	-0.23752	0.01879	-12.639	0.0000
AgeF1[57,70)	-0.37591	0.02221	-16.927	0.0000
AgeF1[70,83)	-0.32665	0.02765	-11.815	0.0000
AgeF1[83,Inf)	-0.16516	0.07270	-2.272	0.0230
VehageF[4,8)	-0.23040	0.01637	-14.075	0.0000
VehageF[8,12)	-0.25376	0.01719	-14.763	0.0000
VehageF[12,16)	-0.42001	0.01974	-21.280	0.0000
VehageF[16,Inf)	-0.65366	0.02855	-22.899	0.0000
DensityF[40,200)	0.09335	0.02243	4.161	0.0000
DensityF[200,500)	0.15264	0.02429	6.284	0.0000
DensityF[500,4e+03)	0.28478	0.02165	13.156	0.0000
DensityF[4e+03,Inf)	0.37834	0.02607	14.510	0.0000
VehpowerF[7,9)	-0.06593	0.01446	-4.559	0.0000
VehpowerF[9,15)	0.09511	0.01842	5.163	0.0000
VehGasRegular	0.06931	0.01272	5.448	0.0000
VehbrandF2other	0.05089	0.01468	3.467	0.0005
Zero-inflation model coefficients	(binomial	with logit	link)	
	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	0.10505	0.04303	2.442	0.0146

TABLE 7 – Estimation des paramètres du modèle regnb

La probabilité prédite d'avoir "zéro" sinistre, est

```
> betapi <- exp(regzip$coefficients$zero[1])
> pi <- betapi/(1 + betapi)
> pi
0.5262383
```

Cette valeur de $\pi = 0.526$, est inférieure à la proportion de zéros (aucun sinistre) dans la base de donnée. Ceci, montre que π n'est probablement pas une constante comme supposée avant, mais qu'elle dépend plutôt des caractéristiques des assurés.
Pour s'assurer du constat fait, on réalise un test du **rapport de vraisemblance** pour vérifier l'hypothèse de nullité de tous les coefficients hors constante du modèle binomial, en opposant les log de vraisemblance du modèle réduit à la constante et le modèle complet (contenant toutes les variables).

```
> regcom=zeroinfl(ClaimNb AgeF1+VehageF+DensityF+VehpowerF+VehGas+VehbrandF2
+offset(log(Exposure))|AgeF1+VehageF+DensityF+VehpowerF+VehGas+VehbrandF2,
data=Contrats,dist = "poisson",link="logit")
> LR_Ref <- 2(regcom$loglik - regzip$loglik)
> pchisq(LR_Ref,df=6,lower.tail=FALSE)
7.716437e-30
```

Au risque 5%, l'hypothèse nulle est rejetée. Il existe des variables pertinentes pour expliquer la valeur structurelle '0' de la variable "ClaimNb".

Dans le modèle ZIP la sélection des variables est difficile, vu que le modèle de comptage et le modèle des inflation de zéros sont enchevêtrés. Après plusieurs tentatives d'estimation avec différents modèles, un modèle était sélectionné comme meilleur (critère d'AIC), il se peut qu'il y est d'autre modèle plus adéquats.

```
> modelzip<-zeroinfl(ClaimNb~AgeF2+DensityF+VehpowerF+VehGas+VehbrandF2+
offset(log(Exposure))|VehageF+VehGas,data=Contrats,dist = "poisson",link="logit")
```

Le modèle ZIP est **meilleur que** celui du binomiale négative, en terme d'AIC.
Pour comparer le modèle ZIP obtenu et le modèle de poisson, on utilise le test de vuong.

```
> pscl::vuong(modelzip,regp2)
```

	Vuong z-statistic	H_A	p-value
Raw	10.152440	model1 > model2	0.0000
AIC-corrected	10.127171	model1 > model2	0.0000
BIC-corrected	9.985674	model1 > model2	0.0000

TABLE 8 – Résultat du test de vuong

Count model coefficients	(poisson	with log	link)	
	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-1.81294	0.04532	-40.003	0.0000
AgeF2[31,44)	-0.35279	0.01920	-18.376	0.0000
AgeF2[44,57)	-0.23010	0.01872	-12.291	0.0000
AgeF2[57,70)	-0.36692	0.02212	-16.589	0.0000
AgeF2[70,Inf)	-0.30742	0.02675	-11.491	0.0000
DensityF[40,200)	0.08470	0.02225	3.806	0.0001
DensityF[200,500)	0.14786	0.02413	6.128	0.0000
DensityF[500,4e+03)	0.27716	0.02147	12.908	0.0000
DensityF[4e+03,Inf)	0.37008	0.02590	14.288	0.0000
VehpowerF[7,9)	-0.05274	0.01436	-3.672	0.0002
VehpowerF[9,15)	0.08781	0.01844	4.762	0.0000
VehGasRegular	0.46448	0.04219	11.009	0.0000
VehbrandF2other	0.05154	0.01465	3.517	0.0004
Zero-inflation model coefficients	(binomial	with logit	link)	
	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-1.04139	0.12219	-8.523	0.0000
VehageF[4,8)	0.54537	0.04288	12.718	0.0000
VehageF[8,12)	0.55771	0.04228	13.192	0.0000
VehageF[12,16)	0.85754	0.04614	18.584	0.0000
VehageF[16,Inf)	1.18533	0.05527	21.446	0.0000
VehGasRegular	0.85583	0.10337	8.279	0.0000

TABLE 9 – Estimation des paramètres du modèle modelzip

On retient le modèle ZIP pour l'estimation de la variable "ClaimNb". En fait, on pouvait tester un autre type de modèle à inflation de zéros, celui qui a comme modèle de comptage la loi binomiale négative, en utilisant

```
> library(gamlss)
> zinbreg<-gamlss(ClaimNb ∼ AgeF2+DensityF+VehpowerF+VehGas+VehbrandF2+
offset(log(Exposure)),nu.fo = ∼ VehageF+VehGas,family=ZINBI, data= Contrats,
control=gamlss.control(n.cyc=100))
```

Modèle de comptage

	Estimate	exp(Estimate)
(Intercept)	-1.81294	0.1631737
AgeF2[31,44)	-0.35279	0.7027248
AgeF2[44,57)	-0.23010	0.7944542
AgeF2[57,70)	-0.36692	0.6928651
AgeF2[70,Inf)	-0.30742	0.7353417
DensityF[40,200)	0.08470	1.0883905
DensityF[200,500)	0.14786	1.1593506
DensityF[500,4e+03)	0.27716	1.3193775
DensityF[4e+03,Inf)	0.37008	1.4478504
VehpowerF[7,9)	-0.05274	0.9486266
VehpowerF[9,15)	0.08781	1.0917807
VehGasRegular	0.46448	1.5911866
VehbrandF2other	0.05154	1.0528913

TABLE 10 – Estimations et exponentiel des estimations des paramètres du modèle de comptage

- (Intercept) : Si les modalités suivantes sont toutes réunies AgeF2[18,31[, DensityF[0,40[, VehpowerF[4,7[, VehGasDiesel, VehbrandF2B1, alors la fréquence prédite de sinistre est 0.1631737.

- AgeF2[31, 44[: Si l'age est entre [31,44[, alors la fréquence prédite de sinistre est 0.7027248 fois la fréquence prédite de sinistre si l'age est entre [18,31[, à condition que les autres variables du modèle ont pour modalités les modalités de référence (DensityF[0,40[, VehpowerF[4,7[, VehGasDiesel, VehbrandFB1).

- AgeF2[44, 57[, AgeF2[57, 70[, AgeF2[70, Inf[ont la même interprétation que la précédente, quitte à remplacer le coefficient par celui correspondant.

- DensityF[40, 200[: si la densité est entre [40,200[hab/km^2, la fréquence prédite de sinistre est 1.0883905 fois la fréquence prédite de sinistre si la densité est entre [0,40[hab/km^2, à condition que les autres variables du modèle ont pour modalités les modalités de référence.

- DensityF[200, 500[, DensityF[500, 4000[, DensityF[4000, Inf[ont la même interprétation que la précédente, quitte à remplacer le coefficient par celui correspondant.

- VehpowerF[7, 9[: si la puissance de la voiture est entre [7,9[, la fréquence prédite de sinistre est 0.9486266 fois la fréquence prédite de sinistre si la puissance de voiture est entre [4.7[, à condition que les autres variables du modèle ont pour modalités les modalités de référence.

- Vehpower[9, 15[: a la même interprétation que la précédente, quitte à remplacer le coefficient par celui correspondant.

- VehGasRegular : si le gaz utilisé est régulier, alors la fréquence prédite de sinistre est 1.5911866 la fréquence prédite de sinistre si le gaz est diesel, à condition que les autres variables du modèle ont pour modalités les modalités de référence.

- VehbrandF2other :si la marque de la voiture est dans la modalité other, alors la fréquence prédite de sinistre est 1.0528913 la fréquence prédite de sinistre si la marque est dans la modalité B1, à condition que les autres variables du modèle ont pour modalités les modalités de référence.

Modèle de zéro inflation

	Estimate	exp(Estimate)
(Intercept)	-1.04139	0.3529637
VehageF[4,8)	0.54537	1.7252466
VehageF[8,12)	0.55771	1.7466680
VehageF[12,16)	0.85754	2.3573545
VehageF[16,Inf)	1.18533	3.2717663
VehGasRegular	0.85583	2.3533268

TABLE 11 – Estimations et exponentiel des estimations des paramètres du modèle de zéro inflation

- (Intercept) : si toutes les modalités de référence sont réunies, alors les chances d'avoir zéro sinistre sont 0.3529637.

- VehageF[4, 8[: si l'age de la voiture est entre [4,8[, alors les chances d'avoir zéro sinistre sont 1.7252466 fois les chances d'avoir zéro sinistre si l'age de la voiture est entre [0,4[, , à condition que les autres variables du modèle ont pour modalités les modalités de référence.

- VehageF[8, 12[, VehageF[12, 16[, VehageF[16, Inf[: ont la même interprétation que la précédente, quitte à remplacer le coefficient par celui correspondant.

- VehGasRegular : si le gaz est régulier, alors les chances d'avoir zéro sinistre sont 2.3533268 fois les chances d'avoir zéro sinistre si gaz diesel, à condition que les autres variables du modèle ont pour modalités les modalités de référence.

Références :

Computational Actuarial Science with R, Arthur Charpentier
Negative Binomial Regression, Joseph M. Hilbe