

Statistique des assurances - Projet

Isabelle Ajtay (41010932)

Smail Chabane (38012939)

Yuxuan Zhang (38019811)

13 mars 2023



Table des matières

1 Contexte et objectifs	2
2. Les données - description et nettoyage	2
2.1 Analyses univariées	3
2.2 Analyses multivariées	5
3. Modélisation des sinistres et des primes pures	10
3.1 Problème d'endogénéité dans les variables	10
3.2 Modélisation de <i>Sinistre0</i>	11
3.4 Modèle pour au moins un des <i>Sinistre1</i> , <i>Sinistre2</i> ou <i>Sinistre3</i>	42
3.5 Modèle pour le prix de Police 1 ou 2 ou 3 (au moins un)	43
3.6 Modèle retenu au final	43
4. Modélisation pour les prix : le nombre de sinistres et la tarification des nouveaux arrivants	43
4.1 Modèle pour le nombre de sinistres, NSin	43
4.2 Méthode de tarification pour les nouveaux arrivants	43
5. Estimation des durées	43
5.1 Estimateur de Kaplan-Meier	44
5.2 Modèle de Cox	44
6. Références	44

1 Contexte et objectifs

Dans le cadre du présent projet, nous sommes une compagnie d'assurance non-vie, qui dispose d'un jeu de données historiques sur des ménages ayant souscrit ses polices d'assurance.

Les objectifs ? Déterminer la prime pure pour un ménage intéressé par l'assurance proposée par notre compagnie. On souhaite construire un modèle qui explique les demandes d'indemnisation (Sinistres 1, 2 ou 3) en utilisant les données qu'on possède. On veut aussi avoir des modèles pour les prix des polices d'assurance précédemment vendues, le nombre de sinistres, ainsi que la durée de vie d'un contrat d'assurance.

Pour ce faire, nous employerons des méthodes et outils vus en cours de *Statistiques des Assurances*, et d'autres cours, et le logiciel *R*.

La mtd train du package caret fait de la cv + boot, et permet d'ajuster des centaines de modèles prédictifs différents, spécifiés facilement avec l'argument method. VerboseIter donne un log du progress, permettant de mesurer le modèle est ajusté.

On va choisir lequel des deux on met dans le modèle: RUC ou full income. Mais pas les 2 car très fortement corrélées. Il y a aussi les quantiles de cet income.

Anat non signif. A suppr.

Police i corresp à sin. i. Il a une assurance de base et rajoute des additionnelles. 0-> le type n'a pas pris cette assurance là.

Sini1 plus facile à modéliser, moins de zeros. 2 et 3 en ont beaucoup.

1283 outlier pê. On doit les enlever.

2. Les données - description et nettoyage

[1] 5352 27

2.1 Analyses univariées

On a utilisé str pour afficher les informations simples concernant les variables, et summary pour afficher les données statistiques pour chaque variable.

Voici les variables:

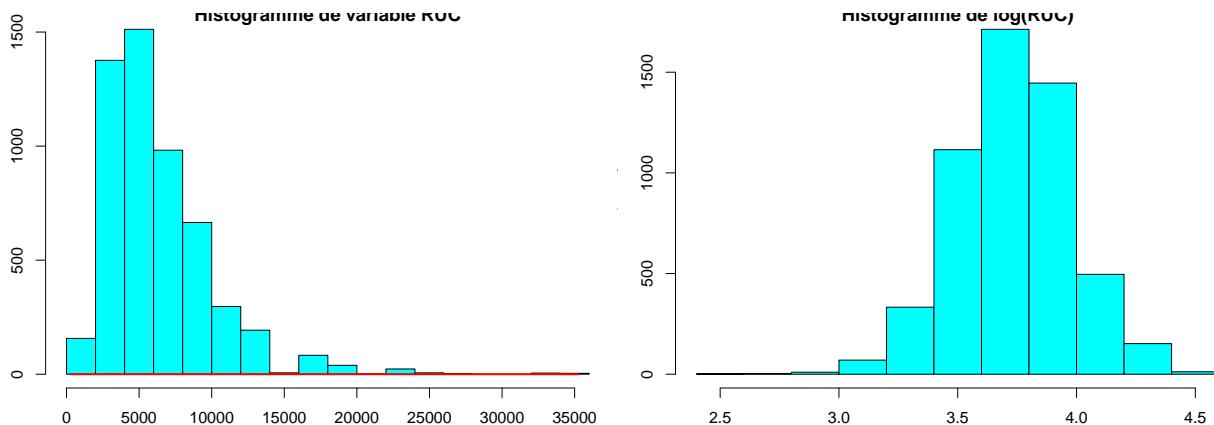
1. pcs
2. RUC
3. cs
4. reves
5. crevpp
6. region
7. habi
8. Ahabi
9. Atyp
10. agecat
11. Acompm
12. nbpers
13. enfants
14. Anat
15. Bauto
16. "Nbadulte"
17. Sinistre1"
18. Sinistre2
19. Sinistre3"
20. Police1
21. "Police2"
22. "Police3"
23. "durPolice1"
24. Durée"
25. NSin"
26. censure"
27. Sinistre0

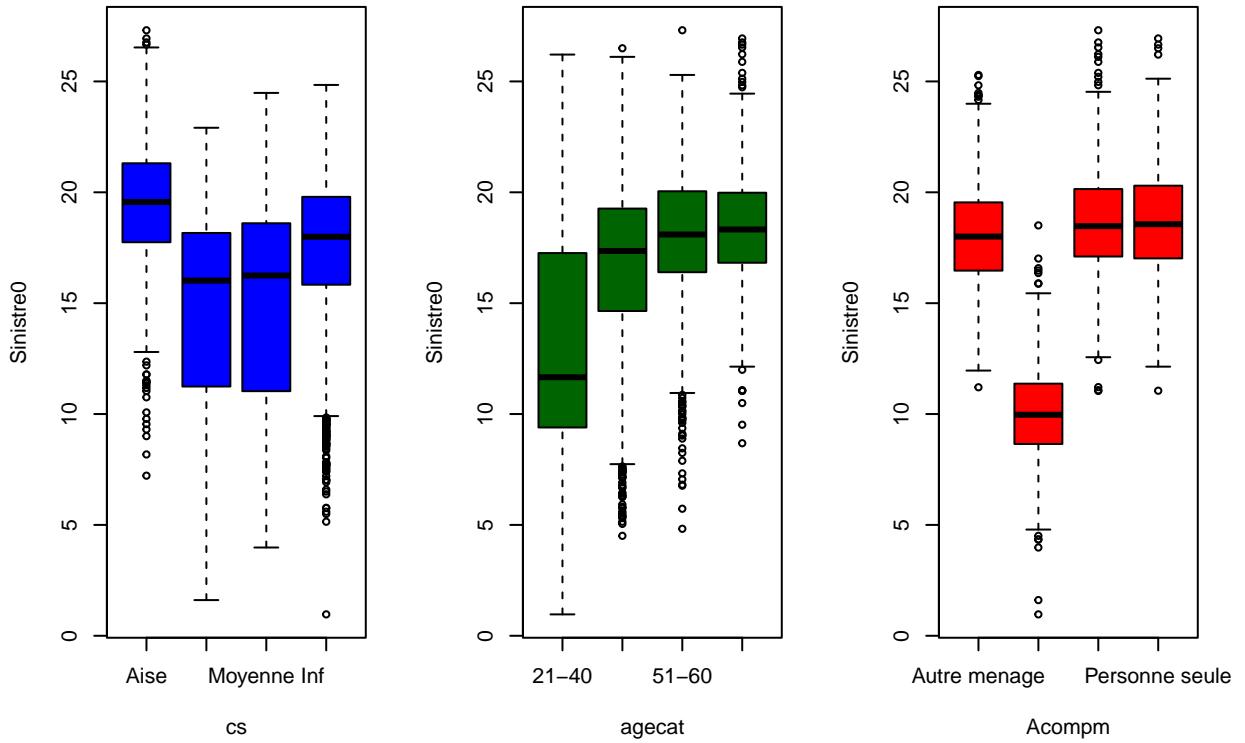
On observe que les variables *pcs*, *cs*, *region*, *crevpp*, *agecat* et *habi* sont qualitatives, malgré leur format caractères ou numérique. On rémedie.

```
## [1] 1 2 3 4 5 7 8 9
```

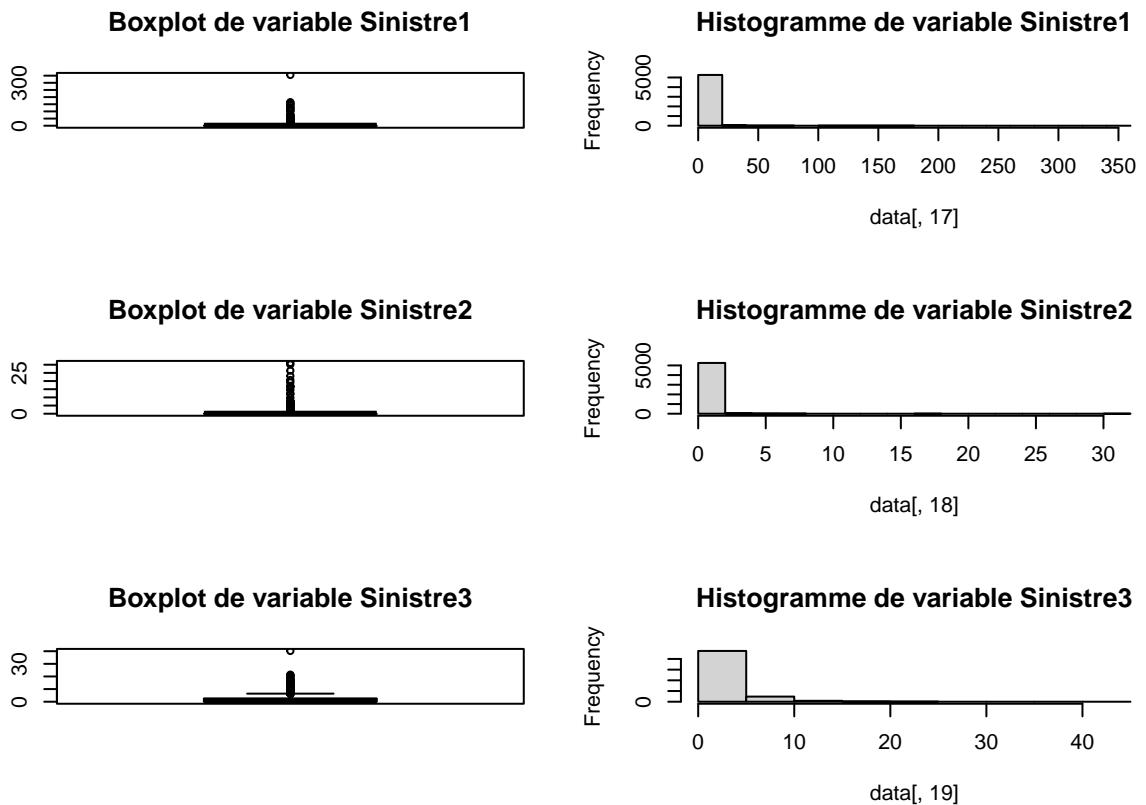
Il y a des sinistres mais aucune police souscrite, pour 13 observations. On les enlève.

On a représenté les boxplots de la variable RUC, et de son log. La transformation log rend la distribution simétrique.





```
par(mfrow=c(3,2))
boxplot(data[,17],main="Boxplot de variable Sinistre1")
hist(data[,17],main="Histogramme de variable Sinistre1")
boxplot(data[,18],main="Boxplot de variable Sinistre2")
hist(data[,18],main="Histogramme de variable Sinistre2")
boxplot(data[,19],main="Boxplot de variable Sinistre3")
hist(data[,19],main="Histogramme de variable Sinistre3")
```



2.2 Analyses multivariées

On peut constater que les variabilités des trois types de *Sinistres* sont toutes grandes.

```
aggregate(data[,c(17,18,19,27)],list(data[,1]),mean)
```

```
##                               Group.1 Sinistre1 Sinistre2 Sinistre3
## 1           Agr. exploitants 0.4860776 0.009051724 2.983017
## 2       Artisans, comm., chefs d'ent. 0.3559116 0.097458564 1.836381
## 3   Autres pers. sans activite prof. 1.4143169 0.107540984 1.761721
## 4 Cadres et prof. intellectuelles sup. 1.6972443 0.184442589 2.253706
## 5                   Employes 1.2374190 0.160204489 1.778940
## 6                   Ouvriers 1.6150491 0.128436556 1.841900
## 7 Professions intermediaires 1.7728983 0.208022599 2.219186
## 8                  Retraites 0.5492878 0.185552958 1.398386
##   Sinistre0
## 1 15.01849
## 2 16.12462
## 3 17.41922
## 4 16.81633
## 5 15.56352
## 6 14.29188
## 7 15.65612
## 8 18.37193
```

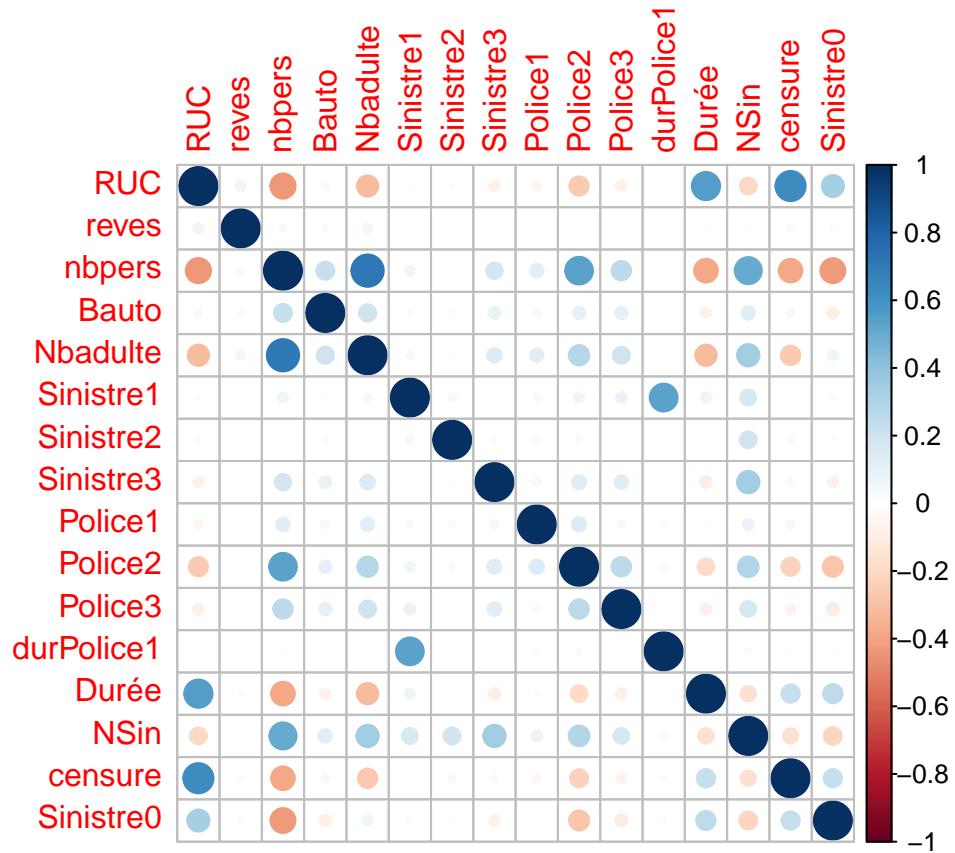
#Representation des corrélations

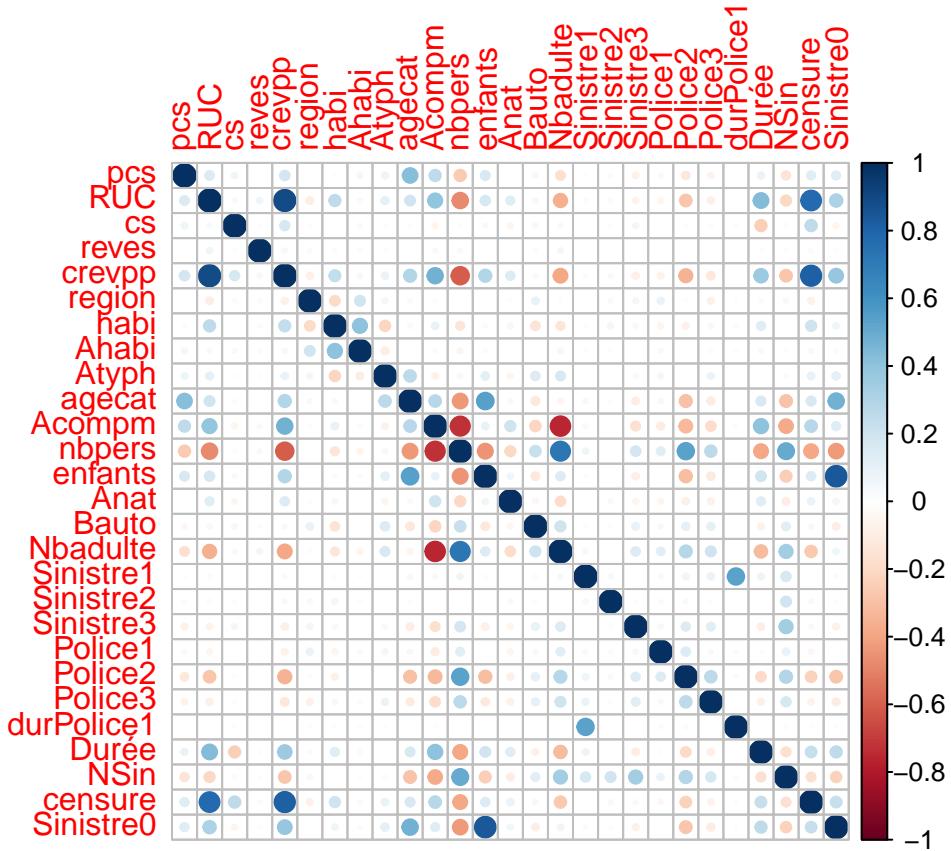
```
variables_quantitatives = data %>% select_if(is.numeric) %>% cor()
```

```
kable(variables_quantitatives, digits=3)
```

	RUC	reves	nbpers	Bauto	Nbadulte	Sinistrel1	Sinistrel2	Sinistrel3	Police1	Police2	Police3	durPolice	Durée	NSin	censure	Sinistre0
RUC	1.000	0.066	-	0.032	-	-	0.018	-	-	-	-	0.005	0.551	-	0.629	0.338
				0.437		0.310	0.018		0.074	0.050	0.256	0.079			0.201	
reves	0.066	1.000	0.031	0.022	0.053	-	-	0.007	-	0.006	0.005	0.001	0.012	0.013	0.029	0.014
						0.001	0.001		0.003							
nbpers	-	0.031	1.000	0.227	0.720	0.064	0.006	0.187	0.127	0.530	0.260	0.009	-	0.507	-	-
				0.437										0.389	0.385	0.428
Bauto	0.032	0.022	0.227	1.000	0.208	0.023	-	0.098	0.039	0.108	0.106	0.004	-	0.126	0.046	-
							0.013							0.073		0.089
Nbadulte	-	0.053	0.720	0.208	1.000	0.042	0.008	0.148	0.123	0.287	0.200	0.008	-	0.345	-	0.067
					0.310									0.313		0.261
Sinistrel1	-	-	0.064	0.023	0.042	1.000	0.042	-	0.030	0.070	0.082	0.534	0.073	0.175	-	-
			0.018	0.001				0.001						0.001	0.021	
Sinistrel2	0.018	-	0.006	-	0.008	0.042	1.000	0.015	0.029	0.030	0.010	-	0.004	0.200	0.037	0.012
					0.001	0.013						0.002				
Sinistrel3	-	0.007	0.187	0.098	0.148	-	0.015	1.000	0.047	0.139	0.129	-	-	0.346	-	-
			0.074				0.001					0.006	0.091		0.031	0.077
Police1	-	-	0.127	0.039	0.123	0.030	0.029	0.047	1.000	0.145	0.033	0.020	0.013	0.086	-	0.001
					0.050	0.003									0.043	
Police2	-	0.006	0.530	0.108	0.287	0.070	0.030	0.139	0.145	1.000	0.264	0.028	-	0.296	-	-
			0.256											0.194		0.228
Police3	-	0.005	0.260	0.106	0.200	0.082	0.010	0.129	0.033	0.264	1.000	-	-	0.171	-	-
			0.079									0.001	0.082		0.068	0.102
durPolice	0.005	0.001	0.009	0.004	0.008	0.534	-	-	0.020	0.028	-	1.000	-	0.021	0.011	0.011
						0.002	0.006					0.001		0.005		
Durée	0.551	0.012	-	-	-	0.073	0.004	-	0.013	-	-	-	1.000	-	0.232	0.256
						0.389	0.073	0.313					0.091	0.194	0.082	0.005
NSin	-	0.013	0.507	0.126	0.345	0.175	0.200	0.346	0.086	0.296	0.171	0.021	-	1.000	-	-
			0.201											0.168		0.162
censure	0.629	0.029	-	0.046	-	-	0.037	-	-	-	-	0.011	0.232	-	1.000	0.231
				0.385		0.261	0.001							0.162		
Sinistre0	0.338	0.014	-	-	0.067	-	0.012	-	0.001	-	-	0.011	0.256	-	0.231	1.000
					0.428	0.089		0.021		0.077		0.279	0.102			0.220

```
corrplot(variables_quantitatives)
```





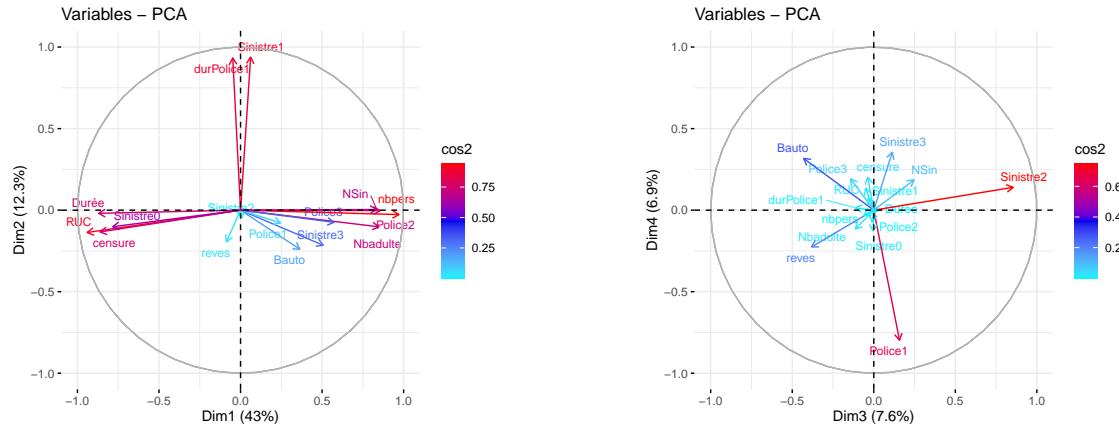
```

## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 16 individuals, described by 16 variables
## *The results are available in the following objects:
##
##      name           description
## 1  "$eig"          "eigenvalues"
## 2  "$var"           "results for the variables"
## 3  "$var$coord"    "coord. for the variables"
## 4  "$var$cor"       "correlations variables - dimensions"
## 5  "$var$cos2"      "cos2 for the variables"
## 6  "$var$contrib"   "contributions of the variables"
## 7  "$ind"           "results for the individuals"
## 8  "$ind$coord"     "coord. for the individuals"
## 9  "$ind$cos2"      "cos2 for the individuals"
## 10 "$ind$contrib"   "contributions of the individuals"
## 11 "$call"          "summary statistics"
## 12 "$call$centre"   "mean of the variables"
## 13 "$call$ecart.type" "standard error of the variables"
## 14 "$call$row.w"    "weights for the individuals"
## 15 "$call$col.w"    "weights for the variables"

##              eigenvalue percentage of variance cumulative percentage of variance
## comp 1        6.89                  43.03                      43.03
## comp 2        1.97                  12.29                      55.32
## comp 3        1.22                   7.60                      62.92
## comp 4        1.10                   6.90                      69.82
## comp 5        1.08                   6.73                      76.55

```

## comp 6	0.86	5.36	81.92
## comp 7	0.81	5.05	86.97
## comp 8	0.69	4.31	91.27
## comp 9	0.43	2.66	93.93
## comp 10	0.33	2.07	96.01
## comp 11	0.29	1.81	97.82
## comp 12	0.17	1.04	98.86
## comp 13	0.13	0.78	99.64
## comp 14	0.04	0.25	99.90
## comp 15	0.02	0.10	100.00

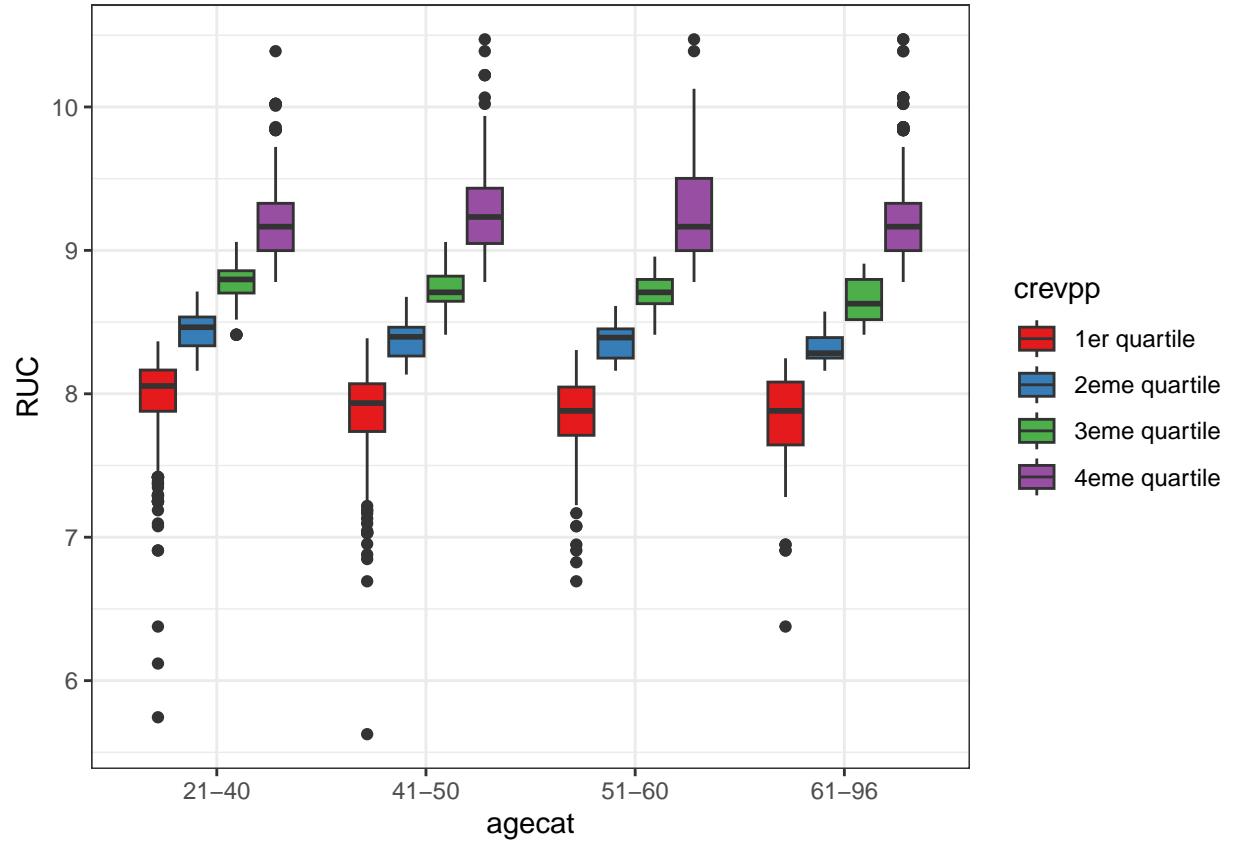


2.3 Traitement des valeurs extrêmes

Les modèles linéaires généralisés dichotomiques sont parfaitement adaptés à notre cas puisque la variable d'intérêt, Classe, est de type binaire et prend deux modalités. Les deux modèles possibles sont logit et probit qui se diffèrent par leur fonction de lien : logistique pour la première et normale pour la seconde. Implémenter ce type de modèle requiert de séparer notre échantillon en deux sous-échantillons. Cette étape s'avère nécessaire sinon l'erreur de prédiction du modèle aurait tendance à être trop optimiste si les données qui ont servi à construire le modèle sont les mêmes sur lesquelles il est testé. De ce fait, nous construisons un sous-échantillon d'apprentissage pour construire les modèles et un souséchantillon de test pour les tester.

La Winsorisation est une bonne technique pour traiter les outliers. Elle consiste à remplacer les valeurs extrêmes par une valeur proche mais plus raisonnable. On remplace les valeurs qui sont inférieures à $Q1 - 1,5 * IQR$ par $Q1 - 1 * IQR$, et les valeurs qui sont supérieures à $Q3 + 1,5 * IQR$ par $Q3 + 1 * IQR$.

Une variable importante est *RUC* (ou *reves*), qu'on voit augmenter avec la catégorie socio-professionnelle, *crevpp* (ou de manière équivalente, *cs* ou *pcs*). A tout groupe d'âge donc, une catégorie supérieure est associé à un revenu supérieur.

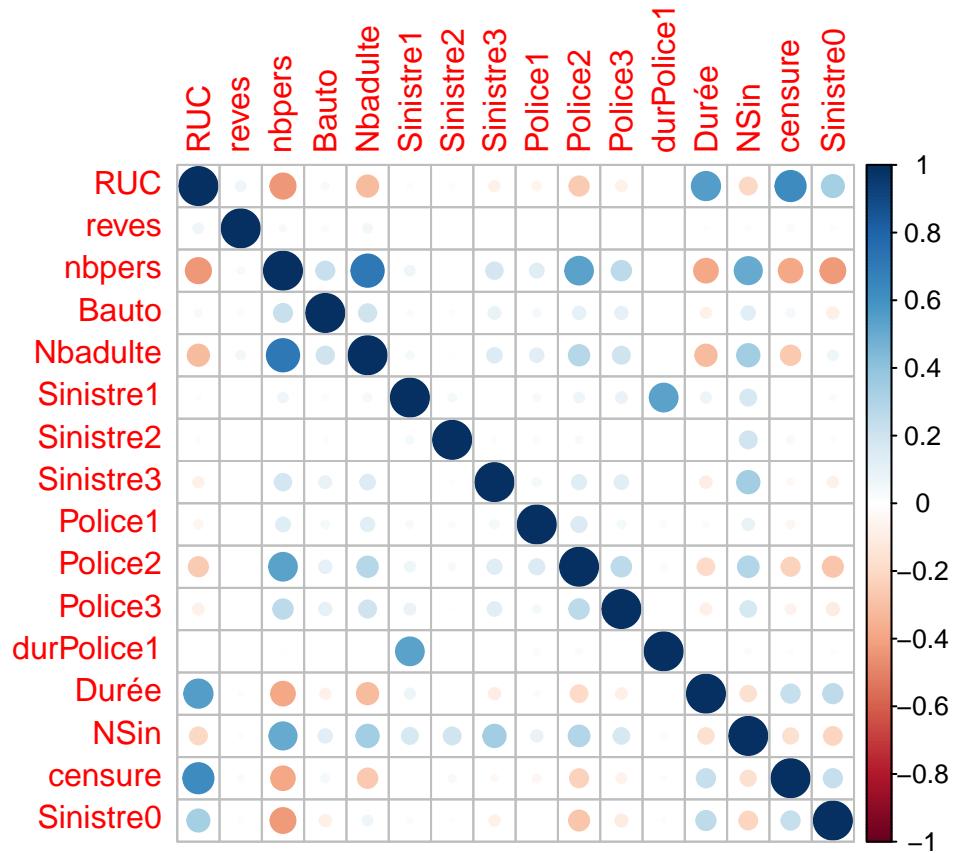


3. Modélisation des sinistres et des primes pures

3.1 Problème d'endogénéité dans les variables

```
# Selection des variables quantitatives
quant_vars <- sapply(data, is.numeric)

# Matrice de corrélation
cor_matrix <- cor(data[, quant_vars])
corrplot(cor_matrix)
```



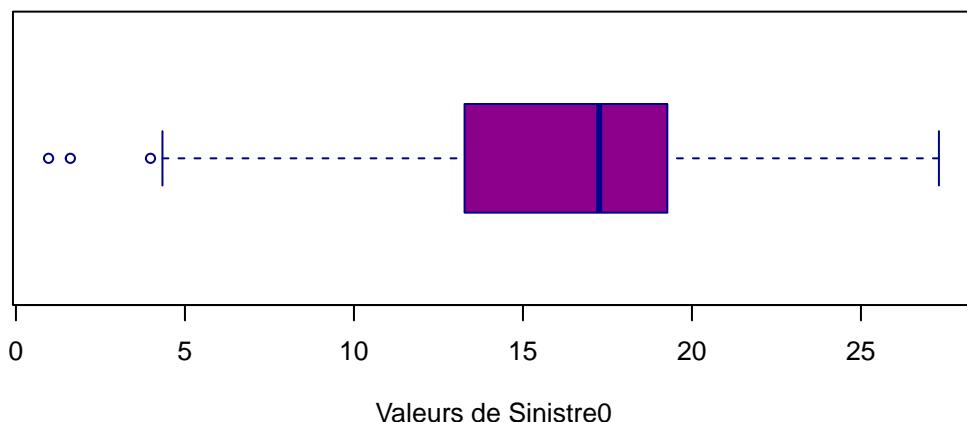
Si on fixe $\alpha = 0.05$, alors il y a une causalité entre Sinistre0 et les variables suivantes : RUC / durPolice1.

3.2 Modélisation de *Sinistre0*

Statistiques descriptives de *Sinistre0*

```
## moyenne de Sinistre0 : 16.16682
## variance de Sinistre0 : 18.46686
```

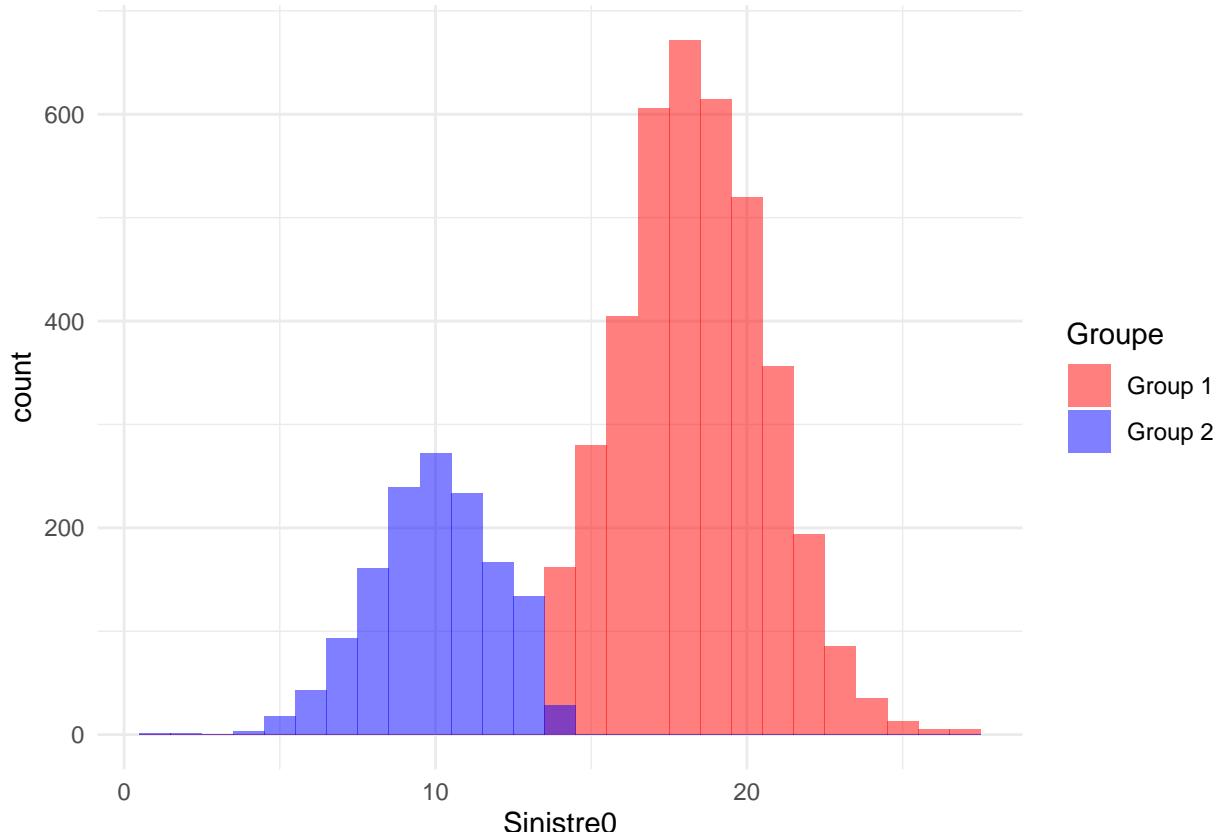
Boxplot de Sinistre0

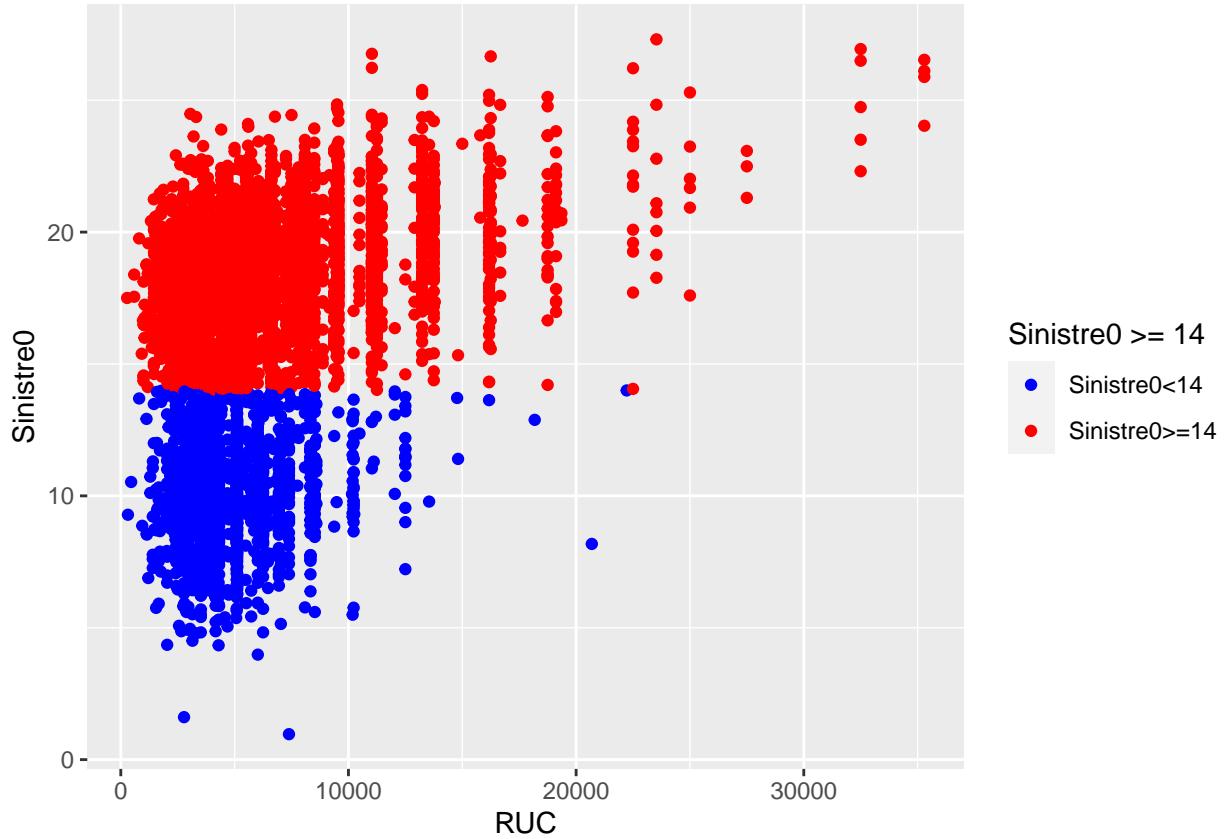


Le boxplot indique seulement 3 valeurs extrêmes dans la partie inférieure de l'intervalle de valeurs. Nous les enlevons car ils peuvent influencer : - les paramètres de la régression (en “tirant” le paramètres de la ligne de régression vers eux), - les résidus - en augmentant la variance résiduelle et en rendant la distribution des résidus non normale - la sensibilité de certains modèles, dont ceux de régression linéaire.

```
##      Min. 1st Qu. Median   Mean 3rd Qu.   Max.
## 0.9652 13.2661 17.2555 16.1725 19.2735 27.3074
```

On observe sur l'histogramme de la variable *Sinistre0* qu'il y a deux sous-populations distinctes, qu'on sépare.





On peut nommer les deux groupes ainsi formés les “bons risques” (Sinistre0 petit, groupe 1) et les “mauvais risques” (Sinistre0 grand, groupe 2). On a 1395 individus dans le premier groupe, et 3954 dans le deuxième. La méthode des MCO donne l'estimateur le plus efficient s'il n'y a pas d'endogénéité.

S'il y a de l'endogénéité, OLS (MCO) va donner des résultats inconsistants. L'estimateur des variables instrumentales va être consistant, mais inéfficient.

```
##
## Call:
## lm(formula = Sinistre0 ~ groupe + RUC + Acompm + nbpers + Anat +
##     Durée, data = Dtrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.2621 -1.3898 -0.0153  1.4253  7.5440
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               8.789e+00  6.368e-01 13.802 < 2e-16 ***
## groupeGroup 2            -4.795e+00  1.993e-01 -24.063 < 2e-16 ***
## RUC                     1.096e+00  6.561e-02 16.697 < 2e-16 ***
## AcompmCouple avec enfant(s) -3.546e+00  2.032e-01 -17.451 < 2e-16 ***
## AcompmCouple sans enfant    1.511e-01  1.022e-01   1.478  0.1394  
## AcompmPersonne seule        4.886e-02  1.455e-01   0.336  0.7370  
## nbpers                   9.418e-02  3.793e-02   2.483  0.0131 *  
## AnatMenage francais       -2.722e-01  2.007e-01  -1.357  0.1750  
## AnatNon declare           -4.159e-01  2.313e-01  -1.798  0.0722 .
```

```

## Durée           3.140e-04  5.705e-05   5.504 3.91e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.036 on 4807 degrees of freedom
## Multiple R-squared:  0.7755, Adjusted R-squared:  0.7751
## F-statistic:  1845 on 9 and 4807 DF,  p-value: < 2.2e-16

On enlève successivement: Anat et nbpers, non significatives

## 
## Call:
## lm(formula = Sinistre0 ~ groupe + RUC + Acompm + nbpers + Durée,
##      data = Dtrain)
##
## Residuals:
##    Min     1Q   Median     3Q    Max
## -9.2573 -1.4033 -0.0206  1.4185  7.5472
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 8.547e+00  6.063e-01 14.097 < 2e-16 ***
## groupeGroup 2 -4.801e+00  1.992e-01 -24.097 < 2e-16 ***
## RUC          1.091e+00  6.555e-02 16.638 < 2e-16 ***
## AcompmCouple avec enfant(s) -3.541e+00  2.031e-01 -17.430 < 2e-16 ***
## AcompmCouple sans enfant    1.571e-01  1.022e-01   1.537 0.12428  
## AcompmPersonne seule        2.388e-02  1.436e-01   0.166 0.86797  
## nbpers         9.803e-02  3.788e-02   2.588 0.00968 ** 
## Durée          3.149e-04  5.706e-05   5.519 3.59e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.037 on 4809 degrees of freedom
## Multiple R-squared:  0.7753, Adjusted R-squared:  0.775 
## F-statistic:  2371 on 7 and 4809 DF,  p-value: < 2.2e-16

## 
## Call:
## lm(formula = Sinistre0 ~ groupe + RUC + Acompm + Durée, data = Dtrain)
##
## Residuals:
##    Min     1Q   Median     3Q    Max
## -9.3417 -1.3941 -0.0217  1.4097  7.5657
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 9.297e+00  5.328e-01 17.450 < 2e-16 ***
## groupeGroup 2 -4.807e+00  1.994e-01 -24.112 < 2e-16 ***
## RUC          1.046e+00  6.324e-02 16.534 < 2e-16 ***
## AcompmCouple avec enfant(s) -3.506e+00  2.028e-01 -17.288 < 2e-16 ***
## AcompmCouple sans enfant    -2.130e-04  8.219e-02  -0.003 0.9979  
## AcompmPersonne seule        -2.314e-01  1.045e-01  -2.215 0.0268 *  
## Durée          3.206e-04  5.705e-05   5.620 2.01e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

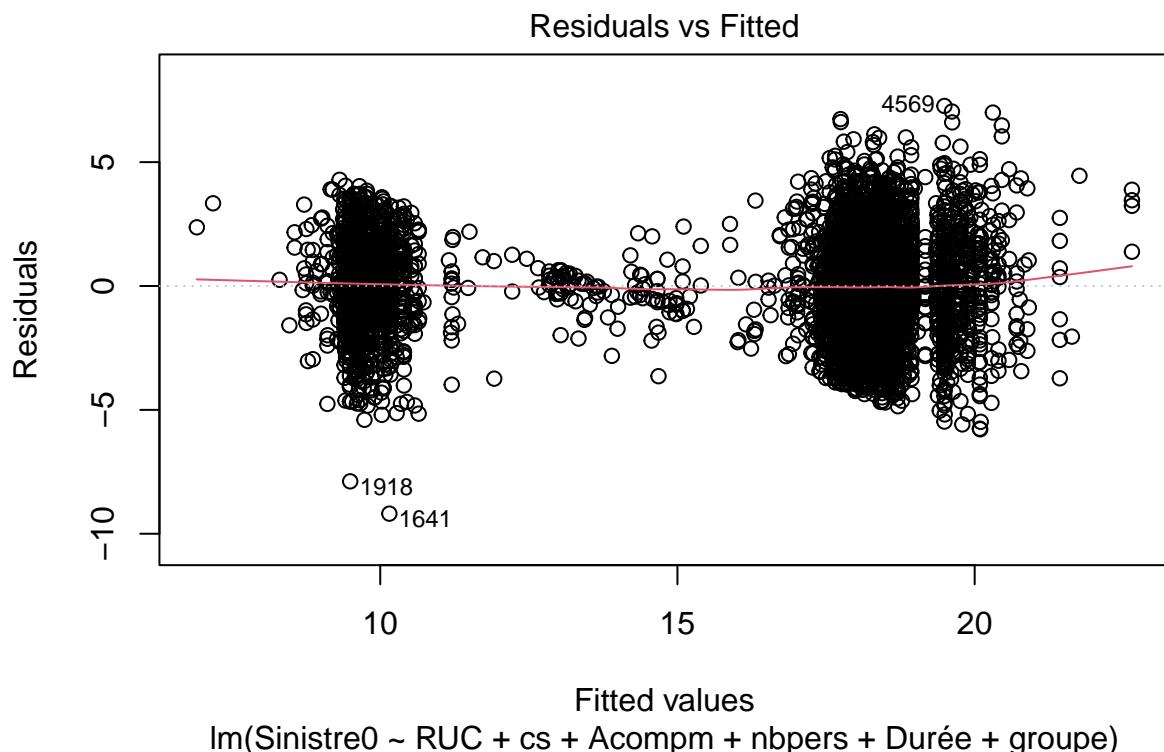
##
## Residual standard error: 2.038 on 4810 degrees of freedom
## Multiple R-squared:  0.775, Adjusted R-squared:  0.7747
## F-statistic:  2761 on 6 and 4810 DF, p-value: < 2.2e-16
## [1] 0

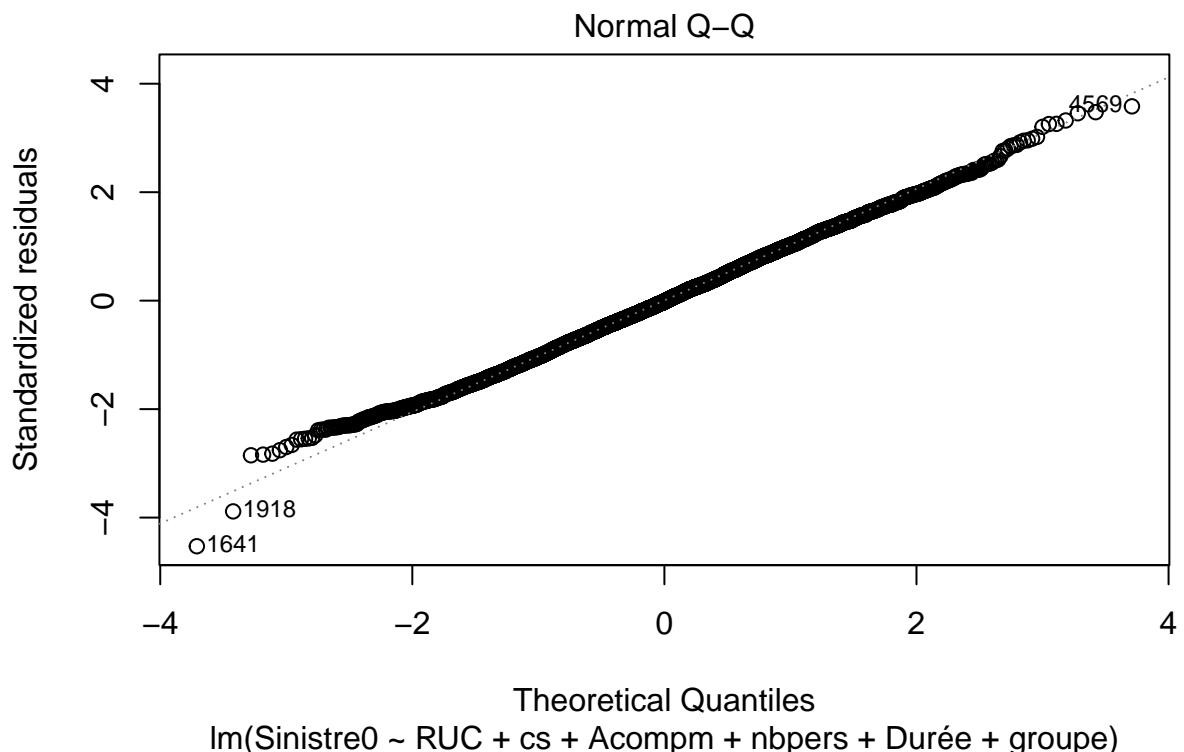
Sinistre0 ~ RUC + cs + Acompm + nbpers + Durée + group

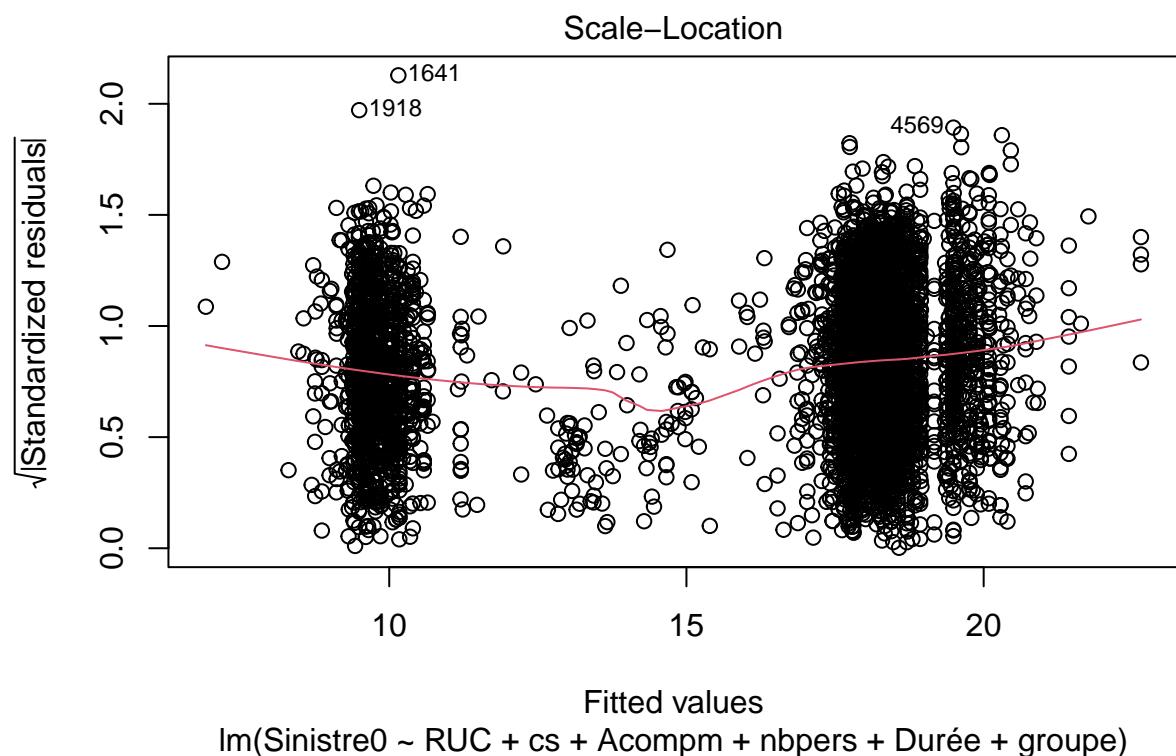
##
## Call:
## lm(formula = Sinistre0 ~ RUC + cs + Acompm + nbpers + Durée +
##     groupe, data = Dtrain[, vars])
##
## Residuals:
##    Min      1Q  Median      3Q      Max
## -9.1883 -1.3989 -0.0303  1.4169  7.2680
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)               7.683e+00  1.445e+00  5.318  1.1e-07 ***
## RUC                      1.221e+00  1.492e-01  8.183  3.5e-16 ***
## csModeste                 8.139e-02  2.692e-01  0.302  0.762435
## csMoyenne Inf             -3.000e-01  1.841e-01 -1.629  0.103287
## csMoyenne Sup             -3.599e-01  1.329e-01 -2.707  0.006809 **
## AcompmCouple avec enfant(s) -3.522e+00  2.030e-01 -17.354 < 2e-16 ***
## AcompmCouple sans enfant   1.401e-01  1.026e-01  1.366  0.172062
## AcompmPersonne seule       -2.729e-03  1.434e-01 -0.019  0.984817
## nbpers                     9.395e-02  3.838e-02  2.448  0.014393 *
## Durée                      2.283e-04  5.939e-05  3.844  0.000123 ***
## groupeGroup 2              -4.810e+00  1.987e-01 -24.204 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.031 on 4806 degrees of freedom
## Multiple R-squared:  0.7767, Adjusted R-squared:  0.7762
## F-statistic:  1671 on 10 and 4806 DF, p-value: < 2.2e-16

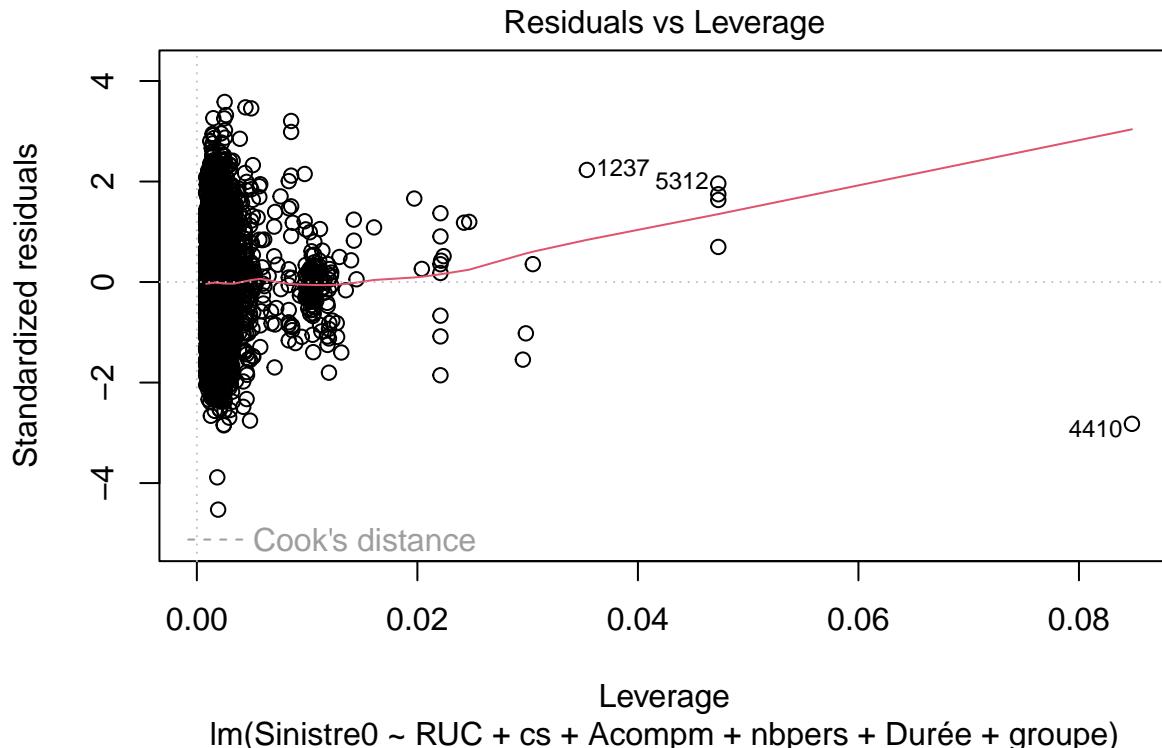
```

On retient donc à cette étape un modèle linéaire où Sinistre0 est expliqué par RUC, Acomp, cs, Durée et le groupe (variable que nous avons introduite).





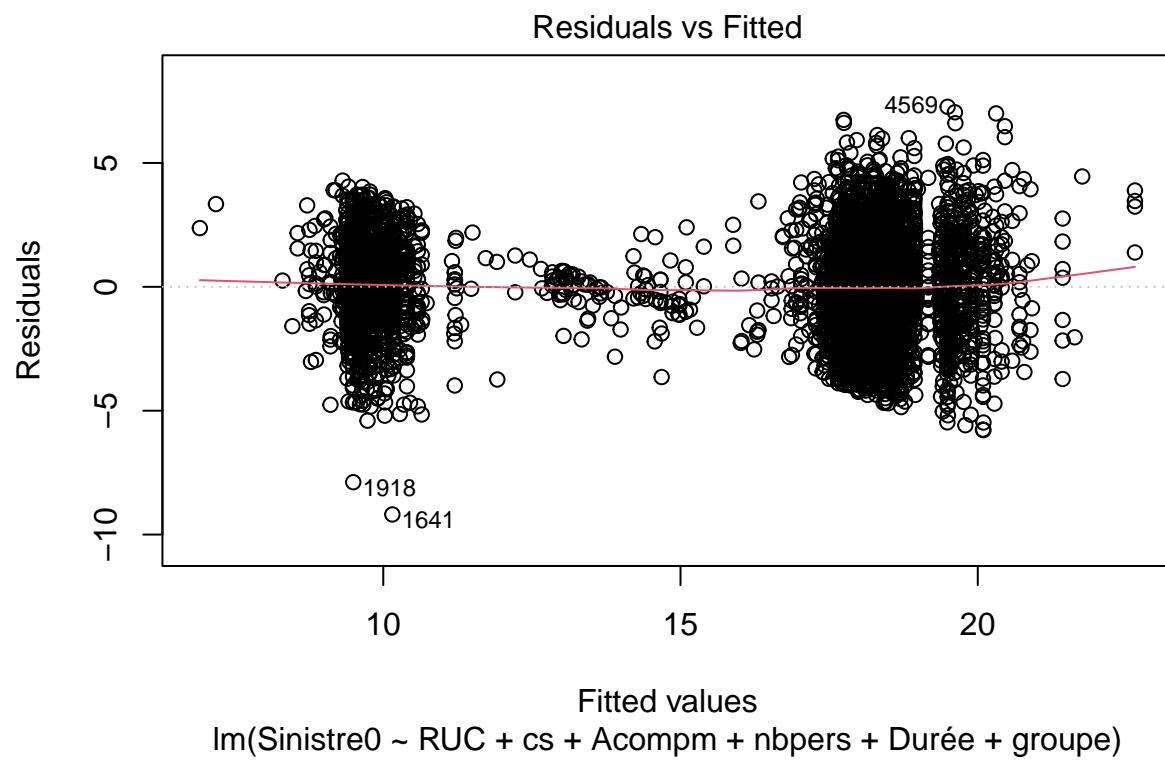


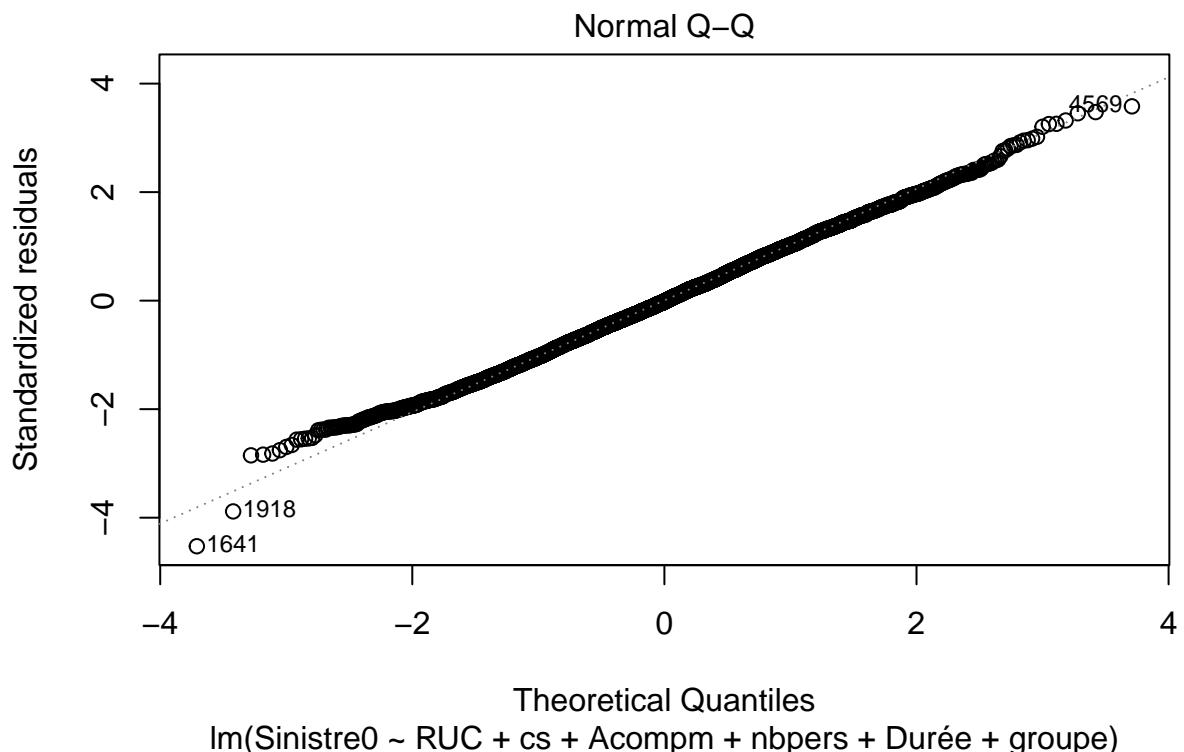


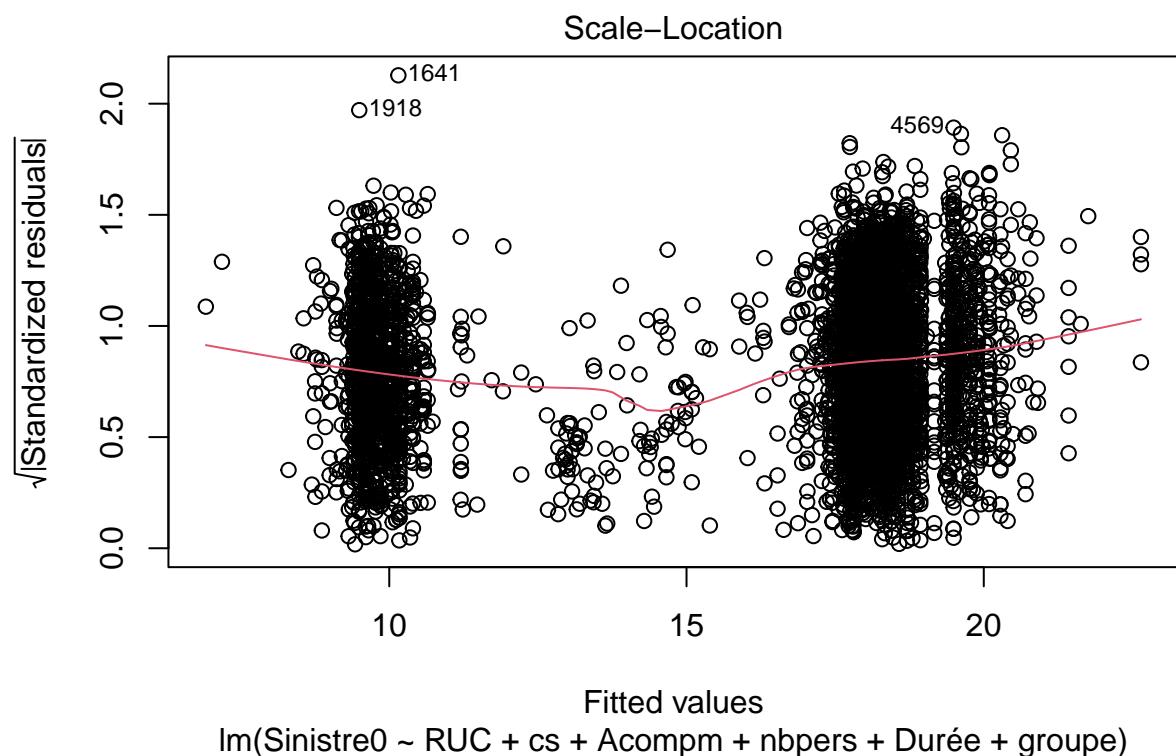
On identifie les outliers: 1918, 1641, 4569, 5290, 894, 4410

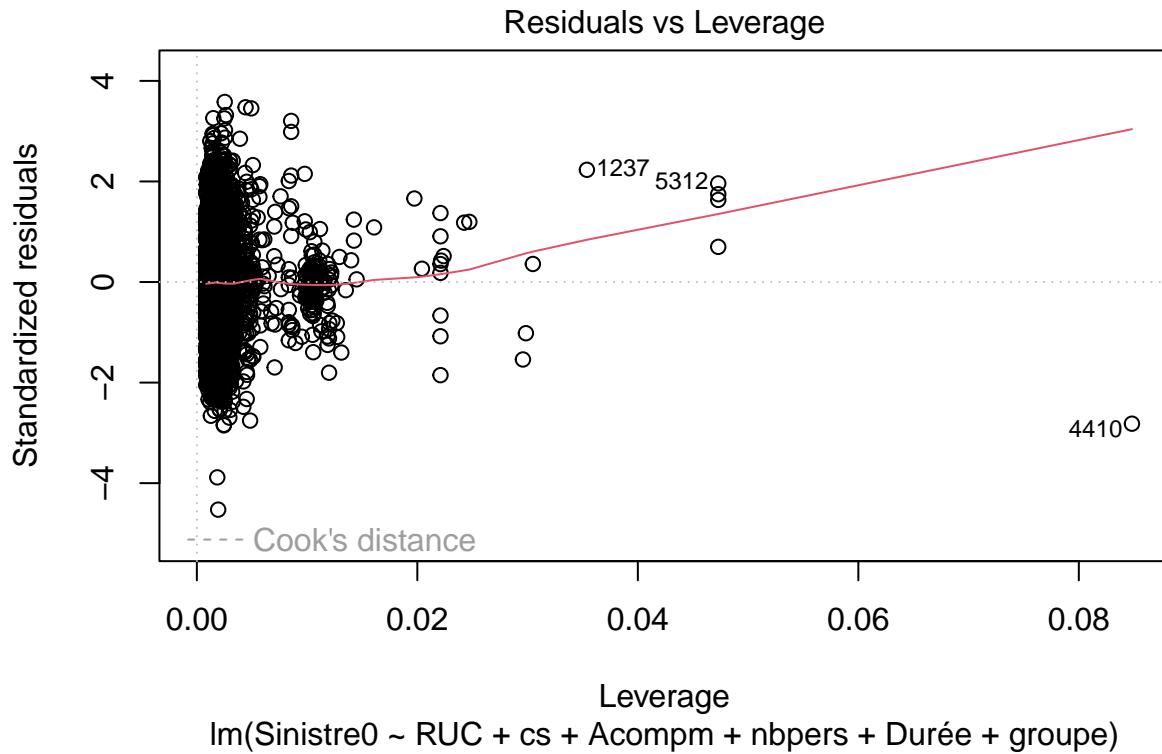
```
##
## Call:
## lm(formula = Sinistre0 ~ RUC + cs + Acompm + nbpers + Durée +
##     groupe, data = data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -9.1876 -1.3993 -0.0306  1.4165  7.2665 
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             7.675e+00  1.446e+00  5.309 1.15e-07 ***
## RUC                     1.221e+00  1.493e-01  8.182 3.54e-16 ***
## csModeste               8.156e-02  2.695e-01  0.303 0.762222  
## csMoyenne Inf            -3.004e-01 1.844e-01 -1.629 0.103382  
## csMoyenne Sup            -3.605e-01 1.332e-01 -2.707 0.006812 ** 
## AcompmCouple avec enfant(s) -3.522e+00  2.030e-01 -17.348 < 2e-16 ***
## AcompmCouple sans enfant    1.425e-01  1.027e-01  1.387 0.165630  
## AcompmPersonne seule      -3.444e-03  1.436e-01 -0.024 0.980867  
## nbpers                   9.458e-02  3.841e-02  2.462 0.013835 *  
## Durée                     2.278e-04  5.942e-05  3.833 0.000128 *** 
## groupeGroup 2            -4.810e+00  1.988e-01 -24.195 < 2e-16 ***
## ---                        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Residual standard error: 2.032 on 4801 degrees of freedom
## Multiple R-squared:  0.7766, Adjusted R-squared:  0.7761
## F-statistic:  1669 on 10 and 4801 DF,  p-value: < 2.2e-16
```









Analyse multivariée

Modèles proposés

Comme la variable *Sinistre0* n'a pas de zéros (toutes les valeurs observées sont positives), nous l'avons modélisée avec un modèle linéaire généralisé avec une fonction de lien appropriée qui garantit que les prévisions soient également positives. Les modèles GLM (Generalized Linear Model), qui étendent le modèle gaussien à la famille de lois exponentielles, ont été "introduits en statistique par Nelder & Wedderburn (1972)", et "permettent de s'affranchir de l'hypothèse de normalité, en traitant de manière unifiée des réponses" ((1) Denuit, 2005, p. 74)

On utilise les variables sélectionnées précédemment : $\text{lm}(\text{formula} = \text{Sinistre0} \sim \text{groupe} + \text{RUC} + \text{Acompm} + \text{Durée}, \text{data} = \text{Dtrain})$

Nous avons ajustés avec des procédures de sélection de variables des modèles:

- GLM : Gamma avec fonction de lien * Logarithmique : $g(x) = \log(x)$
- * Identité : $g(x) = x$
- * Inverse : $g(x) = \frac{1}{x}$
- * Inverse carrée : $g(x) = \frac{1}{x^2}$

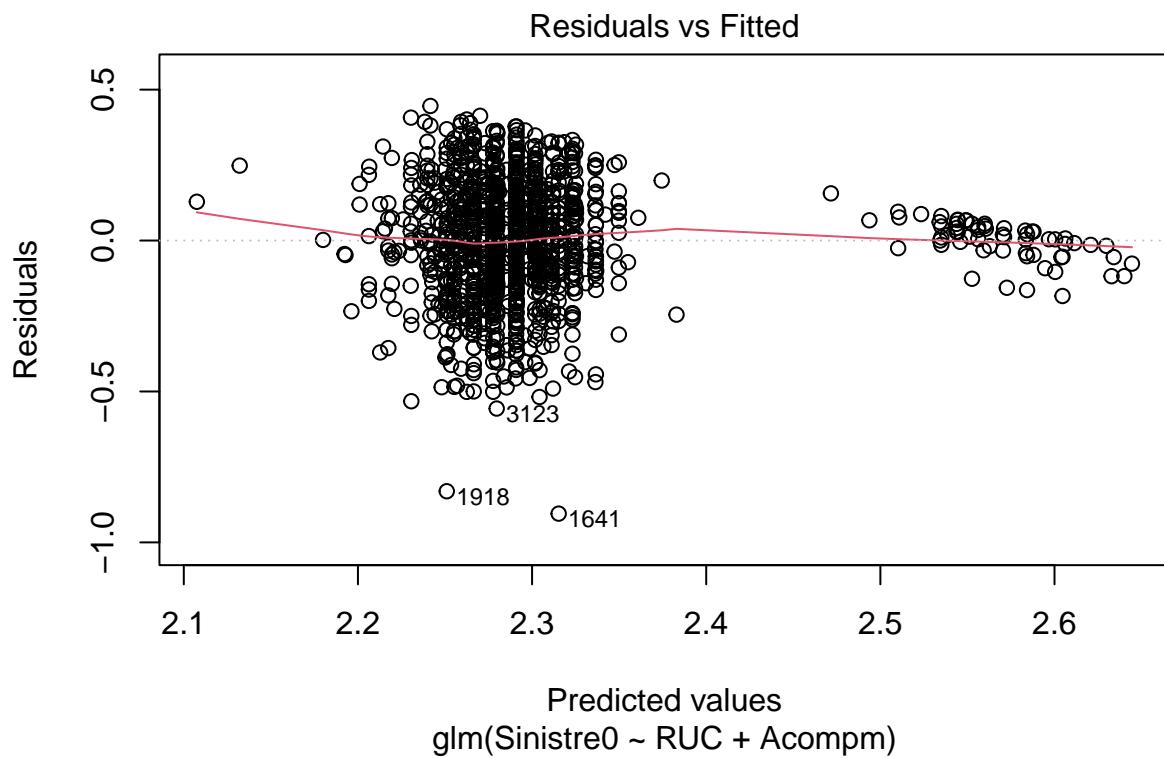
- GLM : Log normal

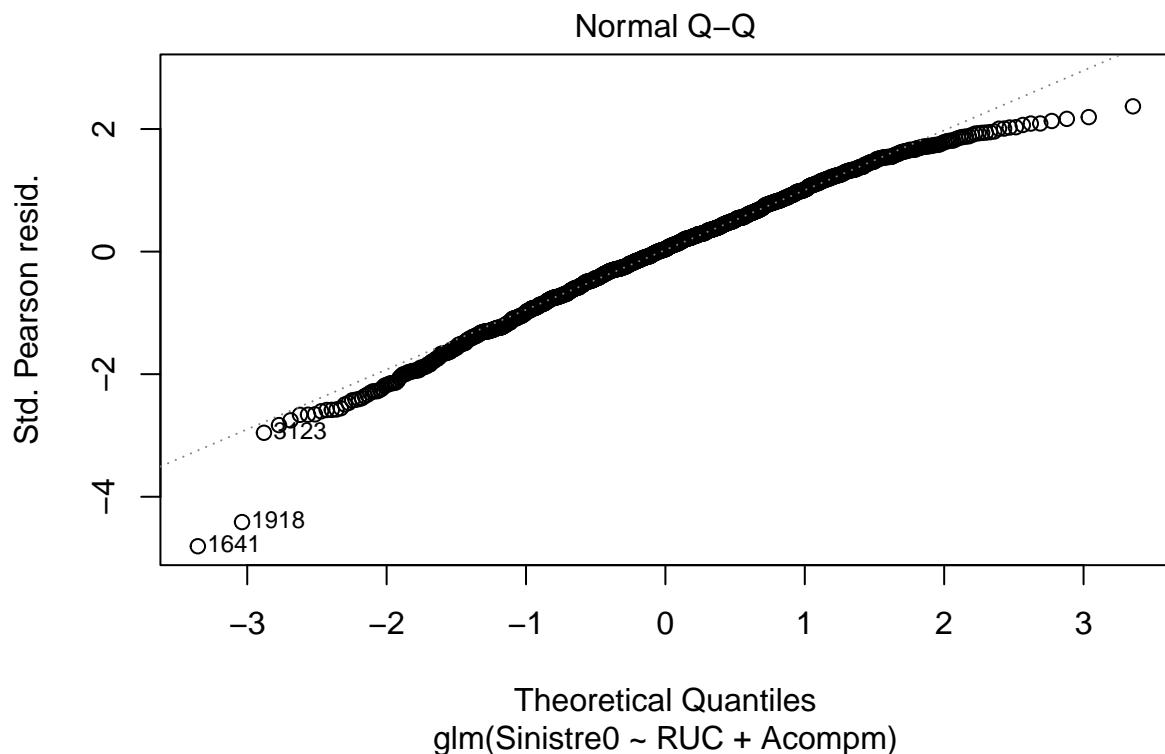
```
##
## Call:
##  $\text{glm}(\text{formula} = \text{Sinistre0} \sim \text{RUC} + \text{Acompm}, \text{family} = \text{Gamma}(\text{link} = \text{"log"}),$ 
##   data = data_Sin0_groupe1)
## 
## Deviance Residuals:
##       Min      1Q      Median      3Q      Max 
## -1.70056 -0.12288  0.00705  0.12337  0.39285 
## 
## Coefficients:
```

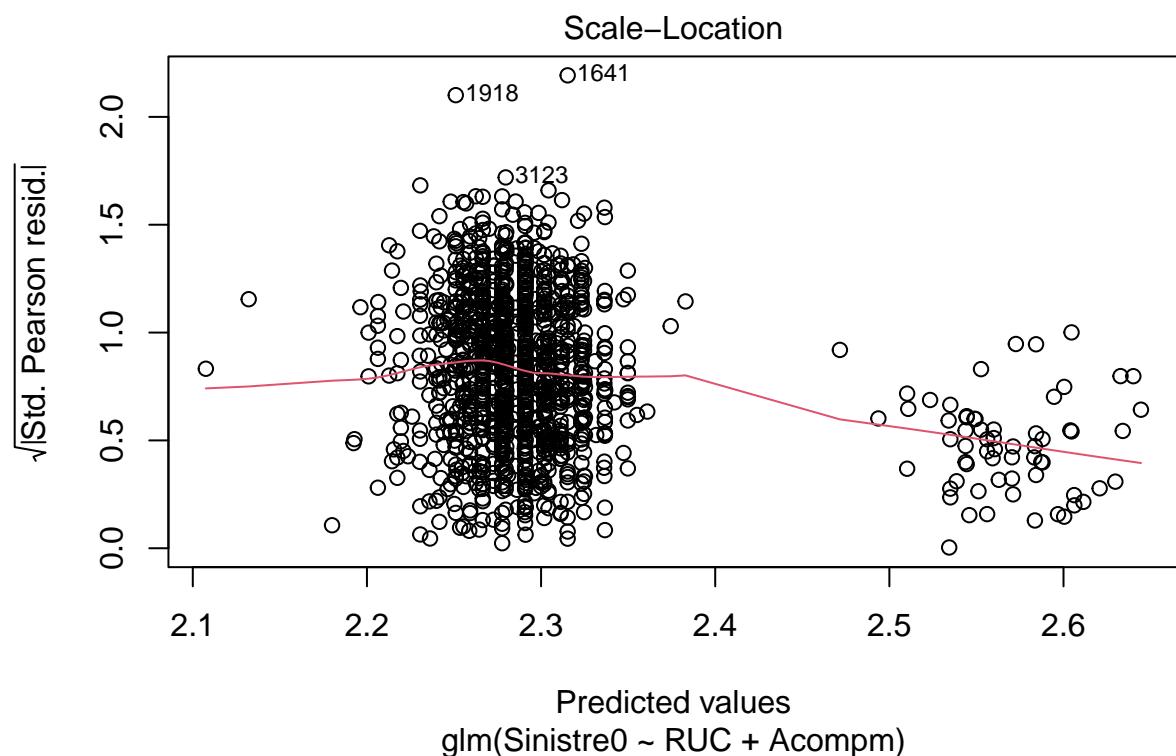
```

##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.03193   0.10335 19.660 < 2e-16 ***
## RUC                           0.06569   0.01201  5.469 5.47e-08 ***
## AcompmCouple avec enfant(s) -0.30182   0.03197 -9.439 < 2e-16 ***
## AcompmCouple sans enfant    -0.03882   0.05283 -0.735   0.463
## AcompmPersonne seule        -0.01188   0.07783 -0.153   0.879
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.035478)
##
## Null deviance: 57.852 on 1256 degrees of freedom
## Residual deviance: 51.557 on 1252 degrees of freedom
## AIC: 5297.5
##
## Number of Fisher Scoring iterations: 4
##
## Call:
## glm(formula = Sinistre0 ~ RUC + crevpp + Acompm + Durée, family = Gamma(link = "log"),
##      data = data_Sin0_groupe2)
##
## Deviance Residuals:
##    Min      1Q      Median      3Q      Max
## -0.32064 -0.08099 -0.00347  0.07696  0.34374
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.217e+00 6.695e-02 33.109 < 2e-16 ***
## RUC                           8.327e-02 8.497e-03  9.800 < 2e-16 ***
## crevpp2eme quartile       -2.490e-02 7.527e-03 -3.308 0.000948 ***
## crevpp3eme quartile       -3.758e-02 9.258e-03 -4.058 5.05e-05 ***
## crevpp4eme quartile       -3.687e-02 1.314e-02 -2.807 0.005033 **
## AcompmCouple avec enfant(s) -2.348e-01 1.717e-02 -13.674 < 2e-16 ***
## AcompmCouple sans enfant    1.600e-03 4.839e-03  0.331 0.741006
## AcompmPersonne seule       -8.218e-03 6.440e-03 -1.276 0.201958
## Durée                         1.019e-05 3.375e-06  3.019 0.002553 **
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.01281537)
##
## Null deviance: 53.967 on 3559 degrees of freedom
## Residual deviance: 45.783 on 3551 degrees of freedom
## AIC: 15303
##
## Number of Fisher Scoring iterations: 4

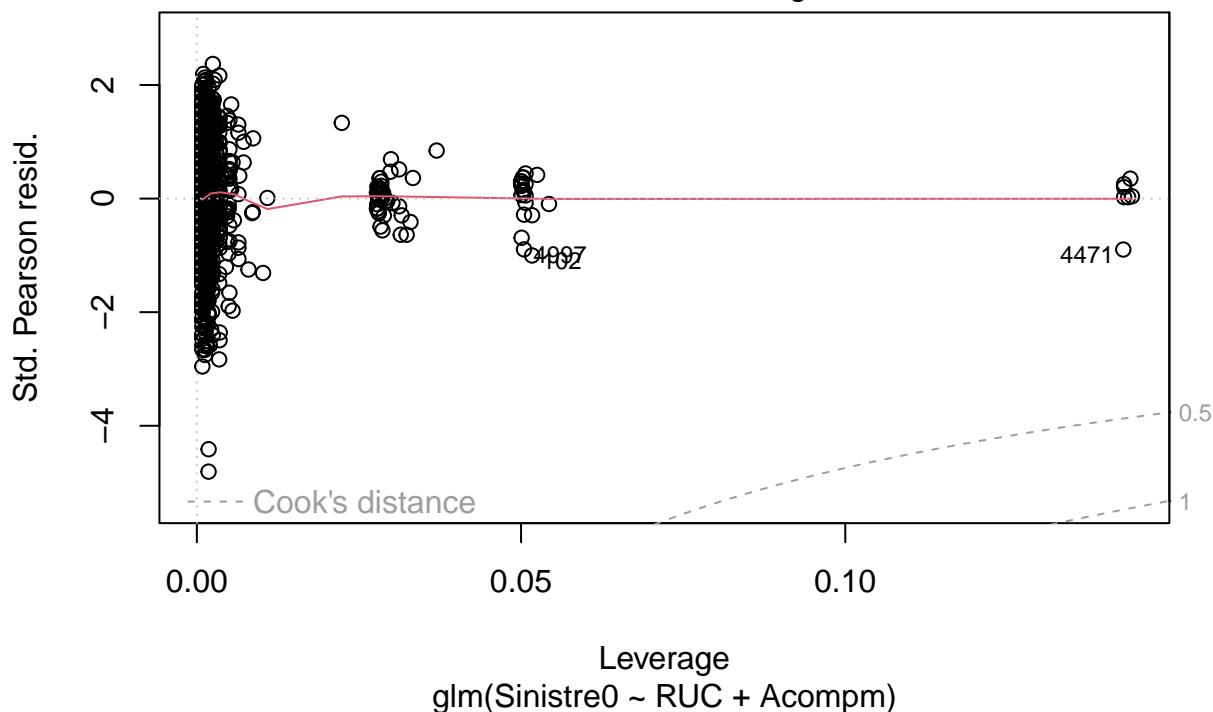
```

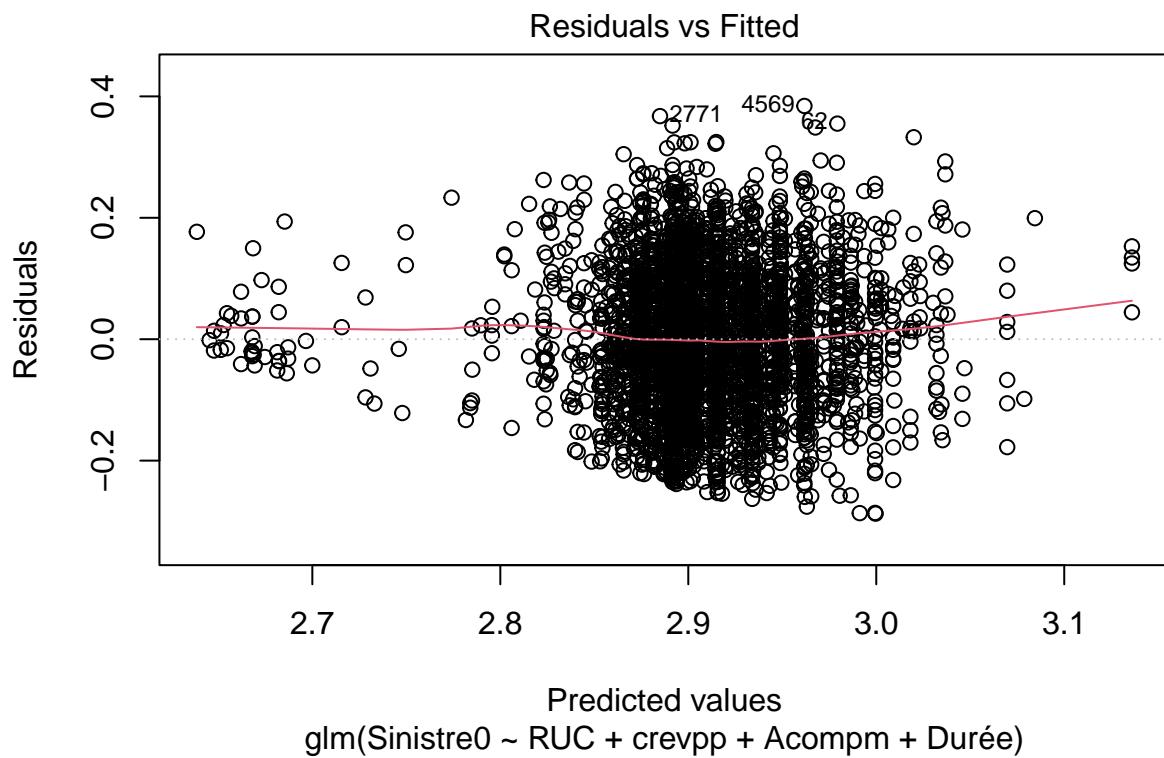


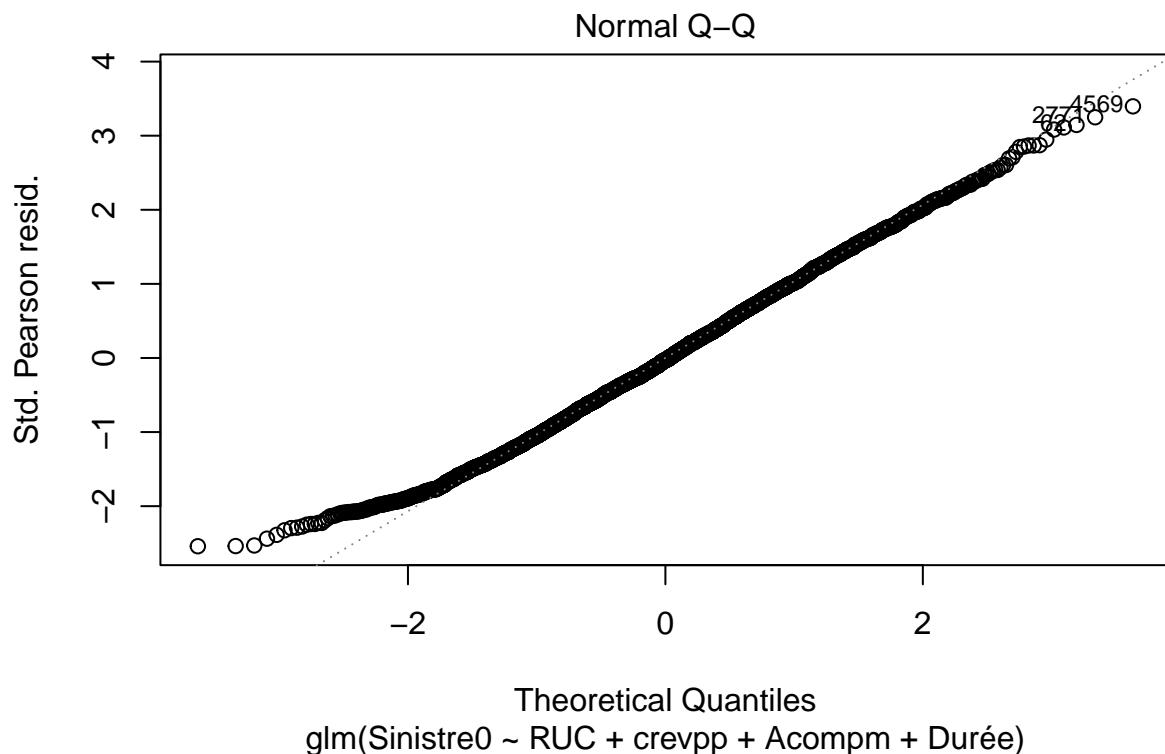


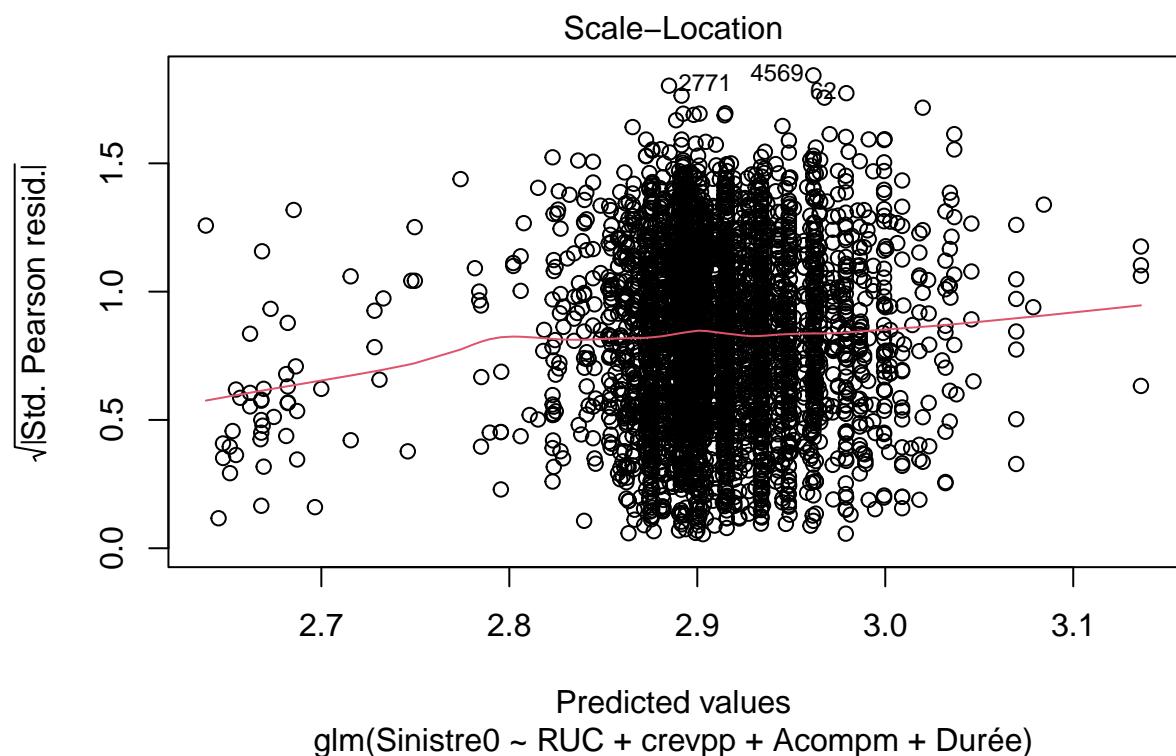


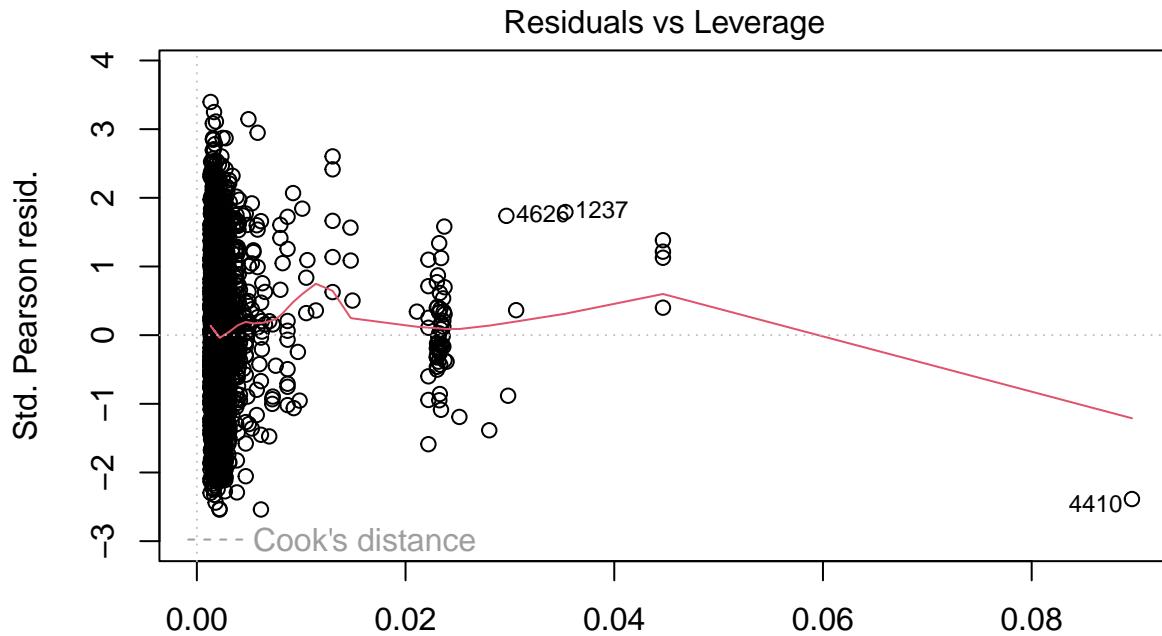
Residuals vs Leverage











Leverage
 $\text{glm}(\text{Sinistre0} \sim \text{RUC} + \text{crevpp} + \text{Acompmm} + \text{Durée})$

```
##
## Call:
## glm(formula = Sinistre0 ~ RUC + Acompmm, family = Gamma(link = "identity"),
##      data = data_Sin0_groupe1)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -1.70065  -0.12263   0.00841   0.12387   0.39390
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 7.7532    1.0457   7.414 2.24e-13 ***
## RUC                         0.6507    0.1179   5.519 4.15e-08 ***
## AcompmmCouple avec enfant(s) -3.4202    0.4145  -8.251 3.95e-16 ***
## AcompmmCouple sans enfant    -0.4164    0.6852  -0.608   0.543
## AcompmmPersonne seule        -0.1359    1.0150  -0.134   0.893
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.03546173)
##
## Null deviance: 57.852 on 1256 degrees of freedom
## Residual deviance: 51.532 on 1252 degrees of freedom
## AIC: 5296.9
##
## Number of Fisher Scoring iterations: 5
```

```

## Sinistre0 ~ RUC + Acompm

##
## Call:
## glm(formula = Sinistre0 ~ RUC + crevpp + Acompm + Durée, family = Gamma(link = "identity"),
##      data = data_Sin0_groupe2)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -0.31741 -0.08102 -0.00329  0.07685  0.34456
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)               6.301e+00  1.221e+00  5.160 2.61e-07 ***
## RUC                      1.447e+00  1.553e-01  9.315 < 2e-16 ***
## crevpp2eme quartile    -4.372e-01  1.361e-01 -3.212 0.001330 **
## crevpp3eme quartile    -6.534e-01  1.695e-01 -3.854 0.000118 ***
## crevpp4eme quartile    -5.955e-01  2.409e-01 -2.472 0.013465 *
## AcompmCouple avec enfant(s) -3.831e+00  2.536e-01 -15.109 < 2e-16 ***
## AcompmCouple sans enfant   2.644e-02  8.909e-02  0.297 0.766658
## AcompmPersonne seule      -1.722e-01  1.192e-01 -1.444 0.148692
## Durée                     2.285e-04  6.861e-05  3.331 0.000875 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.0128504)
##
## Null deviance: 53.967 on 3559 degrees of freedom
## Residual deviance: 45.900 on 3551 degrees of freedom
## AIC: 15312
##
## Number of Fisher Scoring iterations: 4

```

La distribution Gamma est souvent utilisée pour modéliser des variables qui sont strictement positives et ont une grande variabilité. Une fonction de lien appropriée est la fonction de lien exponentielle. Nous ajustons donc un modèle linéaire généralisé avec une distribution Gamma et une fonction de lien exponentielle. Ce modèle garantit que les prédictions sont également positives.

```

##
## Call:
## glm(formula = Sinistre0 ~ RUC + Acompm, family = Gamma(link = "inverse"),
##      data = data_Sin0_groupe1)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -1.7000 -0.1228  0.0076  0.1218  0.3902
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.1293508  0.0101984 12.683 < 2e-16 ***
## RUC                      -0.0064120  0.0011998 -5.344 1.08e-07 ***
## AcompmCouple avec enfant(s) 0.0266159  0.0024751 10.754 < 2e-16 ***
## AcompmCouple sans enfant   0.0034973  0.0040753  0.858   0.391
## AcompmPersonne seule       0.0009116  0.0059653  0.153   0.879
## ---

```

```

## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.0355084)
##
## Null deviance: 57.852 on 1256 degrees of freedom
## Residual deviance: 51.612 on 1252 degrees of freedom
## AIC: 5298.8
##
## Number of Fisher Scoring iterations: 4
##
## Call:
## glm(formula = Sinistre0 ~ RUC + crevpp + Acompm + Durée, family = Gamma(link = "inverse"),
##      data = data_Sin0_groupe2)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -0.32535 -0.08087 -0.00331   0.07693   0.34316
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                9.371e-02  3.607e-03 25.979 < 2e-16 ***
## RUC                         -4.712e-03  4.568e-04 -10.315 < 2e-16 ***
## crevpp2eme quartile        1.384e-03  4.146e-04   3.337 0.000856 ***
## crevpp3eme quartile        2.105e-03  5.012e-04   4.199 2.75e-05 ***
## crevpp4eme quartile        2.176e-03  7.088e-04   3.069 0.002161 **
## AcompmCouple avec enfant(s) 1.439e-02  1.166e-03 12.340 < 2e-16 ***
## AcompmCouple sans enfant    -9.562e-05  2.622e-04  -0.365 0.715405
## AcompmPersonne seule        3.938e-04  3.460e-04   1.138 0.255075
## Durée                      -4.470e-07  1.647e-07  -2.714 0.006684 **
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.01278353)
##
## Null deviance: 53.967 on 3559 degrees of freedom
## Residual deviance: 45.675 on 3551 degrees of freedom
## AIC: 15294
##
## Number of Fisher Scoring iterations: 4
##
## Call:
## glm(formula = Sinistre0 ~ RUC + Acompm, family = Gamma(link = "sqrt"),
##      data = data_Sin0_groupe1)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -1.70066 -0.12250  0.00784   0.12392   0.39356
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.76642   0.16429 16.838 < 2e-16 ***
## RUC                         0.10376   0.01886  5.501 4.57e-08 ***
## AcompmCouple avec enfant(s) -0.50795   0.05754 -8.827 < 2e-16 ***

```

```

## AcompmCouple sans enfant      -0.06378    0.09512   -0.671    0.503
## AcompmPersonne seule        -0.02033    0.14054   -0.145    0.885
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.03546828)
##
## Null deviance: 57.852  on 1256  degrees of freedom
## Residual deviance: 51.541  on 1252  degrees of freedom
## AIC: 5297.1
##
## Number of Fisher Scoring iterations: 4
##
## Call:
## glm(formula = Sinistre0 ~ RUC + crevpp + Acompm + Durée, family = Gamma(link = "sqrt"),
##      data = data_Sin0_groupe2)
##
## Deviance Residuals:
##       Min      1Q      Median      3Q      Max
## -0.31903 -0.08085 -0.00317  0.07690  0.34413
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.836e+00  1.433e-01 19.793 < 2e-16 ***
## RUC                         1.739e-01  1.820e-02  9.554 < 2e-16 ***
## crevpp2eme quartile       -5.233e-02  1.601e-02 -3.268  0.00109 **
## crevpp3eme quartile       -7.861e-02  1.983e-02 -3.963 7.53e-05 ***
## crevpp4eme quartile       -7.459e-02  2.817e-02 -2.648  0.00813 **
## AcompmCouple avec enfant(s) -4.742e-01  3.297e-02 -14.381 < 2e-16 ***
## AcompmCouple sans enfant    3.254e-03  1.038e-02  0.313  0.75405
## AcompmPersonne seule       -1.880e-02  1.386e-02 -1.356  0.17505
## Durée                      2.417e-05  7.614e-06  3.174  0.00152 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.01283249)
##
## Null deviance: 53.967  on 3559  degrees of freedom
## Residual deviance: 45.841  on 3551  degrees of freedom
## AIC: 15307
##
## Number of Fisher Scoring iterations: 4
##
## Call:
## glm(formula = Sinistre0 ~ RUC + Acompm, family = gaussian(link = "log"),
##      data = data_Sin0_groupe1)
##
## Deviance Residuals:
##       Min      1Q      Median      3Q      Max
## -9.152  -1.183   0.085   1.266   4.178
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)

```

```

## (Intercept)          2.049209  0.100030 20.486 < 2e-16 ***
## RUC                  0.063146  0.011788  5.357 1.01e-07 ***
## AcompmCouple avec enfant(s) -0.297404  0.024357 -12.210 < 2e-16 ***
## AcompmCouple sans enfant    -0.036435  0.040070 -0.909   0.363
## AcompmPersonne seule      -0.008782  0.058589 -0.150   0.881
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 3.418751)
##
## Null deviance: 4994.4  on 1256  degrees of freedom
## Residual deviance: 4280.3  on 1252  degrees of freedom
## AIC: 5119.4
##
## Number of Fisher Scoring iterations: 4
##
## Call:
## glm(formula = Sinistre0 ~ RUC + crevpp + Acompm + Durée, family = gaussian(link = "log"),
##      data = data_Sin0_groupe2)
##
## Deviance Residuals:
##       Min      1Q Median      3Q      Max
## -5.8502 -1.4734 -0.0592  1.4510  7.4042
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.149e+00  6.675e-02 32.194 < 2e-16 ***
## RUC                  9.176e-02  8.452e-03 10.856 < 2e-16 ***
## crevpp2eme quartile -2.847e-02  7.655e-03 -3.720 0.000203 ***
## crevpp3eme quartile -4.376e-02  9.256e-03 -4.727 2.37e-06 ***
## crevpp4eme quartile -4.698e-02  1.310e-02 -3.587 0.000339 ***
## AcompmCouple avec enfant(s) -2.379e-01  2.155e-02 -11.039 < 2e-16 ***
## AcompmCouple sans enfant    1.986e-03  4.842e-03  0.410  0.681722
## AcompmPersonne seule      -7.318e-03  6.392e-03 -1.145  0.252317
## Durée                 9.475e-06  3.042e-06  3.115  0.001857 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 4.358483)
##
## Null deviance: 18265  on 3559  degrees of freedom
## Residual deviance: 15477  on 3551  degrees of freedom
## AIC: 15355
##
## Number of Fisher Scoring iterations: 4
##
##           df      AIC
## model_gamma_log_1     6 5297.451
## model_gamma_identity_1 6 5296.851
## model_gamma_inverse_1 6 5298.800
## model_gamma_sqrt_1    6 5297.070
## model_lognorm_1       6 5119.400
##
##           df      AIC

```

```

## model_gamma_log_2      10 15302.60
## model_gamma_identity_2 10 15311.72
## model_gamma_inverse_2   10 15294.17
## model_gamma_sqrt_2      10 15307.08
## model_lognorm_2         10 15354.59

```

Les cinq modèles ont quasiment le même score AIC par groupe, avec un très légèrement mieux pour *Gamma identité* dans le cas du groupe 1, et pour *Gamma inverse* dans le cas du groupe 2.

```

## Sinistre0 ~ RUC + Acompm
## Sinistre0 ~ RUC + crevpp + Acompm + Durée

```

Cependant on voit dans les résidus (“nuage” serré) qu’il reste de l’information à extraire. On régresse les carrés de ces résidus sur les variables du modèle initial.

```

##
## Call:
## lm(formula = residus_carre ~ RUC + Acompm, data = data_Sin0_groupe1)
##
## Residuals:
##       Min     1Q Median     3Q    Max
## -0.05095 -0.03829 -0.02275  0.00611  2.85177
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.044972  0.061766   0.728  0.4667
## RUC                      -0.005138  0.007179  -0.716  0.4743
## AcompmCouple avec enfant(s) 0.041233  0.019109   2.158  0.0311 *
## AcompmCouple sans enfant   0.005326  0.031571   0.169  0.8661
## AcompmPersonne seule        0.002592  0.046512   0.056  0.9556
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1126 on 1252 degrees of freedom
## Multiple R-squared:  0.006103,  Adjusted R-squared:  0.002928
## F-statistic: 1.922 on 4 and 1252 DF,  p-value: 0.1044
##
## Warning: 'rgl:::rgl.viewpoint' est obsolète.
## Utilisez plutôt 'view3d'.
## Voir help("Deprecated")
##
## Warning: 'rgl:::rgl.bg' est obsolète.
## Utilisez plutôt 'bg3d'.
## Voir help("Deprecated")
##
## Warning: 'rgl:::rgl.texts' est obsolète.
## Utilisez plutôt 'text3d'.
## Voir help("Deprecated")
##
## Warning: 'rgl:::rgl.texts' est obsolète.
## Utilisez plutôt 'text3d'.
## Voir help("Deprecated")
##
## Warning: 'rgl:::rgl.texts' est obsolète.
## Utilisez plutôt 'text3d'.
## Voir help("Deprecated")

```

```

## Warning: 'rgl::rgl.points' est obsolète.
## Utilisez plutôt 'points3d'.
## Voir help("Deprecated")

## Warning: 'rgl::rgl.lines' est obsolète.
## Utilisez plutôt 'segments3d'.
## Voir help("Deprecated")

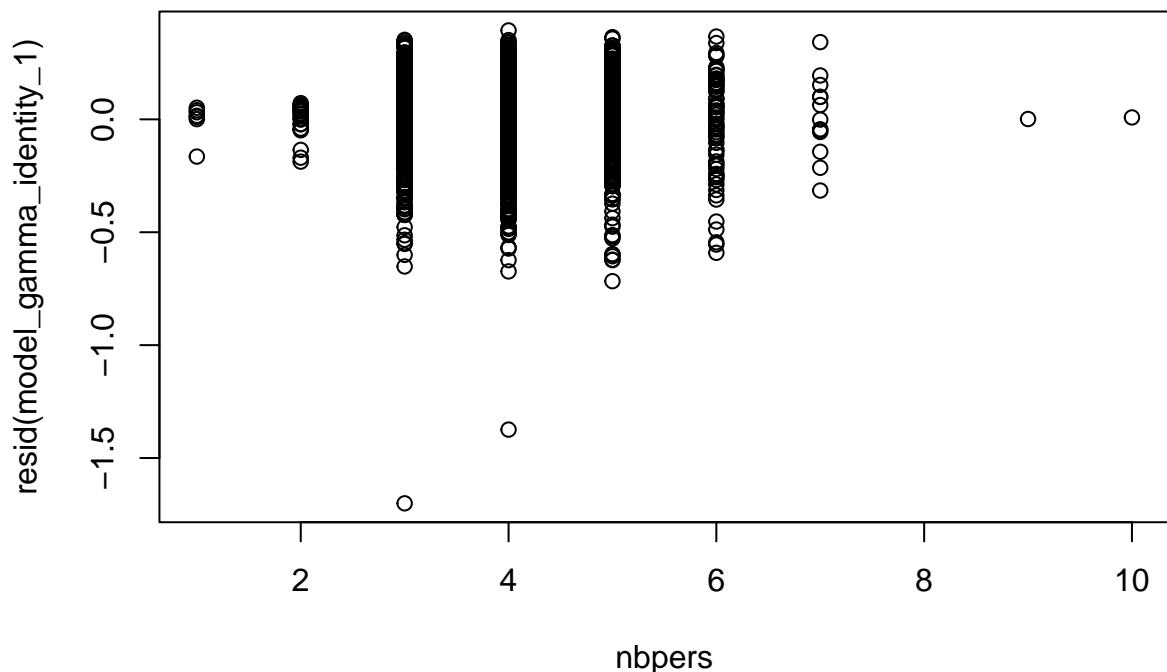
## Warning: 'rgl::rgl.lines' est obsolète.
## Utilisez plutôt 'segments3d'.
## Voir help("Deprecated")

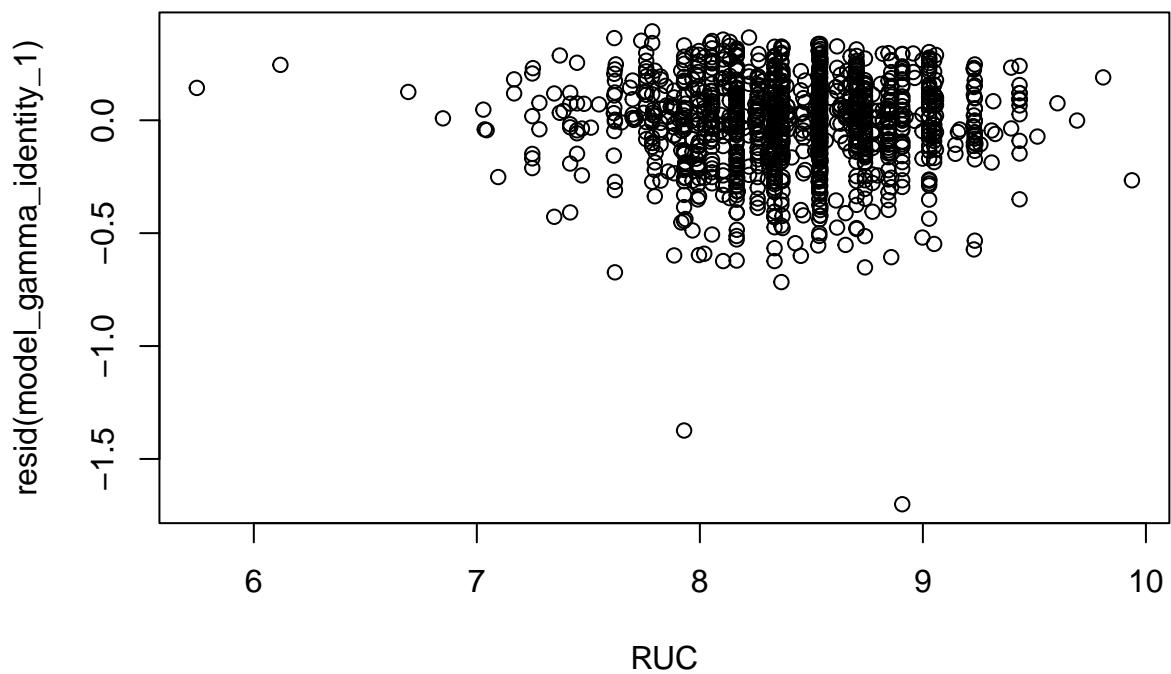
## Warning: 'rgl::rgl.texts' est obsolète.
## Utilisez plutôt 'text3d'.
## Voir help("Deprecated")

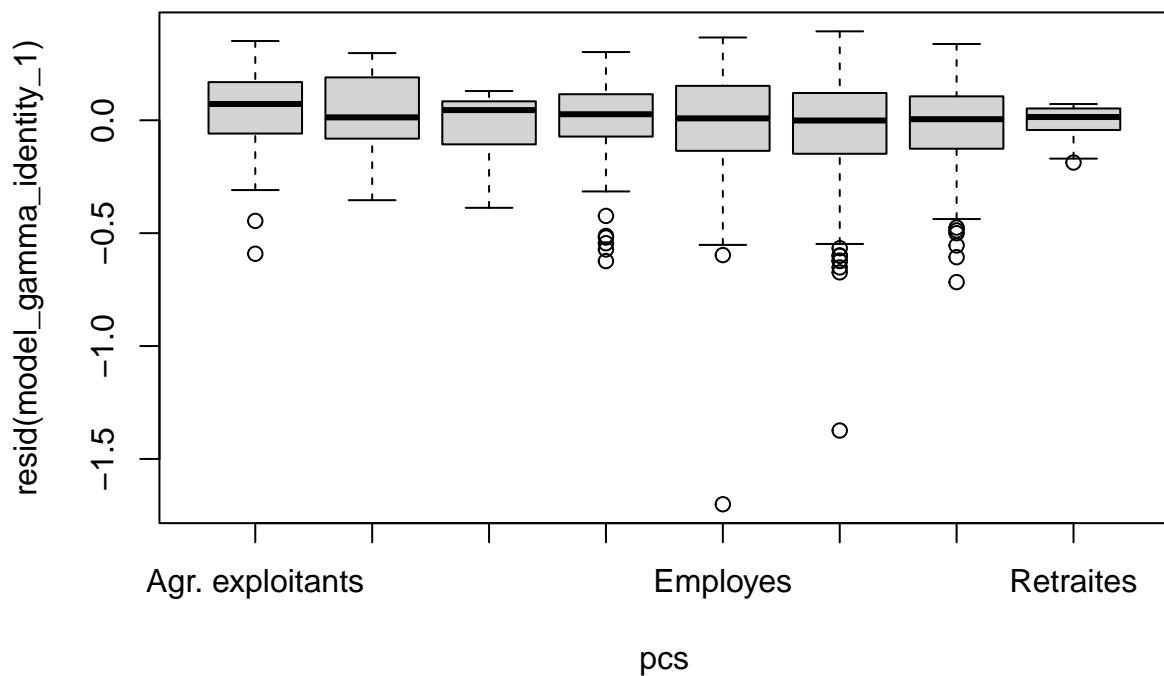
## Warning: 'rgl::rgl.texts' est obsolète.
## Utilisez plutôt 'text3d'.
## Voir help("Deprecated")

## Warning: 'rgl::rgl.texts' est obsolète.
## Utilisez plutôt 'text3d'.
## Voir help("Deprecated")

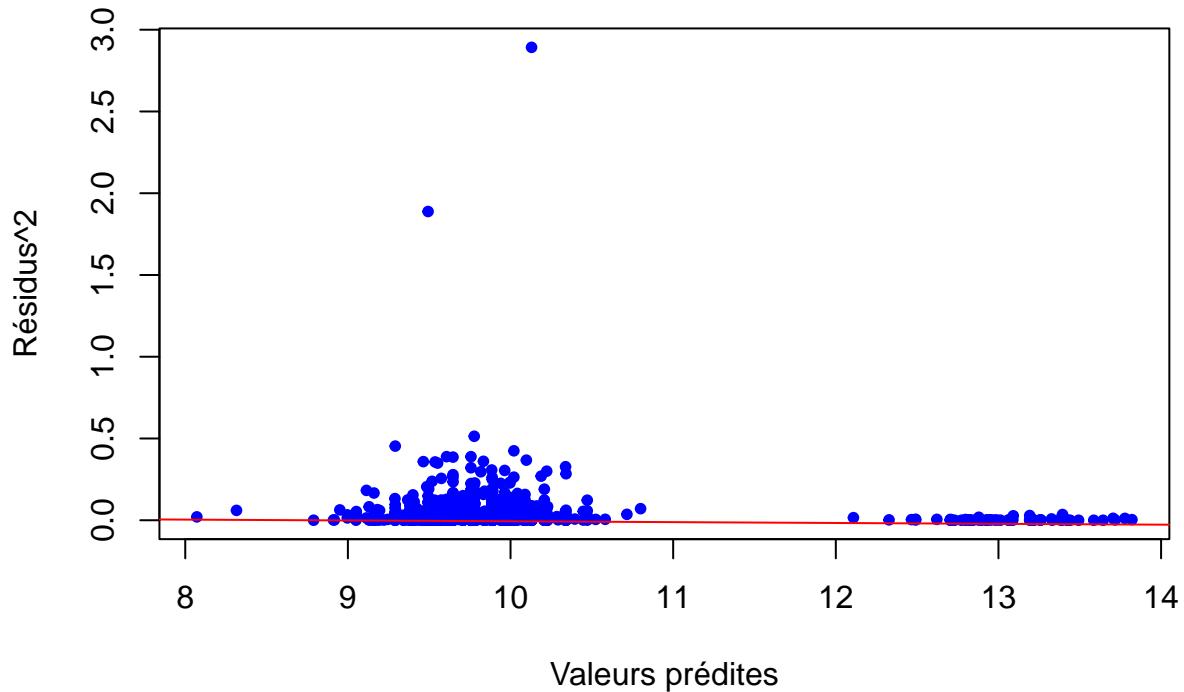
```







```
## Warning in abline(modele_het_1, col = "red"): utilisation des deux premiers des
## 5 coefficients de régression
```



Prof: , on peut essayer le modéliser par un modèle linéaire f. de: compo du ménage, catégorie d'âge, type d'habitation, nationalité, voiture ou pas, catégorie socio-prof, la région, le revenu (un d'eux!!). On ne va pas mettre la variable censure ou pas, qui nous dit si l'individu est dans la base ou pas. Il faut pas la mettre. Durée? Ca peut faire du sens, mais ça risque de compliquer un peu le modèle. Mais durée de Police est celle de la Police1, or je ne sais à quoi c lié le Sinistre0. Les enfants, ça peut être rendondant avec d'autres variables (type de ménage pê). Ca donne un premier modèle.

Plein de choses ne sont pas signif. Couple avec enfant c *très signif*; la compo du ménage, et le revenu aussi, très signif. On va faire le tri, regarder les AIC.

Et si on fait des analyses numériques, notamment des graphes, il y a des phénomènes un peu bizarres: des gros packets. Cad on a des individus dont les fitted values sont très petites; et pour d'autres, très grosses. Donc c un mélange. Il y a vraiment DEUX POPULATIONS la dédans - une certaine **hétérosced**.

Pour modéliser l'hétérosced, il y a qqch de très simple dans un 1er temps: prendre les résidus du modèle et les mettre au carré. Puis regresser sur les variables mises dans le modèles. Car si on regresse et on voit qu'il y a des variables qui sont significatives, càd que les residus dépendent des variables observées. Donc un moyen très simple, lin_modele_1.1, si on plot les residus, on va les mettre au carré + nommer () et on va les regresser (LM) sur les variables que j'ai vu qu'étaient significatives: RUC, Acompm (compo du menage). On voit que RUC est très signif - donc il y a de l'hétérosced. Donc faudra ut. les moindres carrés linéaires généralisés.

GLM: on peut essayer de modéliser. On rajoute une nouvelle var, delta: le fait que le sinistre1 soit >0. Cad j'ai une sinistre, vs. j'ai pas de sinistre. Donc j'ai une nouvelle var, que je vais modéliser par un probit: modèle lin gen, var. delta expliquée par : cs, anat, type... Fam Binomiale, avec modèle soit Probit ou logit. Cloglog (double exponentielle). Rcmd donne le modèle: glm, famille de lien binomial, avec une famille logit. Ca sort tous les estimateurs, et faudra choisir quelles sont pertinentes pour savoir si on aura un sinistre de type 1 ou pas.

On voit que la catégorie d'âge est importante en particulier pour les personnes agées; la région aussi, mais aussi la compon du ménage; et le type être proprio ou pas est légèrement signif. Faudra qu'on choisisse nous les variables.

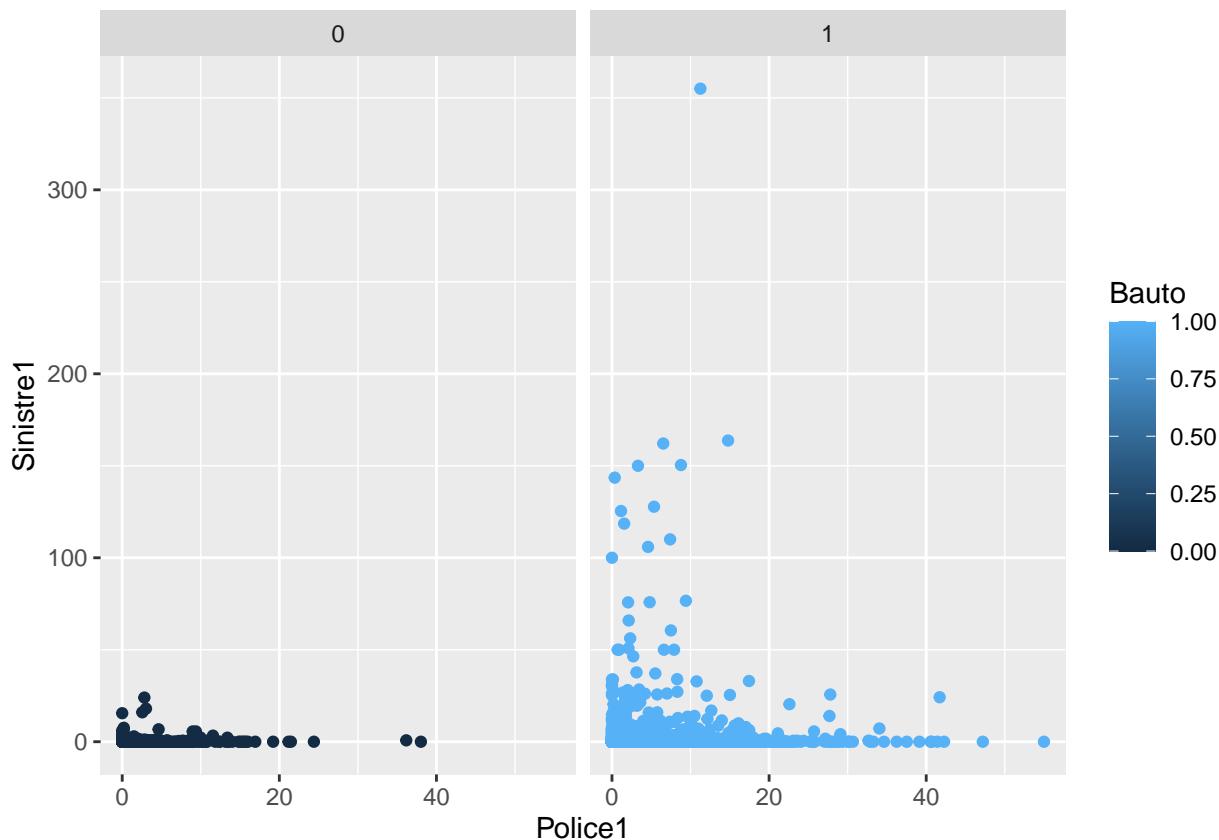
On va voir, très souvent, que les modèles linéaire, gamma, autre modèles, ne sont pas très différents, à la fin. En terme des coefficients ou des résidus. Mais s'il y a beaucoup de zéros, ça va être plus compliqué. Sur Sinistre0 on pourra essayer déjà de faire des choses.

2.3 Modélisation de Sinistre 1 ou 2 ou 3 (au moins un) notamment pour Sinistre1 à 3 on choisira entre modèle gamma combiné à probit/logit, tobit, tobit généralisé ou double hurdle pour des variables bien choisies

3.4 Modèle pour au moins un des *Sinistre1*, *Sinistre2* ou *Sinistre3*

“Modélisation du coût total, importance des contrats sans sinistre

Supposons que la variable d'intérêt Y soit le coût (total) des polices, sur un an, pour l'ensemble des polices du portefeuille. Un très grand nombre de polices n'ayant pas de sinistre, la variable Y sera alors nulle pour la plupart des observations. Une loi Gamma (par exemple) ne permet pas de modéliser ce genre de comportement (voir le Chapitre 6 du Tome 1 sur la différence entre le modèle collectif et le modèle individuel). La loi de Tweedie permet de prendre en compte ce genre de comportement, en rajoutant une mesure de Dirac en 0 à une loi de probabilité de support IR+. La loi de Y est alors une loi Poisson composée,” (Charpentier, p. 98)



Ce graphique montre qu'on a plus de sinistres indemnisés parmi les souscripteurs de Police1 qui possèdent une ou plusieurs voitures.

3.5 Modèle pour le prix de Police 1 ou 2 ou 3 (au moins un)

3.6 Modèle retenu au final

Le choix du modèle retenu au final et les critères choisis devront être justifiés.

IV regressions

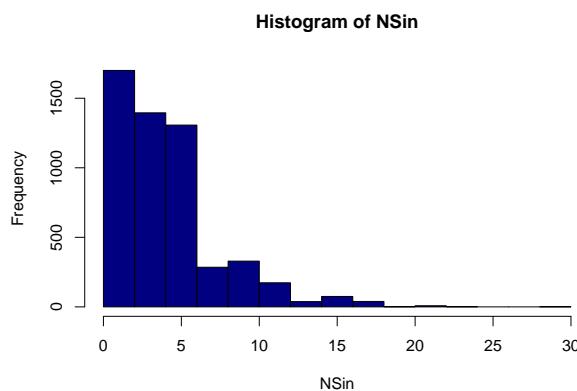
The four kinds of variables in IV

- Y = outcome variables
- X = endogenous, causal variable(s)
- Z = instrument(s): doivent être exogènes, c'ad leur influence sur Y se fait seulement via leur influence sur X, la var endogene
- W = any exogenous variables not including instruments

4. Modélisation pour les prix : le nombre de sinistres et la tarification des nouveaux arrivants

4.1 Modèle pour le nombre de sinistres, NSin

Les modèles de comptage sont utilisés en assurance pour estimer le nombre de sinistres dans une période donnée en fonction des caractéristiques du risque.



4.2 Méthode de tarification pour les nouveaux arrivants

On a deux types de modèles pour la tarification :

- *tarification a priori* : pour une nouvelle police d'assurance souscrite, nous ne savons pas quelles garanties ont été souscrites, et connaissons uniquement les caractéristiques du ménage qui a souscrit le contrat. Concrètement, nous n'utiliserons pas les variables *Police*.
- *tarification a posteriori* : nous savons ici quelles garanties ont été souscrites, et le prix payé pour celles-ci. On souhaite savoir le coût estimé pour l'assureur de ce ménage. Ce modèle est différent car il s'avère qu'une plus grande couverture en assurance est associée à des coûts plus importants pour l'assureur. Ces modèles sont plus compliquées car on aura un souci d'endogénéité entre les variables.

5. Estimation des durées

En assurance, les modèles de survie sont utilisées pour estimer la durée d'un événement, nommé "temps de survie", comme par exemple la durée de d'un contrat (notre cas ici) ou le temps avant la survenance d'un

sinistre.

5.1 Estimateur de Kaplan-Meier

5.2 Modèle de Cox

6. Références

- (1) Denuit, M. et Charpentier, A. - "Mathématiques de l'assurance non-vie", tome II, Economica, 2005.