Projet M2 Statistique des assurances

2025-04-16

1. Analyse descriptive

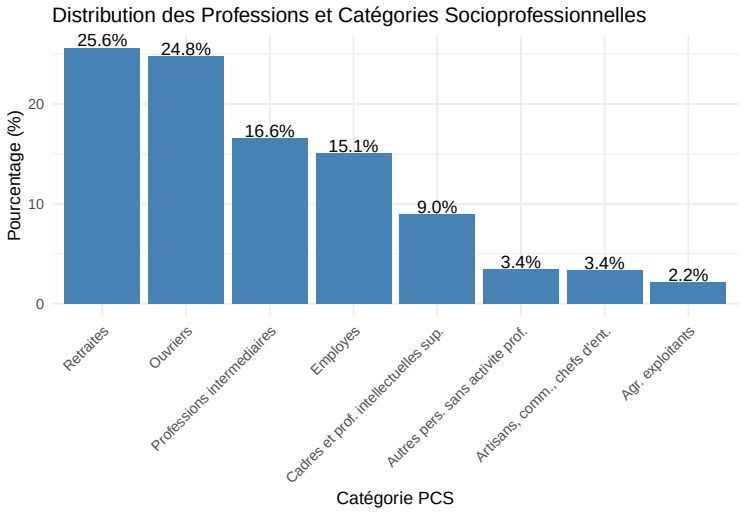
1.1 État du DataFrame

Le dataframe contient 5352 observations, aucune manquante, et 27 variables incluant des caractéristiques socio-démographiques (pcs, cs, reves), des informations sur l'habitation (region, habi, Ahabi), des caractéristiques du ménage (agecat, Acompm, nbpers) ainsi que des données d'assurance (Sinistre0, Sinistre1, Sinistre2, Sinistre3, Police1, Police2, Police3).

1.2 Analyse des variables catégorielles clés

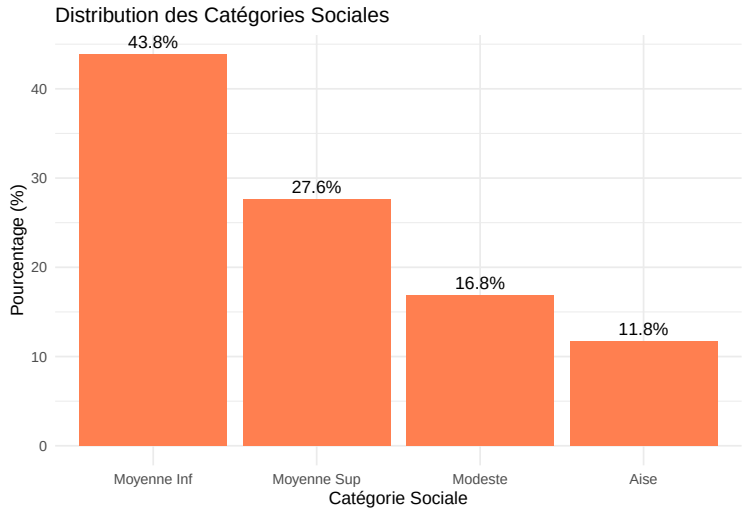
On établit par intuition les variables PCS (Professions et Catégories Socioprofessionnelles), CS (Catégories Sociales) et Ahabi (Types d'habitation) comme importante pour la tarification. On portera donc notre attention sur elles durant cette analyse exploratoire.

a) Professions et Catégories Socioprofessionnelles



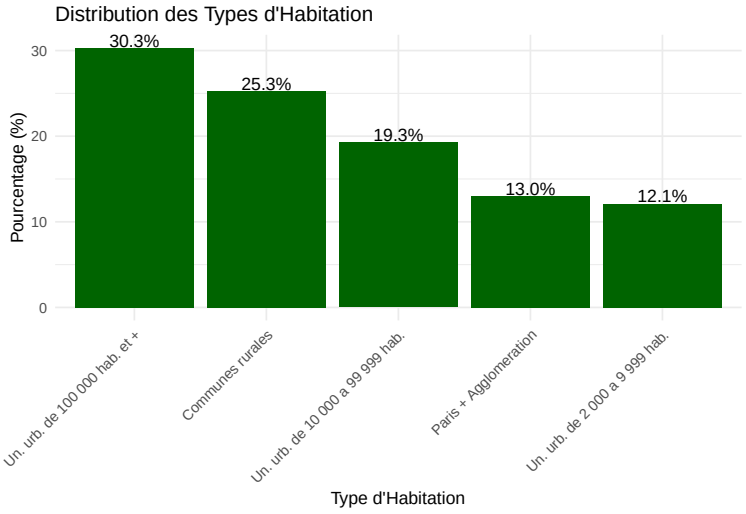
L'échantillon présente une surreprésentation des retraités (25,6%) et des employés (24,8%), suivis par les professions intermédiaires (16,6%) et les ouvriers (15,1%). Les cadres et professions intellectuelles supérieures représentent 9% de l'échantillon. Les personnes sans activités professionnelles ainsi les artisans, commercants, chefs d'entreprises représentent 3,4% alors que les Agr. exploitants représentent eux 2.2% de l'échantillon.

b) Catégories Sociales



La variable cs montre que la classe moyenne inférieure domine nettement l'échantillon avec 43,8% des individus y appartenant. Viens ensuite la classe moyenne supérieure avec 27,6% puis la classe modeste avec 16,8% et en enfin la classe aisée avec 11,8%.

c) Le Type d'habitation

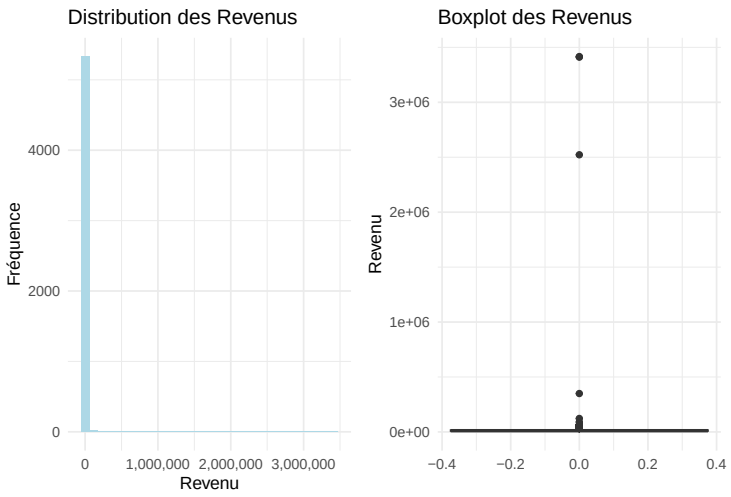


La distribution des types d'habitation suggère une population majoritairement située dans des zones urbaines. En effet on note qu'il existe une prédominance des "Unités urbaines de 100 000 habitants et plus" (30,3%), des "Unités urbaines de 10 000 à 99 999 habitants" (19,3%) et "Paris + Agglomeration" (13%).

1.3 Analyse des variables numériques

a) Revenus

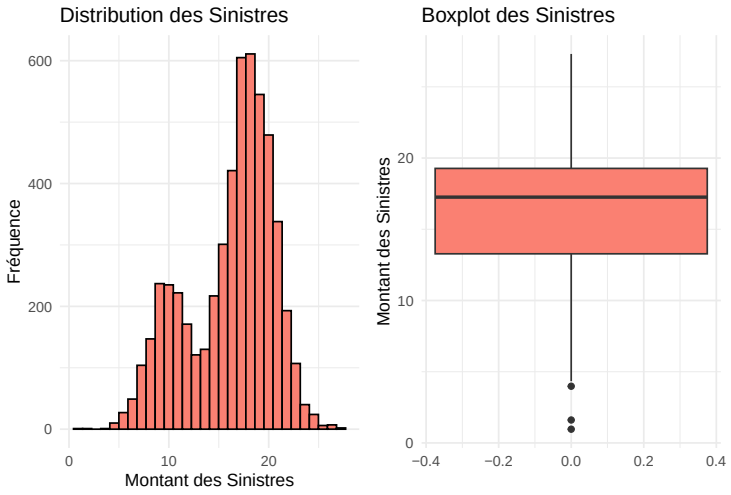
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    8500   11250   14880   16250 3416250
```



La distribution des revenus présente quelques valeurs extrêmement élevées comparées à la médiane, facilement identifiées à l'aide du boxplot, suggérant la présence d'une petite proportion de clients très fortunés dans le portefeuille.

b) Sinistres

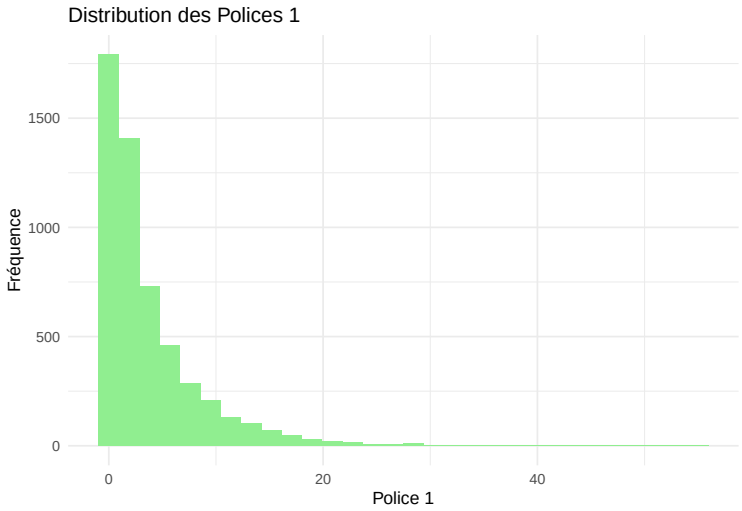
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9652 13.2769 17.2558 16.1732 19.2718 27.3074
```

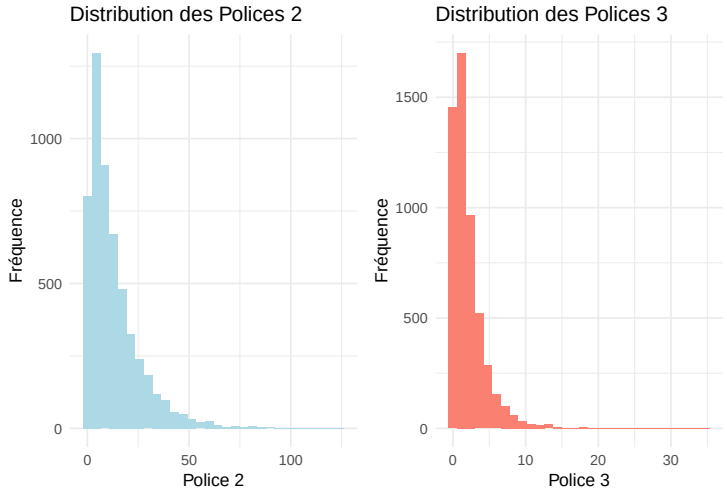


Le montant des sinistres présente une distribution relativement équilibrée.

c) Polices

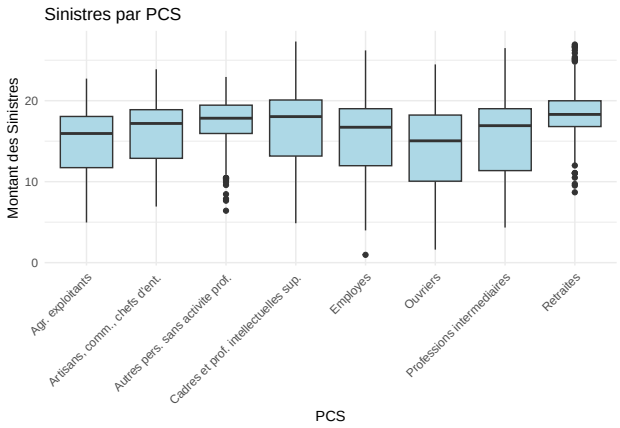
```
##    Police   Moyenne Médiane Écart_type Min     Max
## 1 Police1  3.750700    1.95   5.020503   0  54.985
## 2 Police2 13.017457    9.06  13.261083   0 124.109
## 3 Police3  2.110487    1.42   2.422230   0  34.743
```

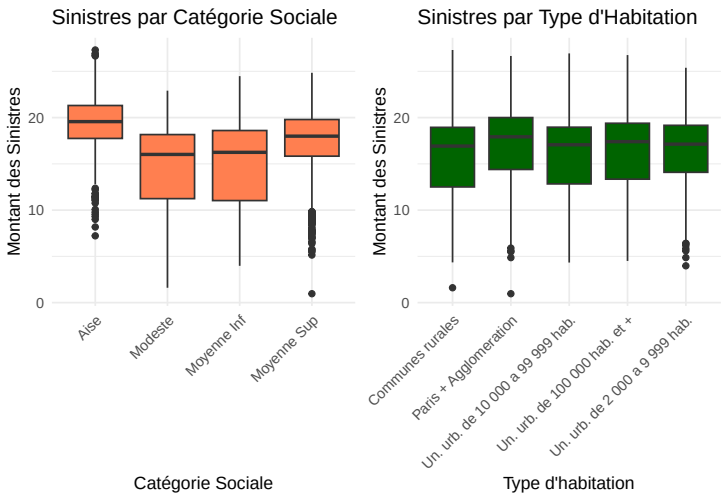




Les trois variables Police présentent des distributions asymétriques. Police2 montre l'écart-type le plus important (13,26) avec un maximum (124,11) très supérieur à la médiane (9,06), indiquant la présence probable de valeurs aberrantes.

1.4 Analyse des relations





L'analyse des relations entre la variable cible Sinistre0 et les variables catégorielles révèle des différences significatives:

Par PCS: Les cadres et professions intellectuelles supérieures et les retraités présentent les montants de sinistres les plus élevés, tandis que les ouvriers ont les sinistres les plus faibles.

Par catégorie sociale: Les ménages de la catégorie "Aisée" ont des sinistres plus importants que ceux de la catégorie "Modeste" et de la classe "Moyenne".

Par type d'habitation: Les habitants de Paris et son agglomération présentent des sinistres légèrement plus élevés aue des zones moins habitées.

```
##           reves Sinistre0 Police1 Police2 Police3  NSin Duree
## reves      1.00      0.01    0.00    0.01    0.01  0.01  0.01
## Sinistre0  0.01      1.00    0.00   -0.28   -0.10 -0.22  0.26
## Police1    0.00      0.00    1.00    0.15    0.03  0.09  0.01
## Police2    0.01     -0.28    0.15    1.00    0.27  0.30 -0.19
## Police3    0.01     -0.10    0.03    0.27    1.00  0.17 -0.08
## NSin       0.01     -0.22    0.09    0.30    0.17  1.00 -0.17
## Duree      0.01      0.26    0.01   -0.19   -0.08 -0.17  1.00
```

Les corrélations faibles, inférieures à |0,15| révèlent que les revenus ne présentent pas de lien linéaire détectable avec les sinistres ou les polices. Cela suggère que le revenu seul ne constitue pas un prédicteur direct des sinistres.

De même, la Police1 présente des corrélations inférieures à 0,15 avec toutes les variables, indiquant un impact marginal sur les autres indicateurs et suggérant qu'elle agit potentiellement de manière indépendante, ce qui pourrait être expliqué par son caractère obligatoire.

Les corrélations modérées à fortes supérieures à |0,15| révèlent des relations plus significatives.

Le montant des sinistres présente une corrélation négative avec la Police2 (-0,28), indiquant que les assurés avec des montants élevés de Police2 ont tendance à avoir des sinistres moins coûteux. Par ailleurs, Sinistre0 présente une corrélation positive avec la durée des contrats (0,26), suggérant que les contrats de longue durée sont associés à des sinistres plus élevés.

La Police2, quant à elle, présente une corrélation positive avec le nombre de sinistres (NSin) à 0,30, indiquant qu'elle est associée à plus de sinistres, mais de moindre coût. Cela renforce l'hypothèse que la Police2 pourrait couvrir des risques à haute fréquence mais faible gravité.

Enfin, la durée des contrats présente une corrélation négative avec le nombre de sinistres (-0,17).

2.Proposition de tarification

On identifie les variables revenus et police2 comme ayant des valeurs extremes pouvant influer sur le modèle donc comme pour la variable cible, sinistre0, on leur applique une transformation logarithmique.

2.1 Tarification a priori

À partir des données de caractéristiques des ménages, on construit et compare plusieurs modèles prédictifs: régression linéaire sur log(Sinistre0) avec sélection de variables par les méthodes forward, backward et stepwise ; modèles GLM Gamma avec sélection de variables forward, backward et stepwise, et évaluation via AIC/BIC et enfin une approche régularisée Lasso pour sélection de variables. Les performances sont mesurées sur l'échantillon test via MSE, R^2.

Table 1: Comparaison des performances des modèles

Modèle	MSE	R^2
Forward	0.023986	0.752725
Backward	0.023586	0.756849
Stepwise	0.023586	0.756849

Les modèles backward et stepwise sont identiques, ils présentent la MSE la plus faible et le R^2 le plus élevé, ce sont donc ceux qu'on préconise.

On retient arbitrairement le modèle stepwise pour la présentation et les tests diagnostics:

```
##
## Call:
## lm(formula = log(Sinistre0) ~ RUC + Acompm + censure, data = dat_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34226 -0.08536  0.00903  0.10136  0.59834
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.818e+00  5.835e-03 482.951   <2e-16 ***
## RUC                          1.138e-05  9.285e-07  12.261   <2e-16 ***
## AcompmCouple avec enfant(s) -6.074e-01  6.614e-03 -91.836   <2e-16 ***
## AcompmCouple sans enfant    -6.733e-03  7.132e-03  -0.944   0.3452
## AcompmPersonne seule        -7.325e-03  8.514e-03  -0.860   0.3897
## censure                      1.210e-02  6.841e-03   1.768   0.0771 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1573 on 3740 degrees of freedom
## Multiple R-squared:  0.7579, Adjusted R-squared:  0.7575
## F-statistic:  2341 on 5 and 3740 DF,  p-value: < 2.2e-16
```

Le modèle explique 75,75 % de la variance de la variable log(Sinistre0) et est globalement significatif au seuil 5% avec une p-value inférieure à 2,2e-16. Parmi les variables retenues, RUC présente un coefficient positif de 1,138e-05 et significatif au seuil de 5% avec une p-value inférieure à 2,2e-16. La variable Acompm révèle que la modalité Couple avec enfant(s) a un coefficient négatif de -0,6074 et est statistiquement significative au seuil de 5% avec une p-value inférieure à 2,2e-16, indiquant que ce type de ménage est associé à une

diminution du log(Sinistre0). En revanche, les autres modalités, Couple sans enfant et Personne seule, ne sont pas statistiquement significatives. La variable censure a un coefficient positif de 0,0121, avec une p-value de 0,0771, ce qui la place au dessus du seuil de signification 5%.

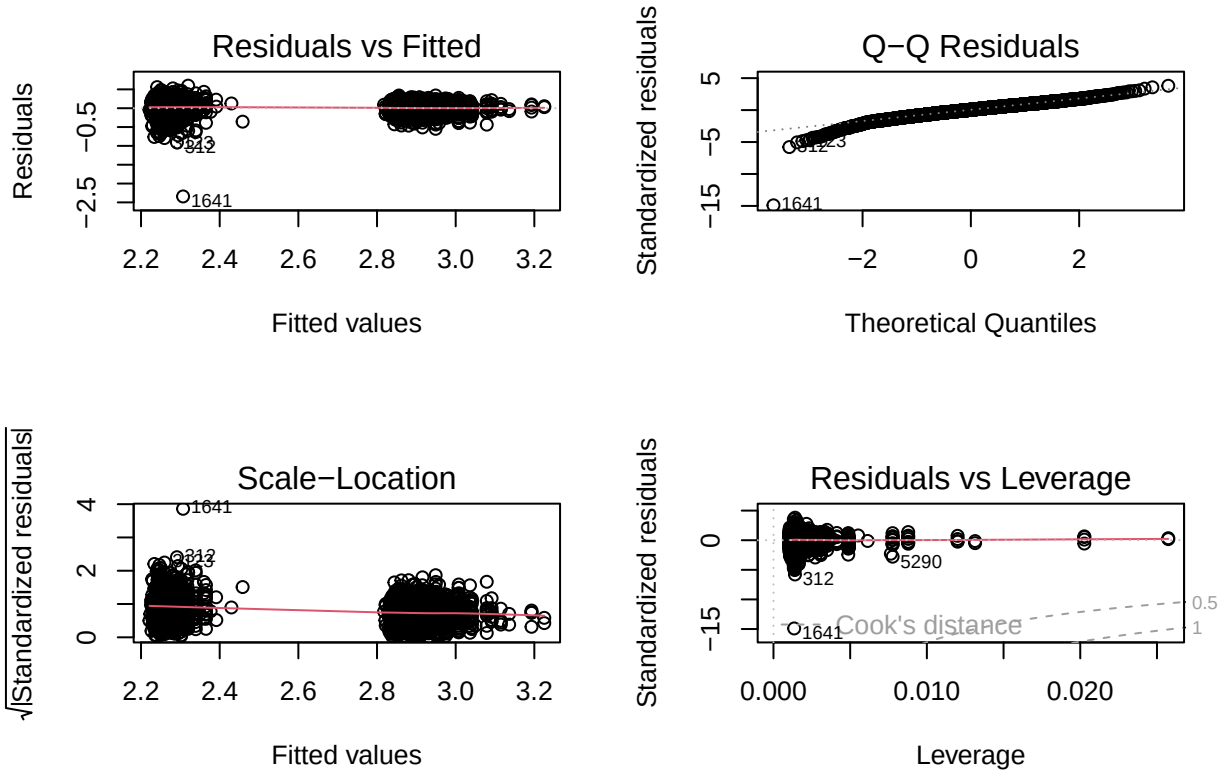
On effectue les tests diagnostics associés aux modèle:

```
##
##  studentized Breusch-Pagan test
##
## data:  modele_stepwise
## BP = 96.108, df = 5, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modele_stepwise)
## W = 0.94343, p-value < 2.2e-16
```

```
##
##  Durbin-Watson test
##
## data:  modele_stepwise
## DW = 2.016, p-value = 0.6879
## alternative hypothesis: true autocorrelation is greater than 0
```

```
##             GVIF Df GVIF^(1/(2*Df))
## RUC     1.816418  1        1.347746
## Acompm  1.216259  3        1.033168
## censure 1.672862  1        1.293392
```



Pour le test de studentized Breusch-Pagan, l'hypothèse nulle d'homoscédasticité est rejetté au seuil de 5%, en effet la p-value inférieure à 2,2e-16 confirme la présence d'hétéroscédasticité. Le test de Shapiro-Wilk vise à vérifier si les résidus suivent une distribution normale. L'hypothèse nulle est l'absence de déviation significative de la normalité. La p-value inférieure à 2,2e-16 conduit à rejeter l'hypothèse nulle au seuil 5% : les résidus ne suivent pas une distribution normale. Le Q-Q plot qui indique une adéquation globale à la distribution normale, les points suivant bien la diagonale, sauf pour la valeur extrême (le point 1641),nous permet de suggérer que l'échec du test de Shapiro-Wilk est du à celui-ci. Le test de Durbin-Watson examine l'absence d'autocorrélation dans les résidus. L'hypothèse nulle est l'absence d'autocorrélation. La p-value de 0,6879 indique que nous ne pouvons pas rejeter l'hypothèse nulle au seuil 5% : il n'y a pas d'autocorrélation détectée. Le tableau des GVIF confirme l'absence de problèmes de multicollinéarité, puisque les valeurs sont proches de 1. Le graphique des résidus contre les valeurs ajustées révèle une structure groupée des résidus, ce qui suggère que le modèle pourrait ne pas capturer une relation non linéaire. Le graphique représentant la distance de Cook souligne que le point 1641 est une observation possiblement influente avec une distance de Cook d'environ 0.05. En conclusion, bien que le modèle présente une bonne performance globale, les hypothèses d'homoscédasticité et de normalité des résidus sont rejetées.

Le point 1641 présente un nombre de sinistres un peu supérieur à la médiane mais le point n'est pas une anomalie, on décide de le conserver.

Afin d'addresser les problèmes d'hétéroscédasticité et de non-normalité résiduelle identifiés, on procedera à une transition vers un modèle GLM Gamma.

Table 2: Comparaison des performances des modèles GLM Gamma

Modèle	AIC	BIC	Déviance	MSE	R^2
Forward GLM Gamma	17081.56	17125.15	87.38382	4.390765	0.750685
Backward GLM Gamma	17081.56	17125.15	87.38382	4.390765	0.750685

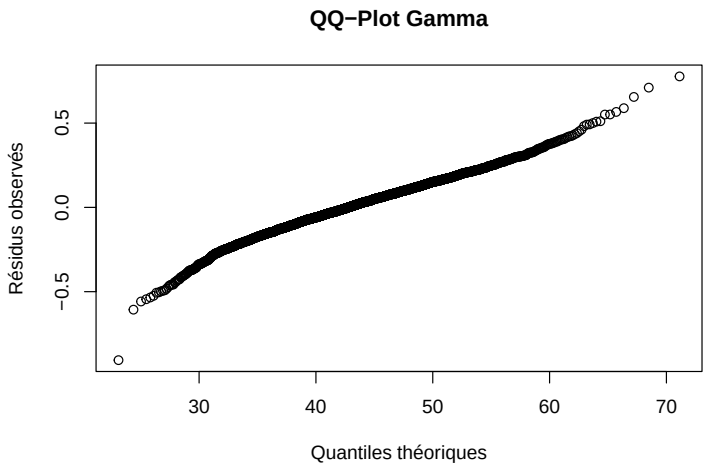
Modèle	AIC	BIC	Déviance	MSE	R^2
Stepwise GLM Gamma	17081.56	17125.15	87.38382	4.390765	0.750685

Les modèles obtenus à l'aide des méthodes stepAIC sont tous identiques. On décide arbitrairement d'étudier le modèle GLM Gamma stepwise.

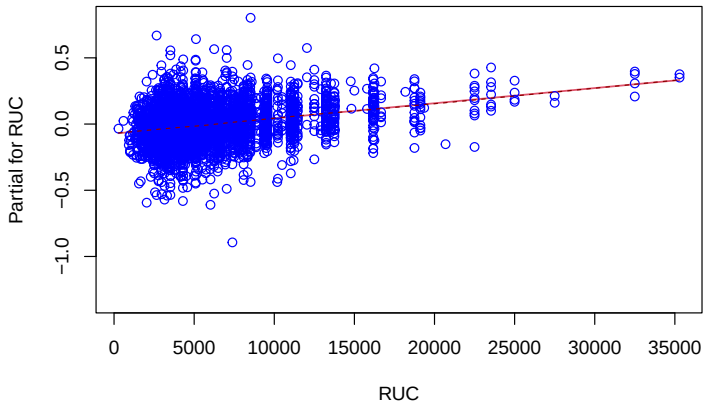
```
##
## Call:
## glm(formula = Sinistre0 ~ RUC + Acompm + censure, family = Gamma(link = "log"),
##     data = dat_train)
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.825e+00  5.491e-03 514.549   <2e-16 ***
## RUC                          1.147e-05  8.737e-07  13.125   <2e-16 ***
## AcompmCouple avec enfant(s) -5.911e-01  6.224e-03 -94.969   <2e-16 ***
## AcompmCouple sans enfant    -5.984e-03  6.711e-03  -0.892   0.3727
## AcompmPersonne seule        -7.451e-03  8.012e-03  -0.930   0.3524
## censure                      1.136e-02  6.437e-03   1.764   0.0778 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.02190342)
##
##     Null deviance: 336.313  on 3745  degrees of freedom
## Residual deviance:  87.384  on 3740  degrees of freedom
## AIC: 17082
##
## Number of Fisher Scoring iterations: 4
```

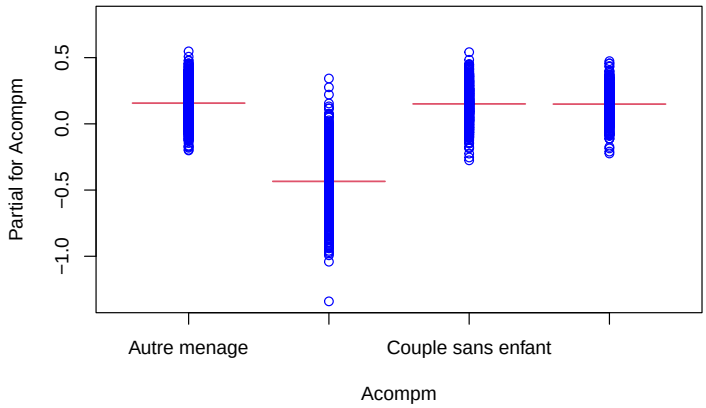
Le modèle ne présente que deux variables, RUC et AcompmCouple avec enfant(s), avec une p-value inférieure à 5%. Ce sont les seules qui sont signififcatives. La déviance du modèle résiduelle du modèle est 87.384 ce qui met en avant un bon ajustement le rapport déviance - degrés de liberté étant d'environ 0.023.

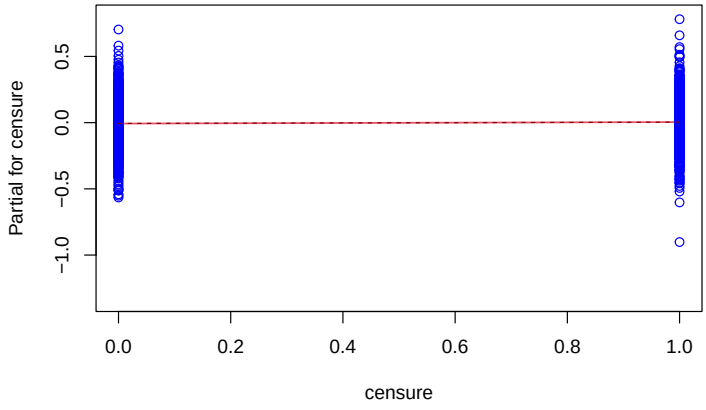
On vérifie ses hypothèses :



Sur le Q-Q Plot avec les résidus de Pearson les points suivent la diagonale, la distribution Gamma est adaptée.





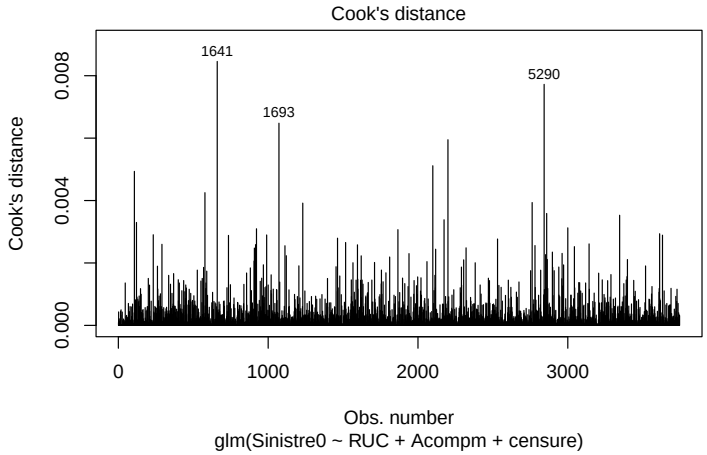


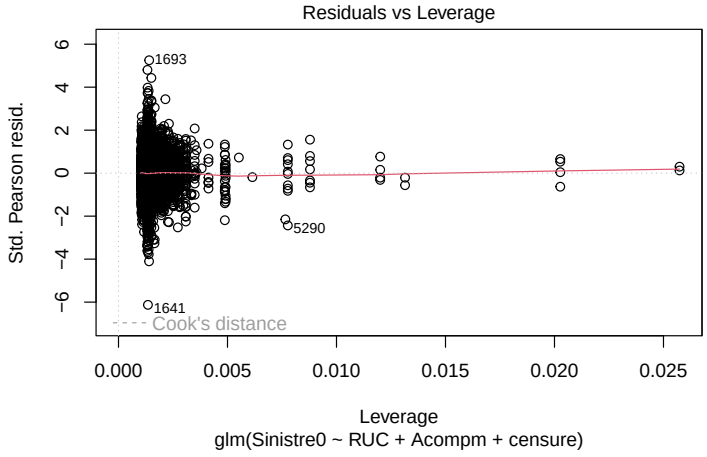
Les résidus partiels mettent en avant la linéarité de la courbe, le lien log est valide.

```
## Statistique de test Cameron & Trivedi : 315.1727
```

```
## p-value : 1.630885e-70
```

Test de Cameron & Trivedi a une p-value inférieure à 0.05, on conclue donc à la présence d'hétéroscédasticité au seuil 5%. Cependant celle-ci est intrinsèque au modèle Gamma et attendue.





Aucun point avec une distance de Cook supérieure à 1. On conclut à l'absence de points influents.

On essaie un GLM avec régularisation Lasso pour distribution Gamma approximer par une transformation logarithmique :

```
##
## MSE: 4.502466
```

```
## R²: 0.7443425
```

Ce modèle ne performe pas mieux que les modèles précédents.

En conclusion on préconise le modèle GLM Gamma StepAIC. En effet il corrige les problèmes rencontrés avec le modèle linéaire, et ses diagnostics ne révèlent aucun problème majeur, tout en présentant des performances similaires en termes de MSE et de R^2.

2.2 Tarification a posteriori

On propose un modèle de régression linéaire à variables instrumentales (IV) afin d'évaluer l'impact des variables Police sur la variable dépendante log(Sinistre0), en tenant compte de la potentialité d'une endogénéité des variables Police.

Les coefficients des variables pcs sont tous proches de zéro et non significatifs, suggérant que la profession n'a pas d'impact significatif sur le log(Sinistre0). Le coefficient positif de RUC (8,873e-06) est significatif (p-value = 0,000641) au seuil 5%, indiquant que RUC a une relation positive avec le log(Sinistre0). Les catégories de cs (Catégorie socioprofessionnelle), log(reves), crevpp (Catégorie de revenu par personne), Ahabi (Type d'habitat), Atyph (Type de proprietary), agecat (Catégorie d'âge), nbpers (Nombre de personnes dans le ménage), Anat (Nationalité), Bauto (Possession de véhicule), Nbadulte (Nombre d'adultes dans le ménage), Duree et censure ont tous des coefficients proches de zéro et non significatifs. La variable Acompm (Composition du ménage) montre que "Couple avec enfant(s)" a un coefficient négatif significatif au seuil de 5% (-6,091e-01, p-value < 2e-16), indiquant que ce type de ménage est associé à une diminution du log(Sinistre0). Les autres modalités de Acompm ne sont pas significatives. Le coefficient de NSin (Nombre de sinistres précédents) est proche du seuil de signification (p-value = 0,066462). Tous les coefficients des variables Police (Police1, log(Police2), Police3) sont non significatifs.

Les tests de diagnostic montrent que le test de Wu-Hausman a une p-value de 0,413, ce qui est supérieur à 0,05. Cela signifie qu'il n'y a pas de preuve suffisante pour rejeter l'hypothèse nulle d'exogénéité des variables Police. Autrement dit, les variables Police ne semblent pas être endogènes dans ce modèle. Le test de Sargan a une p-value de 0,930, ce qui est supérieur à 0,05, suggérant que les instruments utilisés (region et habi) ne sont pas corrélés avec l'erreur du modèle et sont donc valides.

En conclusion, les résultats des tests diagnostiques indiquent qu'il n'y a pas de preuve d'endogénéité pour les variables Police (p-value du test de Wu-Hausman = 0,413). De plus, les tests de Sargan confirment la validité des instruments (p-value = 0,930). Cela nous permet de proposer un modèle a posteriori linéaire étendu incluant les variables Police.

On compare différents modèles:

Table 3: Comparaison des performances des modèles a posteriori

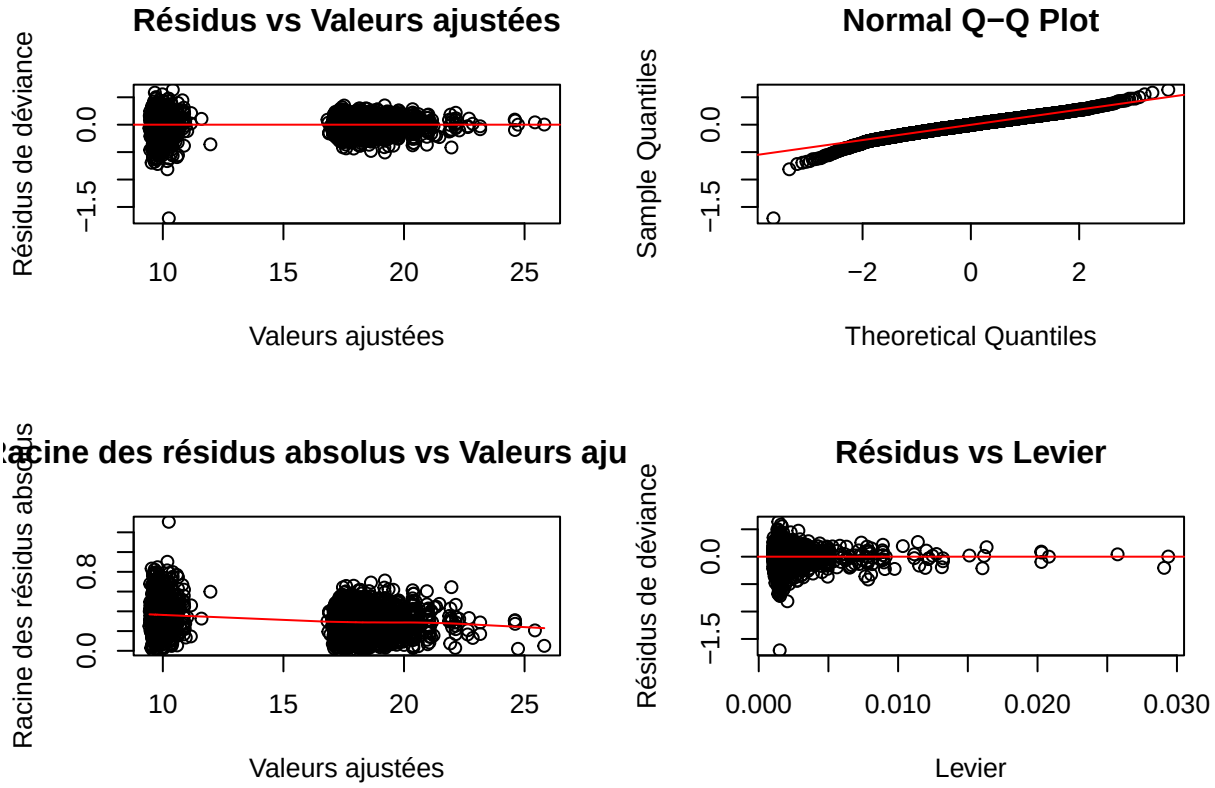
	MAE	MSE	RMSE	R_squared	AIC	BIC
Modèle Linéaire Complet	1.692183	4.455078	2.110705	0.747033	-3174.195	-2925.057
Modèle StepAIC	1.688181	4.436672	2.106341	0.748078	-3220.295	-3170.467
Modèle GLM Gamma	1.683447	4.404143	2.098605	0.749925	17127.384	17376.521
Modèle GLM Gamma StepAIC	1.680565	4.392196	2.095757	0.750604	17080.690	17130.517

On choisit le modèle GLM Gamma StepAIC, en effet, d'une part, ce modèle présente les meilleures performances parmi ceux évalués, avec les plus faibles valeurs de MAE (1,680565), MSE (4,392196) et RMSE (2,095757), ainsi que la plus forte valeur de R carré (0,7506038), témoignant d'une meilleure capacité prédictive. D'autre part, le processus de sélection stepwise a permis de réduire le nombre de variables explicatives aboutissant à un modèle plus simple.

```
##
## Call:
## glm(formula = Sinistre0 ~ RUC + Acompm + censure + Police1, family = Gamma(link = "log"),
```

```
##     data = dat_train)
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.822e+00  5.894e-03 478.681   <2e-16 ***
## RUC                           1.146e-05  8.734e-07  13.126   <2e-16 ***
## AcompmCouple avec enfant(s) -5.907e-01  6.225e-03 -94.890   <2e-16 ***
## AcompmCouple sans enfant    -5.211e-03  6.724e-03  -0.775   0.4384
## AcompmPersonne seule        -6.002e-03  8.054e-03  -0.745   0.4562
## censure                       1.155e-02  6.436e-03   1.794   0.0729 .
## Police1                       8.457e-04  4.837e-04   1.748   0.0805 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.02189052)
##
##     Null deviance: 336.313  on 3745  degrees of freedom
## Residual deviance:  87.317  on 3739  degrees of freedom
## AIC: 17081
##
## Number of Fisher Scoring iterations: 4
```

Les variables retenues sont RUC, Acompm, censure et Police1. RUC présente un coefficient positif significatif au seuil 5% (1,146e-05 et p-value $<$ 2e-16), indiquant que son augmentation est associée à une hausse des sinistres. La variable Acompm montre que les couples avec enfants ont un coefficient négatif significatif au seuil 5% (-0,5907 et p-value $<$ 2e-16), suggérant des sinistres moins élevés pour ces ménages. Censure et Police1 ont des coefficients positifs plûtot faible (respectivement 0,01155 et 8.457e-04) et non significatifs au seuil 5% (p-value = 0.0729 et 0.0805 respectivement).




```
## Analysis of Deviance Table
##
## Model 1: Sinistre0 ~ RUC + Acompm + censure + Police1
## Model 2: Sinistre0 ~ RUC + Acompm + censure + Police1 + I(preds^2) + I(preds^3)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3739     87.317
## 2      3737     87.160  2  0.15709  0.02752 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Deviance Table
##
## Model 1: Sinistre0 ~ RUC + Acompm + censure + Police1
## Model 2: Sinistre0 ~ RUC + Acompm + censure + Police1 + resid_police1 +
##     resid_police2 + resid_police3
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3739     87.317
## 2      3736     87.305  3 0.012009   0.9082
```

Les diagnostics visuels et statistiques du modèle GLM Gamma StepAIC révèlent plusieurs aspects importants. Le graphique des résidus contre les valeurs ajustées montre une structure groupée des données mais sans indication claire d'hétéroscédasticité. Le Normal QQ Plot indique une adéquation raisonnable à une distribution normale. Concernant la non-linéarité, le graphique de la racine des résidus absolus contre les valeurs ajustées révèle une légère tendance, mais c'est le test RESET, avec une p-value de 0,02752, qui souligne une spécification incorrecte du lien, suggérant que la fonction de lien logarithmique pourrait ne pas

être optimale. Le graphique des résidus versus le levier indique l'absence de points influents. Le test de Wu-Hausman, avec une p-value de 0,9082, confirme l'exogénéité des variables Police.

Pour améliorer le modèle, il serait pertinent d'explorer des interactions entre variables pour mieux capturer les relations complexes, ou d'utiliser un modèle GAM afin de modéliser automatiquement les relations non linéaires. Malgré des performances globales satisfaisantes, le modèle pourrait être affiné en adressant les problèmes de spécification du lien identifiés par le test RESET.

```
## Warning: 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")

## Warning: 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
```

Table 4: Comparaison des performances des modèles a posteriori

	MAE	MSE	RMSE	R_squared	AIC	BIC
Modèle Linéaire Complet	1.692183	4.455078	2.110705	0.747033	-3174.195	-2925.057
Modèle StepAIC	1.688181	4.436672	2.106341	0.748078	-3220.295	-3170.467
Modèle GLM Gamma	1.683447	4.404143	2.098605	0.749925	17127.384	17376.521
Modèle GLM Gamma StepAIC	1.680565	4.392196	2.095757	0.750604	17080.690	17130.517
Modèle GLM Gamma avec Interactions	1.680271	4.392977	2.095943	0.750559	17084.521	17146.806
Modèle GAM	1.680625	4.392068	2.095726	0.750611	17089.913	17149.291

Le modèle GAM est avéré le plus performant parmi les modèles évalués, avec les plus faibles valeurs de MAE (1,680625), MSE (4,392068) et RMSE (2,095726), ainsi que la plus forte valeur de R carré (0,7506111).

```
##
## Family: Gamma
## Link function: log
##
## Formula:
## Sinistre0 ~ s(RUC) + s(log(reves)) + Acompm + Nbadulte + s(Police1)
##
## Parametric coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.914011   0.013560 214.898   <2e-16 ***
## AcompmCouple avec enfant(s) -0.596299   0.008100 -73.614   <2e-16 ***
## AcompmCouple sans enfant    -0.003744   0.008397  -0.446    0.656
## AcompmPersonne seule        -0.000806   0.013038  -0.062    0.951
## Nbadulte                    -0.004001   0.004129  -0.969    0.333
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                 edf Ref.df      F p-value
## s(RUC)        1.001  1.001 46.871  <2e-16 ***
## s(log(reves)) 1.001  1.002  1.939   0.164
## s(Police1)    1.532  1.904  1.620   0.148
```

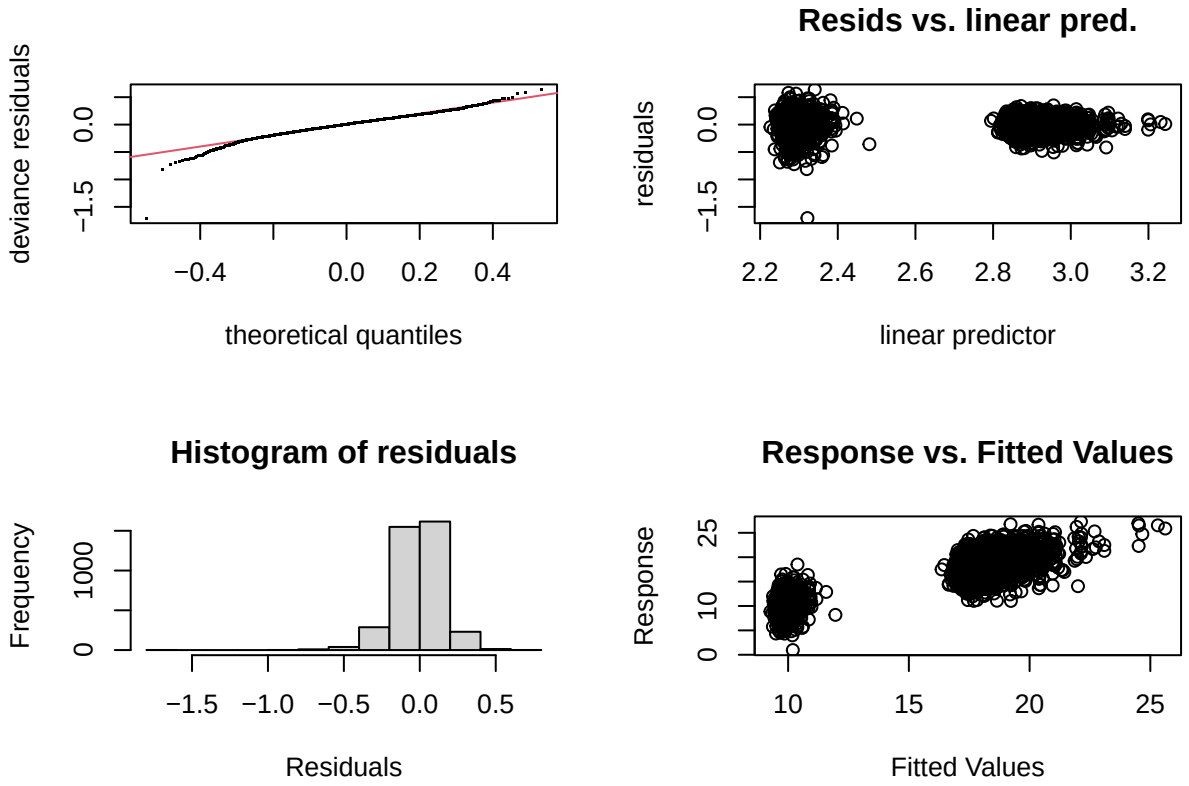
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.749   Deviance explained =   74%
## GCV = 0.023413  Scale est. = 0.021889  n = 3746
```

D'après le modèle la variable AcompmCouple avec enfant(s), avec un coefficient négatif de -0,596299 et une p-value inférieure à 2e-16, est significative au seuil de 5%. Les termes AcompmCouple sans enfant et AcompmPersonne seule n'ont pas d'impact significatif sur les sinistres, avec des p-values respectives de 0,656 et 0,951. Le coefficient de Nbadulte, -0,004001, n'est pas statistiquement significatif (p-value = 0,333).

Les termes lissés montrent que s(RUC), avec un edf de 1,001, indique une relation quasi-linéaire et est très significatif au seuil de 5% (p-value < 2e-16), ce qui signifie que l'augmentation de RUC est associée à une hausse linéaire des sinistres. Le terme s(log(reves)), lui aussi, a un edf de 1,001, suggérant une relation linéaire, avec une p-value de 0,164, indiquant un effet non significatif sur les sinistres. Le terme s(Police1), avec un edf de 1,532, indique une relation légèrement non-linéaire, avec une p-value de 0,148, un effet non significatif sur les sinistres.

Le modèle a un R carré ajusté de 0,75.

On vérifie ses hypothèses:




```
##
## Method: GCV   Optimizer: outer newton
## full convergence after 6 iterations.
## Gradient range [-9.378801e-09,1.512364e-10]
## (score 0.023413 & scale 0.0218893).
```

```
## Hessian positive definite, eigenvalue range [4.382411e-09,2.110912e-06].
## Model rank =  32 / 32
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##                 k'  edf k-index p-value
## s(RUC)        9.00 1.00    1.00    0.54
## s(log(reves)) 9.00 1.00    0.97    0.04 *
## s(Police1)    9.00 1.53    1.00    0.46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On teste l'hypothèse k qui vérifie si les dimensions de base (k) des splines sont suffisantes. Ici les p-values indiquent que, sauf pour s(log(reves)), les dimensions de base k des splines sont suffisantes pour capturer les relations non linéaires dans les données.

```
##
##  Shapiro-Wilk normality test
##
## data:  res_dev[sample(length(res_dev), 2000)]
## W = 0.97913, p-value < 2.2e-16
```

Le test de normalité de Shapiro-Wilk montre une déviation significative de la normalité (p-value $<$ 2,2e-16) au seuil 5%.

```
##                          df      AIC
## model_gam          9.53328 17089.91
## model_gam_inverse 10.01133 17110.96
```

La comparaison des liens confirme que le lien log est préférable, avec un AIC plus faible (17089,91) comparé au lien inverse (17110,96).

```
##               para    s(RUC) s(log(reves)) s(Police1)
## worst    0.9795673 0.9615685     0.9579137 0.05980130
## observed 0.9795673 0.8814061     0.9423738 0.03437499
## estimate 0.9795673 0.5379512     0.6187055 0.01396122
```

Le diagnostic de concurvité révèle des valeurs worst proches de 1 pour certains termes, ce qui pourrait indiquer une forte concurvité.

On préconise donc le modèle GAM.

3. Tobit, Tobit Généralisé et Double Hurdle

Pour créer ces modèles a priori expliquant Sinistre1, le choix des variables a été guidé par des considérations théoriques sur les facteurs susceptibles d'influencer les sinistres.

On a donc choisi le revenu du ménage, le nombre d'adultes dans le ménage, la possession d'un véhicule, la composition du ménage et le type d'habitat.

```
##
## Call:
## censReg(formula = Sinistre1 ~ log_reves + Nbadulte + Bauto +
##     Acompm + agecat + Ahabi, left = 0, data = dat)
##
## Observations:
##          Total  Left-censored     Uncensored Right-censored
##           5352           4085           1267              0
##
## Coefficients:
##                                     Estimate Std. error t value  Pr(> t)
## (Intercept)                         -19.0711     7.8835  -2.419  0.01556 *
## log_reves                            -0.2503     0.8415  -0.297  0.76610
## Nbadulte                              1.6315     0.5922   2.755  0.00587 **
## BautoPas de vehicule                 -1.2084     1.6041  -0.753  0.45125
## AcompmCouple avec enfant(s)           0.3534     1.3152   0.269  0.78817
## AcompmCouple sans enfant             -0.3387     1.3998  -0.242  0.80879
## AcompmPersonne seule                  0.2482     1.8970   0.131  0.89589
## agecat41-50                          -2.5052     1.1476  -2.183  0.02903 *
## agecat51-60                          -3.9261     1.3618  -2.883  0.00394 **
## agecat61-96                          -7.8720     1.3223  -5.953 2.63e-09 ***
## AhabiParis + Agglomeration            3.1766     1.4290   2.223  0.02622 *
## AhabiUn. urb. de 10 000 a 99 999 hab.   2.6252     1.2082   2.173  0.02979 *
## AhabiUn. urb. de 100 000 hab. et +    5.2437     1.0769   4.869 1.12e-06 ***
## AhabiUn. urb. de 2 000 a 9 999 hab.   0.1177     1.4220   0.083  0.93403
## logSigma                              3.0423     0.0211 144.167  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Newton-Raphson maximisation, 6 iterations
## Return code 8: successive function values within relative tolerance limit (reltol)
## Log-likelihood: -7100.459 on 15 Df
```

Le modèle Tobit met en évidence plusieurs coefficients significatifs au seuil de 5%, avec des p-valeurs inférieure à 5%. Nbadulte présente un coefficient de 1,6315, indiquant que chaque adulte supplémentaire dans le ménage est lié à une augmentation de 1,6315 unités de Sinistre1 avant censure. Les individus des catégories d'âge 51-60 et 61-96 ont respectivement 3,9261 et 7,8720 unités moins de sinistres que la catégorie de référence.Concernant Ahabi, les habitants d'habitats urbains, surtout dans les unités urbaines de plus de 100 000 habitants, montrent des taux de sinistres plus élevés. Cependant les variables, log_reves, Bauto et Acompm ne présentent pas de relation significative avec les sinistres, suggérant qu'elles ne sont pas des prédicteurs pertinents ici.

```
##
## Call:
## mhurdle(formula = Sinistre1 ~ log_reves + Nbadulte | log_reves +
##     Bauto + Acompm, data = dat, dist = "ln", h2 = FALSE, corr = FALSE)
##
## Frequency of 0:  0.76327
##
## Coefficients :
##                                Estimate Std. Error  t-value  Pr(>|t|)
## h1.(Intercept)                -1.045322   0.337955  -3.0931  0.001981 **
## h1.log_reves                   0.020512   0.036939   0.5553  0.578692
## h1.Nbadulte                    0.056579   0.018297   3.0923  0.001986 **
```

```
## h2.(Intercept)                  0.673499   0.772711   0.8716  0.383423
## h2.log_reves                   -0.057075   0.081448  -0.7008  0.483456
## h2.BautoPas de vehicule        -0.134785   0.160207  -0.8413  0.400170
## h2.AcompmCouple avec enfant(s) -0.128501   0.097047  -1.3241  0.185468
## h2.AcompmCouple sans enfant    -0.225343   0.111410  -2.0226  0.043111 *
## h2.AcompmPersonne seule        -0.392735   0.138906  -2.8274  0.004693 **
## sd.sd                           1.421212   0.011413 124.5238 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -5166.5 on 10 Df
##
## R^2 :
##  Coefficient of determination : 1
##  Likelihood ratio index       : 0.0021746
```

Le modèle Tobit Généralisé sans corrélation révèle que l'intercept de l'équation de la barrière (h1) est estimé à -1,0453, représentant le log-odds de présenter un sinistre non nul lorsque toutes les variables de h1 sont à zéro. Le coefficient de log_reves dans h1 (0,0205) n'est pas significatif (p-value = 0,5787), indiquant que le revenu n'influence pas significativement la probabilité de sinistre non nul. En revanche, Nbadulte dans h1 (0,0566) est significatif (p-value = 0,002), suggérant que chaque adulte supplémentaire augmente le log-odds de sinistre non nul. Dans l'équation de la partie continue (h2), l'intercept (0,6735) et log_reves (-0,0571) ne sont pas significatifs (p-values respectives de 0,3834 et 0,4835). La variable Bauto n'a pas d'impact significatif (p-value = 0,4002). Pour Acompm, seules les catégories "Couple sans enfant" (-0,2253, p-value = 0,0431) et "Personne seule" (-0,3927, p-value = 0,0047) sont significatives, associées à des montants de sinistres plus faibles. L'écart type des résidus est estimé à 1,4212 et significatif (p-value < 2e-16).

```
##
## Call:
## mhurdle(formula = Sinistre1 ~ log_reves + Nbadulte | log_reves +
##     Bauto + Acompm, data = dat, dist = "ln", h2 = FALSE, corr = TRUE,
##     robust = TRUE)
##
## Frequency of 0:  0.76327
##
## Newton-Raphson maximisation method
## 0 iterations, 0h:0m:0s
## g'(-H)^-1g =    NA
##
##
## Coefficients :
##                                 Estimate Std. Error t-value  Pr(>|t|)
## h1.(Intercept)                 -1.038508   0.338049 -3.0721  0.002126 **
## h1.log_reves                    0.018916   0.036961  0.5118  0.608809
## h1.Nbadulte                     0.059902   0.018448  3.2471  0.001166 **
## h2.(Intercept)                  1.096572   0.864739  1.2681  0.204764
## h2.log_reves                   -0.063912   0.082048 -0.7790  0.436004
## h2.BautoPas de vehicule        -0.131392   0.160211 -0.8201  0.412149
## h2.AcompmCouple avec enfant(s) -0.112337   0.098382 -1.1419  0.253516
## h2.AcompmCouple sans enfant    -0.207227   0.112850 -1.8363  0.066313 .
## h2.AcompmPersonne seule        -0.363205   0.141975 -2.5582  0.010521 *
## sd.sd                           1.442892   0.048879 29.5198 < 2.2e-16 ***
## corr12                         -0.197751   0.173704 -1.1384  0.254938
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -5166.1 on 11 Df
##
## R^2 :
##  Coefficient of determination : -0.01543
##  Likelihood ratio index       : 0.0022528
```

Le modèle Tobit Généralisé avec corrélation, l'intercept de h1 est légèrement différent (-1,0385), et Nbadulte reste significatif (0,0599, p-value = 0,0012). Les autres coefficients de h1 restent non significatifs. Dans h2, l'intercept est estimé à 1,0966 (non significatif, p-value = 0,2048), et log_reves (-0,0639) reste non significatif (p-value = 0,4360). Bauto n'est toujours pas significatif (p-value = 0,4121). Pour Acompm, seule "Personne seule" (-0,3632, p-value = 0,0105) est significative. sd.sd est estimé à 1,4429 (significatif, p-value < 2e-16). La corrélation entre les erreurs des deux équations (corr12) est estimée à -0,1978 mais n'est pas significative (p-value = 0,2549).

```
##
## Call:
## mhurdle(formula = Sinistre1 ~ log_reves + Nbadulte + Bauto |
##     Acompm + agecat + Ahabi, data = dat, dist = "ln", method = "bfgs")
##
## Frequency of 0:  0.76327
##
## Coefficients :
##                                        Estimate Std. Error  t-value
## h1.(Intercept)                        -0.9993400  0.3493024  -2.8610
## h1.log_reves                           0.0162854  0.0378252   0.4305
## h1.Nbadulte                            0.0551887  0.0184922   2.9844
## h1.BautoPas de vehicule               -0.0381560  0.0729723  -0.5229
## h2.(Intercept)                         0.1264270  0.1293031   0.9778
## h2.AcompmCouple avec enfant(s)        -0.1866981  0.1133522  -1.6471
## h2.AcompmCouple sans enfant           -0.1244059  0.1210832  -1.0274
## h2.AcompmPersonne seule               -0.3595271  0.1391315  -2.5841
## h2.agecat41-50                        -0.0089237  0.1093186  -0.0816
## h2.agecat51-60                        -0.0709352  0.1344967  -0.5274
## h2.agecat61-96                        -0.3294722  0.1338989  -2.4606
## h2.AhabiParis + Agglomeration          0.0018045  0.1381238   0.0131
## h2.AhabiUn. urb. de 10 000 a 99 999 hab. -0.0119104  0.1249678  -0.0953
## h2.AhabiUn. urb. de 100 000 hab. et +  0.2852013  0.1100112   2.5925
## h2.AhabiUn. urb. de 2 000 a 9 999 hab. -0.2997492  0.1498250  -2.0007
## sd.sd                                  1.4058378  0.0115380 121.8442
##                                        Pr(>|t|)
## h1.(Intercept)                         0.004224 **
## h1.log_reves                           0.666801
## h1.Nbadulte                            0.002841 **
## h1.BautoPas de vehicule                0.601055
## h2.(Intercept)                         0.328195
## h2.AcompmCouple avec enfant(s)         0.099545 .
## h2.AcompmCouple sans enfant            0.304213
## h2.AcompmPersonne seule                0.009764 **
## h2.agecat41-50                         0.934941
## h2.agecat51-60                         0.597907
## h2.agecat61-96                         0.013870 *
## h2.AhabiParis + Agglomeration          0.989577
```

```
## h2.AhabiUn. urb. de 10 000 a 99 999 hab.  0.924071
## h2.AhabiUn. urb. de 100 000 hab. et +     0.009529 **
## h2.AhabiUn. urb. de 2 000 a 9 999 hab.    0.045429 *
## sd.sd                                    < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -5152.6 on 16 Df
##
## R^2 :
##  Coefficient of determination : 1
##  Likelihood ratio index       : 0.0048625
```

Dans notre modèle le premier hurdle inclut les variables log_reves (revenu), Nbadulte (nombre d'adultes), Bauto (possession de véhicule) et Acompm (composition du ménage) tandis que le second hurdle retient les variables Acompm, agecat (catégorie d'âge), Ahabi (type d'habitat) et région.

Ainsi, dans l'équation de participation (Hurdle 1), seul le nombre d'adultes dans le ménage (Nbadulte) influence significativement la probabilité de déclaration d'un sinistre, avec une augmentation de 5,5% par adulte supplémentaire. Le revenu (log_reves) et la possession de véhicule (BautoPas de vehicule) ne montrent pas d'effet significatif. Pour l'équation de montant (Hurdle 2), les ménages de type "Personne seule" et les seniors (agecat61-96) présentent des montants de sinistres réduits de 36% et 32,9% respectivement. À l'opposé, les habitants des grandes agglomérations (plus de 100k habitants) ont un montant de sinistre supérieur de 28,5%. Le modèle présente une log-vraisemblance de -5152,6, indiquant une qualité d'ajustement modeste, avec un très faible pouvoir explicatif (Likelihood ratio index = 0,0049). La forte fréquence de zéros (76,3%) justifie l'utilisation d'un modèle à double seuil.

Avec ces résultats on conclue que le modèle Double Hurdle est le plus approprié pour analyser les données de sinistres présentant une forte proportion de zéros (76,3%). Il présente une meilleure log-vraisemblance de -5152,6, largement supérieure au Tobit standard (-7100,5) et au Tobit généralisé (-5166,5), indiquant un ajustement nettement supérieur aux données. Son R^2, bien que faible (0,0049), est supérieur aux autres modèles, suggérant une meilleure explication de la variance.

4. Nombre de Sinistres

```
## [1] 0.2600897
```

La proportion de zéros (26%) est modérée. On choisit de ne pas utiliser un modèle Zero-Inflated.

Les variables sélectionnées pour le modèle de régression de Poisson ont été choisies en fonction de leur pertinence théorique : D'un point de vue socio-démographique, les variables retenues telles que Acompm (composition du ménage), agecat (catégorie d'âge) et Nbadulte (nombre d'adultes) sont des paramètres reconnus pour leur influence sur le risque de sinistres en assurance. Sur le plan économique, la variable log_reves a été incluse car elles reflète la capacité financière des ménages. En effet, un revenu plus élevé peut aller de pair avec une plus grande exposition aux risques, les ménages aisés possédant souvent plus de biens à assurer. Enfin, les variables géographiques region et Ahabi ont également été sélectionnées. Elles offrent la possibilité de modéliser des spécificités régionales telles que les conditions climatiques, la densité de population ou encore les pratiques d'assurance propres à chaque zone.

D'après le modèle seul une partie des variables sont significatives au seuil 5%.

Le revenu, représenté par la variable log_reves, a un coefficient estimé à 0,2116 et une p-value inférieure à 2e-16, ce qui indique une relation positive et extrêmement significative avec le taux de sinistres. Une augmentation de 1% du revenu se traduit par une multiplication du taux de sinistres par e^0,2116, soit environ 1,235, correspondant à une augmentation de 23,5%. Cette corrélation positive est compréhensible puisque les ménages avec des revenus plus élevés possèdent généralement plus de biens à assurer, comme

plusieurs véhicules ou une habitation de valeur, ce qui augmente leur exposition aux risques et donc le nombre de sinistres déclarés.

Le nombre d'adultes dans le ménage (Nbadulte) présente également une relation positive significative avec le taux de sinistres, avec un coefficient de 0,1668 et une p-value inférieure à 2e-16. Cela signifie qu'un adulte supplémentaire dans le ménage est associé à une augmentation d'environ 18,1% du taux de sinistres. Cette relation parait logique car un foyer avec plus d'adultes comporte plus de personnes susceptibles de générer des sinistres.

À l'inverse, la variable AcompmPersonne seule, qui indique les ménages composés d'une seule personne, affiche un coefficient négatif de -0,6092 et une p-value inférieure à 2e-16. Cette diminution peut être expliquée par le fait que les ménages individuels disposent généralement de moins de biens à assurer et ont moins d'occasions de faire face à des sinistres.

Le revenu moyen par ménage de la commune (RUC) a un coefficient de -0,0000325 et une p-value de 2,72e-08, indiquant une relation négative significative. Une augmentation de 1 unité de RUC est liée à une réduction du taux de sinistres de 0,00325%. Cela suggère que dans les communes avec un revenu moyen plus élevé les ménages déclarent légèrement moins de sinistres.

Les catégories d'âge, agecat41-50, agecat51-60 et agecat61-96, ont tous des coefficients négatifs et extrêmement significatifs, respectivement -0,1325 (p-value=9,35e-12), -0,2021 (p-value=4,85e-16) et -0,2468 (p-value=4,39e-09). Par exemple, la catégorie agecat51-60 correspond à une réduction de 18,4% du taux de sinistres par rapport à la référence (eˆ-0,2021environ égale à0,816). Cela reflète probablement une plus grande prudence et une moindre exposition aux risques avec l'âge, ainsi qu'une meilleure gestion des risques par les ménages plus âgés.

La variable AcompmCouple avec enfant(s), qui concerne les ménages composés de couples avec enfants, a un coefficient de 0,1979 et une p-value inférieure à 2e-16. Cela indique une augmentation du taux de sinistres de 21,9% pour ces ménages par rapport à la référence. Cette hausse peut être attribuée à une activité domestique plus intense, à des occasions accrues de sinistres domestiques ou à une plus grande exposition aux risques liée à la présence d'enfants.

Les variables régionales region4, region8 et region9 présentent des coefficients positifs et significatifs, allant de 0,1131 (p-value=0,024969) à 0,2485 (p-value=9,48e-07). Cela peut être expliqué par des facteurs géographiques tels que les conditions climatiques, la densité de population ou des pratiques d'assurance locales particulières.

Enfin, la variable AhabiUn. urb. de 100 000 hab. et +, qui concerne les unités urbaines de plus de 100 000 habitants, a un coefficient de 0,07184 et une p-value de 0,000144. Cela signifie que les ménages situés dans ces zones urbaines ont un taux de sinistres 7,4% plus élevé (eˆ0,07184environ égale à1,074). Cette augmentation pourrait être due à la densité de population plus élevée, à des conditions de circulation plus complexes ou à des risques spécifiques liés à l'environnement urbain.

Le paramètre de dispersion est de 2.89 supérieur à 1, indiquant une surdispersion marquée.

On passera donc à une régression Binomiale Négative afin d'améliorer le modèle.

```
##
## Call:
## glm.nb(formula = NSin ~ pcs + RUC + cs + log_reves + region +
##     Ahabi + agecat + Acompm + Nbadulte, data = dat, init.theta = 2.025279052,
##     link = log)
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        -2.047e+00  4.918e-01  -4.163 3.14e-05
## pcsArtisans, comm., chefs d'ent.    5.965e-02  1.045e-01   0.571 0.568190
## pcsAutres pers. sans activite prof.  2.464e-01  1.078e-01   2.286 0.022276
```

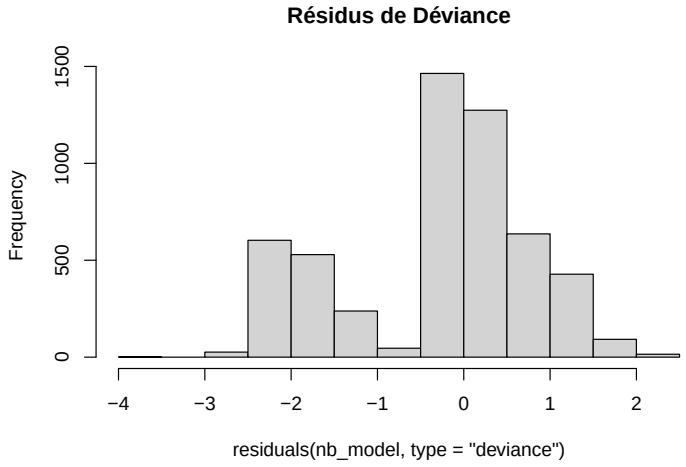
```
## pcsCadres et prof. intellectuelles sup.  2.090e-01  9.507e-02   2.199 0.027899
## pcsEmployes                               1.712e-01  8.943e-02   1.915 0.055515
## pcsOuvriers                               1.245e-01  8.581e-02   1.451 0.146737
## pcsProfessions intermediaires             1.996e-01  8.936e-02   2.234 0.025501
## pcsRetraites                              3.318e-02  1.027e-01   0.323 0.746642
## RUC                                      -4.087e-05  9.914e-06  -4.123 3.74e-05
## csModeste                                 1.124e-01  1.085e-01   1.037 0.299907
## csMoyenne Inf                             5.533e-02  8.686e-02   0.637 0.524175
## csMoyenne Sup                            -5.703e-02  6.675e-02  -0.854 0.392895
## log_reves                                 3.415e-01  5.394e-02   6.331 2.44e-10
## region2                                   5.681e-02  8.670e-02   0.655 0.512292
## region3                                  -1.538e-02  9.399e-02  -0.164 0.870013
## region4                                   1.331e-01  8.990e-02   1.480 0.138898
## region5                                  -1.698e-02  8.785e-02  -0.193 0.846741
## region7                                  -2.229e-03  9.072e-02  -0.025 0.980395
## region8                                   1.615e-01  8.824e-02   1.830 0.067265
## region9                                   2.669e-01  9.046e-02   2.950 0.003173
## AhabiParis + Agglomeration                1.441e-01  9.081e-02   1.586 0.112626
## AhabiUn. urb. de 10 000 a 99 999 hab.     5.293e-02  3.711e-02   1.426 0.153796
## AhabiUn. urb. de 100 000 hab. et +        8.615e-02  3.434e-02   2.509 0.012119
## AhabiUn. urb. de 2 000 a 9 999 hab.       1.782e-02  4.229e-02   0.421 0.673559
## agecat41-50                              -1.330e-01  3.610e-02  -3.685 0.000229
## agecat51-60                              -2.189e-01  4.445e-02  -4.926 8.41e-07
## agecat61-96                              -2.628e-01  6.947e-02  -3.784 0.000155
## AcompmCouple avec enfant(s)               1.692e-01  4.246e-02   3.984 6.77e-05
## AcompmCouple sans enfant                 -1.610e-01  4.568e-02  -3.526 0.000422
## AcompmPersonne seule                     -5.382e-01  6.849e-02  -7.858 3.92e-15
## Nbadulte                                  1.512e-01  2.073e-02   7.296 2.96e-13
##
## (Intercept)                              ***
## pcsArtisans, comm., chefs d'ent.
## pcsAutres pers. sans activite prof.      *
## pcsCadres et prof. intellectuelles sup. *
## pcsEmployes                              .
## pcsOuvriers
## pcsProfessions intermediaires            *
## pcsRetraites
## RUC                                      ***
## csModeste
## csMoyenne Inf
## csMoyenne Sup
## log_reves                                ***
## region2
## region3
## region4
## region5
## region7
## region8                                  .
## region9                                  **
## AhabiParis + Agglomeration
## AhabiUn. urb. de 10 000 a 99 999 hab.
## AhabiUn. urb. de 100 000 hab. et +       *
## AhabiUn. urb. de 2 000 a 9 999 hab.
## agecat41-50                              ***
```

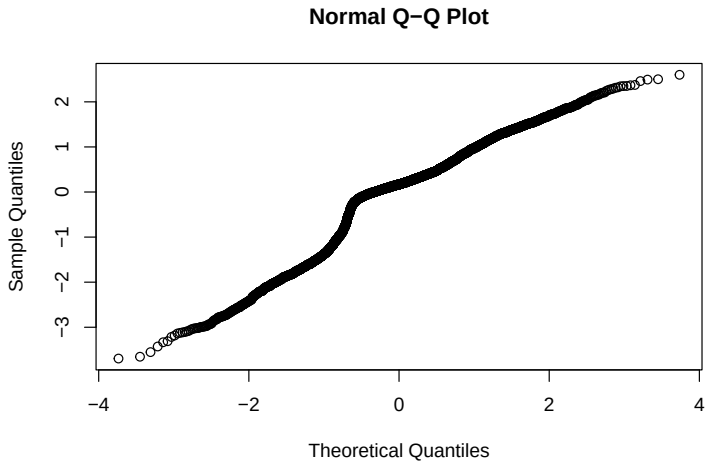
```
## agecat51-60                              ***
## agecat61-96                              ***
## AcompmCouple avec enfant(s)              ***
## AcompmCouple sans enfant                 ***
## AcompmPersonne seule                     ***
## Nbadulte                                 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(2.0253) family taken to be 1)
##
##     Null deviance: 8575.5  on 5351  degrees of freedom
## Residual deviance: 6860.2  on 5321  degrees of freedom
## AIC: 25972
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  2.0253
##           Std. Err.:  0.0741
##
##  2 x log-likelihood:  -25907.7540
```

Certaines des variables significatives, au seuil de 5%, selon le modèle de Poisson le sont ici aussi, en effet celles-ci présentent des p-values inférieures à 0.05.

Le revenu du ménage, par exemple, se révèle être un facteur extrêmement significatif (p-value = 2.44e-10), avec un coefficient positif. De même, le nombre d'adultes dans le ménage présente un coefficient positif et extrêmement significatif (p-value = 2.96e-13). D'autres variables notables incluent AcompmPersonne seule (p-value = 3.92e-15), qui présente un coefficient négatif, et AcompmCouple avec enfant(s) (p-value = 6.77e-05), qui a un coefficient positif. Le RUC, quant à lui, présente un coefficient négatif et extrêmement significatif (p-value = 3.74e-05), une augmentation de 1 unité du revenu moyen par ménage de la commune est associée à une réduction du taux de sinistres. Les catégories d'âge agecat41-50, agecat51-60 et agecat61-96 ont également des coefficients négatifs extrêmement significatifs (p-values respectives de 0.000229, 8.41e-07 et 0.000155). Enfin, region9 (p-value = 0.003173) et AhabiUn. urb. de 100 000 hab. et + (p-value = 0.012119) présentent des coefficients positifs significatifs.

On cherche à valider le modèle:



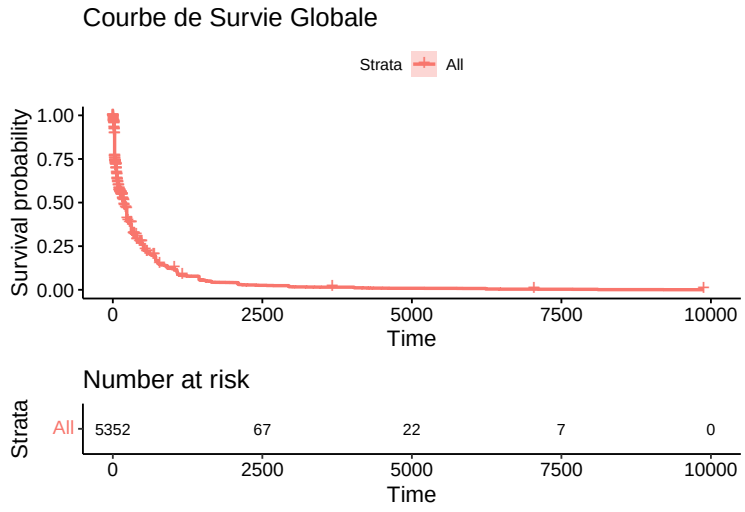


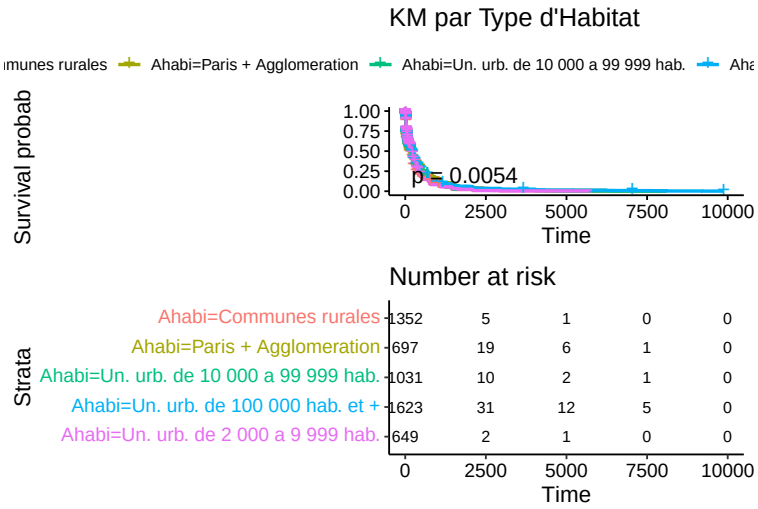
L'histogramme des résidus de déviance associé au QQ-Plot, qui n'indique qu'une très légère déviation de la diagonale, laisse suggérer que les résidus sont conformes à une distribution normale.

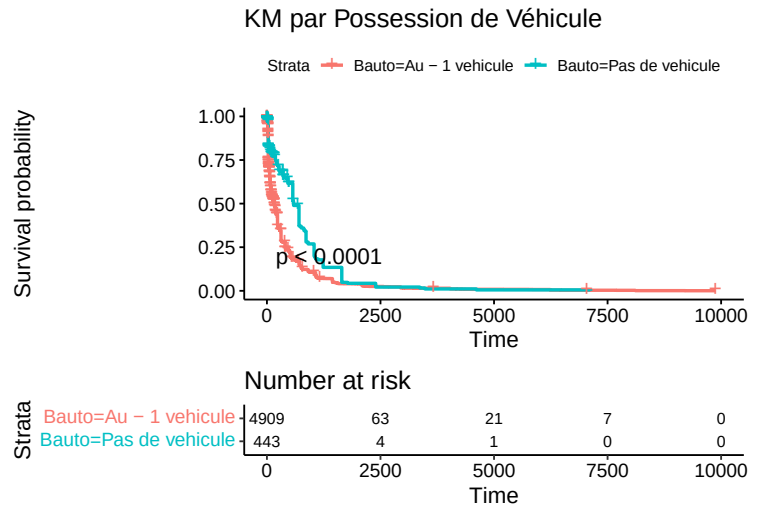
```
##                GVIF Df GVIF^(1/(2*Df))
## pcs        7.102602  7        1.150312
## RUC        8.182058  1        2.860430
## cs         8.583960  3        1.430918
## log_reves 5.446819  1        2.333842
## region     6.864900  7        1.147518
## Ahabi      7.037635  4        1.276228
## agecat     7.694233  3        1.405058
## Acompm     7.364517  3        1.394839
## Nbadulte   3.358009  1        1.832487
```

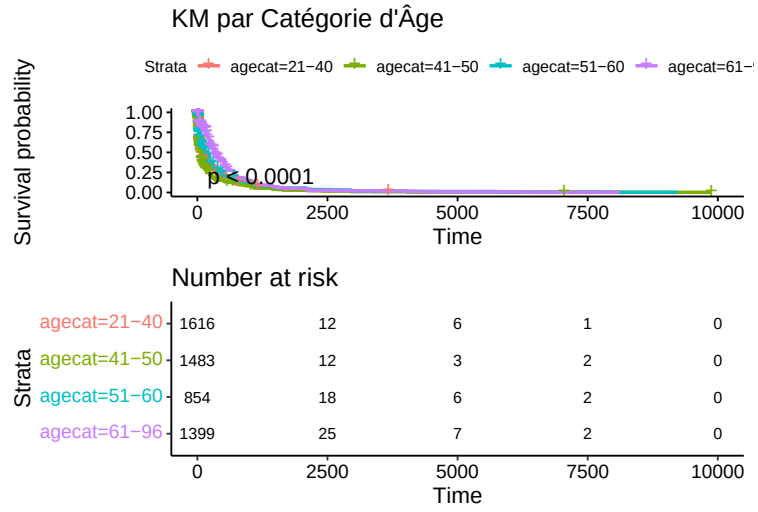
Les facteurs de gonflement de variance sont globalement acceptables, le plus élevé étant celui de RUC avec 2.86. Cependant, les VIF élevés de RUC et log_reves (GVIF >5) suggèrent une corrélation potentielle entre ces deux variables.

5. Modélisation de Durée









Les tests du log-rank révèlent des différences significatives dans la distribution des durées de souscription selon le type d'habitat (Ahabi), la possession de véhicule (Bauto) et la catégorie d'âge (agecat).

```
## Call:
## survdiff(formula = surv_obj ~ Ahabi, data = dat)
##
##                                          N Observed Expected (O-E)^2/E
## Ahabi=Communes rurales                1352      663      599     6.920
## Ahabi=Paris + Agglomeration            697      590      615     1.004
## Ahabi=Un. urb. de 10 000 a 99 999 hab. 1031      644      634     0.147
## Ahabi=Un. urb. de 100 000 hab. et +    1623     1061     1130     4.191
## Ahabi=Un. urb. de 2 000 a 9 999 hab.    649      349      329     1.172
##                                       (O-E)^2/V
## Ahabi=Communes rurales                    9.113
## Ahabi=Paris + Agglomeration               1.341
## Ahabi=Un. urb. de 10 000 a 99 999 hab.    0.194
## Ahabi=Un. urb. de 100 000 hab. et +       6.854
## Ahabi=Un. urb. de 2 000 a 9 999 hab.      1.392
##
##  Chisq= 14.7  on 4 degrees of freedom, p= 0.005
```

Pour Ahabi, le statistique du chi-deux vaut 14,7 pour 4 degrés de liberté, ce qui correspond à une valeur p de 0,005. Cela indique que les courbes de survie diffèrent significativement selon le type d'habitat. Les unités urbaines de plus de 100 000 habitants montrent une survie inférieure à celle attendue, tandis que les communes rurales montrent une survie supérieure.

```
## Call:
## survdiff(formula = surv_obj ~ Bauto, data = dat)
##
##                            N Observed Expected (O-E)^2/E (O-E)^2/V
## Bauto=Au - 1 vehicule 4909     3065     2887      11.0        93
## Bauto=Pas de vehicule  443      242      420      75.5        93
##
##  Chisq= 93  on 1 degrees of freedom, p= <2e-16
```

Pour Bauto, la statistique du chi-deux est de 93 pour 1 degré de liberté, avec une valeur p $<$ 2e-16, ce qui souligne une différence majeure entre les détenteurs de véhicules et ceux qui n'en possèdent pas. Les clients sans véhicule présentent une survie inférieure à celle attendue.

```
## Call:
## survdiff(formula = surv_obj ~ agecat, data = dat)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## agecat=21-40  1616      885      728    34.060     46.92
## agecat=41-50  1483      804      552   114.770    149.15
## agecat=51-60   854      591      607     0.411      0.54
## agecat=61-96  1399     1027     1420   108.949    208.40
##
##  Chisq= 283  on 3 degrees of freedom, p= <2e-16
```

En ce qui concerne agecat, la statistique du chi-deux est de 283 pour 3 degrés de liberté, avec également une valeur p $<$ 2e-16. Les catégories d'âge 41-50 et 61-96 montrent des écarts significatifs par rapport aux attentes, avec une survie respectivement inférieure et supérieure.

Les estimateurs de Kaplan-Meier par type d'habitat montrent que les clients résidant dans des unités urbaines de plus de 100 000 habitants ont une probabilité de résiliation plus élevée au cours du temps. À l'opposé,

les clients vivant dans des communes rurales présentent une probabilité de résiliation plus faible. Pour la possession de véhicule, les clients ne possédant pas de véhicule ont une probabilité de résiliation plus élevée que ceux qui en possèdent un. En termes de catégorie d'âge, les clients âgés de 61 à 96 ans ont une probabilité de résiliation plus faible, alors que les clients plus jeunes ont une probabilité de résiliation plus élevée, particulièrement pour les 21-40 ans.

Les résultats des modèles de Cox montrent que les trois modèles ont tous une performance statistique significative, mais avec des niveaux de concordance (C-index) différents.

Le Modèle 1, qui inclut un ensemble complet de variables socio-démographiques, économiques et géographiques, présente un C-index de 0,769, indiquant une bonne capacité prédictive.

Le Modèle 2, plus restreint, a un C-index de 0,678, tandis que le Modèle 3, le plus simple, affiche un C-index de 0,638. Cependant les trois modèles de Cox violent l'hypothèse de proportionnalité des risques au seuil 5%, leurs p-values étant inférieure à 0.05. Ceci limite leur validité.

Ces résultats suggèrent que le Modèle 1 est le plus performant en termes de capacité prédictive, malgré sa complexité et les violations de l'hypothèse de proportionnalité des risques, c'est donc celui que l'on propose.