Mémoire de Master 1 Mathématiques Appliquées de l'Université de Paris Dauphine

Titre : Tarification en assurance IARD avec les GLM en intégrant les données issues de la sécurité routière

Par : Victoire BELVEZE ; Elodie LIU ; Hadrien LEFLOCH

Directeur de Mémoire : DUTANG Christophe
Numéro de groupe : 10

Confidentialité : ☑ Non ☐ Oui (Durée : ☐ 1 an ☐ 2 ans)

Université Paris-Dauphine – MIDO – Bureau B536, Place du Maréchal de Lattre de Tassigny, 75016 PARIS

Table des matières

Introduction

Une prime d'assurance est la somme que paie le souscripteur d'un contrat à un assureur en échange de garanties définies. Il est nécessaire que cette prime reflète le risque associé au contrat. En effet, pour chaque police d'assurance, la prime est une fonction de variables de tarification qui permet alors de segmenter la population en fonction de son risque d'avoir un accident ou non.

La réalisation d'un tarif en assurance IARD s'appuie sur l'analyse de la prime pure dans le cadre d'un modèle fréquence/sévérité dans lequel l'effet des variables explicatives sur le niveau du risque est modélisé par des modèles de régression de type modèles linéaires généralisés (GLM).

On utilisera ici une approche fréquence/sévérité afin d'estimer le coût annuel d'une police d'assurance.

Nous tenterons d'estimer la prime pure avec des GLM à l'aide de la méthode Forward-Backward puis ajouterons les données de la sécurité routière afin de tester la pertinence de cet apport.

La prime pure est définie par :

$$\mathbb{E}[X] = \mathbb{E}[N] \times \mathbb{E}[B], \tag{1}$$

où X est la variable aléatoire représentant les coûts monétaires aux risques, N la fréquence des sinistres (le nombre de sinistre pour une période donnée) et B la sévérité des sinistres (les montants des sinistres).

Il va alors falloir estimer la loi de la fréquence et de la sévérité à l'aide des données `pg17trainpol` et `pg17trainclaim` obtenues du package R **CASdatasets**. Ce sont des données d'assurance mobile utilisées pour le Actuarial Pricing Game de 2017. Nous invitons le lecteur à se référer aux descriptions des données disponibles dans le package.

Chapitre 1

Exploration des données et corrélations

Tout d'abord, commençons par analyser les jeux de données que nous avons : `pg17trainpol` et `pg17trainclaim` afin de sélectionner les variables intéressantes pour la création du modèle et de déterminer le meilleur possible.

1.1 Analyse descriptive

L'analyse descriptive permet, entre autres, de déterminer les caractéristiques d'un individu moyen afin de connaître la population assurée et de vérifier la pertinence des variables tout en étudiant de façon plus ou moins succincte la corrélation entre les variables, notion primordiale lors de la modélisation.

1.1.1 Analyse des variables qualitatives

Les variables qualitatives que nous allons analyser sont présentées dans le tableau 1.1

variable	intitulé de la variable
`drv_drv2`	s'il y a un deuxième conducteur
`drv_sex1`	sexe du conducteur
`vh_fuel`	type de carburant du véhicule
`vh_type`	type du véhicule

TABLE 1.1: variables qualitatives

- `drv_drv2`
  Nous observons, à l'aide du graphique 1.1, que le nombre de polices n'ayant pas de deuxième conducteur est deux fois plus élevé que celui n'en ayant qu'un seul.

- `drv_sex1`
  Nous constatons d'après l'histogramme 1.2 qu'il y a plus d'hommes que de femmes dans notre portefeuille, mais cette dominance n'est pas écrasante.

- `vh_fuel`
  Sur ces trois types de carburant des véhicules, les `Diesel` et les `Gasoline` représentent la majorité des polices étudiées comme nous pouvons le voir en 1.3.

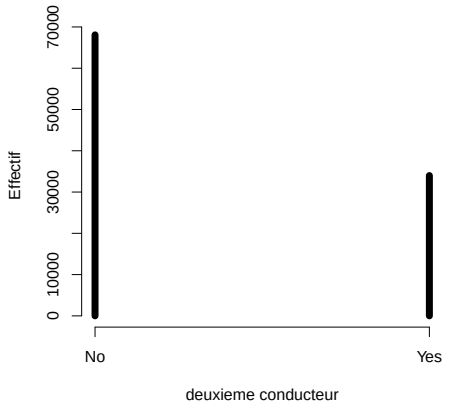


FIGURE 1.1: Présence d'un deuxième conducteur

FIGURE 1.2: Répartition des données par sexe du conducteur

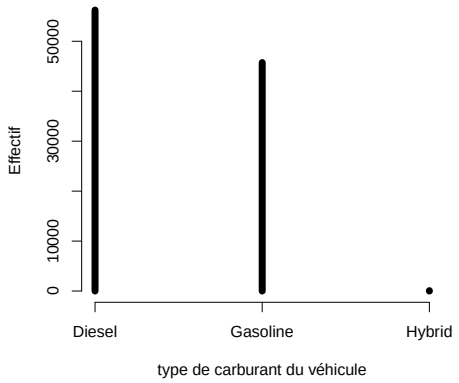


FIGURE 1.3: Répartition des données par type de carburant

- `vh_type`
  Cette variable n'a que deux modalités qui sont `Commercial` et `Tourism`, nous observons grâce à l'histogramme 1.4 que `Tourism` représente plus de 80% des données.

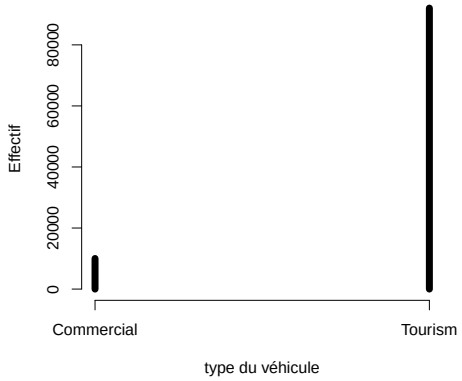


FIGURE 1.4: répartition des données par le type du véhicule

1.1.2 Analyse des variables quantitatives

Nous allons analyser les variables quantitatives listées ci-dessous en utilisant des histogrammes.

Variable	Intitulé de la variable	Histogramme
`drv_age1`	âge du conducteur	3.13a
`drv_age_lic1`	temps du permis	3.13c
`pol_bonus`	coefficient du bonus	1.7
`pol_duration`	durée de la police	1.8
`pol_sit_duration`	durée actuelle de la police	3.15c

TABLE 1.2: Variables quantitatives

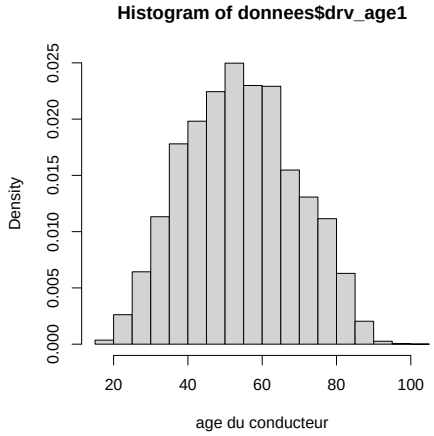


FIGURE 1.5: Distribution de l'âge du conducteur

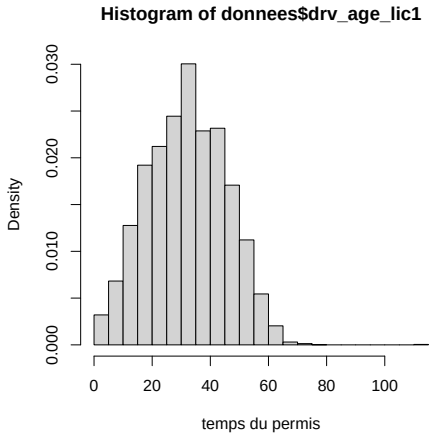


FIGURE 1.6: Distribution de l'ancienneté du permis du conducteur

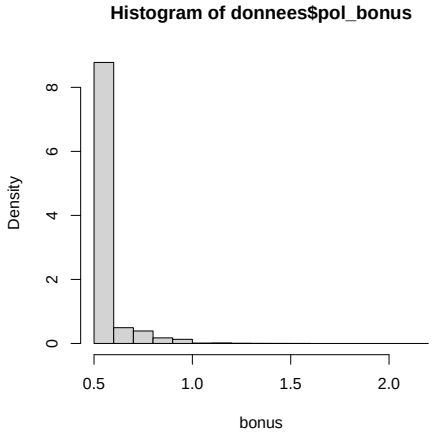


FIGURE 1.7: Distribution du coefficient du bonus

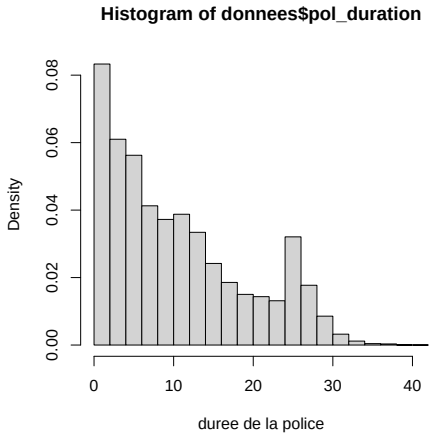


FIGURE 1.8: Distribution de la durée des polices

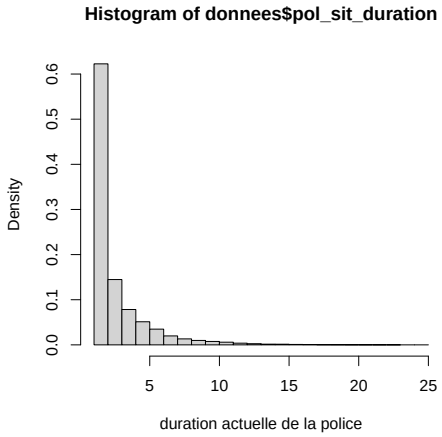


FIGURE 1.9: Distribution de la durée actuelle des polices

Nous remarquons qu'il y a une asymétrie à droite dans les histogrammes des variables quantitatives `drv_age_lic1`, `pol_bonus`, `pol_duration`, `pol_sit_duration`. Il nous faut alors créer des classes de modalités pour celles-ci ainsi que pour `drv_age1` afin d'avoir une relation linéaire avec les variables à expliquer dans le GLM.

1.2 ACP et corrélation entre des variables qualitatives

Pour la sélection des variables, nous effectuerons une ACP. Nous utiliserons également un tableau de corrélation pour repérer les variables qualitatives redondantes et les retirer du modèle.

ACP On remarque sur 1.10 plusieurs groupes de variables : un premier concernant les caractéristiques techniques du véhicule de l'assuré et un second regroupant les valeurs liées au conducteur assuré et enfin un troisième regroupant les variables qui concernent l'ancienneté du véhicule.

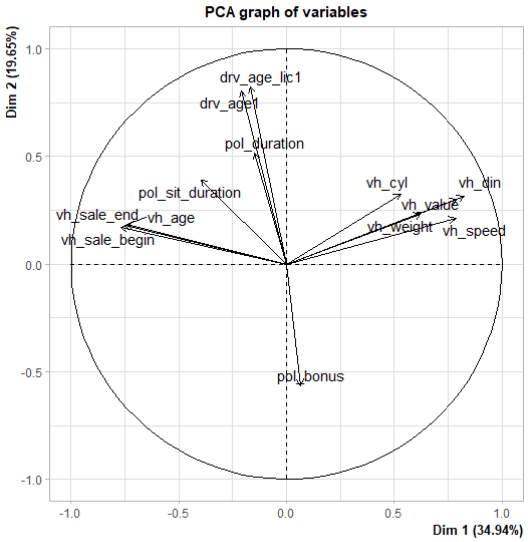


FIGURE 1.10: ACP des variables

Etude de la corrélations entre certaines variables qualitatives Nous allons étudier l'indépendance entre les variables `vh_fuel`, `vh_make`(la marque du véhicule), `vh_model`(le modèle du véhicule) et `vh_type`, en utilisant le test du Khi-deux d'indépendance. L'hypothèse nulle du test est de supposer que deux variables sont indépendantes, et si la p-valeur obtenues par la commande `chisq.test()` sur R est supérieure à 5%, on rejette l'hypothèse nulle, sinon elles ne sont pas indépendantes. Nous appliquons ce test à ces 4 variables qualitatives, on obtient alors les p-valeurs présentées dans le tableau 1.3.

	vh_make	vh_model	vh_type
vh_fuel	$< 2.2e^{-16}$	$< 2.2e^{-16}$	1
vh_make		$< 2.2e^{-16}$	$< 2.2e^{-16}$
vh_model			$< 2.2e^{-16}$

TABLE 1.3: Résultat des tests de Khi-deux

Cependant, on observe que quasiment toutes les p-valeurs sont inférieures à 5%. On peut donc essayer de trouver un lien entre ces variables en utilisant le V de Cramer. En effet, plus le V de Cramer est proche de 0, plus il y a l'indépendance entre deux variables. Ainsi, nous allons déterminer les V de Cramer pour chaque paire de variables en appliquant la commande `cramer.v()` sur R et nous avons le tableau 1.4

	vh_make	vh_model	vh_type
vh_fuel	0.756	0.192	0.205
vh_make		0.951	0.231
vh_model			0.796

TABLE 1.4: Résultat des V de Cramer

Nous constatons que ces paires de variables sont fortement corrélées :

- `vh_fuel` et `vh_make`
- `vh_model` et `vh_make`
- `vh_type` et `vh_model`

Nous concluons que la marque du véhicule et le modèle du véhicule ont des fortes liens de corrélations entre eux et avec les deux autres variables, nous pouvons donc de ne pas les mettre dans la modélisation.

1.3 Graphique de corrélation

Nous pouvons observer les groupes de données ayant des corrélations fortes avec 1.11.

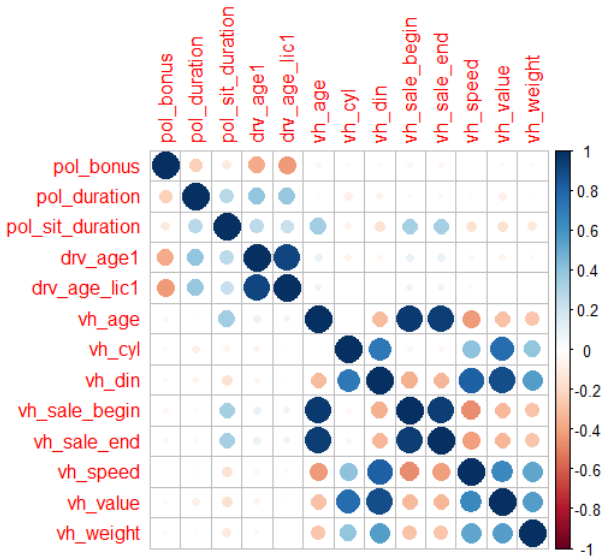


FIGURE 1.11: Graphique de corrélation

Nous avons ensuite regroupé les deux jeux de données principales basées sur notre sujet pour s'intéresser à la sévérité et à la fréquence des indemnités.

Pour calculer des modèles plus rapidement, nous avons décidé de former des groupes représentatifs de population. Par exemple, nous avons formé des groupes sur l'âge des conducteurs et sur des catégories de polices d'assurances qui avaient beaucoup de facteurs.

Ces aspects constituant une première approche et un préliminaire à la réalisation de la modélisation du risque automobile que nous allons à présent décrire, nous pouvons alors commencer à s'interroger sur les lois suivies par ces données.

1.4 Modification des variables

Afin d'avoir des résultats de modèles plus génériques, nous regroupons des modalités de variables. Nous créons des catégories d'âge :

- [18 ;25[
- [25 ;40[
- [40 ;60[
- [60 ;80[
- [80 ;103[

Ces catégories ont été déterminées de manière humainement logique. En utilisant les quantiles de nos données, nous regroupons les modalités de `drv_age_lic1` qui donne l'âge de permis de l'assuré. Nous faisons de même pour les variables `pol_bonus`, `pol_duration` et `pol_sit_duration` qui déterminent la police d'assurance souscrite.

Nous classons donc les assurés dans des cases plus homogènes.

Chapitre 2

GLM : fréquence et sévérité

2.1 Modélisation de la fréquence des sinistres

À l'aide de nos données, nous allons alors pouvoir proposer des modèles linéaires généralisés. Nous cherchons ici à modéliser la fréquence des sinistres. On définit alors N la variable aléatoire représentant le nombre de sinistres, à valeurs dans les entiers naturels.

Pour écrire un GLM, nous devons d'abord choisir une loi de probabilité pour N au sein de la famille exponentielle naturelle.

Pour rappel, la variable N appartient à la famille exponentielle naturelle si sa densité de probabilité s'écrit sous la forme :

$$f_N(x) = \exp(\frac{1}{\gamma(\phi)}(x\theta - b(\theta) + c(x, \phi)),$$

où c est une fonction dérivable, b une fonction trois fois dérivable et la dérivé première de b est inversible.

Ainsi, d'après le cours d'Actuariat 1 et de part cette approche, nous avons le choix entre entre une loi Binomiale Négative et une loi de Poisson pour modéliser N.

2.1.1 Étude de la corrélation entre la fréquence des sinistres et les variables explicatives

On souhaite tout d'abord étudier le lien entre la variable à expliquer `freq`, qui est quantitative, et les variables ci-dessous :

- `drv_age1`
- `drv_drv2`
- `vh_cyl`
- `vh_speed`
- `drv_age_lic1`
- `drv_age_lic2`

L'étude sera effectuée pour chaque variable grâce à la fonction `tapply` qui permet d'observer la relation entre la fréquence et une certaine variable. Nous pouvons alors visualiser ces corrélations à l'aide des graphiques correspondant 2.1.

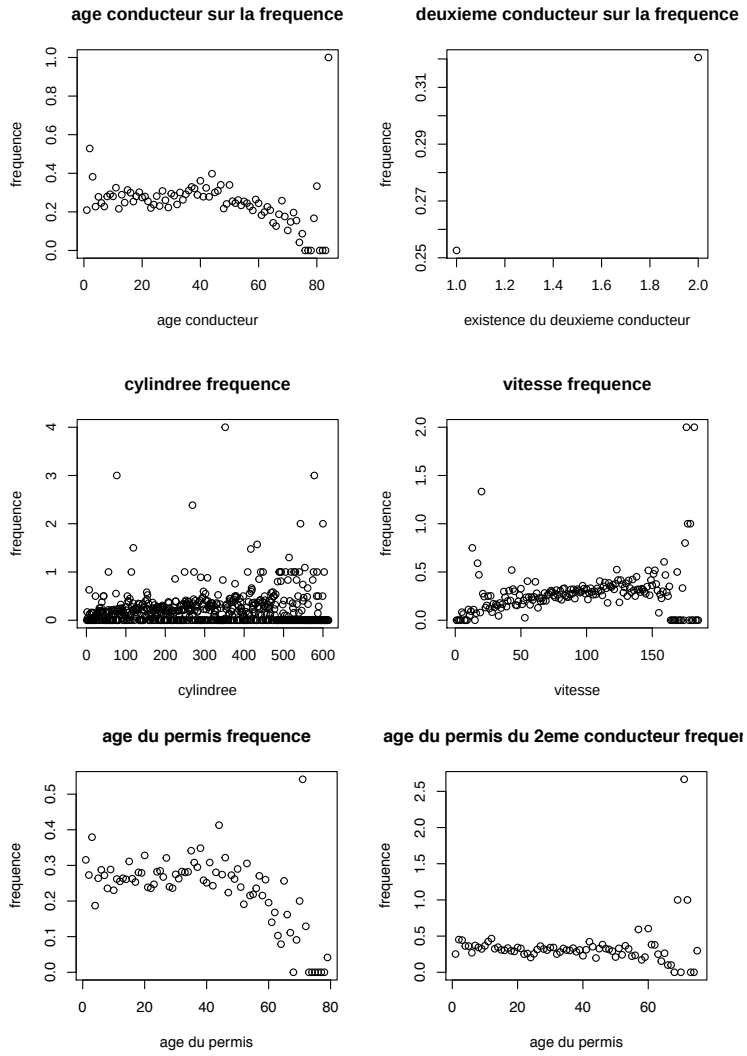


FIGURE 2.1: Corrélation entre la fréquence et les variables explicatives

Conclusion : D'après les figures obtenues, nous observons en particulier que :

- L'âge du conducteur impacte sur la fréquence. En effet, à partir de l'âge de 50 ans, les conducteurs ont moins de sinistres.
- Un seconde conducteur peut conduire à plus de sinistres.
- La cylindrée du véhicule n'influence pas la fréquence, il n'y a pas de tendance particulière.
- La fréquence est croissante en fonction de la vitesse.
- Plus le permis du conducteur est ancien, moins il aura de sinistres.
- L'ancienneté du permis du deuxième conducteur n'a pas d'impact sur le nombre de sinistres.

Ainsi, pour modéliser la fréquence des sinistres, nous n'allons pas considérer les variables `vh_cyl` et `drv_age_lic2`.

2.1.2 Choix de la fonction de lien

À présent, nous modélisons le lien entre l'espérance des N_i et les variables explicatives au travers d'une fonction g inversible :

$$g(\mathbb{E}[x_i]) = x_i\beta.$$

Par défaut, nous choisissons la fonction de lien canonique qui est identique pour la loi de Poisson et la loi Binomiale Négative : $g(\mu) = \log(\mu)$ et obtenons alors :

$$\mathbb{E}[N_i] = g^{-1}(x_i\beta) = \exp(x_i\beta).$$

2.1.3 Estimation des coefficients de la fréquence

Nous allons estimer les coefficients β_j par maximum de vraisemblance.

Une solution pour calculer l'estimateur du maximum de vraisemblance est d'utiliser des procédures itératives d'optimisation mais ici les estimations seront réalisées par la fonction `glm` pour la loi de Poisson et la fonction `glm.nb` pour la loi Binomiale Négative, dans R.

Tout d'abord, nous incluons toutes les variables dans les modèles et obtenons ainsi pour la loi de Poisson :

```
> summary(modelfullnb)

Call:
glm(formula = freq ~ pol_bonus + pol_coverage + pol_pay_freq +
    vh_sale_end + vh_value + vh_age + vh_cyl + vh_din + vh_fuel +
    vh_sale_begin + vh_speed + vh_value + vh_weight + drv_sex1 +
    drv_drv2 + drv_age1 + drv_age_lic1 + pol_sit_duration + pol_duration,
    family = poisson(), data = Train)
```

L'estimation du coefficient β pour chaque variable est présentée dans le tableau 2.1 dans la colonne Estimate.

Tandis que pour la loi Binomiale Négative, nous pouvons lire l'estimation des coefficients dans le tableau 2.2.

```
> summary(modelfullnb)

Call:
glm.nb(formula = freq ~ pol_bonus + pol_coverage + pol_pay_freq     +
    vh_sale_end + vh_value + vh_age + vh_cyl + vh_din + vh_fuel +
    vh_sale_begin + vh_speed + vh_value + vh_weight + drv_sex1 +
    drv_drv2 + drv_age2 + drv_age1 + drv_age_lic1 + pol_sit_duration +
    pol_duration, data = Train, init.theta = 0.2329253796, link = log)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.6607	0.1083	-6.10	0.0000
pol_bonus	0.0511	0.0105	4.87	0.0000
pol_coverage	-0.1581	0.0096	-16.52	0.0000
pol_pay_freq	-0.0081	0.0054	-1.50	0.1329
vh_sale_end	-0.0127	0.0040	-3.16	0.0016
vh_value	0.0000	0.0000	6.25	0.0000
vh_age	0.0099	0.0044	2.24	0.0249
vh_cyl	0.0000	0.0000	0.93	0.3503
vh_din	0.0015	0.0007	2.26	0.0237
vh_fuel	-0.1952	0.0191	-10.22	0.0000
vh_sale_begin	-0.0113	0.0040	-2.83	0.0047
vh_speed	-0.0019	0.0006	-3.27	0.0011
vh_weight	-0.0001	0.0000	-6.16	0.0000
drv_sex1	-0.0050	0.0145	-0.35	0.7284
drv_drv2	0.1700	0.0140	12.12	0.0000
drv_age1	-0.0635	0.0152	-4.16	0.0000
drv_age_lic1	0.0590	0.0089	6.65	0.0000
pol_sit_duration	-0.0408	0.0065	-6.27	0.0000
pol_duration	-0.0156	0.0025	-6.28	0.0000

TABLE 2.1: Coefficients de la régression avec la loi de Poisson

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.6144	0.1633	-3.76	0.0002
pol_bonus	0.0390	0.0160	2.44	0.0147
pol_coverage	-0.1551	0.0134	-11.57	0.0000
pol_pay_freq	-0.0052	0.0080	-0.65	0.5159
vh_sale_end	-0.0140	0.0057	-2.44	0.0147
vh_value	0.0000	0.0000	3.91	0.0001
vh_age	0.0125	0.0063	1.99	0.0461
vh_cyl	0.0000	0.0000	0.69	0.4927
vh_din	0.0019	0.0010	1.81	0.0699
vh_fuel	-0.1837	0.0288	-6.38	0.0000
vh_sale_begin	-0.0118	0.0056	-2.11	0.0353
vh_speed	-0.0023	0.0009	-2.57	0.0103
vh_weight	-0.0001	0.0000	-3.58	0.0003
drv_sex1	-0.0097	0.0217	-0.45	0.6543
drv_drv2	0.1862	0.0213	8.75	0.0000
drv_age1	-0.0902	0.0225	-4.01	0.0001
drv_age_lic1	0.0689	0.0132	5.24	0.0000
pol_sit_duration	-0.0457	0.0092	-4.98	0.0000
pol_duration	-0.0167	0.0037	-4.47	0.0000

TABLE 2.2: Coefficients de la régression avec la loi Binomiale Négative

2.1.4 Sélection des variables du modèle

La sélection de modèle peut être vue comme la recherche du modèle optimal parmi toutes les possibilités.

Pour sélectionner le meilleur modèle, nous allons nous appuyer sur un critère qui permet de comparer les modèles entre eux : le critère AIC.

Nous utilisons alors une méthode pas-à-pas. Trois méthodes sont souvent utilisées :

- **Méthode Forward** : cette méthode part du modèle réduit à l'intercept et on le compare, en utilisant le critère AIC, à les modèles contenant une variable explicative. Nous choisissons alors le meilleur modèle ayant la plus petite valeur d'AIC. On ajoute ensuite une variable parmi les autres covariables et on choisit de nouveau le meilleur modèle selon le critère. On s'arrête quand l'ajout d'une variable n'améliore pas la valeur de l'AIC.
- **Méthode Backward** : cette méthode a une stratégie inverse de la précédente, elle consiste à partir du modèle complet et on enlève une à une les variables en comparant les modèles deux à deux avec le critère AIC.
- **Méthode Both (Forward-Backward)** : cette méthode est un mélange des deux précédentes. À chaque étape, on ajoute ou enlève une variable et on choisit le meilleur modèle, ensuite on recommence.

Ici, nous décidons d'appliquer la méthode Forward-Backward en utilisant la commande `step` de R, fonction choisissant un modèle par AIC dans un algorithme pas à pas. Cependant, nous avons aussi besoin de réaliser des analyses de la variance (ANOVA). En effet, la fonction `anova` avec `test='Chisq'` sur R permet d'étudier si chaque covariable a un effet significatif pour expliquer la variable réponse.

Ainsi le résultat de plusieurs sélections des variables pour le modèle avec la loi de Poisson est (2.3) :

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.9332	0.0812	-11.49	0.0000
pol_bonus	0.4532	0.0648	6.99	0.0000
pol_coverage	-0.1541	0.0083	-18.58	0.0000
vh_sale_end	-0.0142	0.0014	-10.08	0.0000
vh_value	0.0000	0.0000	5.79	0.0000
vh_din	0.0022	0.0005	4.33	0.0000
vh_fuel	-0.2346	0.0146	-16.10	0.0000
vh_speed	-0.0022	0.0005	-4.70	0.0000
vh_weight	-0.0001	0.0000	-6.42	0.0000
drv_drv2	0.1756	0.0125	14.07	0.0000
drv_age_lic1	0.0037	0.0005	7.05	0.0000
pol_sit_duration	-0.0224	0.0033	-6.71	0.0000
pol_duration	-0.0072	0.0008	-9.12	0.0000

TABLE 2.3: Coefficients de la régression après la sélection

Tandis que pour le modèle avec la loi Binomiale Négative nous obtenons le résultat de la sélection dans 2.4 :

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.0793	0.0975	-11.07	0.0000
vh_sale_end	-0.0239	0.0046	-5.21	0.0000
vh_value	0.0000	0.0000	11.19	0.0000
pol_coverage	-0.1552	0.0119	-13.08	0.0000
vh_fuel	-0.2150	0.0200	-10.76	0.0000
drv_drv2	0.1822	0.0190	9.57	0.0000
pol_sit_duration	-0.0250	0.0047	-5.29	0.0000
pol_duration	-0.0075	0.0012	-6.30	0.0000
vh_age	0.0121	0.0044	2.76	0.0058
vh_weight	-0.0001	0.0000	-3.78	0.0002
drv_age_lic1	0.0098	0.0018	5.42	0.0000
drv_age1	-0.0064	0.0016	-3.90	0.0001
pol_bonus	0.3841	0.1003	3.83	0.0001

TABLE 2.4: Coefficients de la régression après la sélection

2.1.5 Sélection du meilleur modèle GLM

Après avoir obtenu deux modèles GLM pour la fréquence, nous allons modéliser alors la fréquence des sinistres grâce au meilleur modèle GLM choisi selon le critère AIC ou le critère BIC.

Comparaison de l'espérance avec la variance Nous pouvons tout d'abord utiliser un critère basé sur les moments de la fréquence.

Dans le cas d'une loi de Poisson :

$$\mathbb{V}[N] = \mathbb{E}[N].$$

Tandis que dans le cas d'une loi Binomiale Négative :

$$\mathbb{V}[N] > \mathbb{E}[N].$$

En définissant une fonction `dispersion_test` qui permet de calculer les moments de la fréquence des sinistres, nous pouvons tester si la variable suit une distribution de la loi de Poisson :

```
> dispersion_test(TabNA$freq)

Mean : 0.2751861
Variance : 0.6430191
Probability of being drawn from Poisson distribution : 0
```

TABLE 2.5: Test de dispersion

D'après ce critère, on devrait alors choisir la loi Binomiale négative pour modéliser la fréquence.

Critère AIC et BIC : Ce critère consiste à calculer et comparer les valeurs d'AIC et BIC de ces deux modèle en notant L la vraisemblance maximisée, k le nombre de paramètre du modèle et n le nombre d'individus :

$$AIC = -2log(L) + 2k.$$

$$BIC = -2log(L) + klog(n).$$

Celui ayant des valeurs plus petite est le meilleur modèle pour modéliser la fréquence.

	AIC	BIC
Poisson	148421	148544.9
Binomiale négative	126776.1	126909.6

TABLE 2.6: Tableau des valeurs AIC et BIC

Nous observons que le modèle avec la loi Binomiale négative admet le plus petit AIC et BIC, il est donc le meilleur modèle.

Avec ces deux critères, nous pourrons alors conclure que le meilleur modèle GLM pour la fréquence des sinistres est celui ayant appliquer la loi Binomiale négative 2.4.

2.1.6 Validité du modèle

Nous allons donner les anomalies que nous allons repérer sur le modèle fait ainsi que les limites du tarificateur que nous avons créée.

Nous pouvons observer la pertinence du modèle créé avec un diagramme Quantile-Quantile, qui permet de présenter l'adéquation des résidus à une loi normale centrée réduite, en regardant les log-résidus :

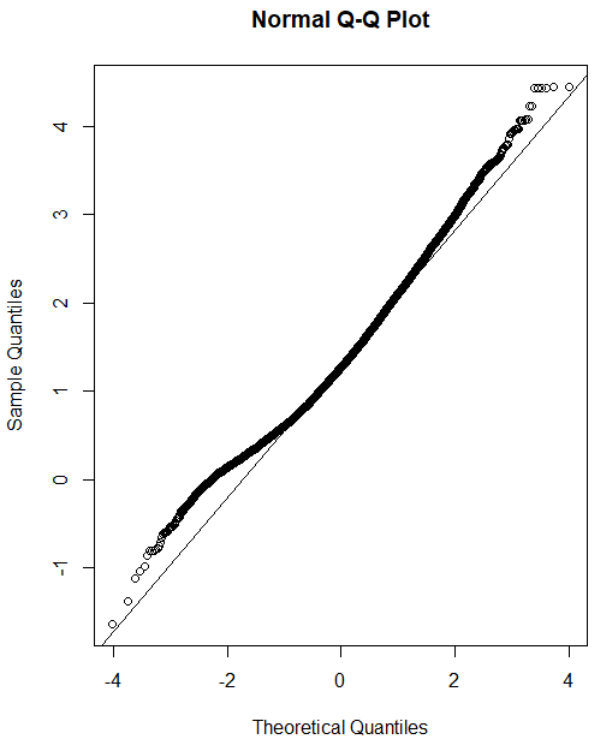


FIGURE 2.2: Diagramme Quantile-Quantile

Le graphique 2.2 montre que les log-résidus suivent une loi Normale car les points semblent alignés le long de la première bissectrice.

Conclusion : Nous avons établi un modèle pour la fréquence en utilisant la loi Binomiale Négative et la fonction de lien log.

2.2 Modélisation de la sévérité des sinistres

Nous allons emprunter des références et nous allons choisir la modélisation de la sévérité en fonction de ce que nous allons y trouver. Notons B les charges des sinistres.

La sévérité étant une variable continue, on a le choix entre la loi Gamma et la loi inverse-gaussienne comme loi du coût de sinistres. En effet, ces deux lois appartiennent à la famille exponentielle. De plus, ces deux lois étant à support dans $]0, \infty[$, nous décidons d'extraire les polices ayant un coût de sinistre strictement positif.

Avant de commencer la modélisation, nous observons des sinistres ayant un montant exceptionnellement élevé, ceci pouvant affecter le résultat. Nous les appelons **les sinistres graves**.

Nous décidons de traiter ces sinistres graves par la **méthode par écrêtement** consistant à plafonner les sinistres à un montant maximum, la charge résiduelle sera alors répartie uniformément sur l'ensemble des montants. Autrement dit, les charges sinistres deviendrons

$$\bar{B_i} = min(B_i, m) + \frac{S}{n}, i = 1, \ldots, n,$$

où n est le nombre de réclamations et S est la charge surcrête

$$S = \sum_{i=1}^{n} max(B_i - m, 0).$$

Ici, nous avons choisi $m = 60000$. En effet, si on visualise les montants de sinistres en faisant `plot(tab_sev$claim_amount)` sur R, nous constatons des points aberrants illustrés dans 2.3, ce sont les sinistres graves et ces observations sont supérieurs à 60000.

Nous regardons ensuite la forme de nos données de sévérité triées à l'aide du graphique 2.4. Les montants ici sont donc devenus similaires pour mieux modéliser la sévérité.

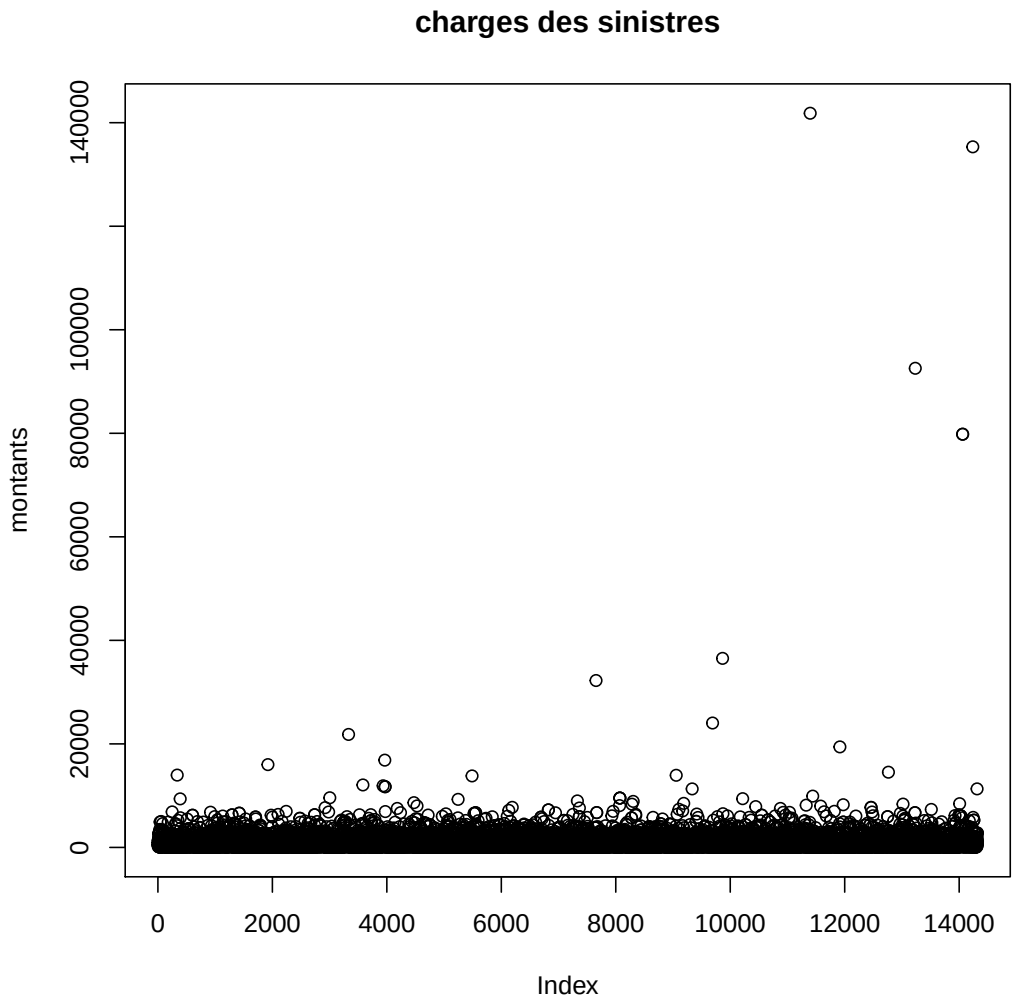


FIGURE 2.3: Montants des sinistres

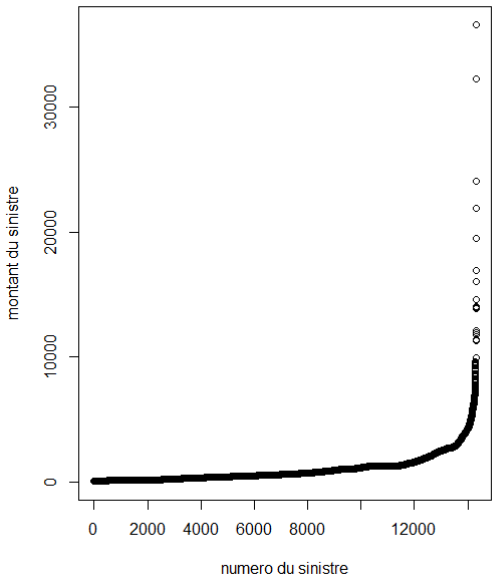


FIGURE 2.4: Montants des sinistres

2.2.1 Étude de la corrélation entre les coûts de sinistres et les variables explicatives

Nous nous intéressons à présent aux corrélations de certaines variables avec la sévérité.

Avec ce premier graphique ci-dessous (2.5), nous voyons que l'âge du conducteur est une variable qui a sûrement un rôle à jouer dans l'explication de la sévérité d'un accident. Ainsi, nous pouvons déduire que les variables corrélées à l'âge du conducteur seront indispensables dans le modèle que nous allons proposer. De même, nous constatons que l'apparition d'un deuxième conducteur a un effet significatif sur la sévérité, également la vitesse du véhicule et l'ancienneté du permis des deux conducteurs. Cependant, la cylindrée ne l'impacte pas donc nous n'allons pas l'inclure dans la modélisation de la variable à expliquer.

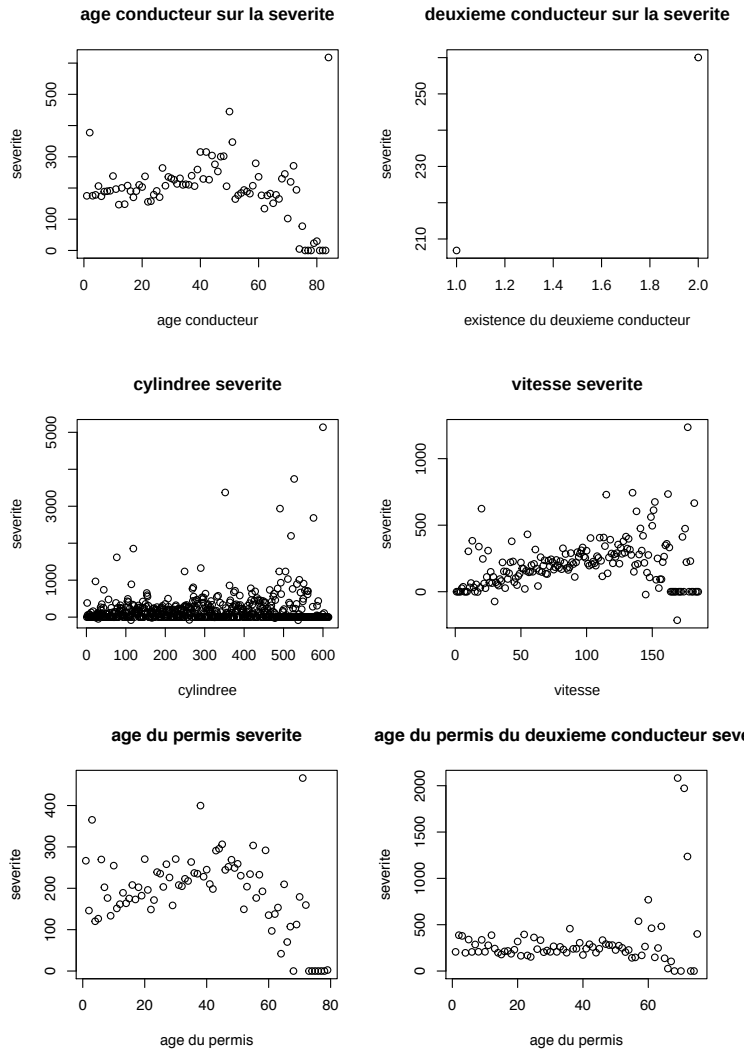


FIGURE 2.5: corrélation entre la sévérité et les variables explicatives

2.2.2 Estimation des coefficients du coût de sinistres

Nous choisissons la fonction logarithme comme fonction de lien pour la loi Gamma et la loi inverse-gaussienne. Bien que la fonction de lien canonique soit la fonction inverse pour la loi Gamma et $x \longmapsto \frac{1}{x^2}$ pour l'autre loi, la fonction logarithme est utilisée plus fréquemment pour ces deux lois.

Les coefficients de la régression β_j sont inconnus et doivent être estimés. Ceci sera effectué de la même façon que pour la fréquence.

Nous allons tout d'abord mettre toutes les variables dans le modèle, nous obtenons ainsi le résultat pour la loi Gamma suivant 2.7 :

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.5171	0.1436	45.38	0.0000
pol_coverage	-0.0928	0.0174	-5.34	0.0000
pol_pay_freq	0.0026	0.0092	0.28	0.7822
pol_bonus	0.3154	0.1242	2.54	0.0111
pol_duration	0.0000	0.0015	0.01	0.9884
pol_sit_duration	-0.0023	0.0067	-0.34	0.7346
drv_drv2	-0.0609	0.0358	-1.70	0.0893
drv_sex1	0.0075	0.0250	0.30	0.7636
drv_age_lic1	0.0017	0.0011	1.62	0.1048
drv_age_lic2	0.0011	0.0010	1.14	0.2534
vh_fuel	0.0991	0.0281	3.52	0.0004
vh_type	0.0718	0.0429	1.67	0.0945
vh_din	-0.0002	0.0008	-0.22	0.8244
vh_sale_end	-0.0090	0.0063	-1.43	0.1536
vh_value	0.0000	0.0000	2.33	0.0198
vh_age	0.0004	0.0062	0.06	0.9502

TABLE 2.7: Coefficients de la régression retournés par R

Pour la loi inverse-gaussienne nous avons 2.8 :

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.5015	0.1430	45.47	0.0000
pol_coverage	-0.0877	0.0163	-5.38	0.0000
pol_pay_freq	0.0055	0.0093	0.59	0.5531
pol_bonus	0.3148	0.1285	2.45	0.0143
pol_duration	0.0000	0.0015	0.02	0.9808
pol_sit_duration	-0.0032	0.0065	-0.49	0.6242
drv_drv2	-0.0485	0.0357	-1.36	0.1748
drv_sex1	0.0047	0.0250	0.19	0.8507
drv_age_lic1	0.0019	0.0011	1.78	0.0745
drv_age_lic2	0.0009	0.0010	0.89	0.3720
vh_fuel	0.1024	0.0286	3.57	0.0004
vh_type	0.0587	0.0408	1.44	0.1500
vh_din	-0.0001	0.0008	-0.17	0.8676
vh_sale_end	-0.0091	0.0061	-1.49	0.1356
vh_value	0.0000	0.0000	2.17	0.0301
vh_age	0.0014	0.0060	0.24	0.8141

TABLE 2.8: Coefficients de la régression retournés par R

2.2.3 Sélection des variables

Après une première sélection de variables réalisée précédemment, il est possible qu'il y ait encore des variables qui ne soient pas significatives dans le modèle. Nous décidons alors d'effectuer les mêmes procédures de sélection des variables pour le coût des sinistres que nous avons pratiqué pour modéliser la fréquence.

Le résultat est obtenu grâce à la commande `step`, donnant alors les variables explicatives pour la loi Gamma suivantes :

- `pol_coverage`
- `vh_fuel`
- `vh_sale_end`
- `pol_bonus`
- `drv_age_lic1`
- `vh_value`
- `vh_type`
- `drv_drv2`
- `drv_age_lic2`

Tandis que pour l'autre loi nous avons les covariables suivantes :

- `pol_coverage`
- `vh_fuel`
- `vh_sale_end`
- `pol_bonus`
- `drv_age_lic1`
- `vh_value`

En appliquant ensuite la commande `anova`, nous pouvons retirer des variables qui ne sont pas significatives dans les modèles. Nous obtenons ainsi le modèle avec la loi Gamma :

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.7918	0.0811	83.74	0.0000
pol_coverage	-0.0793	0.0150	-5.29	0.0000
vh_sale_end	-0.0085	0.0024	-3.59	0.0003
vh_type	0.0963	0.0367	2.63	0.0086
vh_value	0.0000	0.0000	4.05	0.0001

TABLE 2.9: Coefficients de la régression après la sélection

Tandis que pour la loi inverse-gaussienne on a :

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.7979	0.0533	127.57	0.0000
pol_coverage	-0.0759	0.0140	-5.40	0.0000
vh_fuel	0.1022	0.0234	4.37	0.0000
vh_sale_end	-0.0092	0.0023	-4.00	0.0001
vh_value	0.0000	0.0000	4.92	0.0000

TABLE 2.10: Coefficients de la régression après la sélection

2.2.4 Sélection du meilleur modèle GLM et la validité du modèle

Nous obtenons un GLM pour la loi Gamma prenant en compte 4 variables et un GLM pour la loi inverse-gaussienne contenant aussi 4 variables. Pour choisir le meilleur modèle, nous allons comparer leur QQ-plot des résidus estimés (les graphiques en haut de 2.6) et on va s'intéresser également aux graphes des résidus estimées en fonction des valeurs prédites (les graphiques en bas de 2.6).

Nous constatons d'après le QQ-plot des deux modèles que les résidus sont répartis normalement. Ensuite, nous observons que les résidus du modèle ayant choisi la loi Gamma restent globalement répartis autour de 0. Cependant, les résidus de l'autre modèle sont répartis en-dessous de 0 ce qui traduit une adéquation plutôt mauvaise du modèle. Ainsi le modèle avec la loi Gamma est plus légitime.

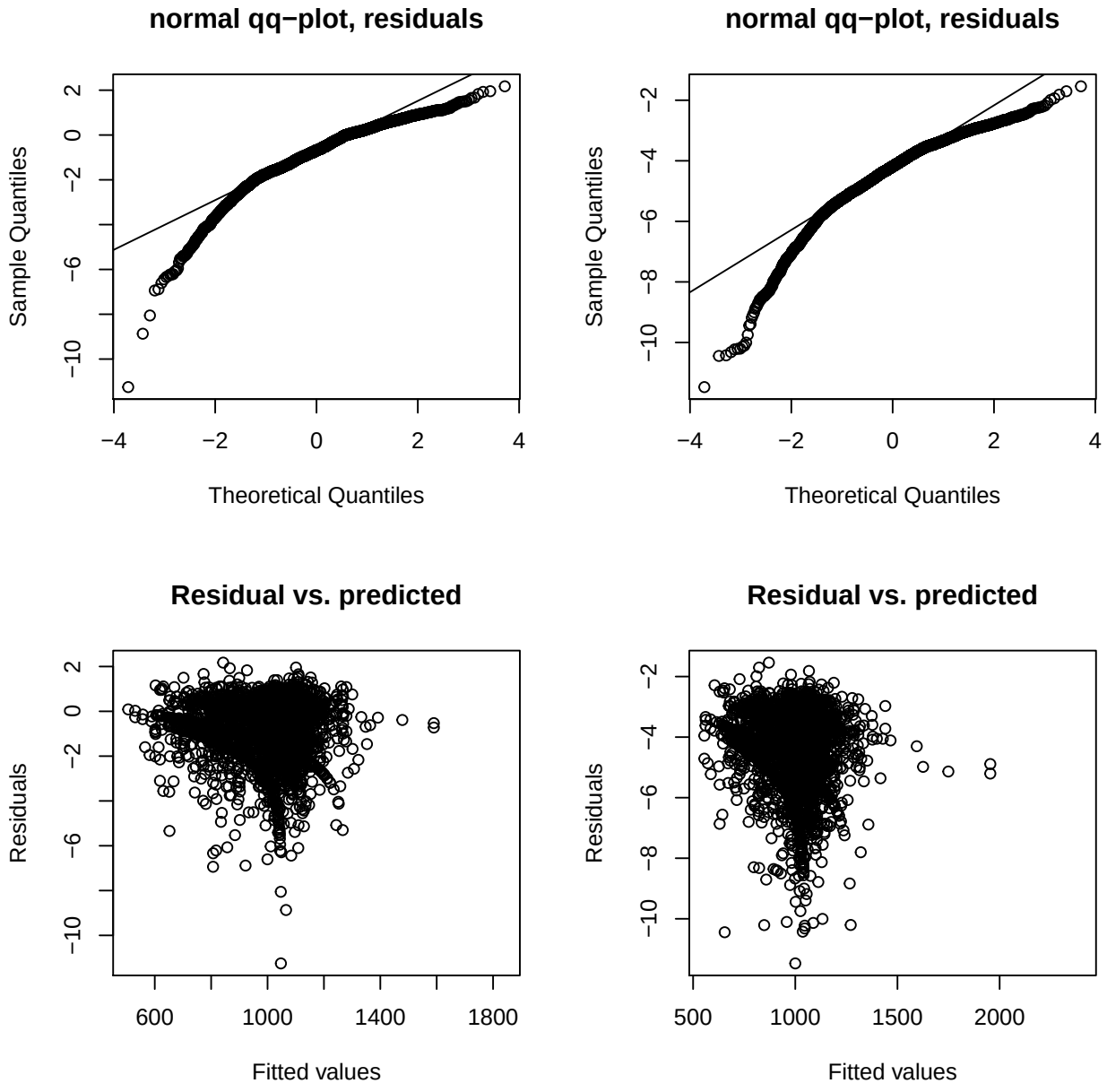


FIGURE 2.6: QQ-plot et Residual vs. predicted des deux modèles

2.3 Calcul de la prime pure pour les polices étudiées

Pour le calcul de la prime pure, nous allons utiliser les valeurs des espérances estimées de la fréquence des accidents (notée N) et de leur sévérité (notée B). Nous utiliserons les coefficients de la fréquence et de la sévérité que nous avons obtenus avec les modèles construits précédemment.

Les variables explicatives des deux modèles sont :

En utilisant la fonction logarithme que nous avons appliqué comme fonction de lien pour expliquer la variable N, nous obtenons la relation de son espérance suivante :

i	x_i
1	vh_sale_end
2	vh_value
3	pol_coverage
4	vh_fuel
5	drv_drv2
6	pol_sit_duration
7	pol_duration
8	vh_age
9	vh_weight
10	drv_age_lic1
11	drv_age1
12	pol_bonus

TABLE 2.11: Variables explicatives pour modéliser la fréquence

i	x_i
1	pol_coverage
2	vh_sale_end
3	vh_type
4	vh_value

TABLE 2.12: Variables explicatives pour modéliser la sévérité

$$\mathbb{E}[N] = \exp\{\beta_0 + \sum_{i=1}^{12} \beta_i \times x_i\}. \tag{2.1}$$

De même, on a choisi la même fonction de lien pour expliquer la variable B :

$$\mathbb{E}[B] = \exp\{\gamma_0 + \sum_{j=1}^{4} \gamma_j \times x_j\}. \tag{2.2}$$

En appliquant les relations (2.1) et (2.2), nous avons la formule du calcul de la prime pure suivante :

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[N] \times \mathbb{E}[B] \\ &= \exp\{\beta_0 + \sum_{i=1}^{12} \beta_i \times x_i\} \times \exp\{\gamma_0 + \sum_{j=1}^{4} \gamma_j \times x_j\}. \end{aligned} \tag{2.3}$$

Nous avons donc le résultat suivant :

```
> summary(primepure$prime_pure)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.56  180.46  260.92  273.23  351.18 2299.30
```

2.3.1 Visualisation

Nous pouvons aussi regarder la répartition de primes pures des polices étudiés dans l'histogramme 2.7, on peut constater que la plupart des montants de prime pure est inférieur à 500.

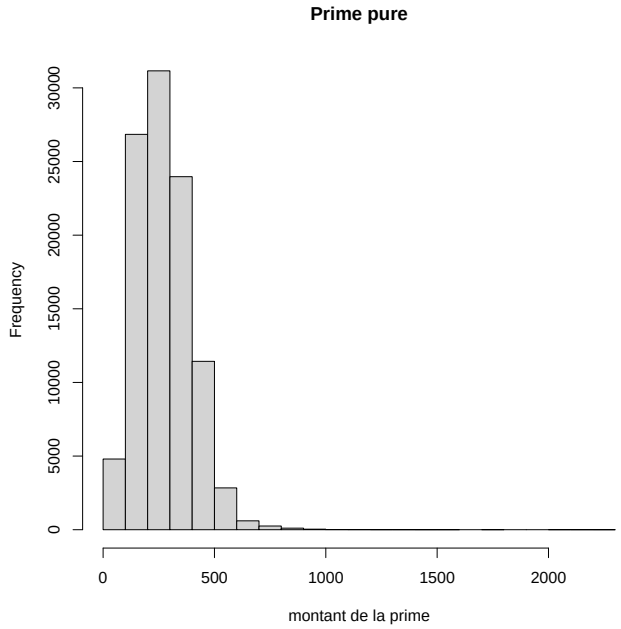


FIGURE 2.7: Histogramme des montants de prime pure

2.3.2 Influences sur la prime pure

Nous allons étudier les impacts de certaines variables sur la prime pure que nous avons déterminé. Les variables que nous avons choisi de présenter sont celles ayant des tendances significatives sur la prime pure.

L'assuré D'après les graphiques obtenus dans la Figure 2.8, nous pouvons remarquer que la différence de la moyenne des primes pures entre Femme et Homme est assez petite (un écart à peu près de 8). En effet la moyenne pour Homme est légèrement supérieure, ainsi nous pouvons dire que le sexe n'affecte pas la prime pure.

Ensuite, si on regarde sur le graphique concernant le nombre de conducteurs, il y a un écart significatif des primes pures entre un seul conducteur et deux conducteurs. Ainsi le nombre de conducteurs a un impact sur le montant de la prime pure.

Nous remarquons aussi que les jeunes conducteurs et les usagers ayant une courte durée de détention du permis ont des montants de prime pure plus élevés. En particulier, plus l'usager est âgé, moins le montant de la prime pure sera élevé. Cependant, plus le client a une durée de détention du permis longue, plus le montant sera élevé.

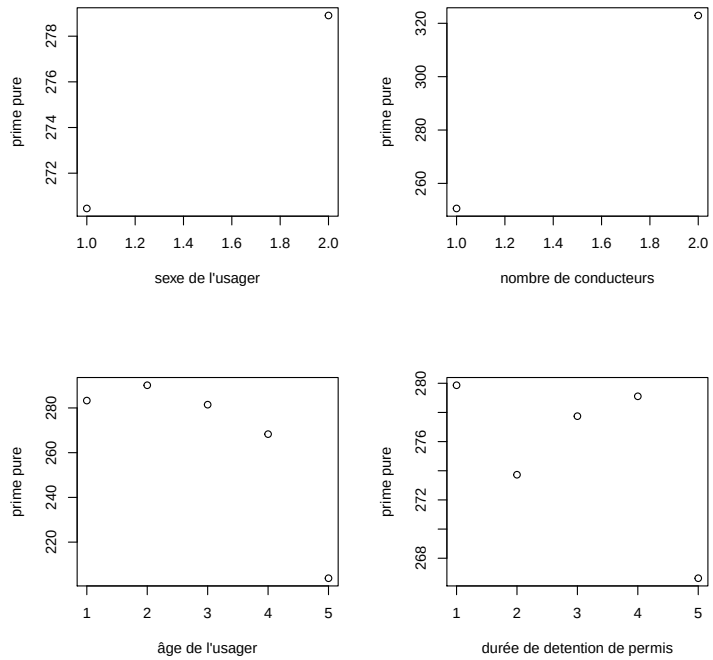


FIGURE 2.8: Données sur l'usager expliquant le montant de la prime pure

Véhicule Concernant l'âge du véhicule et l'année de fin de vente, nous pouvons observer, d'après 2.9, que plus le véhicule est récent, plus la moyenne des montants de la prime pure sera importante. De plus, nous voyons une décroissance significative de la prime pure en fonction de l'âge du conducteur.

Pour le type de carburant, on remarque que les véhicules diesel (1) et hybrides (3) ont des montants beaucoup plus élevés que le véhicule essence (2).

Nous constatons aussi que le véhicule de tourisme a une moyenne de primes pures plus élevées que le véhicule commercial.

Ensuite, nous pouvons remarquer que dans les graphiques sur la vitesse maximale, la cylindrée du véhicule et la valeur du véhicule, il y a des tendances croissantes. Ainsi, nous pouvons dire que plus ces valeurs sont grandes, plus la prime pure sera élevée.

Police d'assurance Dans la Figure 2.10, le premier graphique nous indique que les polices d'assurance ayant une couverture maximale peuvent atteindre un montant de prime pure beaucoup plus important que les autres types de couverture.

Finalement, plus la la police d'assurance est récente, plus l'assuré aura un montant de prime pure élevé, ce qui est illustré dans les deux autres graphiques de 2.10 concernant l'âge de la police d'assurance.

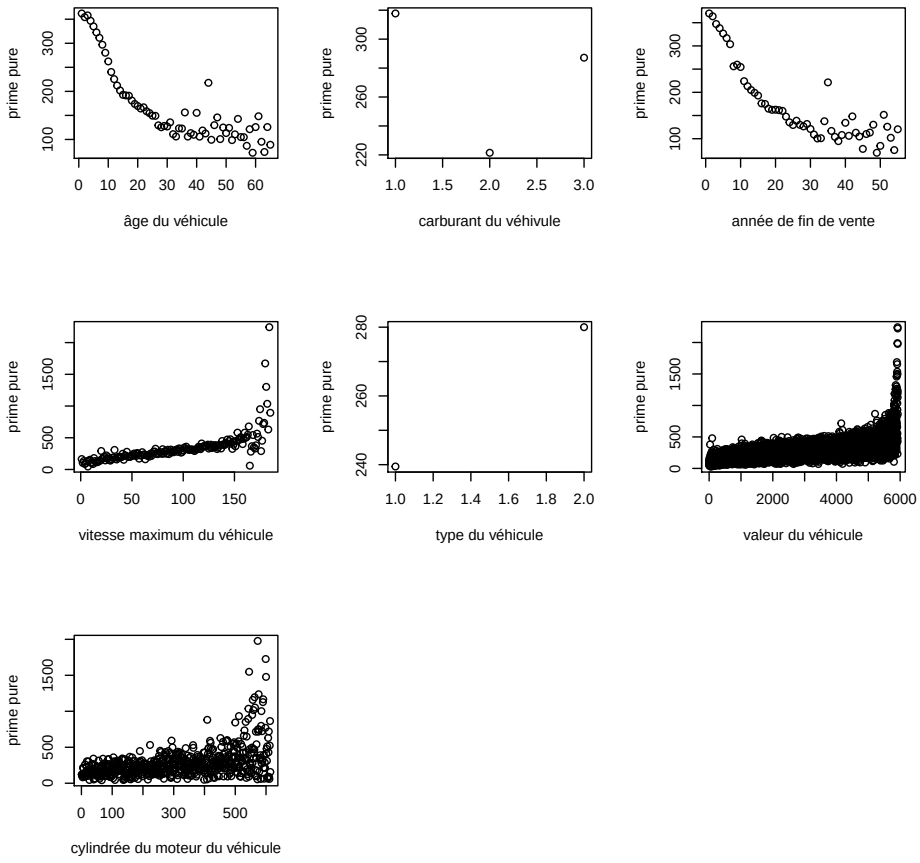


FIGURE 2.9: Données sur le véhicule expliquant le montant de la prime pure

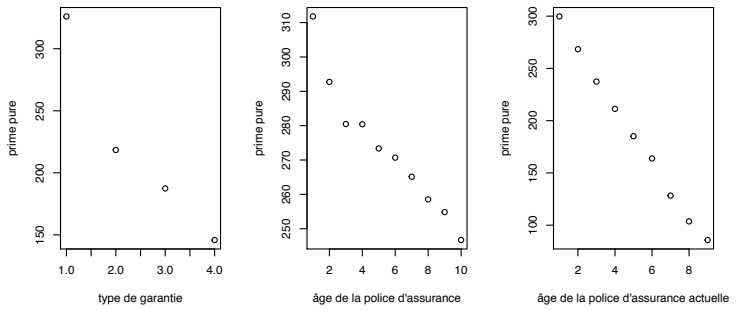


FIGURE 2.10: Données sur les polices d'assurance expliquant le montant de la prime pure

Chapitre 3

Ajout des données ONISR

Les données ONISR sont les données gouvernementales de l'Observatoire national interministériel de la sécurité routière, ONISR (2018) où tout accident corporel survenant en France entre 2005 et 2017 y est répertoriée. L'observatoire met à disposition 4 fichiers annuels sur les caractéristiques de l'accident, le lieux des accidents, les véhicules impliquées et les usagers impliqués.

Nous testerons dans cette partie la pertinence de l'apport de ces données nationale à notre GLM. Nous allons employer les mêmes méthodes que dans les sections précédentes puis comparerons, à la fin, les tarifs obtenus.

3.1 Exploration des données ONISR et premières remarques

Nous observons en premier lieu que peu de variables sont communes à nos deux bases de données utilisées dans ce projet. Nous essayons donc de trouver des variables qui s'en rapprochent comme l'année de naissance `An_nais` ou le sexe de l'usager `sexe`. Nous prendrons aussi des variables qui diffèrent de nos modèles afin d'avoir une finesse supplémentaire dans nos modèles.

Tout d'abord nous cherchons les données qui peuvent expliquer soit la fréquence d'accidents, soit la sévérité d'un sinistre ou les deux à la fois. Ainsi, intuitivement, on se dirige vers des données évidentes telles que les variables se rapportant à la route sur laquelle est survenu le sinistre, mais aussi les données concernant les accidentés.

Pour les données de la base `véhicules`, nous avons tout d'abord vu que nous n'allons pas utiliser toutes les catégories de véhicules. En effet, nous ne gardons que les catégories de véhicules présentes dans la base `pg17testyear1`. Nous effectuons donc une sélection dans la base `véhicules` et obtenons alors des données homogènes entre les deux bases de données.

Nous avons fais le choix de garder seulement les variables que nous pouvons considérer comme propres aux individus. Par exemple, les données sur la manoeuvre principale lors de l'accident `manv` ne nous semble pas intéressante a étudier car pas intrinsèque à des groupes de population.

La gravité est une variable qui nous semble être intéressante pour approcher la sévérité d'un sinistre. La gravité de la blessure de l'usager est un signe sur le montant de dégâts d'un accident. Nous avons la répartition de la gravité des sinistres avec l'histogramme 3.1.

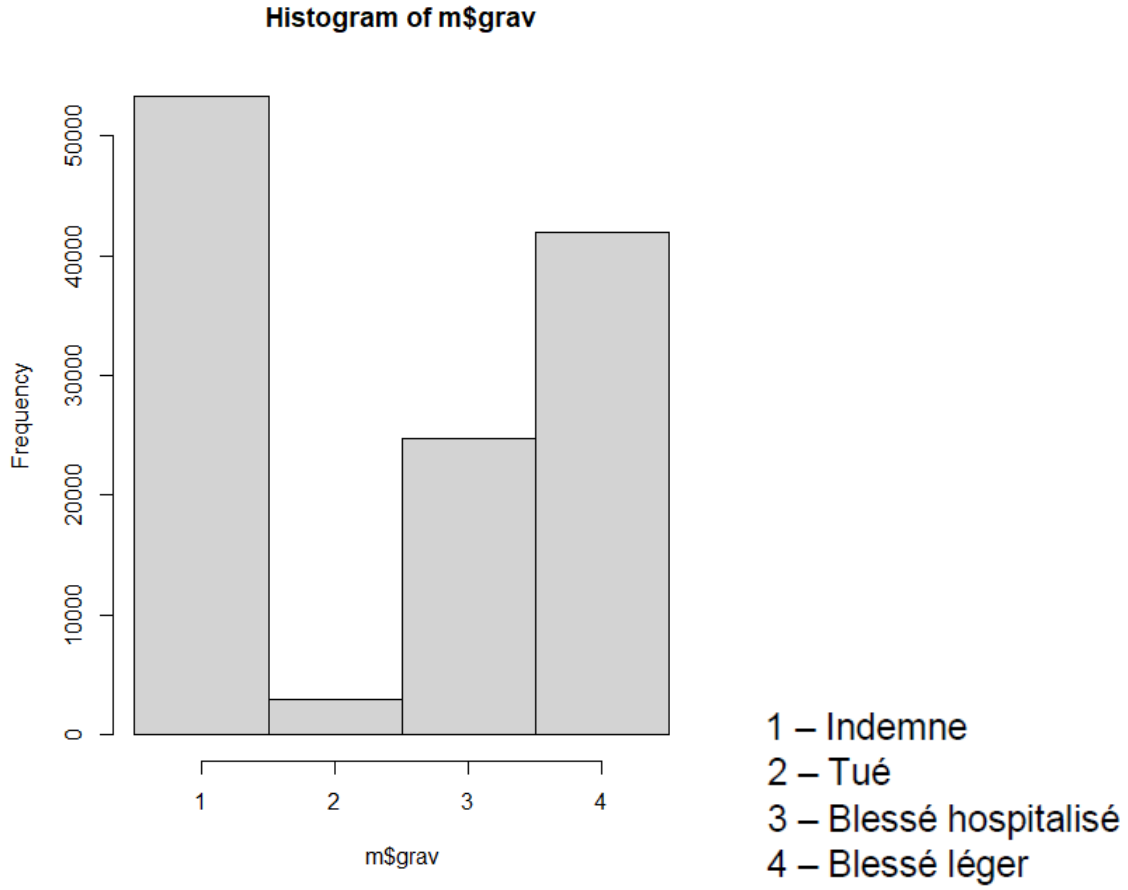


FIGURE 3.1: Gravité des sinistres de la base de données

3.2 Modélisation de la sévérité et de la fréquence sur les données ONISR

L'idée que nous allons mettre en place pour incorporer les nouvelles données apportées par la base ONISR est le partitionnement des données. Il s'agit de définir des groupes homogènes à l'aide des informations dont on dispose sur ces données. Comme nous disposons des codes postaux des communes des assurés dans la base de données `pg17testyear1`, nous allons organiser les départements métropolitains en groupes, aussi appelés "clusters". Nous aurons donc de nouvelles variables qui affineront nos GLM exposeés dans la première partie.

Le but va être de regrouper les départements en fonction de deux scores différents. Le premier sera un score reflétant la fréquence de sinistres et le second reflétera la sévérité de ces sinistres. Pour créer ces scores, nous prenons différentes variables disponibles dans les données ONISR qui expliquent soit la fréquence de sinistres, soit leur gravité.

3.2.1 Pour la fréquence

Pour lisser les différences entre les départements, notamment au niveau de la densité de population, nous avons importé les recensements de population par département afin d'observer de manière plus objective les départements où sont comptés le plus de sinistres. Nous avons donc importé une base de

données comportant le nombre d'habitant par département.

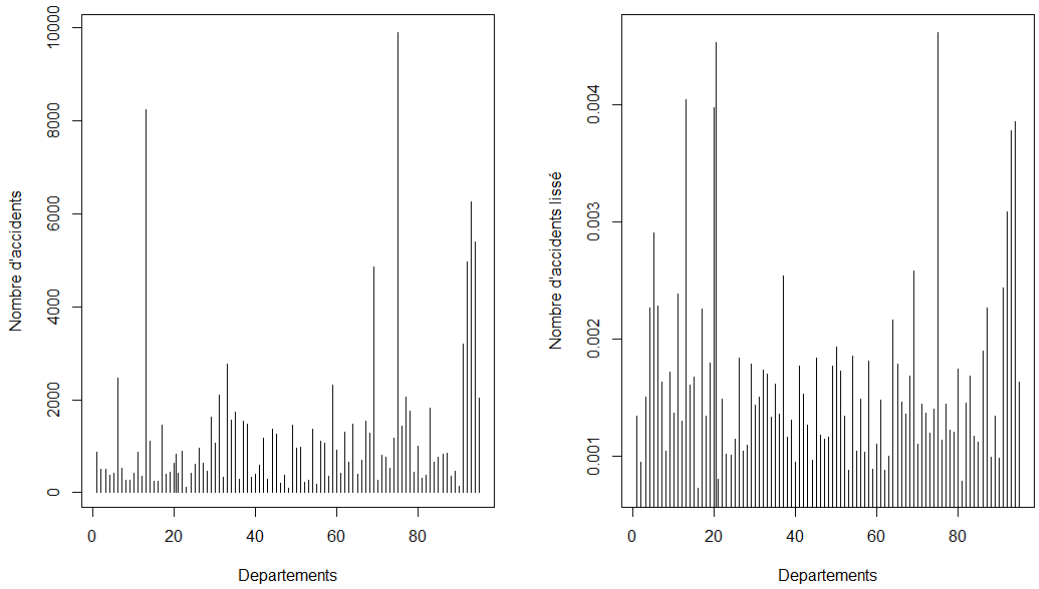


FIGURE 3.2: Nombre d'accidents par départements (non lissé et lissé)

On observe notamment le cas de la Corse : comparé à sa population, le nombre d'accidents y est très élevé. Aussi, on observe une nette corrélation entre les zones très urbaines et le nombre d'accidents, ce qui semble assez cohérent. D'une façon plus visuelle, nous avons la carte 3.3.

Comme le nombre d'habitant par département joue un rôle important dans le nombre d'accidents dénombrés, il est intéressant de créer un score de population nommé `scorepop` qui permet d'attribuer une valeur dans [0,1]. On remarque que les zones très urbaines ont tendance à avoir un nombre d'accidents plus élevé que les départements plus ruraux. On crée donc un score lié aux sinistres en agglomération que l'on a nommé `scoreagg`. Plus ce score est proche de 1, plus le nombre de sinistres arrivés en agglomération représente une grosse part dans le nombre de sinistres total. Cela va permettre d'avoir une certaine représentation de l'urbanisation des départements.

Nous prenons également en compte les conditions de surface de sol lorsque le sinistre a eu lieu. En effet, après avoir retiré les données en conditions "normales", on s'intéresse aux conditions liées à la pluie et à la neige. On a donc un score de surface nommé `scoresurf`.

Nous utilisons les données sur les intersections. Le score lié aux intersections est nommé `scoreint`. Nous avons regroupé tous les types d'intersections en un seul groupe par soucis de simplicité. Les formes d'intersections ne semblaient pas importantes.

Nous normalisons les scores afin de ne pas avoir de problème de poids dans les scores. En sommant puis normalisant les scores, on obtient un score `scorefreq` qui représente la fréquence de sinistres par département. Plus un département est proche de 1 plus il a de facteurs pouvant augmenter le nombre de sinistres.

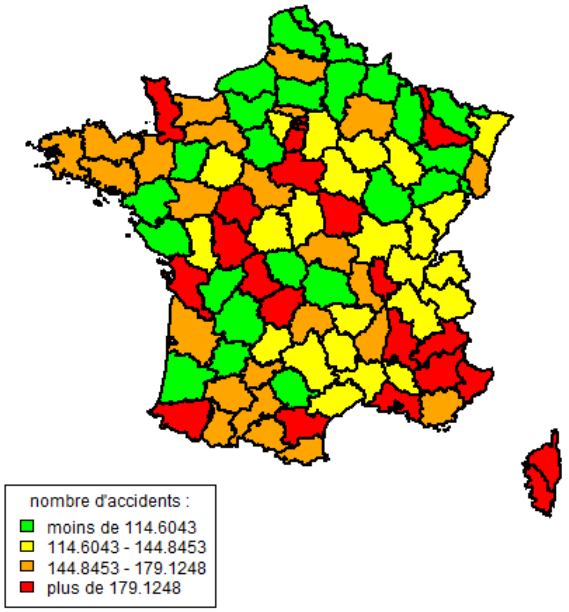


FIGURE 3.3: Carte du nombre d'accidents par départements (lissé)

3.2.2 Pour la sévérité

Dans les données ONISR, cinq variables expliquent la sévérité des sinistres : le type d'obstacle touche `obsm`, le type de collision `col`, le point de choc initial `choc`, la localisation de l'accident (en ou hors agglomération) `agg` et enfin la gravité de blessure de l'usager `grav`.

De la même manière que pour la fréquence, nous créons des scores en fonction des données trouvées dans la base. Pour le type d'obstacle heurté, le fait de s'intéresser aux données qui concernent les piétons et les véhicules nous permet d'expliquer un haut montant de sévérité. Pour le type de collision, nous ne gardons que les collisions entre deux véhicules et nous ne les différencions pas. Pour les chocs, nous observons seulement les chocs les plus graves que sont les chocs multiples. Pour la variable de gravité de l'accident, nous ne conservons que les données concernant les décès et les hospitalisations. Ces choix nous permettent d'avoir une certaine idée sur la sévérité des sinistres par département.

Nous obtenons le score de sévérité `scoresev`, par somme et normalisation des scores : `scorecoll`, `scoreobsm`, `scorechoc`, `scoreagg`, `scoregrav`.

3.2.3 Groupe de départements

Ainsi, pour chaque département nous avons un score de fréquence et de sévérité. Grâce à ces scores, nous pouvons créer des clusters de départements, on peut y distinguer trois groupes clairs d'après 3.4

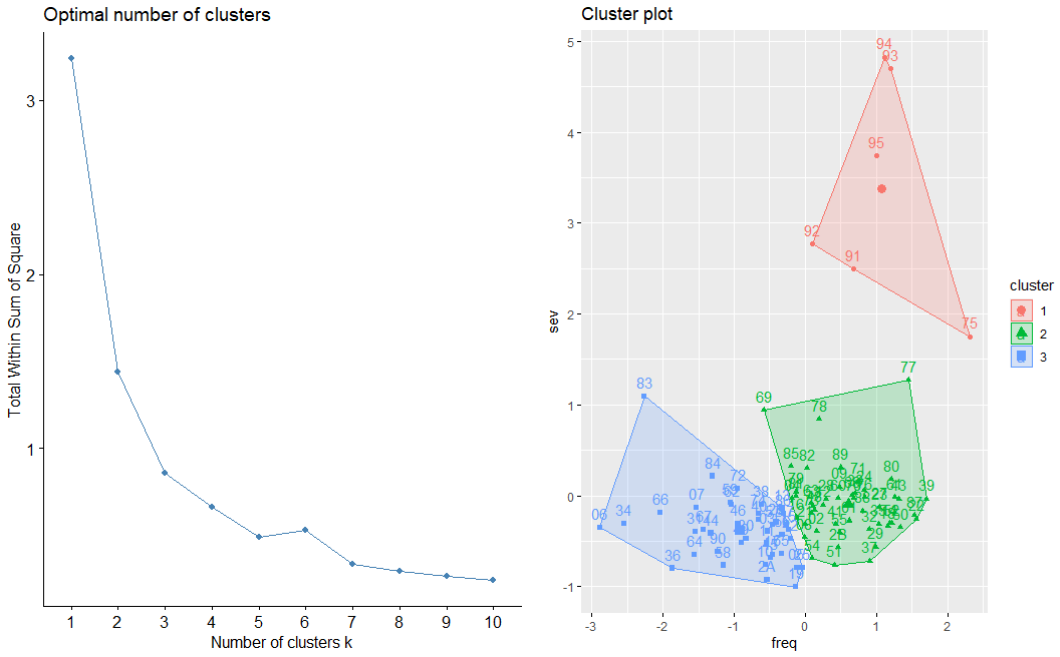


FIGURE 3.4: Groupes de départements

Nous utilisons la technique des "coudes" pour obtenir le nombre de groupes `k` optimal. Parfois, comme c'est le cas pour les groupes de fréquence 3.5, le choix du nombre optimal peut ne pas être clair. Dans ce cas nous essayons les différentes valeurs probables de `k` et nous sélectionnons la valeur `k` qui permet d'avoir une stabilité des groupes malgré les différents points de départ de l'algorithme.

Pour inclure ces groupes dans les modèles linéaires généralisés on va créer des clusters de fréquence et des cluster de sévérité. Cela permettra d'avoir une meilleure précision. L'avantage de séparer les groupes est de pouvoir créer peu de groupe de départements pour la fréquence et pour la sévérité mais d'avoir une plus grande diversité de clusters.

Cluster de fréquence Grâce aux différents scores expliquant la fréquence, on peut créer trois groupes de départements. Ces groupes sont visibles dans 3.5. En essayant avec cinq groupes, les clusters ne sont pas stables lorsque nous changeons de département de départ de l'algorithme contrairement a nos essais avec trois groupes. Nous prenons donc un nombre de clusters égal à trois.

Cluster de fréquence Grâce aux différents score expliquant la sévérité, on peut créer trois groupes de départements. Ces groupes sont visibles dans 3.6.

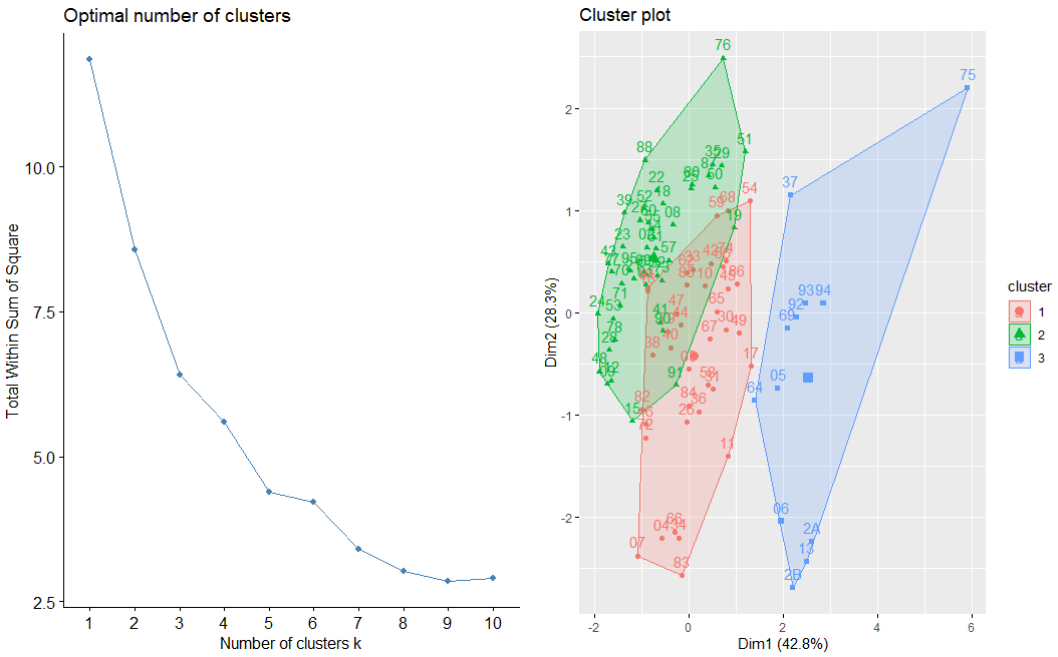


FIGURE 3.5: Groupes de départements (fréquence)

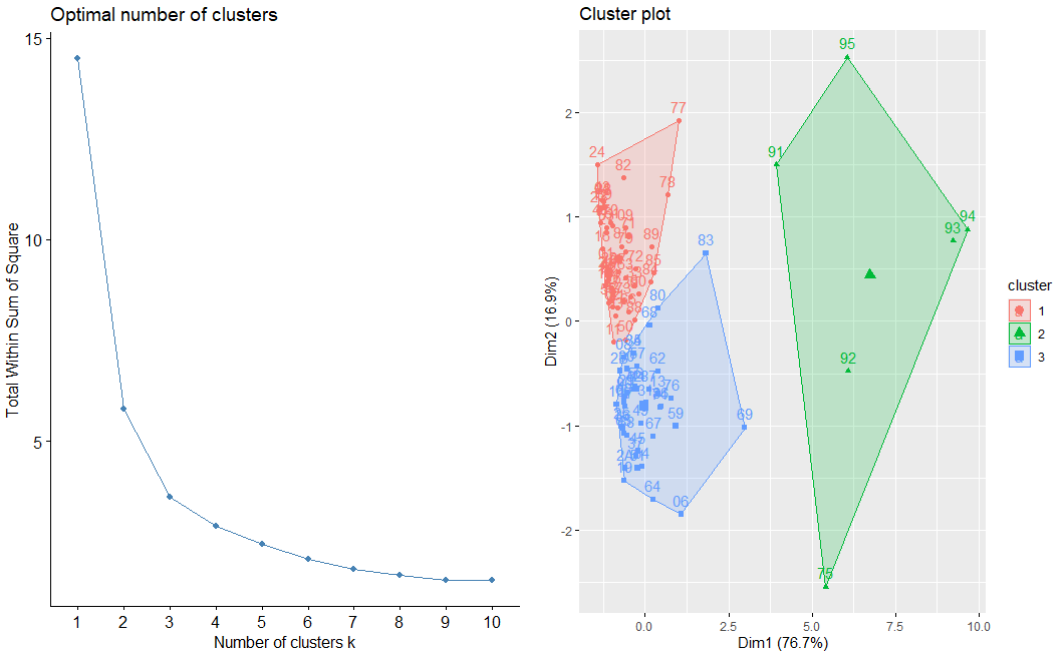


FIGURE 3.6: Groupes de départements (sévérité)

Résumé des clusters Ainsi, chaque département possède 2 groupes : un groupe que l'on utilisera dans le GLM de fréquence et un groupe pour le modèle de sévérité.

Dep	clusterfreq$cluster	clustersev$cluster
01	1	3
02	1	3
03	2	3
04	2	3
05	3	2
06	3	2
07	2	3
08	1	2
09	1	3
10	2	2
11	2	3
12	1	3

FIGURE 3.7: Exemple de clusters par département

3.2.4 Création des nouveaux modèles

On a créé deux variables explicatives supplémentaires : `ClusterFreq` et `ClusterSev` à partir des données ONISR. Ces variables contiennent les groupes auxquels appartiennent tous les départements métropolitains. Ces nouvelles variables ont pour objectif d'affiner la précision de notre modèle précédemment créé.

Modèle de fréquence

On intègre `ClusterFreq` dans le modèle de fréquence obtenu précédemment et on a donc les variables explicatives 3.1.

On vérifie que le modèle nouvellement créé est toujours correct grâce à des tests ANOVA 3.8 et des diagrammes Quantile-Quantile 3.9.

Modèle de sévérité

On intègre `ClusterSev` dans le modèle de sévérité obtenu précédemment et on a donc les variables explicatives 3.2.

Par des tests ANOVA, on voit bien que le modèle est plus fin que le précédent grâce aux nouvelles variables créées à partir des données ONISR.

i	x_i
1	vh_sale_end
2	vh_value
3	pol_coverage
4	vh_fuel
5	drv_drv2
6	pol_sit_duration
7	pol_duration
8	vh_age
9	vh_weight
10	drv_age_lic1
11	drv_age1
12	pol_bonus
13	ClusterFreq

TABLE 3.1: Variables explicatives pour modéliser la fréquence

i	x_i
1	pol_coverage
2	vh_sale_end
3	vh_type
4	vh_value
5	ClusterSev

TABLE 3.2: Variables explicatives pour modéliser la sévérité

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			102079	48983		
vh_sale_end	1	721.58	102078	48261	< 2.2e-16	***
vh_value	1	274.23	102077	47987	< 2.2e-16	***
pol_coverage	1	120.78	102076	47866	< 2.2e-16	***
vh_fuel	1	128.16	102075	47738	< 2.2e-16	***
drv_drv2	1	91.89	102074	47646	< 2.2e-16	***
pol_sit_duration	1	50.52	102073	47595	1.178e-12	***
pol_duration	1	26.55	102072	47569	2.566e-07	***
vh_age	1	10.63	102071	47558	0.0011150	**
vh_weight	1	15.62	102070	47543	7.738e-05	***
drv_age_lic1	1	12.86	102069	47530	0.0003362	***
drv_age1	1	25.15	102068	47505	5.316e-07	***
pol_bonus	1	8.18	102067	47496	0.0042254	**
ClusterFreq	1	24.90	102066	47472	6.047e-07	***

FIGURE 3.8: ANOVA pour la fréquence

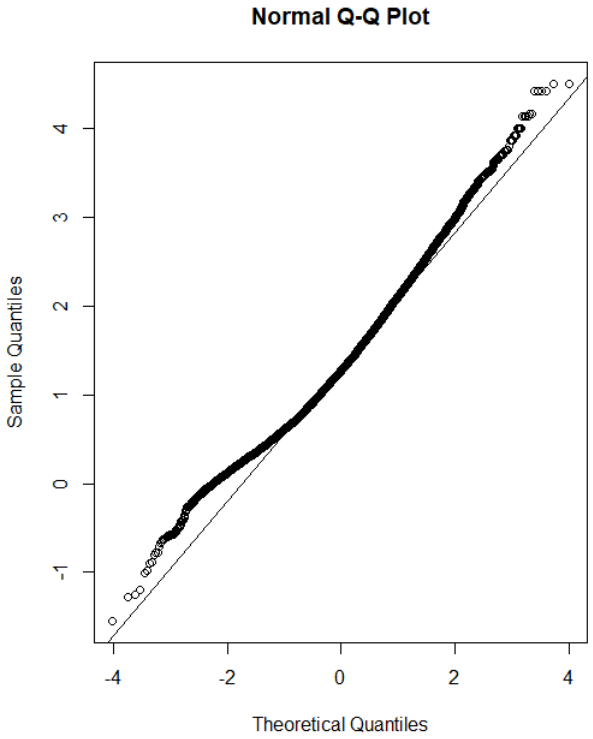


FIGURE 3.9: QQ-plot du modèle de fréquence

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			14314	14467		
pol_coverage	1	160.946	14313	14306	< 2.2e-16	***
vh_sale_end	1	30.790	14312	14275	6.792e-06	***
vh_type	1	9.901	14311	14265	0.01072	*
vh_value	1	25.932	14310	14239	3.629e-05	***
clusterSev	1	29.791	14309	14209	9.577e-06	***

FIGURE 3.10: ANOVA pour la sévérité

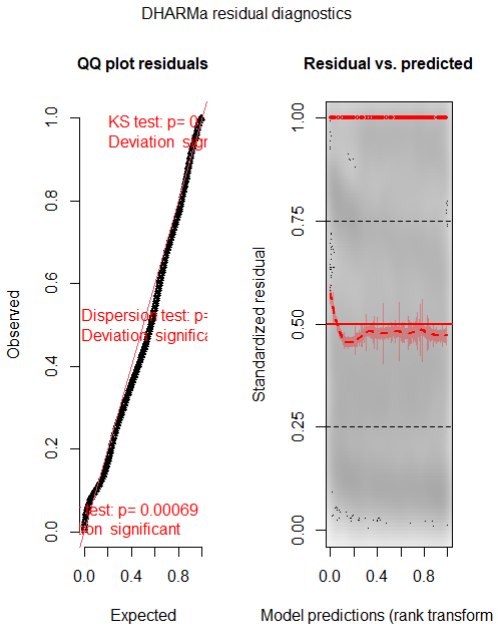


FIGURE 3.11: QQ-plot du modèle de sévérité

3.3 Prime pure ONISR

3.3.1 Visualisation

En utilisant les formules vues dans la section 2.3, nous pouvons calculer la prime pure ONISR, nous obtenons la répartition 3.12

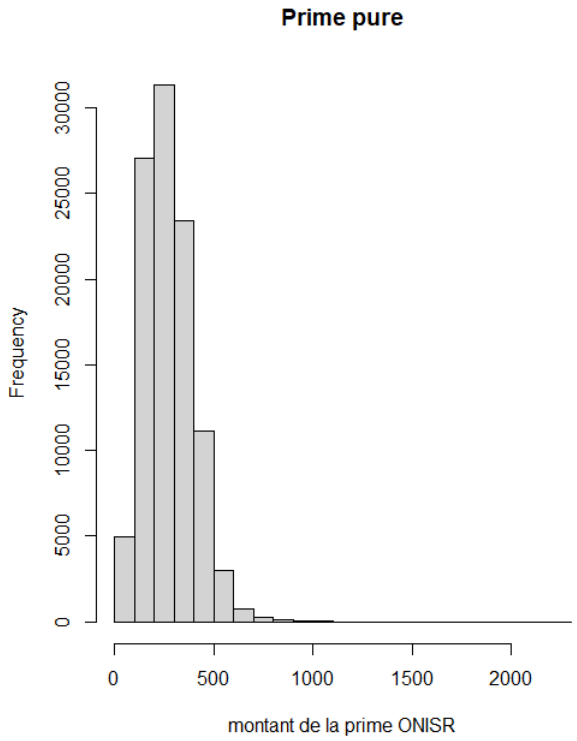
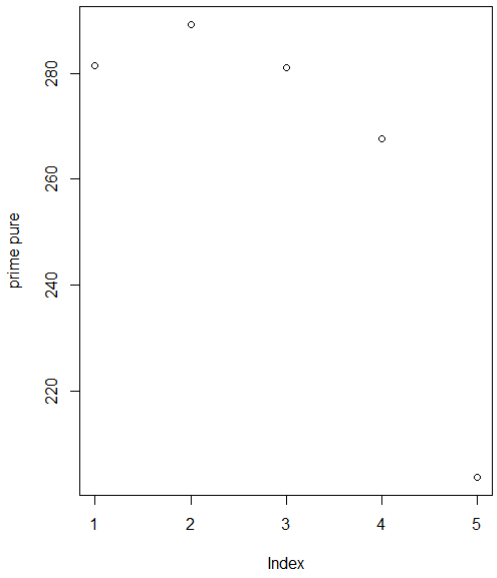


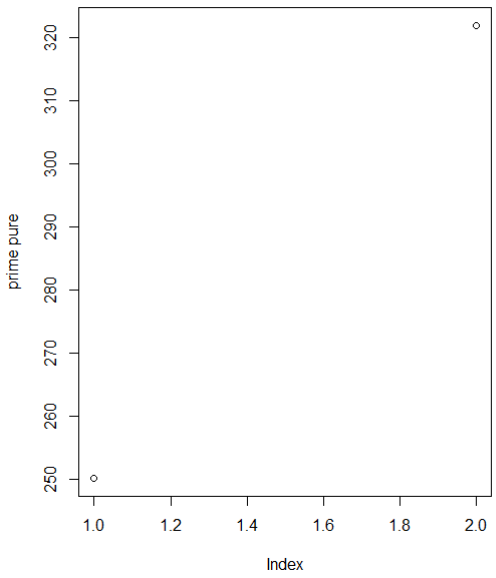
FIGURE 3.12: Histogramme des montants de prime pure ONISR

3.3.2 Influences sur la prime pure ONISR

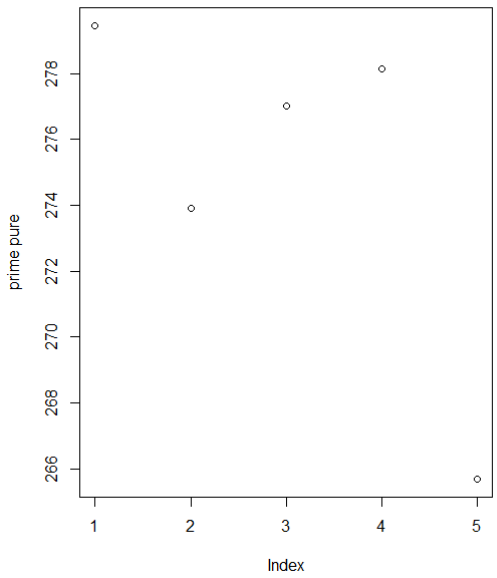
Nous pouvons observer les influences des différentes variables sur la prime pure obtenue. Nous ne montrons que les variables qui ont une forte influence. Les autres ne semblent pas être déterminantes dans les variations des primes pures.



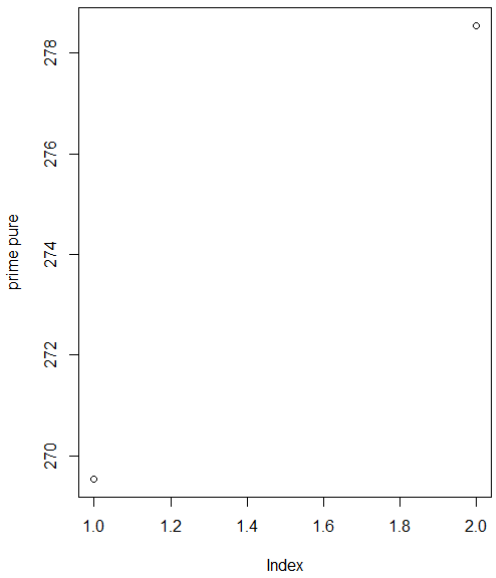
(a) Âge de l'usager



(b) Nombre de conducteurs

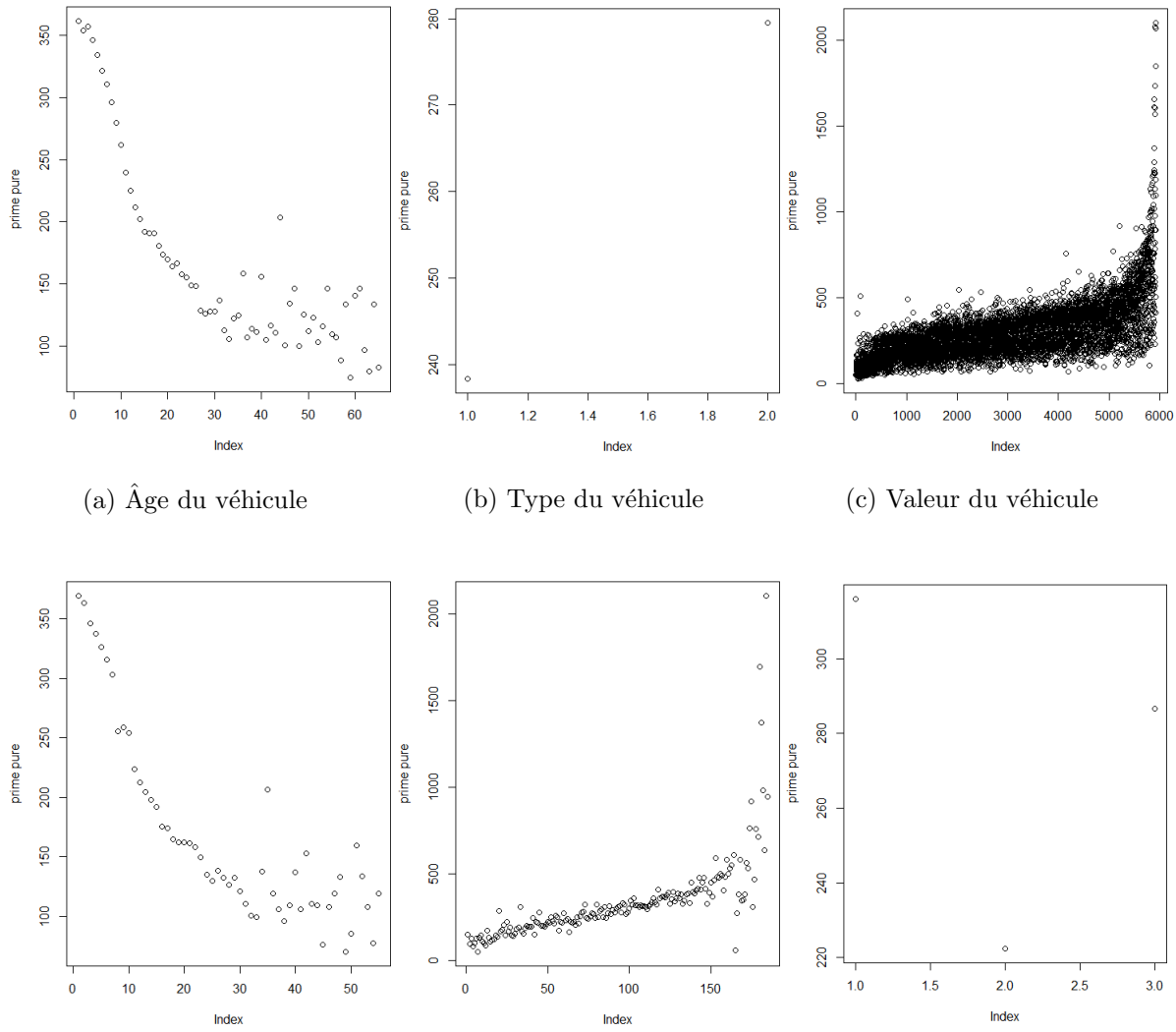


(c) Durée de détention du permis de conduire



(d) Sexe

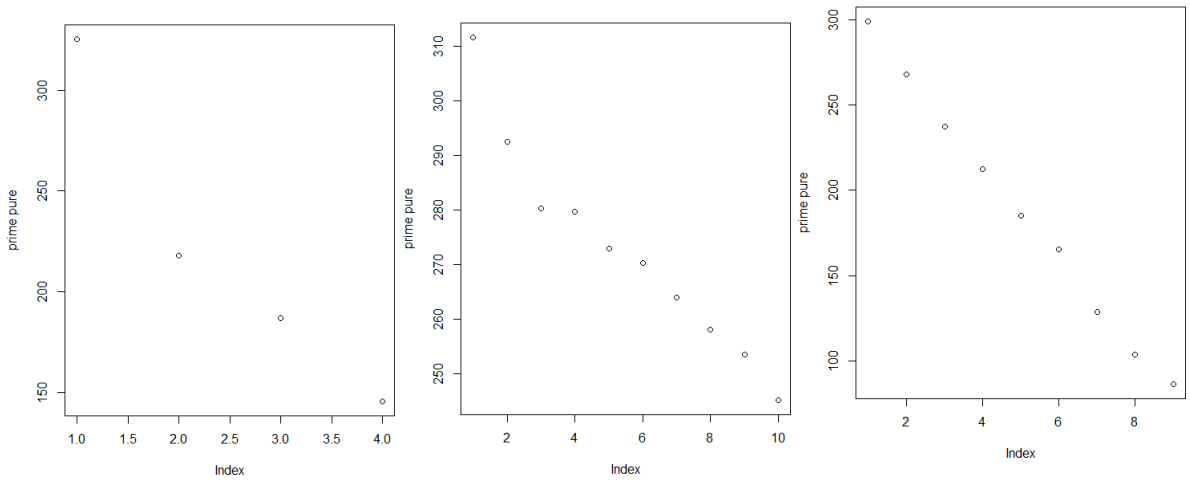
FIGURE 3.13: Données sur l'usager expliquant le montant de la prime pure usager



(a) Âge du véhicule (b) Type du véhicule (c) Valeur du véhicule

(d) Année de fin de vente (e) Vitesse maximum du véhicule (f) Carburant du véhicule

FIGURE 3.14: Données sur le véhicule expliquant le montant de la prime pure



(a) Types de garanties (b) Âge de la police d'assurance (c) Âge de la police d'assurance actuelle

FIGURE 3.15: Données sur les polices d'assurance

L'assuré Certaines variables relatives à l'assuré peuvent expliquer certaines valeurs de primes pures. En effet, la durée de détention du permis de conduire 3.13c a une influence notable sur le montant. Ce qu'on appelle "les jeunes permis" ont, en moyenne, un montant de prime pure plus élevé que les détenteurs de permis plus anciens. Cependant, on remarque qu'en dehors des cas de "jeunes permis", plus un usager a un permis depuis longtemps, plus sa prime pure, en moyenne, sera élevée.

Pour le sexe de l'assuré, la différence de prime pure moyenne entre Homme et Femme n'est pas si élevée d'après 3.13d. On peut donc dire qu'il ne s'agit pas, ou peu, d'un facteur pouvant augmenter ou diminuer le montant de la prime pure.

En regardant le nombre de conducteurs sur le véhicule 3.13b, on remarque qu'il s'agit d'un facteur qui a une influence élevée sur le montant de prime pure. On y remarque une différence moyenne de plus de 70 entre un et deux conducteurs ; c'est équivalent à plus de 20% de différence.

Véhicule La prime pure calculée à partir des données ONISR est notamment expliquée par certaines variables intrinsèques au véhicule de l'assuré.

On observe en 3.14 que l'âge du véhicule 3.14a (qui est par ailleurs très fortement corrélé à l'année de fin de vente 3.14d du véhicule comme le rappelle le tableau de corrélations 1.11) influe de manière "décroissante" sur la prime pure. En effet, plus le véhicule de l'usager est récent, plus la prime pure associée est élevée.

Certaines autres caractéristiques du véhicules influent de manière "croissante". C'est le cas de la vitesse maximale 3.14e. Plus le véhicule de l'assuré peut atteindre une vitesse élevée, plus l'assuré a tendance à avoir une prime pure élevée. On observe également un résultat similaire avec la valeur du véhicule 3.14c.

Un usager avec un véhicule de tourisme aura tendance à avoir une prime pure plus élevée qu'un usager avec un véhicule de type commercial. Cette différence est assez notable avec 3.14b.

En observant le type de carburant en 3.14f, un véhicule essence (2) aura tendance à baisser la prime pure, au contraire d'un véhicule diesel (1). Pour les véhicules hybrides (3), le marché étant limité en 2017, nous ne faisons pas d'analyse car nous ne possédons pas assez de données pour en généraliser les effets.

Police d'assurance En regardant les variables se rapportant aux polices d'assurance souscrites 3.15c et 3.15b, nous observons que plus la police d'assurance est ancienne, plus le souscripteur aura, en moyenne, un montant faible de prime pure. Pour le type de garantie de la police d'assurance 3.15a, c'est étonnant d'observer que plus l'assuré est couvert au niveau des risques, moins sa prime pure associée sera, en moyenne, réduit.

3.4 Comparaison avec les tests `pg17testyear1`

Pour la prime pure des données `pg17testyear1` :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18.86	180.44	260.06	273.35	351.62	2242.97

Pour la prime pure des données ONISR :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18.72	179.23	259.08	272.98	350.44	2290.08

Ainsi, en ajoutant les données ONISR, nous avons réduit la somme des montants de prime pure des assurés. En moyenne, les primes pures ont baissé de 0,37 avec cet ajout de données. On remarque que pour les plus hauts montants, la prime pure a augmenté. On peut en déduire que l'ajout des scores à partir des données ONISR a permit d'affiner les prédictions des modèles de fréquence et de sévérité.

Chapitre 4

Conclusion

Au cours de ce mémoire, nous avons pu modéliser la prime pure à l'aide de deux GLM, un avec et un sans les données ONISR en choisissant le modèle Binomial Négatif pour la fréquence et le modèle Gamma pour la sévérité.

Avec l'ajout de nouvelles données, nous avons pu affiner nos modèles et nos estimations. Nous avons utilisé les départements et leurs aspérités dans lesquels vivent les assurés afin de calculer la prime pure avec plus de précision. Cependant, nous pensons que notre modèle ne prend pas assez en compte d'autres facteurs importants qui peuvent influer sur l'apparition de sinistres et leur gravité. Par exemple, nous ne prenons en compte que les départements, ce qui est assez vague et ne permet pas de distinguer des souscripteurs frontaliers à certains lieux qui pourraient avoir une grande influence sur la prime pure.

A partir des données que nous possédons, nous avons pu observer les différentes variables relatives au souscripteur de la police d'assurance, au véhicule de l'assuré et à la police d'assurance souscrite elle même qui influent sur le montant de prime pure de chaque usager.