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

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

Elliot MULLER
Hugo SALLEZ
Nicolas WAGNER

Directeur de Mémoire : Christophe DUTANG
Sujet : A8

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 assurance automobile ne coûte pas le même prix pour tous les conducteurs et tous les véhicules. Les assurances doivent donc discriminer les clients en fonction de certaines caractéristiques (âge, puissance du véhicule, etc.).

Dans notre sujet, nous traitons des données issues du package CASDATASET. Nous allons utiliser les jeux de données `pg17trainpol` et `pg17trainclaim`. Le premier regroupe 100 000 polices d'assurances, pour divers clients et diverses caractéristiques, aussi bien au niveau des individus (âge, code postal, sexe, expérience...) que pour le véhicule (poids, modèle, valeur...). Le second regroupe des informations quant aux accidents survenus sur l'année 2017, en particulier *claim_amount* qui indique la valeur des dommages engendrés par l'accident.

L'objectif, à l'aide de ces deux jeux de données, est de déterminer une prime pure pour un client en fonction de ses caractéristiques et de celles de son véhicule. Cette prime sera déterminée sur la base d'une approche fréquence-sévérité. On note X la variable aléatoire qui représente l'ensemble des coûts des dégâts. N la variable aléatoire qui représente la loi de fréquence (nombre de sinistres susceptibles de se produire) et B la variable aléatoire qui représente loi de sévérité (coût engendré pour un sinistre).

$$X = \sum_{i=1}^{N} B_i \implies \mathbb{E}[X] = \mathbb{E}[N] \times \mathbb{E}[B]$$

Pour déterminer une prime pure, il faut donc déterminer $\mathbb{E}[N]$ et $\mathbb{E}[B]$. Il sera cependant impossible de les déterminer de manière exacte, on devra donc les estimer pour un nouveau client en fonction de ses caractéristiques. Pour cela nous allons avoir recourt à l'usage des modèles linéaires généralisés (GLM) sur nos jeux de données.

Notre objectif est donc de d'utiliser les GLM pour déterminer le nombre de sinistres et leurs coûts.

Nous allons proposer des GLM pertinents. Nous utiliserons la méthode Forward-Backward pour déterminer les meilleurs GLM. Cette méthode demandant un certain temps d'exécution, il sera préférable tout d'abord de supprimer des covariables jugées inutiles.

Afin de réaliser l'intégralité de nos méthodes statistiques ainsi que la manipulation de nos données, nous utiliserons le logiciel R.

Chapitre 1

Premières manipulations des données

1.1 Exploration des données

Avant tout chose, nous allons ajouter certaines colonnes à nos jeux de données. Dans le tableau `PGTRAIN17POL` nous ajoutons 3 colonnes.

- La première appelée `id_policy` regroupera `id_client` et `id_vehicle` pour des raisons de simplicité et de maniement des données.
- La deuxième sera appelée `accident`, elle prendra deux valeurs TRUE ou FALSE selon que le client avec son véhicule ait eu au moins un accident ou non.
- La dernière sera appelé `claim_nb`, elle indiquera simplement le nombre d'accidents pour un couple client-véhicule donné.

Nous allons également créer un nouveau tableau. Ce tableau aura l'avantage d'avoir les données combinées des deux tableaux, mais uniquement quand il y a eu accident. Son intérêt sera développé plus tard, quand il faudra faire des calculs sur la sévérité.
Tout d'abord nous avons commencé par regarder ce que signifiait chacune des variables à l'aide du manuel de référence du package **Casdataset**. Ainsi, on a pu savoir à quoi correspondait chacune des variables et comprendre ce que les différentes valeurs quelles prenaient représentaient.

Ces premières observations nous ont permis d'avoir un premier regard sur les variables qui nous seront utiles et celles qui ne le seront pas.

Nous avons également stocké les tableaux de données dans des variables nommées `tc` et `tp` afin de pouvoir y faire appel plus facilement et d'aérer un peu notre code. De plus, si on doit modifier ces tableaux, on conservera toujours les données initiales et nous pourrons ainsi réutiliser des données ultérieurement.

Afin d'enlever certaines covariables, plusieurs méthodes peuvent être envisagées :

- Faire une ACP, pour relever les covariables qui apportent la même information
- Utiliser des outils graphiques, pour voir si une variable est répartie de la même manière dans les 100 000 polices et dans celles qui ont eu un accident (indiquant dans ce cas que la covariable n'a pas de réel impact)
- Le bon sens

1.2 Statistiques descriptives :

Jeu de données pg2017trainpol : Pour les statistiques descriptives, nous avons commencé par regarder les colonnes qui nous intéressaient. Ainsi, nous avons décidé de ne pas regarder les variables suivantes, pour lesquelles des statistiques n'apporterait rien :

- id_client
- id_vehicule
- id_policy
- id_year
- drv_sex2
- drv_age2
- drv_age_lic2
- pol_insee_code

il ne sert à rien d'avoir des statistiques sur les identifiants puisqu'ils servent seulement à identifier de quelle police on parle. De plus nous avons choisi de ne pas regarder les composantes relatives au deuxième conducteur puisque nous n'avons pas moyen de voir lequel des deux conducteurs a été la cause de l'accident. Enfin nous avons jugé que les codes INSEE des communes ne nous servaient pas pour cette première partie.

Nous avons ensuite regardé les statistiques à l'aide de la fonction `summary`. Pour les valeurs quantitatives cette fonction nous donne le nombre d'occurrence de chaque valeur.

Jeu de données `pg2017trainclaim` : Nous avons alors réalisé les mêmes étapes pour le jeu de donnée `pgtrain17claim`. Un audit rapide des données nous a permis de nous rendre compte que certains montants demandés étaient négatifs. Après une interrogation de notre part nous avons décidé dans un premier temps avec notre enseignant de ne pas prendre en compte ces valeurs.

Nous avons donc enlevé certaines variables pour lesquelles nous les statistiques ne nous apporteront rien :

- id_client
- id_vehicule
- id_year

Après un `summary` nous nous sommes rendu compte que la variable `claim_nb` était toujours égale à 1. Nous l'avons donc supprimé également.

1.3 Première AFP

Pour notre première AFC, nous avons fusionné les 2 tableaux `tc` et `tp` à l'aide de la fonction `merge`. Puis nous avons conservé uniquement les colonnes qui contenaient des valeurs numériques.

On remarque que toutes les variables ne sont pas toutes représentées de la même manière. Ici, plus le cos2 est élevé, mieux la variable est représentée. De plus, on peut également se rendre compte que les 2 premiers axes factoriels représentent environ 50% de l'information. On remarque également que deux groupes se dégagent :
On remarque que ces deux groupes de vecteurs sont (presque) orthogonaux. Cela implique qu'un vecteur du groupe rouge n'est pas corrélé à un vecteur du groupe bleu (il peut cependant exister une dépendance puisque nous ne sommes pas en présence de vecteurs gaussien). En effet, il semble assez logique que l'âge d'un conducteur n'est pas corrélé avec le poids de son véhicule.

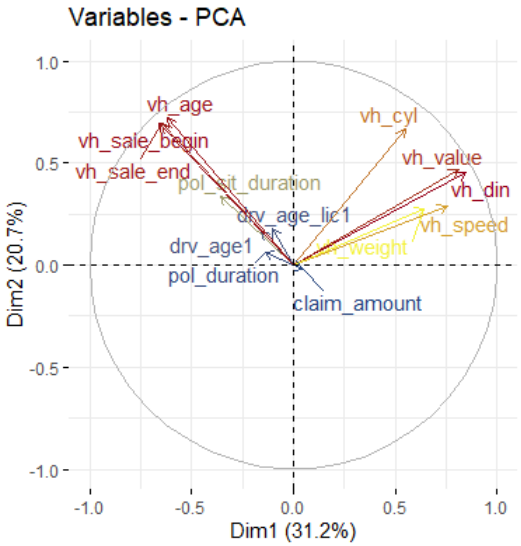

FIGURE 1.1: Nuage des variables

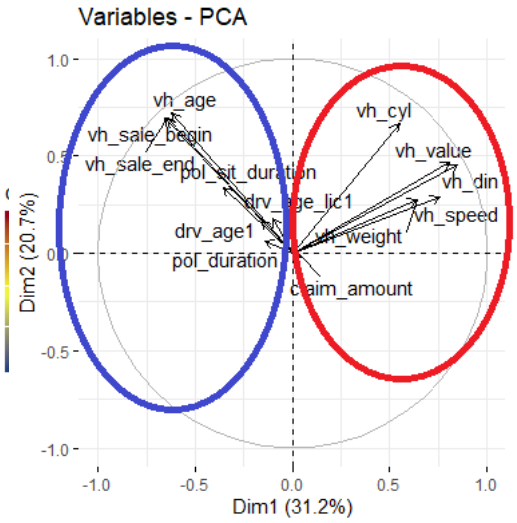

FIGURE 1.2: Deux groupes dde variables se dégagent

1.4 Graphique des corrélations

Nous nous sommes alors posé la question de savoir si certaines variables étaient corrélées. En effet il semblerait logique que l'âge du conducteur ainsi que l'âge de permis puissent être fortement liées. Il est plus probable qu'un homme de 50 ans ait 20 ans de permis qu'un homme de 40 ans.

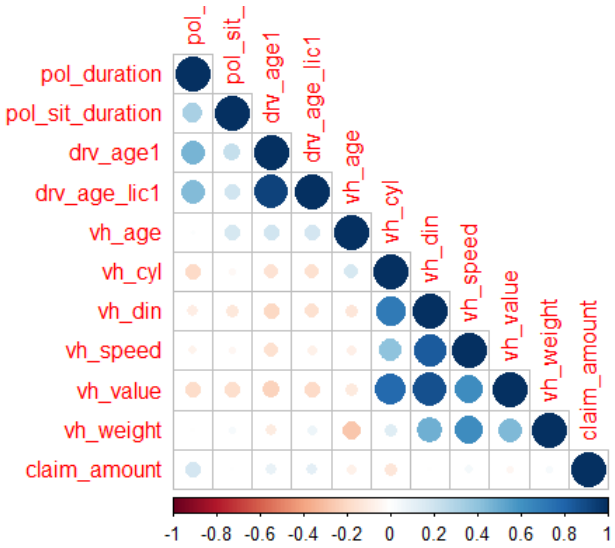

FIGURE 1.3: Graphique des corrélations de nos variables

Chapitre 2

Choix des covariables

2.1 Manipulation des tableaux de données

Nous avons créé une variable **tc2** qui pour chaque couple client-véhicule ayant eu un sinistre (`id_policy` du tableau `pg17trainclaim` appelé maintenant tc) renvoie le nombre d'accidents associé (que nous avons noté `claim_nb`). On fusionne cette colonne avec le tableau **tp** (la fonction `merge` va donc s'effectuer sur la colonne `id_policy`) afin de créer un tableau que nous nommons **TAB**, en utilisant l'argument « all.x = TRUE » dans la fonction `merge` qui signifie que pour les couples `id_policy` qui n'ayant pas eu de sinistres (ie ceux présents dans **tp** mais pas dans **tc**), la fonction `merge` va renvoyer NA. On remédie à cela en remplaçant tous les NA de cette colonne par des 0.

TAB est donc notre principal tableau d'étude, celui contenant toutes les informations pour chacune des 100 000 polices d'assurances. Notre tableau d'intérêt pour déterminer la loi de fréquence est donc le tableau **TAB**. Ce tableau contient plus de 30 covariables, et il convient naturellement d'en supprimer pour proposer un modèle robuste.

2.2 Approche générale et procédure

Un point de départ de notre procédure sera de répartir les variables selon qu'elles influeront sur la fréquence ou la sévérité. On va se servir de l'approche présentée dans l'ouvrage "Mathématiques de L'Assurance Non-Vie" de Michel Denuit et Arthur Charpentier : dans une analyse de tarification automobile, la fréquence est une variable liée essentiellement au conducteur, tandis que le coût moyen est lié essentiellement aux caractéristiques du véhicule (Chapitre 9.1).
Ce faisant, nous pouvons d'ores et déjà supprimer de l'analyse de la fréquence les variables suivantes :

- TAB$vh_make
- TAB$vh_speed
- TAB$vh_sale_begin
- TAB$vh_sale_end
- TAB$vh_weight
- TAB$vh_type
- TAB$vh_age
- TAB$vh_din
- TAB$vh_cyl
- TAB$vh_model

On supprime ensuite les covariables inutiles :

- La variable `id_year` est une variable qualitative ne prenant qu'une seule modalité dans l'entièreté du tableau. Elle n'apporte aucune information.

- `id_client` et `id_vehicle` : ces deux colonnes ne sont plus utiles maintenant que l'on les a fusionnées en 1 seule colonne : `id_policy`.
- On prend le choix de supprimer également toutes les informations relatives à un éventuel deuxième conducteur donc `drv_age2 , drv_sex2 , drv_age_lic2` comme nous l'avons évoqué dans le premier chapitre.
- `pol_insee_code` : il s'agit d'un découpage par code postal, on la convertit en département pour s'en servir plus tard avec les données de l'ONISR

Un premier exemple est l'étude de la fréquence de la sinistralité par âge sur nos données. On obtient le résultat sur la figure 2.1.

Ici, on observe clairement une tendance : les accidents semblent être moins probables entre 60 et 80 ans (à 80 ans la moyenne est de 0 car il n'y a simplement plus assez d'individus). Cette tranche d'âge semble moins exposée au risque et donc se distingue des autres, ce qui nous amène à une méthode supplémentaire à mettre en œuvre : la discrétisation des variables.

2.3 Numérisation et discrétisation

Parmi les variables restantes dans l'explication de la fréquence, les plus simples à discrétiser sont les variables relatives :

1. Le sexe du conducteur (`drv_sex1`)
2. Contrat basé sur le kilométrage (`pol_payd`)
3. L'existence éventuelle d'un second conducteur (`drv_drv2`)

Remarque : Selon les informations du `CASDatasets`, il est interdit légalement de discriminer des polices d'assurance selon le sexe. Ceci est toutefois permis dans notre cas, à vertu purement analytique.

Ici, on parlera plutôt de numérisation plutôt que de discrétisation : l'assuré aura une valeur de 1 si c'est une femme, 0 si c'est un homme. Une valeur de 1 s'il a souscrit, 0 sinon (selon le même ouvrage que cité précédemment (Chapitre 9.2), par exemple. On mettra toujours la modalité la plus représentée en valeur 0, car elle servira de référence).
Dans notre cas, on a les résultats en figure 2.2, d'où notre choix.

En ce qui concerne les autres covariables, on va combiner les résultats de l'ACP à des études graphiques, pour identifier celles que nous allons conserver.
Sauf cas particuliers, nous allons répartir les données des variables continues en classes de déciles. Après une étude graphique des comportements de ces classes, si une tendance apparaît clairement, nous saurons que la variable a une influence sur la fréquence.
Nous commençons donc par les cas particuliers :

1. En tarification automobile, l'âge des individus est souvent arbitrairement séparé par les classes $[18, 25],]25, 45],]45, 60],]60, \infty[$, nous faisons de même ici.
2. Le bonus qui ne contient pas assez de valeurs distinctes sera découpé selon les classes $[0.5, 0.505],]0.505, 0.55],]$

On rappelle qu'un principe important dans la discrétisation est d'avoir une représentation suffisamment significative dans chaque classe créée.

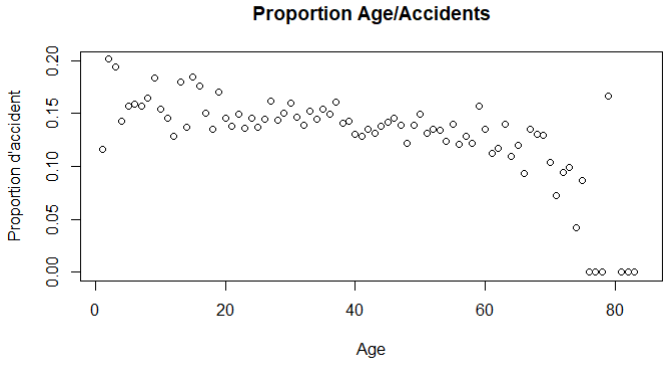

FIGURE 2.1: Proportion d'accidents par rapport à l'âge

```
> summary(TAB$drv_sex1)
    F     M
39799 60200
> summary(TAB$pol_payd)
   No   Yes
95847  4152
```

FIGURE 2.2: Statistiques sur le sexe ainsi que pol_paid

Les autres covariables continues sont découpées par déciles :

1. `pol_duration`
2. `pol_sit_duration`
3. `drv_age_lic1`

Pour le moment, on se contente de garder les variables qualitatives telles quelles, pour essayer de dresser des graphiques. On obtient, les graphiques suivants :

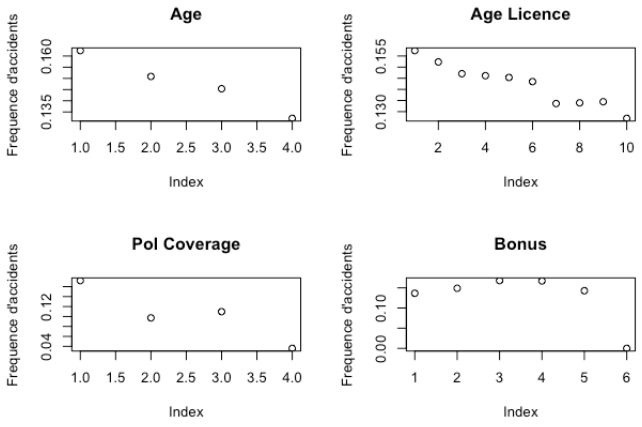

FIGURE 2.3: Variables qualitatives

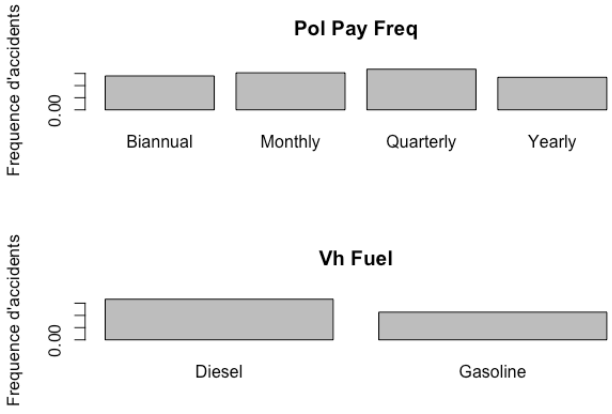

FIGURE 2.4: Variables qualitatives

2.4 Interprétations et décisions

2.4.1 Age et Age du permis

Une première observation est la présence de tendances semblables entre l'âge du conducteur et l'âge du permis. Cette similarité est confirmée par les résultats de l'ACP : on voit que l'age et l'age du permis sont fortement corrélés. Par ailleurs, après des tests de GLM contenant les deux covariables, on voit que selon la base d'apprentissage sélectionnée, l'âge du conducteur n'est pas forcément explicatif. On

prend donc le choix de supprimer cette covariable.
On discrétise l'âge du permis en trois classes :

1. Entre 1 et 20 ans
2. Entre 20 et 36 ans (modalité prise comme référence car elle est la plus représentée)
3. Plus de 36 ans

2.4.2 Pol_pay_freq

On constate une tendance selon que le contrat soit payé avec une fréquence élevée (Mensuellement ou Trimestriellement) ou non (annuellement ou Bi Annuellement). On discrétise donc cette variable : 0 ou 1 selon un cas ou l'autre.

2.4.3 Bonus

Les deux derniers points sont à omettre de l'étude (respectivement 7 et 1 individus). On aurait donc une tendance de croissance fréquence/bonus. Néanmoins, le bonus étant une variable liée à la sinistralité passée, nous ne l'incluons pas dans le modèle. En effet, l'assureur reverra dans ce cas le montant de la prime en fonction des sinistres causés par l'assuré à l'aide de la théorie de la crédibilité ou des systèmes bonus-malus.

2.4.4 Pol_Usage

On pré-suppose qu'un véhicule à fins professionnelles parcourra plus de kilomètre qu'un véhicule personnel, et donc aura une plus grosse exposition au risque. Il y a une différence claire entre les modalités, mais cela est dû à la sous-représentation d'une des deux modalités. On prend le choix d'exclure cette covariable par manque de données pertinentes.

2.4.5 Vh_fuel

Un véhicule diesel est souvent caractérisé par des distances couvertes plus importantes (en raison du prix du carburant, plus faible que celui de l'essence) et donc une plus grosse exposition au risque

$$X = \begin{cases} 0 & \text{si c'est un diesel} \\ 1 & \text{sinon} \end{cases}$$

2.4.6 Pol_Coverage

On discrétise en trois catégories :

1. Couverture Minimale
2. Couverture Médiane
3. Couverture Maximale (prise comme référence)

2.4.7 Pol_duration et Pol_sit_duration

L'exposition au risque est un facteur important en tarification automobile : un assuré couvert depuis 4 mois sur l'année et déclarant 2 sinistres sur la période sera plus à risque qu'un assuré couvert depuis un an et déclarant le même nombre de sinistres. Cependant dans notre cas, nos deux covariables correspondent à l'ancienneté des assurés en années. Ne permettant pas de mesurer l'exposition au risque (qu'on supposera durer l'année entière pour tout le monde dans notre cas) et possédant un modèle suffisamment fourni, nous décidons de supprimer ces deux variables.

2.5 Répartitions des données en classes

Les résultats obtenus en figure 2.3 et en figure 2.4 nous donnent des exemples de comment traiter les données efficacement. Si des tendances graphiques se démarquent nous regroupons plusieurs classes ensemble afin de réduire les dégrés de liberté (qui contribuent à la variabilité du modèle). Par exemple, pour la variable `pol_pay_freq`, on constate que la sinistralité est sensiblement la même selon que le paiement soit biannuel ("bi-annual") ou annuel ("yearly"). La sinistralité, plus élevée dans le cas de paiements fréquents, peut s'expliquer par une position précaire chez l'assuré, et donc notamment un véhicule moins bien entretenu. Nous pourrons donc regrouper ces deux classes entre elles pour notre variable `pol_pay_freq`.
Ces regroupements peuvent aussi se faire par un test Khi-deux, pour vérifier si deux variables ou non sont indépendantes. On fait un exemple de ce test avec `drv_age_lic1` (l'ancienneté du permis). On considère les 3 catégories :

- 25 à 29 ans d'ancienneté
- 29 à 33 ans d'ancienneté
- 45 à 50 d'ancienneté

Les tests du Khi-deux nous donnent :

```
Nb_Sinistres            0    1    2    3    4
  age de 25 a 29 ans 7720 1096  113   11    1
  age de 45 a 50 ans 8285 1006  120   14    0
> chisq.test(Nb_Sinistres,TAB$claim_nb[t3])

        Pearson's Chi-squared test

data:  Nb_Sinistres and TAB$claim_nb[t3]
X-squared = 12.623, df = 4, p-value = 0.01327
```

FIGURE 2.5

```
Nb_Sinistres            0    1    2    3    4
  age de 25 a 29 ans 7720 1096  113   11    1
  age de 29 a 33 ans 10036 1324  157   13    1
> chisq.test(Nb_Sinistres,TAB$claim_nb[t3])

        Pearson's Chi-squared test

data:  Nb_Sinistres and TAB$claim_nb[t3]
X-squared = 3.2856, df = 4, p-value = 0.5112
```

FIGURE 2.6

On constate que dans le premier cas, notre p_value est inférieure à 0.05. On rejette donc l'hypothèse d'indépendance, et on conjecture qu'il y a une différence de sinistralité selon que le permis soit plus ou moins ancien. Dans le second cas la p_value vaut 0.51, on ne rejette donc pas l'hypothèse d'indépendance et on conjecture qu'avoir un permis ancien de 25 ou de 33 ans ne fait pas une grande différence. On peut donc regrouper ces deux classes. Ces résultats sont bien confirmés graphiquement, et on se basera à l'avenir sur les représentations graphiques, plus intuitives.

2.6 Traitement relatif des covariables liées au véhicule

On termine cette section en nuançant le résultat admis en début de chapitre selon lequel la fréquence ne s'explique que par les variables relatives au conducteur. En effet, on constate que la valeur du véhicule représente une réelle tendance dans la fréquence. On est en droit de se demander s'il en est de même pour les autres covariables relatives au véhicule. On recourt pour cela à un test de V-Cramer : il s'agit d'une quantité que l'on peut calculer pour mesurer la corrélation entre deux variables. La formule est la suivante :

$$V_{Cramer} = \sqrt{\frac{\chi^2}{N \times k}}$$

où N représente la taille de l'échantillon et k le degré de liberté de notre χ^2.

Le résultat nous donne la corrélation dont l'interprétation est donnée dans le tableau suivant :

Valeur du V de Cramer	Intensité de la relation entre les variables
$v < 0.1$	Relation nulle ou très faible
$0.1 \leq v < 0.2$	Relation faible
$0.2 \leq v < 0.3$	Relation moyenne
$v > 0.3$	Relation forte

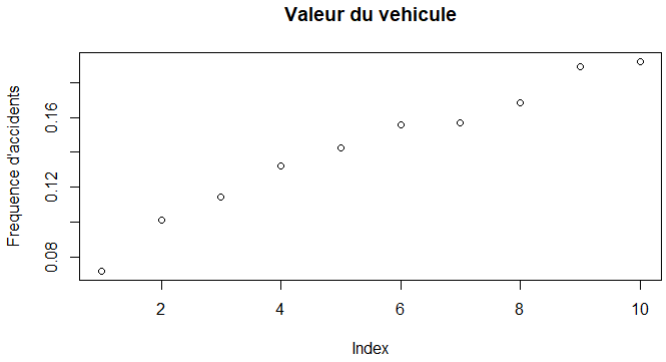

FIGURE 2.7: Fréquence d'accident en fonction de la valeur du véhicule

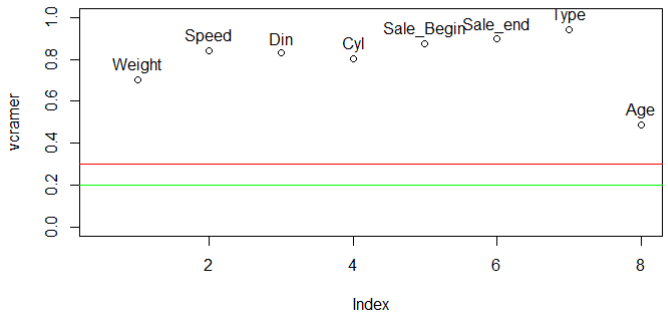

FIGURE 2.8: Position des différents test de V-Cramer

Le V-Cramer est bien supérieur à 0.3 (ligne rouge) pour l'ensemble des variables. Leurs fortes corrélations avec la valeur du véhicule nous dispensent de leur intégrer au modèle de la fréquence (on n'utilisera néanmoins pas cette méthode pour le calcul de la sévérité avec autant d'insistance, pour éviter d'avoir un modèle trop imprécis).

Chapitre 3

GLM : coefficients de la fréquence

3.1 Famille exponentielle et fonction de lien

Maintenant que nous avons supprimé les covariables les moins utiles, nous pouvons commencer à proposer des GLM. L'approche qui sera faite sera celle du forward-backward. Cette méthode permet, à chaque étape, d'ajouter ou de supprimer une covariable selon celles déjà intégrées au modèle. Le but est de minimiser un critère, ici celui d'Akaïke (ce critère mesure un compromis entre la précision du modèle et sa variance).

Les familles de modèles linéaires généralisés sont définies par une classe de lois appelée "Famille Exponentielle". Une loi appartient à cette famille si sa densité peut se mettre sous la forme :

$$f(y|\theta,\phi) = e^{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)}$$

La fréquence des sinistres est à valeurs dans les entiers naturels. On peut donc utiliser soit un modèle binomial, soit un modèle de Poisson. D'après "Mathématiques de l'Assurance Non-Vie" (Chapitre 9.8.14), le nombre de sinistres se modélise usuellement en tarification automobile par une loi de Poisson, appartenant à la famille exponentielle (preuve omise). Dans notre cas, on envisagera les deux lois énoncées (naturellement, la loi binomiale négative appartient aussi à la famille exponentielle).
En réalité, le modèle binomial négatif est utilisé en cas de surdispersion, c'est-à-dire lorsque l'on a :

$$\mathbb{V}(N) > \mathbb{E}[N]$$

Où représente N la loi du nombre de sinistres (où l'on noterait n_i la réalisation pour l'individu i).
Tandis que le modèle de Poisson est utilisé dans le cas d'une équidispersion, c'est-à-dire lorsque :

$$\mathbb{V}(N) = \mathbb{E}[N]$$

On retrouve bien ici la caractéristique de la loi de Poisson selon laquelle l'espérance et la variance sont les mêmes.

Une analyse préalable des données nous donne :

$$\mathbb{E}[N] = 0.14243 \qquad \text{et} \qquad \mathbb{V}[N] = 0.1578853$$

Par ailleurs, si on choisit de comparer l'espérance et la variance au sein de chaque classe, on observe bien une surdispersion qui nous dirigerait donc vers un modèle binomial négatif (voir figure 3.1). On s'attardera néanmoins sur les deux cas.

Pour s'assurer que notre modèle soit performant, on va le déterminer sur une base d'apprentissage (80% des données) et on cherchera à le valider sur les 20% restants. On prend le choix d'utiliser la fonction de lien canonique dans ce modèle, ainsi on a :

$$g(u) = log(u) \iff g^{-1}(u) = e^u$$

Or, on a

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

Remarque : Il peut y avoir plusieurs choix de fonction de lien. Selon l'ouvrage "Mathématiques de l'Assurance Non Vie" (Chapitre 9.8.14), il est souvent d'usage d'utiliser la fonction de lien logarithmique, puisqu'elle présente l'avantage de donner un modèle multiplicatif, et les coefficients β_j ont alors une interprétation simple, en termes de multiplicateurs.

3.2 Forward-Backward et détermination des coefficients

L'espérance étant bien le paramètre que l'on souhaite trouver pour déterminer la prime pure, il faut simplement récupérer les coefficients données par le GLM, l'appliquer aux données via $x_i \times \beta$, et composer par la fonction exponentielle. On obtient les résultats R en en figure 3.2 :

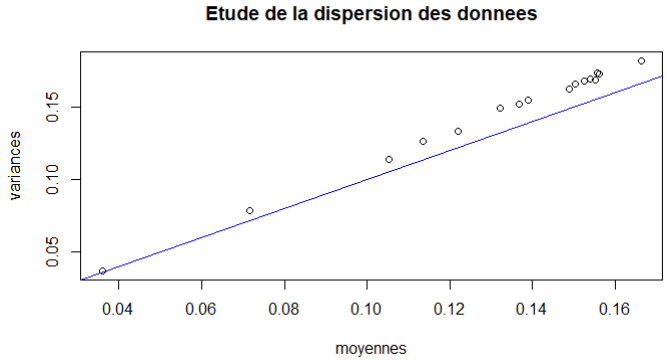

FIGURE 3.1: Variance en fonction de la moyenne

V1	V1
Min. :0.02071	Min. :0.02063
1st Qu.:0.10436	1st Qu.:0.10444
Median :0.14737	Median :0.14728
Mean :0.14240	Mean :0.14241
3rd Qu.:0.18568	3rd Qu.:0.18542
Max. :0.23255	Max. :0.23296

FIGURE 3.2: Fréquences des accidents

C'est un premier résultat convaincant qui dit que selon le client et son véhicule, ce dernier aura en moyenne une probabilité de 0.14 d'avoir un accident. Les deux types de résultats sont par ailleurs sensiblement proches. On constate en effet que sur un exemple de base d'apprentissage, nous obtenons :

$$\sum_{i=1}^{n} |\lambda_i - \lambda_i^{'}| = 7.244944$$

$$\sum_{i=1}^{n} (\lambda_i - \lambda_i^{'})^2 = 0.001025676$$

Où λ_i sont les prédictions du modèle de Poisson et $\lambda_i^{'}$ sont celles du modèle binomial négatif.

3.3 Tests des modèles

Les coefficients ont été déterminés dans les deux modèles, nous devons maintenant nous assurer que le modèle est robuste, d'où l'intérêt d'avoir gardé une base de test sous la main. On va donc tester les résultats des modèles sur les 20% de données restantes et observer par une régression si les prédictions sont proches des réalisations.

3.3.1 Le modèle de Poisson

Une première approche naïve serait de lancer 100 000 tirages de Poisson avec les paramètres trouvés et de calculer la somme des moindres carrés :

$$\sum_{i=1}^{n}(Y_i - x_i\beta)^2$$

Cette approche est cependant beaucoup trop aléatoire et on optera pour une méthode de validation basée sur la vraisemblance :

$$\sum_{i=1}^{n}\lambda_i = \sum_{i=1}^{n} n_i$$

Cette égalité est censée nous dire qu'en moyenne, le nombre moyen de sinistres prédits sur la base de test doit être égal au nombre de sinistres qui se sont effectivement produits (dans le cas de la base d'apprentissage, on a égalité parfaite puisque les coefficients ont été justement calculés sur cette base, pour avoir une vraisemblance parfaite). Dans le cas d'une simulation de huit procédures Forward-Backward successives dans le cas d'un modèle Poisson, on obtient l'outil de validation en figure 3.3 :

Les estimations du nombre de sinistres (points) sont proches en valeur absolue et relative des sinistres ayant effectivement eu lieu sur la base de test (en rouge). Ceci diffère évidemment à chaque nouveau tirage (comparaison faite avec une base d'apprentissage correspondant à 80% des données).

Néanmoins, lorsque l'on veut s'assurer que les données de la base de test ont une distribution semblable à une loi de Poisson (avec un test du Khi-deux notamment), on constate une statistique de test et une p-valeur totalement inadaptées en figure 3.4

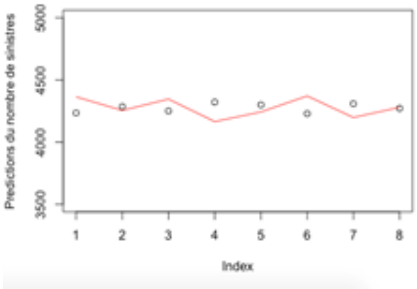

FIGURE 3.3: Vérification du modèle

```
> chisq.test(cpt2,p=cpt1,rescale.p = TRUE)

        Chi-squared test for given probabilities

data:  cpt2
X-squared = 68.303, df = 3, p-value = 9.853e-15
```

FIGURE 3.4: Test du Khi-2

Une explication est que le nombre de sinistres par police possède la fameuse propriété de surdispersion annoncée (présence de polices à 5 ou 6 sinistres). De plus, le nombre de polices à un seul sinistre est le seul à être surestimé, attestant que le modèle prédit ne varie peut-être pas suffisamment.

```
tirage
    0     1     2     3
17403  2417   164    16
> cpt2

    0     1     2     3
17416  2298   261    25
```

FIGURE 3.5: Simulation des Sinistres

Nombre de sinistres simulés (tirage) et nombre de sinistres ayant eu lieu dans la base de test de validation (cpt2).

3.3.2 Le modèle binomial négatif

Le modèle binomial négatif présente donc l'avantage, par rapport au modèle de Poisson, de représenter la surdispersion des données. Heureusement, comme nous l'avons vu, les distances entre les prédictions des deux modèles sont proches et la préférence d'un tel modèle ne faussera pas les prédictions, mais respectera néanmoins davantage le comportement des données. On observe d'ailleurs que, de même que dans le modèle de Poisson, la prédiction globale est relativement juste. On termine cette section avec une analyse des résidus pour identifier une éventuelle source majeure d'erreurs.

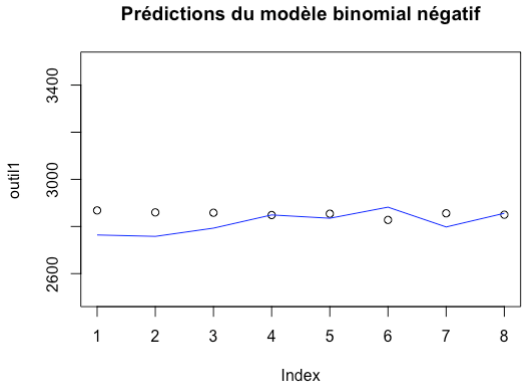

FIGURE 3.6: Prédictions du modèle binomial négatif

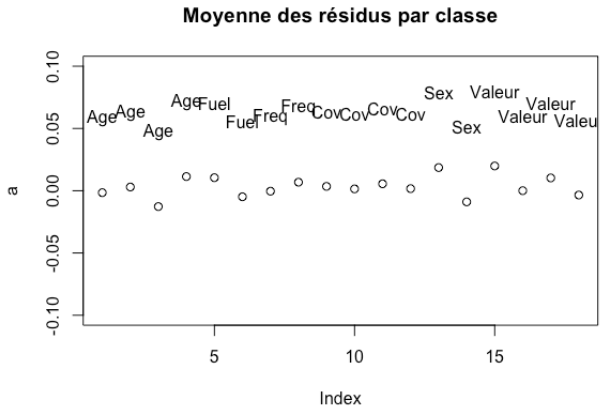

FIGURE 3.7: Résidus par classes

Pour chacune des classes (que le conducteur soit âgé ou non, qu'il ait un véhicule avec forte valeur ou non etc...), on ne constate pas de tendance drastique et on peut étudier les résidus de deux manières différentes : résidus de Pearson et résidus de déviance. Ces résidus sont respectivement donnés par les formules suivantes :

- Résidus de pearson

$$r_i^p = \frac{n_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$$

- Résidus de déviance

$$r_i^D = sign(n_i - \hat{\lambda_i}) \times \sqrt{2\left\{n_i \ln \frac{n_i}{\hat{\lambda_i}} - (n_i - \hat{\lambda}_i)\right\}}$$

$$\text{De plus on pose} \quad y\ln(y) = 0 \quad \text{si} \quad y = 0$$

On constate qu'une partie des résidus se situe en dehors des valeurs "non aberrantes" (supérieures à 2). Cela s'explique par le fait que pour une police ayant eu 2 accidents ou plus, l'estimation donnée sera éloignée de la valeur effective et cela justifie que l'étude de tels résidus dans le cas où la variable à expliquer ne prend que quelques valeurs (comme c'est le cas ici) ne présente qu'un intérêt limité. Néanmoins, cela nous conforte dans le fait que les valeurs "mal prédites" sont celles correspondant à un grand nombre de sinistres, erreur qui est cependant largement compensée par la mutualisation des contrats au sein du portefeuille d'assurances.

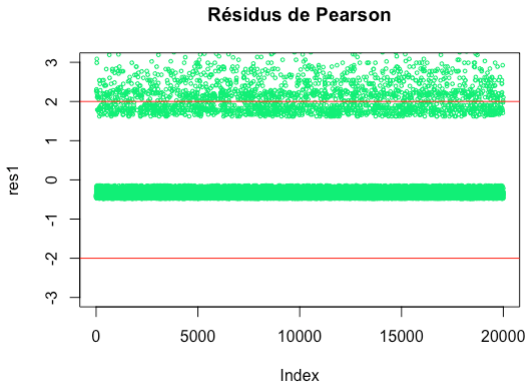

FIGURE 3.8: Résidus de Pearson

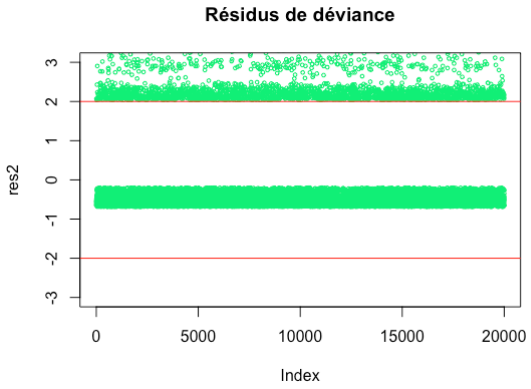

FIGURE 3.9: Résidus de déviance

3.3.3 Synthèse du modèle

Notre modèle de fréquence étant construit, nous pouvons désormais caractériser nos différentes classes d'usagers par la probabilité de ne déclencher aucun accident sur l'exercice. En effet, le paramètre r de notre binomiale négative étant donné par le GLM, on obtient pour un assuré i la probabilité de non accident suivante :

$$\mathbb{P}(\text{assuré i ne cause aucun accident}) = p_i$$

Et :

$$\begin{aligned} \lambda_i = \mathbb{E}[Y_i] = \frac{r(1-p_i)}{p_i} \\ \implies \quad \lambda_i p_i + r p_i = r \\ \implies \quad p_i(\lambda_i + r) = r \\ \implies \quad p_i = \frac{r}{\lambda_i + r} \end{aligned}$$

La probabilité de non-accident est fonction décroissante des prédictions du GLM. On obtient dès lors le tableau suivant, pour les 3 "meilleures" et "pires" classes d'assurés.

Carburant	Fréquence paiement	Sexe	Couverture	Permis	Valeur véhicule	Probabilité
Essence	Peu Élevée	Homme	Minimale	Ancien	Peu cher	0.979
Essence	Peu Élevée	Femme	Minimale	Ancien	Peu cher	0.976
Essence	Peu Élevée	Homme	Minimale	Récent	Peu cher	0.976
...	...	...	...	...	...	...
Diesel	Élevée	Femme	Maximale	Intermédiaire	Cher	0.809
Diesel	Élevée	Homme	Maximale	Récent	Cher	0.806
Diesel	Élevée	Femme	Maximale	Récent	Cher	0.798

TABLE 3.1: Probabilité de non accident en fonction des classes d'assurés

On constate que le classe la plus à risque dans notre portefeuille d'assurés est la suivante : Une conductrice dont le permis est récent, avec une couverture de risques maximale dont le paiement est fragmenté et roulant dans un véhicule au diesel et à la valeur élevée. Ces observations sont bien en accord avec nos premières intuitions et inversement, la classe d'assurés la moins à risque est celle présentant toutes les caractéristiques opposées. On observe par ailleurs que dans la plupart des cas (voir Remarque ci-dessous), les classes consécutives se distinguent par un unique changement, attestant d'influences inégales des caractéristiques sur la probabilité d'accident.

Exemple : On voit que le sexe est la caractéristique la moins influente. Un usager qui se révèle être un homme, mais qui présente toutes les autres caractéristiques "propices au risque" demeure très nettement une des classes à la probabilité de non-accident la plus basse.

Remarque : On constate toutefois que dans certains cas, les probabilités de non-accident sont égales entre deux classes différentes et donc que les risques "se compensent". Il faut donc garder en tête qu'une probabilité de non-accident ne détermine par une classe d'usagers de manière unique.

Chapitre 4

GLM : coefficients de la sévérité et prime pure

Notre point de départ pour traiter la sévérité est le deuxième jeu de données `pg17trainclaim` qui contient les polices d'assurances de 14 243 individus ayant eu 1 sinistre.

Dans ce jeu de données, nous avons 6 variables : l'identifiant du client (`id_client`), l'identifiant du véhicule (`id_vehicule`), l'année de couverture de cette police d'assurance (`id_year`) qui est toujours l'année 0 (on a donc décidé de supprimer cette variable), le numéro d'identification de la police d'assurance qui indique à quel type d'assurance l'individu a souscrit (`id_claim`), le nombre d'accident (`claim_nb`) qui vaut toujours 1 (une même voiture peut avoir plusieurs accidents mais chacun d'entre eux est "unique" d'où la colonne de 1 ; on a donc également supprimé cette variable) et enfin le montant de l'accident (`claim_amount`).

Le premier problème rencontré était de réunir les données de notre dataframe `pg17trainclaim` avec celles de `pg17trainpol`, car ces dernières ne sont pas homogènes (les données liées aux polices d'assurances sont caractérisées par les numéros de polices, celles des accidents par des numéros de véhicules). Ce problème est contourné à l'aide la fonction `aggregate` de R permettant des définir des couples assuré-véhicule. Ainsi, chaque accident peut être clairement caractérisé son coût engendré (`pg17trainclaim`), et par les caractéristiques de l'assuré et de son véhicule impliqué dans l'accident (`pg17trainpol`).

Nous nous retrouvons ainsi 14 243 couples client-véhicule sinistrés et 34 variables. Il est à noter que pour certains accidents, le montant de la prime est négatif ou nul : nous ne les étudierons pas ici, et nous généraliserons l'étude des 12 931 accidents restants à l'ensemble des accidents.

4.1 Pré-sélection des variables :

4.1.1 Variables que nous avons supprimé d'office

On supprime les 7 variables relatives au conducteur car, comme vu dans le livre de référence de Charpentier, celles-ci ne permettent pas d'expliquer la sévérité mais seulement la fréquence. De même, les 8 variables relatives à la police d'assurance n'expliquent pas la sévérité ; on les supprime donc. Les 4 variables : identifiants client, véhicule, année et claim sont désormais inutiles.

Enfin la variable `claim_nb` (qui vaut toujours 1) est également supprimée.

Après cette pré-sélection, nous avons 12 variables dont on distingue 2 types :

- 4 variables de type "character" que nous allons numériser : `vh_fuel, vh_make, vh_model` et `vh_type`.

- Les 8 variables numériques restantes (`vh_age, vh_cyl, vh_din, vh_sale_begin, vh_sale_end, vh_speed, vh_value et vh_weight`) prenant chacune un grand nombre de valeurs différentes, nous les considérons comme de type "continues".

Pour continuer cette sélection de variables, nous allons étudier nos 8 variables continues et les discrétiser.

Ici, comme pour la fréquence, nous avons choisi de découper chacune de nos variables numériques continues en classe de déciles. Une fois celles-ci discrétisées, nous avons affiché pour chacune d'entre elles la sévérité moyenne de chacune de leurs 10 classes.

4.1.2 Variables corrélées

D'après les informations données par le package `CasDataset`, les variables `vh_din, vh_cyl, vh_speed` et `vh_value` sont hautement corrélées. Ces résultats sont d'ailleurs confirmés par l'ACP(voir Figure 1.3). Nous avons cependant voulu voir si ces 4 variables affichaient les mêmes "tendances" concernant la sévérité des sinistres. Nous avons observé les plots donnant la sévérité moyenne de chacune des 10 classes de ces 4 variables. Ces 4 plot montrent la même tendance : plus la classe de la variable est élevée, plus le montant moyen l'est aussi. Les variables étant donc fortement liées, nous décidons de ne garder que la variable `vh_din`.

L'ACP a également indiqué que les variables `vh_sale_begin` indiquant le début de commercialisation du véhicule et `vh_sale_end` indiquant la fin de commercialisation du véhicule sont très corrélées.
Ceci est confirmé par les 2 plots. Sur chacun d'eux , 3 mêmes groupes se dégagent. On constate une globale décroissance.
On décide alors de ne conserver que la variable `vh_sale_end`. Cette variable correspond au nombre d'années depuis la fin de campagne marketing du véhicule. Une campagne ancienne traduirait un véhicule âgé, obsolète, plus lent, éventuellement plus résistant, moins propice à de gros sinistres.

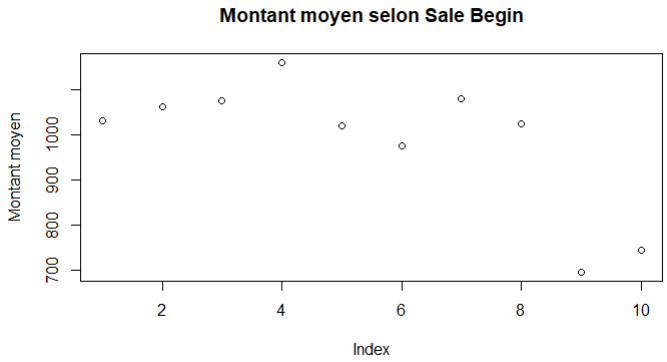

FIGURE 4.1: Montant moyen en fonction du début de vente du véhicule

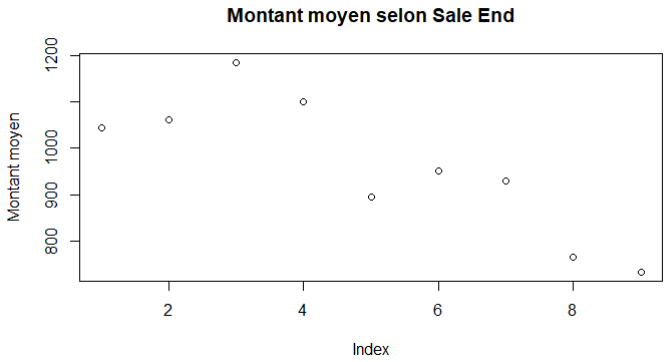

FIGURE 4.2: Montant moyen en fonction de la fin de vente du véhicule

Ces mêmes plot pour la variable `vh_weight` ne montre aucune tendance particulière. On supprime donc cette variable puisqu'elle n'est pas explicative.

En revanche pour le même plot avec la variable `vh_age`, on distingue les 8 premières classes des 2 dernières, on la conserve donc.

4.1.3 Variables de type "character"

Ces 4 variables prenant chacune peu de valeurs différentes, nous n'avons pas eu besoin de les discrétiser. Nous avons également affiché pour chacune d'entre elles la sévérité moyenne de chacune de leurs classes (qui ne sont donc plus au nombre de 10). Pour les variables `vh_make` et `vh_model`, on n'observe pas de tendance particulière entre leurs différentes modalités : elles sont donc peu significatives.

En revanche, pour la variable `vh_fuel` qui prend les 3 modalités Diesel, Gasoline et Hybrid, on observe une tendance : les véhicules Diesel ont en moyenne une sévérité plus faible que les véhicules Gasoline et Hybrid. On découpe alors `vh_fuel` en 2 modalités : 0 pour les véhicules Diesel, 1 pour les véhicules Gasoline et Hybrid.
De plus, nous allons distinguer la modalité "Commercial" qui présente une sévérité moyenne significativement plus faible que la modalité "Tourisme" pour la variable `vh_type`.
Nous numérisons alors cette variable en codant par 0 le type "Tourisme" et par 1 le type "Commercial".

4.1.4 Récapitulatif

Après cette pré-sélection de variables, il nous reste donc 5 variables (2 binaire et 3 continues).
Nous créons alors pour les 3 variables continues restantes (`vh_din, vh_age`, et `vh_sale_end`) de nouvelles variables correspondant aux classes déterminées par les plots.

Nous noterons que pour une variable dont nous avons distingué x modalités, nous ne créons que $x-1$ nouvelles variables. En effet, comme étudié en cours de Modèles linéaires, R supprimera automatiquement une des nouvelles variables crées par soucis de degré de liberté (vient du fait qu'autrement la matrice X contenant les variables n'est plus inversible).

Pour `vh_din`, 2 classes : les puissances moteurs strictement supérieures 120 (codées par un 1 et qui est donc la nouvelle variable) et les autres (codées par un 0).
Pour `vh_age`, 2 classes : les véhicules âgés de plus de 11 ans (codés par un 1 et qui est donc la nouvelle variable) et les autres (codés par un 0).
Pour `vh_sale_end`, 3 classes : les véhicules donc la commercialisation s'est terminée il y a plus de 11 ans (codés par un 1, nouvelle variable), entre 6 ans et 11 ans (codés par un 1, nouvelle variable) et il y a moins de 6 ans (codé par un 0).

Avant de finir notre sélection de variables, nous cumulons donc 12 391 polices d'assurance sinistrés pour 6 variables explicatives.

	id_policy	vh_age	vh_din	vh_fuel	vh_type	claim_amount	sale_end_6_11	sale_end_11_plus
1	A00000009-V01	0	0	0	0	927.16	1	0
2	A00000016-V01	0	0	0	0	555.48	0	0
3	A00000026-V01	0	1	0	0	478.01	1	0
4	A00000040-V01	0	0	0	0	512.83	0	0
5	A00000056-V01	0	0	0	0	1236.00	1	0
6	A00000070-V01	0	0	1	0	158.28	0	0
9	A00000092-V01	0	1	0	0	1429.82	1	0
10	A00000103-V01	0	1	0	0	1389.39	0	0
11	A00000103-V01	0	1	0	0	419.63	0	0
12	A00000107-V01	0	0	0	0	405.90	0	0
13	A00000125-V01	0	0	1	0	496.00	0	0
14	A00000136-V01	0	0	0	0	2671.61	1	0
15	A00000140-V01	0	0	0	0	544.38	0	0

FIGURE 4.3: Extrait de notre base de travail

4.2 Gestion des sinistres extrêmes

Souvent, une faible part des sinistres cause la plupart des dépenses de la compagnie. Ceci nécessite un traitement particulier de ses "sinistres graves", sur lesquels on ne segmente pas ou peu en général. C'est le cas ici, certains sinistres rares dépassent les dizaines de milliers d'euros. Les garder dans nos GLM biaiseraient totalement nos résultats.

Pour résoudre ce problème, nous allons stocker les 1% de nos sinistres les plus graves à part, et la segmentation se fera sur les 99% restants. Plus précisément, la charge totale des sinistres produit par une police s'écrira comme suit :

$$S = \sum_{k=1}^{N} C_k + I \times L$$

Où :

- N est le nombre de sinistres standards supposé pour une loi binomiale négative
- C_k est le coût du k-ième sinistre standard
- I indique si la police a généré un sinistre grave
- L est le coût cumulé de ces sinistres graves

Nous segmenterons sur $\mathbb{E}[N]$ et $\mathbb{E}[C_k]$ pour calculer la sévérité espérée. Mais pas sur $\mathbb{E}[I]$ et encore moins $\mathbb{E}[L]$, les sinistres graves étant trop nombreux pour autoriser une personnalisation des montants.

4.3 Sélection finale des variables par approche Backward-Forward

Nous sommes alors partis du modèle de sévérité sans variable et avons lancé dessus une approche Backward-Forward avec les 6 variables restantes.
La procédure Backward Forward garde les variables minimisant le critère AIC.
Exécutée sur notre base d'apprentissage, elle a gardé nos 6 variables : puissance du moteur, âge du véhicule, type de véhicule, fin de vente du véhicule datant d'entre 6 et 11 ans et fin de vente du véhicule datant d'au moins 11 ans. Nous allons donc modéliser la sévérité avec `vh_din, vh_age, vh_type, sale_end_6_11, sale_end_11_plus` et `vh_fuel`.

4.4 Modélisation de la sévérité

On a supposé comme cela était écrit dans le livre de référence de Charpentier, que la sévérité suivait une loi exponentielle (voir Figure 4.4). Lors de la modélisation de notre MLG pour la sévérité, nous voulions choisir comme fonction de lien la fonction inverse qui est la fonction canonique d'une loi exponentielle. Cependant, l'algorithme IRLS vu en Modèle linéaire ne convergeait pas. Nous avons alors opté pour la fonction `log`, dont les propriétés de multiplicités et d'additivité sont désirables.

De même que pour la fréquence, nous allons créer une base d'apprentissage (80% de nos données) sur laquelle nous allons estimer nos paramètres. Les 20% de données restant (base de vérification) nous permettrons de tester nos résultats obtenus sur cette base d'apprentissage.
Pour cela, nous tirons aléatoirement 0.8*12391 lignes de TabSin. Nous calculons donc les coefficients de notre MLG sur la sévérité (variable `claim_amount`) avec la commande `glm` que nous stockons dans un vecteur "coeffsev".

Nous estimons ainsi $\mathbb{E}(B_i)$ pour tout i appartenant à notre base d'apprentissage.
Nous vérifions ensuite nos résultats sur la base d'apprentissage (voir Figure 4.4).

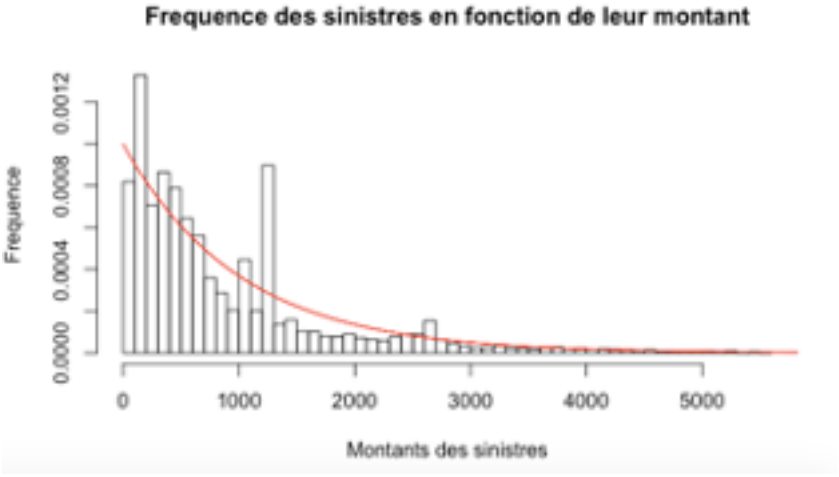

FIGURE 4.4: Fréquence sinistre en fonction des montants

Comme pour la fréquence, nous pouvons segmenter nos classes d'assurés en fonction des coûts de sinistres moyens engendrés, si sinistre il y a.

Moteur	Age véhicule	Type Véhicule	Fin de campagne marketing	Carburant	Sinistre moyen
Peu puissant	Élevé	Commercial	Lointaine	Diesel	616.81
Peu puissant	Élevé	Commercial	Intermédiaire	Diesel	659.37
Peu puissant	Élevé	Commercial	Lointaine	Essence	667.68
...	...	...	...	...	...
Puissant	Peu élevé	Tourisme	Récente	Diesel	996.92
Peu puissant	Peu élevé	Tourisme	Récente	Essence	1000.79
Puissant	Peu élevé	Tourisme	Récente	Essence	1079.14

TABLE 4.1: Montant moyen des sinistres en fonction des classes de véhicules

Les remarques sont globalement les mêmes que pour la fréquence. Ici on constate que les facteurs ayant le moins d'influence sont la puissance du moteur et la date de la fin de campagne marketing du véhicule. Pour rappel, la moyenne des coûts est inférieure à celle de nos accidents initiaux car les sinistres graves ont été omis. Leur présence sera comptabilisée dans le calcul de la prime pure.

Remarque : On aurait pu envisager un modèle log-normal dans cette séquence. La différence avec le modèle de Gamma est la distribution des résidus, distribués de façon gaussienne ou exponentielle. On garde le modèle Gamme pour représenter l'occurrence possible de catastrophes et donc une distribution dense aux valeurs modérées, qui s'écrase par la suite.

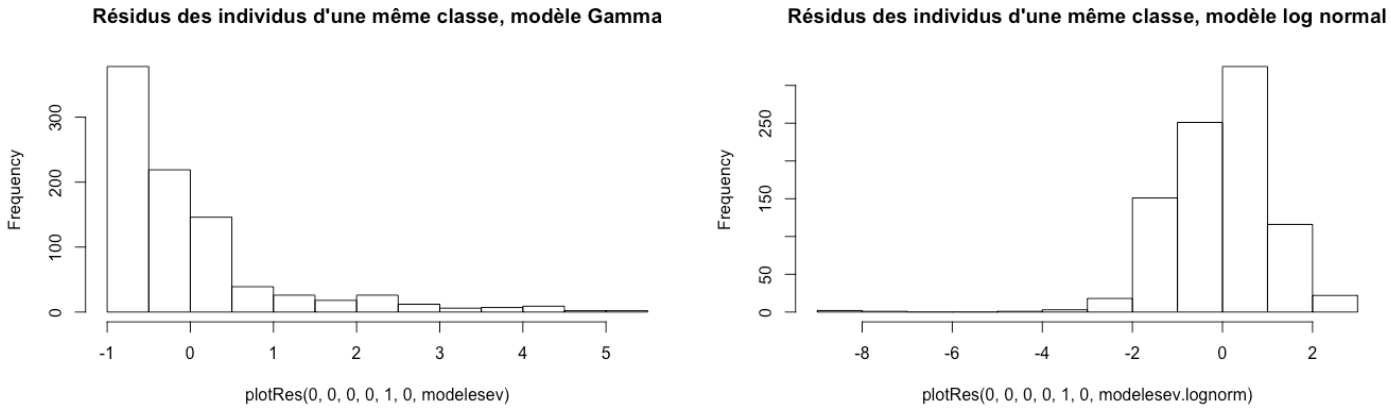

FIGURE 4.5

4.5 Calcul de la prime pure

Maintenant que nous avons estimé $\mathbb{E}[B]$ et $\mathbb{E}[N]$, nous pouvons désormais estimer la prime pure $\mathbb{E}[X]$.

Nous avions au préalable anticipé le fait que la prime pure se calculait avec les variables utilisées pour estimer la fréquence ainsi que celles utilisées pour estimer la sévérité. Ainsi nous n'avions supprimé aucune variable relative au véhicule dans le tableau `TAB` afin de pouvoir utiliser ce dernier pour calculer la prime pure. En outre, on choisit d'estimer $\mathbb{E}[N]$ avec la loi binomiale négative plutôt qu'avec la loi de Poisson.

On crée le vecteur `COEFF` qui contient tous les coefficients de la fréquence et ceux de la sévérité.

On a alors :

$$\begin{aligned}\mathbb{E}[X'] &= \mathbb{E}[N] \times \mathbb{E}[B] \\ &= \exp\{\beta_0 + \beta_1 \times \text{pol\_sit\_duration} + ... + \beta_9 \times \text{value9600\_16200}\} \times \\ &\exp\{\alpha_0 + \alpha_1 \times \text{vh\_din} + ... + \alpha_6 \times \text{vh\_fuel}\} \\ &= \exp\{\beta_0 + \beta_1 \times \text{pol\_sit\_duration} + ... + \beta_9 \times \text{value9600\_16200}\alpha_0 + \alpha_1 \times \text{vh\_din} + ... + \alpha_6 \times \text{vh\_fuel}\}\end{aligned}$$

Comme annoncé, il ne s'agit que d'une prime qui ne considère que les sinistres hors graves. Pour prendre en compte les catastrophes et permettre à l'assureur d'être solvable, nous allons quelque peu modifier cette première prime pure.

En négligeant les sinistres graves, la prime est de la forme :

$$\exp\left\{(\beta_{freq} + \beta_{cost})^t x_i\right\}$$

Notre prime pure doit alors s'élever à :

$$\exp\left\{(\beta_{freq} + \beta_{cost})^t x_i\right\} + q_i \mathbb{E}[L]$$

Où $q_i = \mathbb{E}[I_i]$ est la probabilité que l'assuré cause au moins un sinistre grave.

Ainsi chaque assuré paiera un excédent de prime pour couvrir les sinistres graves, excédent qui dépendra de sa propension à causer un accident.

Note : nous aurions pu aller plus loin et différencier le facteur q selon les classes d'assurés. Nous le supposons universel ici.

Chapitre 5

Ajout des données de l'ONISR

5.1 Préambule

5.1.1 Présentation

Maintenant que nous avons une première analyse de notre portefeuille de polices d'assurances, nous allons y inclure des données sur l'ONISR de 2017. Ces données concernent aussi bien les usagers impliqués dans l'accident, que les véhicules, les lieux des accidents, ou bien les caractéristiques liées aux accidents en eux-mêmes.

Une première observation est de mise : les données ne sont pas homogènes. En effet, par souci de confidentialité, les données de l'ONISR sont caractérisées uniquement par un numéro d'accident. Pour chacun de ces accidents, nous ne connaissons ni les polices d'assurances ni les montants de sinistres associés.

Ainsi afin d'établir un lien entre ces données de l'ONISR et nos jeux de données `pg17trainpol` et `pg17trainclaim`, nous allons utiliser la variable code postal `pol_insee_code` commune à tous nos tableaux de données. Nous allons à partir des données de l'ONISR déterminer des variables expliquant la fréquence et la sévérité, puis étudier ces variables dans chaque département. Cela permettra de déterminer de manière empirique les départements qui seraient "plus à risque" que d'autres. Nous pourrons ainsi enrichir notre premier modèle.

Cette influence de département n'est pas fortuite : il est globalement avéré qu'en zone urbaine, en raison des limitations de vitesse, la sévérité moyenne d'un sinistre sera plus faible qu'en milieu rural, et que la fréquence sera probablement plus élevée en raison du trafic important (le département 75, par exemple, sera caractérisé par une très forte fréquence de sinistres, en raison de la concentration exceptionnelle de population et de trafic).

Notre seconde observation est que, dans un cadre purement pratique, les données de l'ONISR ne sont pas des données observables a priori et ne devraient pas être utilisées dans la grille de tarification d'un assureur (les données relatives à l'année 2017 ne peuvent pas être utilisées pour la tarification en début de cette même année). Les données de l'année 2017 permettent donc d'affiner l'établissement de la prime pure pour l'année 2018, en supposant que les caractéristiques du zonier à maille spatiale restent globalement les mêmes d'une année sur l'autre.

5.1.2 Nettoyage des données :

Maintenant que les objectifs d'utilisation de ces données sont définis, nous passons à leur nettoyage. Nos polices d'assurances ne concernent que des assurés présents en France métropolitaine et en Corse. On exclut donc tous les accidents ayant été déclarés en Outre-Mer des 4 tableaux de l'ONISR (caractéristiques, lieux, véhicules, usagers). Par ailleurs, comme notre portefeuille initial ne concerne que des accidents de véhicules quatre-roues, il nous faut exclure tous les accidents n'impliquant que

des deux roues, ou des véhicules publics (comme les trains). Cette tâche est facilitée par la présence d'une covariable indicée indiquant la catégorie du véhicule (il suffit de supprimer les accidents qui n'impliquent pas au moins un véhicule à quatre roues).

Nous allons ensuite par des méthodes similaires à la première partie chercher dans les données de l'ONISR les variables ayant une influence sur la fréquence et/ou sur la sévérité. Une fois cela-fait, nous avons décidé de tester 2 approches différentes :

1. Attribuer à chaque département un score de fréquence (respectivement de sévérité) à partir des covariables jugées explicatives de la fréquence (respectivement explicatives de la sévérité). Nous aurons donc une unique nouvelle covariable pour la fréquence, de même que pour la sévérité, à inclure dans notre premier modèle.

2. Inclure toutes les variables jugées explicatives de la fréquence (respectivement de la sévérité) dans notre premier GLM expliquant la fréquence (respectivement expliquant la sévérité). Nous aurons donc plusieurs nouvelles covariables pour la fréquence et pour la sévérité à inclure dans notre premier modèle.

On convient bien évidemment que notre seconde méthode donnera un modèle plus précis (les coefficients du GLM seront plus nombreux), mais certainement aussi moins robuste. Il conviendra donc d'évaluer quelle méthode privilégier.

5.2 Variables de l'ONISR pour l'approche fréquence/sévérité

5.2.1 Variables de fréquence

Comme dans la première partie de nos travaux, il va s'agir de sélectionner les variables issues des données de l'ONISR permettant de caractériser les départements selon leurs nombres d'accidents, dans un cadre tout d'abord purement fréquentiel. Nous avons pour cela des variables relatives aux lieux des accidents, aux usagers, aux véhicules, et d'autres liées aux accidents eux-mêmes. Cependant, comme expliqué en préambule, les données ne sont pas homogènes par souci de confidentialité. Nous ne pouvons donc pas utiliser nos jeux de données initiaux pour caractériser la fréquence de chaque département. Nous allons donc contourner le problème de la manière suivante : utiliser les données fournies pour déterminer les conditions les plus dangereuses (qu'elles soient météorologiques, structurelles etc. . .) puis, par le biais de recherches ultérieures ou non, déterminer les départements présentant ces conditions. La variable expliquant la fréquence la plus naturelle à extraire des données de l'ONISR est celle du nombre d'accidents par département.

Il est intuitif de penser que les départements les plus représentés sont les plus dangereux. Cependant, cette intuition est biaisée par le fait que les populations par département ne sont pas égales. Un département particulièrement peuplé comme Paris a une exposition globale au risque bien plus importante et donc une fréquence d'accidents nettement supérieure aux autres départements. En normalisant par les populations dans chaque département, on obtient une représentation plus fidèle à la réalité.

Malgré la normalisation, on constate tout de même une domination des départements à grosses agglomérations (région parisienne, Marseille, Corse, Lyon) et donc forte population. On décide alors d'établir une relation entre la quantité d'accidents et la population dans chaque département. Cette relation n'est pas linéaire.

La relation en échelle logarithmique nous montre bien qu'une augmentation de population accroit le nombre de sinistres de manière exponentielle. Nous choisissons donc de pénaliser les départements à forte population, sujette à de nombreux accidents réguliers. Nous créons donc la variable `score_pop`.

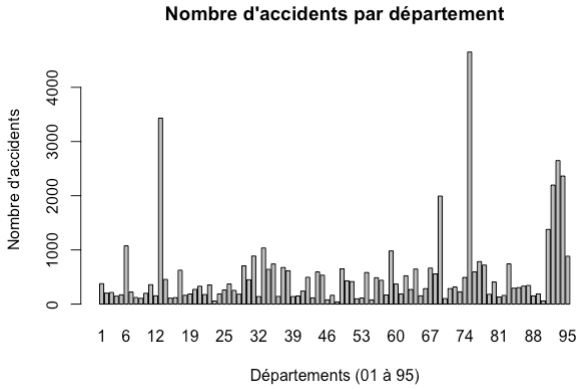

FIGURE 5.1: Nombre d'accident par département

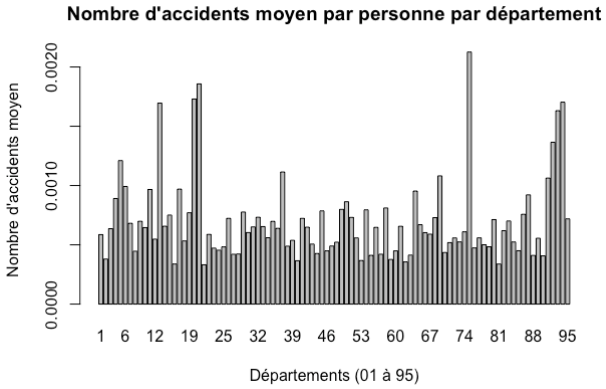

FIGURE 5.2: Proportion d'accident par département

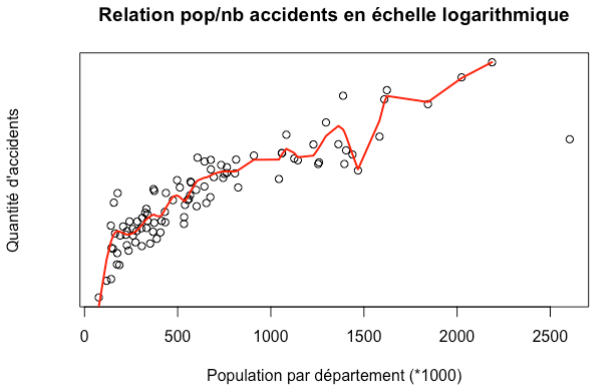

FIGURE 5.3: Accidents par population

Nous passerons la partie consistant à l'élimination des variables inutiles pour la fréquence, car les arguments restent semblables à ceux de la première partie et nous viendrons directement à expliciter les variables que nous retiendrons, et comment les traiter.

`Surf` : l'état de la surface lors de l'accident. Nous avons constaté que les conditions surface enneigée et surface verglacée étaient propices à la sinistralité de manière générale. Un département présentant une forte proportion d'accidents dans ces conditions sera in fine un département aux conditions météorologiques propices à la fréquence et sera à pénaliser.

`Env1` : proximité ou non d'une école. Nous avons dans l'idée qu'un département présentant une proportion élevée d'accidents ayant lieu à proximité d'une école traduit une agressivité au volant plus marquée et un manque de prudence.

`Int` : Si l'accident a eu lieu dans une intersection, et si oui dans quel type d'intersections. Les intersections sont propices aux accidents. Un grand nombre d'intersections dans un département augmente considérablement son nombre d'accidents.

`Atm` : Les conditions atmosphériques au moment de l'accident. Nous avons étudié l'influence de l'ensoleillement sur le nombre d'accidents. Comme `atm` ne renseignait pas cette condition météorologique, nous avons récupéré le nombre d'heures d'ensoleillement dans chaque département en 2017.

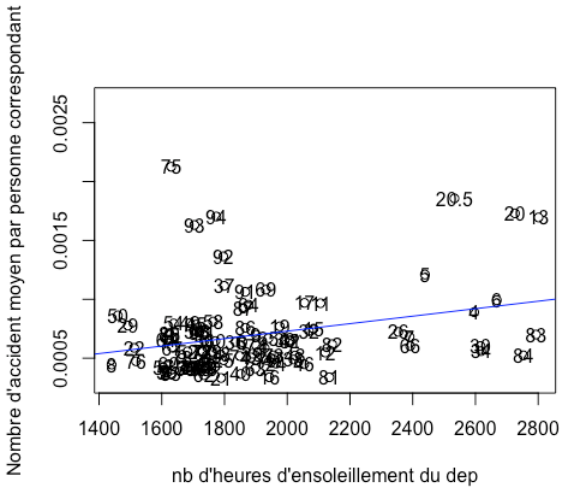

FIGURE 5.4: Accident par Ensoleillement

5.2.2 Variables de sévérité

Les variables de sévérité présentent un grand avantage analytique par rapport aux variables de fréquence : elles ne sont pas biaisées par leurs proportions de présence dans les accidents. En effet, alors que la fréquence avait son caractère ambigu (difficile de déterminer les conditions dangereuses sachant que les accidents arrivent le plus souvent en conditions non pas dangereuses mais fréquentes), il est bien plus facile d'observer parmi les accidents ayant eu lieu ceux particulièrement violents et démarquer les départements recensant des sinistres à fortes valeurs (en observant par exemple la proportion de tués par département, qui serait à priori un bon indicateur pour la sévérité).

Une première intuition que l'on peut émettre (et qui a déjà été énoncée auparavant) est qu'en milieu rural, la sévérité sera souvent plus élevée du fait des vitesses plus élevées.

`Col` : Le type de collision lors de l'accident. Une collision frontale implique en général des dégâts non seulement matériels mais aussi corporels dans au moins l'un des véhicules impliqués dans l'accident. Il est ainsi naturel de considérer que ce type de collision va fortement augmenter la sévérité.

`Obsm` : L'obstacle mobile heurté lors de l'accident. Les dégâts corporels impliquant une sévérité bien plus élevée que les dégâts matériels (et donc un piéton heurté peut coûter bien plus cher à l'assureur).

`Choc` : Le point de choc initial du véhicule lors de l'accident. De multiples chocs/tonneaux lors d'accidents impliquent de fortes vitesses. Les dégâts matériels voir corporels ont donc de fortes chances d'être élevés.

`Agg` : Si l'accident a eu lieu hors ou en agglomération. Comme la limitation de vitesse est plus élevée hors agglomération, nous estimons que les accidents y sont plus coûteux.

`Grav` : La gravité de l'accident. On suppose naturellement qu'elle doit être fonction croissante de la sévérité.

5.2.3 Première approche : score de fréquence et de sévérité uniques

On commence avec une première idée simple : ajouter à nos données initiales un score de fréquence unique, et un score de sévérité unique.

Fréquence

De manière générale, chaque score a ainsi été déterminé par un rapport entre la proportion d'occurrence dans la modalité dans le département par la proportion maximale de chaque département : un score est donc compris entre 0 et 1 à chaque fois (Exemple : un département ayant un `score_surf` de 1, est celui qui présente le plus souvent des conditions enneigées/verglacées, si un autre un score de 0.5, il présente 2 fois moins ces conditions que le premier).

Le score de fréquence utilisé dans notre première approche est donc le vecteur de taille 96 (94 départements métropolitains + les 2 Corses) suivant (que nous normalisons une dernière fois) :

$$score_{freq} = score_{acc\_par\_pers} + score_{pop} + score_{surf} + score_{ecole} + score_{int} + score_{soleil}$$

Afin de discrétiser notre variable, nous voulons faire apparaitre des classes homogènes de département ayant un score de fréquence proches. Nous utilisons pour cela la méthode des k-means. Cependant cette méthode comporte 2 inconvénients :

- Nous devons spécifier nous-mêmes le nombre de classes que nous souhaitons. Pour cela, nous nous aider d'un dendogramme. Nous observons que l'amplitude de la branche permettant de scinder le score de fréquence en 3 classes est relativement grande. Nous optons donc pour un 3 classes.
- La méthode des k-means s'initialise de manière aléatoire. Les classes crées d'une exécution à l'autre du code ne sont pas toujours les mêmes.

Pour remédier à cela, nous choisissons nous même les valeurs d'initialisations de la méthode comme suit : la première classe (classe 1 : scores faibles) part de min(`score_freq`), la seconde (classe 2 : scores intermédiaires) part de la médiane de `score_freq` et la dernière (classe 3 : scores élevés) part de max(`score_freq`). Ainsi la méthode reste stable d'une exécution du code à l'autre.

Nous obtenons donc 3 classes : la classe 1 est de taille 27, la classe 2 de taille 56 et la classe 3 de taille 13. Une fois nos clusters définis au sein des départements, nous ajoutons à `TAB` l'information des scores (qui s'obtient par une fusion de tableaux selon le département). Cette information est représentée par deux colonnes : Risque Faible et Risque Fort, valant 0 ou 1 (et toutes les deux 0 si le département est de Risque Modéré). Un nouveau Forward-Backward (appliqué dans un modèle binomial négatif) conserve toutes nos variables. Ainsi le score de fréquence constitue une information supplémentaire (selon que le département soit de risque faible, modéré, ou fort).

Sévérité

Chaque score par département est ainsi déterminé par une moyenne des modalités.

Exemple :

$$Gravite = \begin{cases} 1 & \text{si l'usager est indemne} \\ 2 & \text{s'il est blessé légèrement} \\ 3 & \text{s'il est hospitalisé ou mort} \end{cases}$$

Pour un département en particulier, une moyenne élevée traduirait une forte proportion d'accidents résultant en la mort ou l'hospitalisation, augmentant de fait la sévérité. On crée un score de sévérité `score_sev` en sommant tous les autres scores de sévérité. Nous le normalisons ensuite comme pour la fréquence.

Après application de la méthode des k-means de la même manière que pour la fréquence, on obtient 3 classes : la classe 1 est de taille 32, la classe 2 de taille 54 et la classe 3 de taille 10 (par ordre croissant de risque). Une fois nos clusters définis au sein des départements, nous ajoutons à TabSin (toujours privé des sinistres extrêmes) l'information des scores (qui s'obtient par une fusion de tableaux selon le département).

Un nouveau Forward-Backward (appliqué dans un modèle de loi Gamma) conserve toutes nos variables. Ainsi le score de sévérité constitue une information supplémentaire (selon que le département soit de risque faible, modéré, ou fort).

5.3 Seconde approche : traitement des variables au cas par cas

Motivation : Un défaut majeur de la première approche est que par construction tous les coefficients multiplicateurs des nouvelles variables sont égaux et en particulier de même signe. Ainsi, cela suppose que chaque variable impacte la sinistralité de façon égale et dans le même sens, ce qui n'est pas forcément le cas, comme on le verra dans cette partie.

Nous allons faire le même choix de variables pour la sévérité et la fréquence dans cette partie. La différence est que nous allons appliquer la méthode des k-means à chaque variable (dont le score pour chacune a été défini de la même façon que pour la première approche). Pour chacune d'elle, nous allons définir 2 clusters, en prenant bien soin de faire en sorte que la modalité égale à 0 soit la plus représentée (cf fonction `discrete` du code R). Si la méthode donne des barycentres différents selon les essais, on relève la séparation la plus fréquente, et on rentre les barycentres en paramètres "à la main" pour obtenir cette configuration de manière certaine. On obtient ainsi pour la fréquence :

$$Ensoleillement = \begin{cases} 0 & \text{si l'ensoleillement est modéré ou faible} \\ 1 & \text{sinon} \end{cases}$$

$$Surface = \begin{cases} 0 & \text{si les conditions neiges et verglas sont très rares} \\ 1 & \text{sinon} \end{cases}$$

$$Ecoles = \begin{cases} 0 & \text{s'il la proportion d'accidents près des écoles est élevée} \\ 1 & \text{sinon} \end{cases}$$

$$Intersection = \begin{cases} 0 & \text{si les intersections à caractères dangereux sont rares} \\ 1 & \text{sinon} \end{cases}$$

$$Population = \begin{cases} 0 & \text{si le département est relativement peu peuplé} \\ 1 & \text{sinon} \end{cases}$$

Pour la sévérité on obtient :

$$Collision = \begin{cases} 0 & \text{si les Collisions impliquant de nombreux véhicules sont moins fréquentes} \\ 1 & \text{sinon} \end{cases}$$

$$Gravite = \begin{cases} 0 & \text{si les accidents provoquent plus souvent des blessés} \\ 1 & \text{sinon} \end{cases}$$

$$Choc = \begin{cases} 0 & \text{s'il y a plus souvent des chocs frontaux ou multiples avec tonnaux} \\ 1 & \text{sinon} \end{cases}$$

$$Obstacle\_mobile = \begin{cases} 0 & \text{si peu de piétons sont heurtés} \\ 1 & \text{sinon} \end{cases}$$

$$Hors\_agglo\_faible = \begin{cases} 0 & \text{si les accidents hors agglomérations sont modérés ou fréquents} \\ 1 & \text{sinon} \end{cases}$$

$$Hors\_agglo\_forte = \begin{cases} 0 & \text{si les accidents hors agglomérations sont rares ou fréquents} \\ 1 & \text{sinon} \end{cases}$$

Remarque 1 : Dans le cas particulier de la variable Agg, on a pris le choix de faire trois classes pour mieux distinguer les départements particulièrement urbains (typiquement le 75).

Remarque 2 : Nous ne verrons pas en détails la façon dont les modalités ont été regroupées en classes (passage de 9 modalités à 3 pour la variable Choc par exemple), pour permettre de faire des scores plus pertinents avant d'appliquer la méthode des k-means. Les regroupements sont expliqués en annotations dans le code R. Nous avons maintenant nos variables définies. Nous pouvons lancer nos nouveaux GLM, en conservant respectivement un modèle binomial négatif et un modèle de Gamma.

Il est a noté que nous faisons cette fois-ci nos GLM sur 100% des données. En effet, l'objectif est de définir une prime pure pour l'exercice 2018, et comme la cohérence de nos modèles a déjà été prouvée, il ne sert à rien de recourir à une base de test. Au contraire, nous voulons une prime qui ne soit pas déterminée par le hasard de la composition d'une base d'apprentissage, elle doit donc être déterminée à partir de l'entièreté des données.

Les résultats obtenus sont les suivants :

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.59036    0.02999 -53.028  < 2e-16 ***
vh_fuel           -0.21456    0.02081 -10.313  < 2e-16 ***
pol_pay_freq      -0.09098    0.01903  -4.782 1.73e-06 ***
drv_sex1           0.05326    0.01861   2.862  0.00421 **
pol_coverage_Median -0.42675  0.02222 -19.204  < 2e-16 ***
pol_coverage_Mini  -1.43290    0.05886 -24.343  < 2e-16 ***
age_lic_1_20       0.06022    0.02314   2.603  0.00924 **
age_lic_36etplus  -0.05404    0.02035  -2.656  0.00791 **
value9600         -0.45113    0.04377 -10.307  < 2e-16 ***
value9600_16200   -0.20494    0.02057  -9.965  < 2e-16 ***
Population         0.14786    0.01935   7.643 2.12e-14 ***
Ensoleillement    -0.06666    0.02824  -2.361  0.01824 *
Ecoles            -0.00522    0.01916  -0.272  0.78526
Surface            0.01602    0.02628   0.610  0.54212
Intersection      -0.02885    0.01893  -1.524  0.12743
```

FIGURE 5.5: Coefficient de la fréquence

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.87670    0.02925 235.140  < 2e-16 ***
vh_age           -0.12717    0.04217  -3.016 0.002570 **
vh_din            0.06572    0.02398   2.741 0.006136 **
vh_fuel           0.07199    0.02022   3.560 0.000372 ***
vh_type          -0.10376    0.03463  -2.997 0.002735 **
sale_end_6_11    -0.11039    0.02280  -4.842  1.3e-06 ***
sale_end_11_plus -0.17139    0.04839  -3.542 0.000399 ***
Collision        -0.04101    0.02190  -1.872 0.061183 .
Gravite          -0.06632    0.02596  -2.555 0.010632 *
Choc             -0.08540    0.02401  -3.557 0.000376 ***
ObstacleMobile   -0.01076    0.02868  -0.375 0.707517
HAfaible          0.12545    0.03756   3.340 0.000839 ***
HAinter           0.05402    0.02467   2.190 0.028536 *
```

FIGURE 5.6: Coefficient de la sévérité

Dans le cas de la fréquence, les p-values pour la présence d'écoles et d'intersections sont supérieures à 0.05. On les retire donc du modèle. De même pour l'état des surfaces. Toutes nos autres variables sont conservées, on peut donc déduire deux observations :

1. Les données relatives à une police d'assurance sont bien plus significatives que les données relatives à un département ; et même si elles n'entrent pas nécessairement en contradiction, elles n'apportent qu'une correction partielle. Il est tout de même naturel de constater que la sinistralité d'un assuré dépend bien plus de l'assuré lui-même que de son environnement.

2. On constate que les données topographiques (i.e. les écoles ou les intersections) ne permettent pas d'expliquer la sinistralité. Les seules données l'expliquant sont celles qui expliquent la densité du trafic en lui-même : une forte population implique un fort trafic, et donc plus d'accidents ; une région ensoleillée traduirait de plus nombreux déplacements aux activités en plein air, augmentant les chances de causer un sinistre.

Pour la sévérité, on constate l'importance de l'agglomération, de la nature du choc, et la gravité en générale, mais pas dans le sens auquel on pourrait s'y attendre. Nous constatons en effet, les sévérités moyennes espérées selon la classe d'assurés :

On utilisera les abréviations suivantes :

- Vh pour Véhicule
- FM pour fin de campagne marketing
- PAHA pour proportion d'accident hors agglomération
- GG pour Gravité Générale
- GC pour Gravité choc

Moteur	Age Vh	Type Vh	FM	Carburant	PAHA	GG	GC	Montant
Peu Puissant	Élevé	Commerciale	Lointaine	Diesel	Forte	Faible	Forte	583.18
Peu Puissant	Élevé	Commerciale	Lointaine	Diesel	Intermédiaire	Forte	Forte	585.27
...	...	...	...	...	...	...	...	...
Puissant	Bas	Tourisme	Récente	Essence	Intermédiaire	Faible	Faible	1134
Puissant	Bas	Tourisme	Récente	Essence	Faible	Faible	Faible	1137

TABLE 5.1: Montant espéré en fonction de nos variables

Contrairement à notre intuition, un choc jugé violent a un impact négatif sur la sévérité, il en va de même pour la gravité générale et la proportion d'accidents hors agglomération. La raison est que dans la segmentation que nous avons effectuée, nous avons omis les sinistres extrêmes. Or, en dressant une relation de la gravité moyenne générale et des chocs en fonction de la moyenne des coûts des sinistres par département (et de même pour la proportion d'accidents hors agglomérations), nous obtenons :

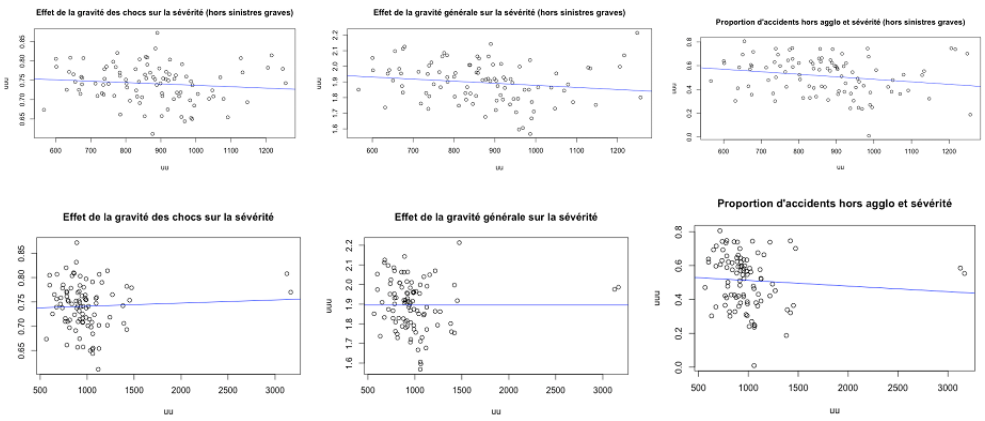

FIGURE 5.7

Les chocs violents ont en fait une importance dans la survenance de sinistres extrêmes. La gravité en revanche, par son caractère probablement trop général, n'impacte pas directement les coûts des sinistres. De plus, la proportion d'accidents hors agglomération conserve le même impact qu'auparavant. L'interprétation que nous pourrions donner est qu'une proportion faible d'accidents hors agglomération traduirait un département très urbain, avec un pouvoir d'achat par habitant plus élevé, et donc des véhicules plus chers, plus fragiles, et peut-être plus d'usagers par véhicules. Ce sont des facteurs qui augmentent le coût d'un sinistre (on peut peut-être également suggérer qu'en milieu urbain, les sinistres légers sont moins souvent déclarés, augmentant de fait le coût moyen d'un sinistre).

Nous ne conserverons ainsi que cette dernière variable dans notre modèle de sévérité. Nous allons marquer l'importance des différents types de chocs en agissant sur la probabilité de causer un sinistre grave (que nous supposions universelle avant d'utiliser les données de l'ONISR).On remarque en effet qu'en ajoutant les sinistres graves, la tendances s'inverse. On obtient ainsi le calcul de prime pure suivant :

$$\exp\left\{(\beta_{freq} + \beta_{cost})^t x_i\right\} + q_i \mathbb{E}[L]$$

Où q_i dépend de la police d'assurance. Notre prime pure sera alors affectée par 3 facteurs :

- Le changement de fréquence pour chaque police
- Le changement de sévérité (hors sinistres extrêmes) pour chaque police
- Le changement de la probabilité de causer un sinistre extrême (qui était unique jusqu'alors).

Chapitre 6

Comparaison des résultats

Les résultats offerts par l'ajout des données de l'ONISR sont très semblables aux résultats initiaux. Il est naturel d'observer qu'avec plus de variables, le spectre des primes possibles s'étend en offrant plus de caractéristiques possibles. On constate les résultats sur les tableaux 7.1 et 7.2 en annexe.

Les données de l'ONISR apportent une information supplémentaire sur la tarification : si un individu pour lequel la tarification est maximale (sans ONISR) a des caractéristiques particulières, ces dernières se retrouveront dans les tarifications maximales de l'ONISR, et les nuances se feront en fonctions des modalités des variables issues de l'ONISR. On notera également que l'individu présentant une tarification maximale est simplement celui cumulant les caractéristiques propices à une forte fréquence et une forte sévérité (pour peu que la classe de polices en question existe bel et bien).

Néanmoins, dans l'assurance, les primes ne sont pas aussi personnalisées et sont segmentées en classes. La méthode des k-means est donc particulièrement utile ici puisqu'elle regroupera différentes classes sous le couvert d'un seul et même tarif. On prend le parti de supposer 10 niveaux de tarification, nous allons simplement dans ce cas essayer de trouver une condition suffisante pour qu'un assuré se situe aux niveaux de tarification les plus élevés. Pour ce faire, nous revenons à nos GLM : les p-values nous donnent l'idée de l'importance des covariables. En prenant les quelques covariables les plus importantes, et les modalités adaptées, nous devrions trouver un "début de profil" qui assurerait une grosse tarification.

Exemple : Un assuré vivant dans un département fortement peuplé, avec une couverture maximale, un véhicule récent et roulant au diesel est assuré de se trouver dans les 5 niveaux de tarification les plus élevés (voir fin du code R).

Cet exemple montre toutefois qu'en raison du nombre important de covariables, les suppositions de classe à faire sont nombreuses, pour un niveau de déduction relatif, et nous n'irons pas plus loin ici.

Avec des primes pures variant de 15 à 294€, la méthode des k-means nous donnerait la grille de tarification suivante :

Risque 1	29.42 €	Risque 6	156.27 €
Risque 2	63.59 €	Risque 7	177.61 €
Risque 3	88.21 €	Risque 8	198.35 €
Risque 4	111.93 €	Risque 9	221.44 €
Risque 5	134.61 €	Risque 10	251.34 €

On observe ensuite à l'aide de boxplots les différences entre primes prédites avec et sans données de l'ONISR.

Nous remarquons que les primes pures prédites selon les modalités des variables sont souvent très proches. Nous observons cependant quelques différences : Un individu ayant une couverture médiane, ou dont la valeur du véhicule est faible, ou encore dont le véhicule est en vente depuis longtemps sur le marché, vont payer une prime pure plus élevée avec notre second modèle (cf les figures 6.1 et 6.2 du rapport et ceux en fin de code R). Cela rentre en accord avec notre intuition : les départements sont caractérisés (en plus de leurs densités de population) par les gammes de véhicules que l'on peut y trouver, et qui expliquent par exemple la sévérité. Il est donc naturel d'observer des écarts plus importants dans les classes d'assurés qui se démarquent par la qualité d'un véhicule.

En revanche, un individu dont le véhicule a une valeur moyenne ou dont le véhicule est peu puissant vont payer une prime pure moins élevée avec notre second modèle.

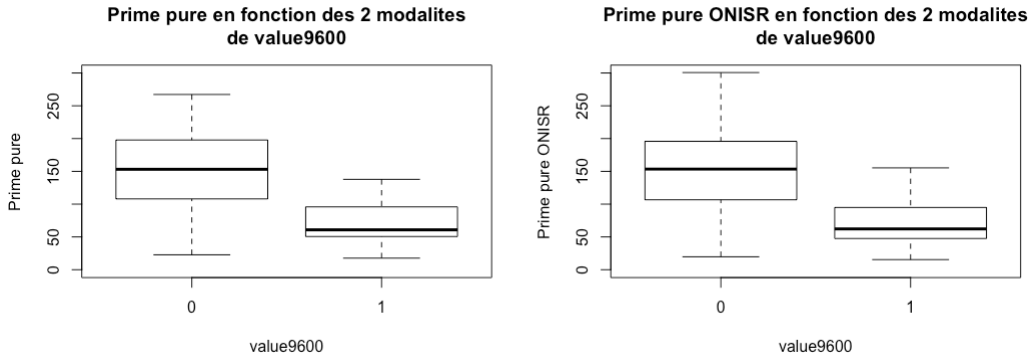

FIGURE 6.1

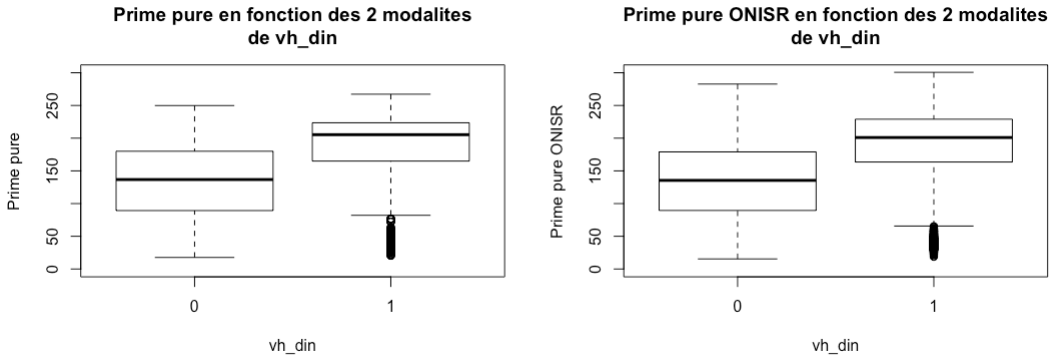

FIGURE 6.2

Néanmoins, les différences restent minimes, et confirment que l'information apportée par l'ONISR reste presque anecdotique.

On observe les différences entre les deux primes pures selon les différents départements. Nous remarquons que la prime pure avec les données de l'ONISR est plus contrastée.

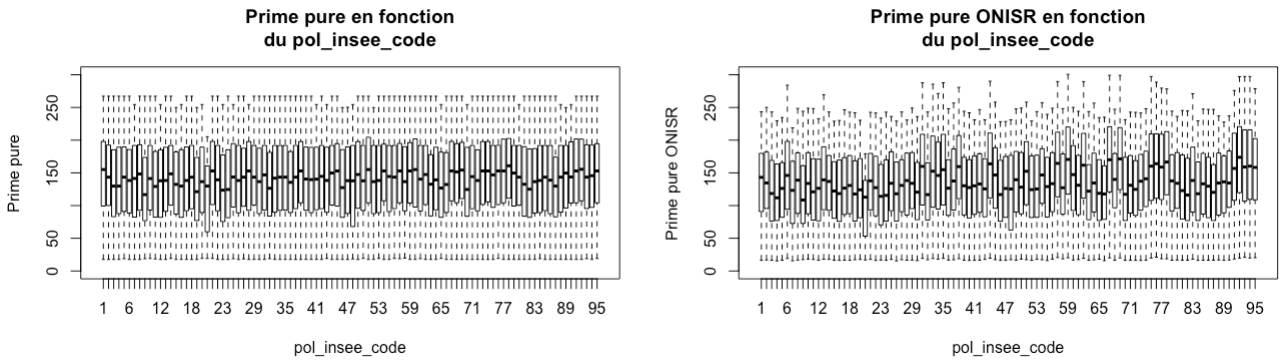
Prime pure en fonction
du pol_insee_code
Prime pure
0
50
150
250
1 6 12 18 23 29 35 41 47 53 59 65 71 77 83 89 95
pol_insee_code
Prime pure ONISR en fonction
du pol_insee_code
Prime pure ONISR
0
50
150
250
1 6 12 18 23 29 35 41 47 53 59 65 71 77 83 89 95
pol_insee_code

FIGURE 6.3

Chapitre 7

Conclusion

En choisissant un modèle binomial négatif pour la fréquence et un modèle de Gamma pour la sévérité nous avons donc trouvé des primes pures prédites pour chacun de nos 2 modèles.

Notre second modèle avec les données de l'ONISR vient confirmer les conclusions de notre premier modèle :

- Les primes pures prédites par nos 2 modèles sont très proches, bien que la variance de celles du second modèle soit plus élevée.
- Le profils type de conducteur présentant des primes pures élevées/faibles sont les mêmes dans les 2 modèles : individu habitant dans un département fortement peuplé avec une couverture maximale, un véhicule récent et roulant au diesel, constituent ceux qui auront une prime plus chère dans les deux cas.
- De manière générale, les primes pures prédites pour chaque modalité de chaque variable utilisée dans nos 2 modèles sont très similaires.

En revanche, les données de l'ONISR apportent une précision plus fine pour la tarification de la prime pure selon les départements.

Cependant ces données donnent lieu à de nouvelles variables pour notre modèle, ce qui en réduit donc la robustesse (cf la variance de nos échantillons). D'après nos différentes comparaisons de modèles, l'ajustement des primes pures grâce à l'ONISR est trop peu significatif par rapport à la perte de robustesse que ces données engendrent. Nous choisissons donc de retenir notre premier modèle sans les données de l'ONISR.

Nous calculons donc les primes pures prédites sur le jeu de données `g17testyear1` (cf annexe du code), que l'on peut à nouveau segmenter avec la méthode des k-means pour obtenir la grille de tarification pour l'année 2018.

Annexes

Carburant	Fréq Pai	Sexe	Couverture	Age Permis	Valeur Vh	Age Vh	Pui véhicule	Type véhicule	FM.	Prime
Essence	Peu Élevée	Homme	Minimale	Élevé	Peu Cher	Vieux	Faible	C	Lointaine	17.62 €
Essence	Peu Élevée	Femme	Minimale	Élevé	Peu Cher	Vieux	Faible	C	Lointaine	18.51 €
Essence	Peu Élevée	Homme	Minimale	Moyen	Peu Cher	Vieux	Faible	C	Lointaine	18.79€
...	...	...	...	...	...	...	...	...	...	...
Diesel	Élevée	Femme	Maximale	Intermédiaire	Cher	Récent	Forte	T	Récente	250.32€
Diesel	Élevée	Homme	Maximale	Faible	Cher	Récent	Forte	T	Récente	254.44 €
Diesel	Élevée	Femme	Maximale	Faible	Cher	Récent	Forte	T	Récente	267.32€

Table 7.1: Tableau récapitulatif sans les données de l'ONISR

Carburant	Fréq Pai	Sexe	Couv.	AgePermis	ValeurVh	ValeurVh	PuiVh	TypeVh	FM	Pop Départ.	Soleil	HA	Prime
Essence	Peu Élevée	Homme	Mini	Élevé	Peu Cher	Vieux	Faible	Comm.	Loint.	Faible	Fort	Fort	15.22
Essence	Peu Élevée	Homme	Mini	Élevé	Peu Cher	Vieux	Faible	Comm.	Loint.	Faible	Fort	Moyenne	15.91
Essence	Peu Élevée	Homme	Mini	Élevé	Peu Cher	Vieux	Faible	Comm.	Loint.	Faible	Faible	Fort	15.93
...	...	...	...	...	...	...	...	...	...	...	...	...	
Diesel	Élevée	Homme	Max	Faible	Cher	Récent	Forte	Tour.	Réc.	Forte	Faible	Faible	281.03
Diesel	Élevée	Femme	Max	Faible	Cher	Récent	Forte	Tour.	Réc.	Forte	Faible	Moyenne	286.93
Diesel	Élevée	Femme	Max	Faible	Cher	Récent	Forte	Tour.	Réc.	Forte	Faible	Faible	293.87

Table 7.2: Tableau récapitulatif avec les données de l'ONISR