IRM | institut du risk management

Mémoire présenté devant l'Institut du Risk Management pour la validation du cursus à la Formation d'Actuaire de l'Institut du Risk Management et l'admission à l'Institut des Actuaires

le

Par : **Benoit Lamarsaude**

Titre : **Construction de variables de coût de sinistre décorrelées pour le pilotage des experts en automobile**

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

Les signataires s'engagent à respecter la confidentialité ci-dessus

Membres présents du jury de l'Institut des Actuaires :

Entreprise :
Nom : Prim'Act
Signature :

Directeur de Mémoire en entreprise :
Nom : Frédéric Planchet
Signature :

Membres présents du Jury de l'Institut du Risk Management :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels (après expiration de l'éventuel délai de confidentialité)

Secrétariat :

Bibliothèque :

Signature du responsable entreprise

Signature du candidat

Résumé

Dans la gestion des sinistres dommages l'assureur sollicite l'intervention d'un expert qui agit comme un agent de la maîtrise des coûts pour le compte de l'assureur. L'assureur sélectionne et anime ses réseaux d'experts pour s'assurer que ses prévisions seront respectées. A cet effet, il est nécessaire de disposer de métriques lui permettant de mesurer la performance des experts. Pour cela on compare la capacité de chacun d'entre eux à produire des évaluations du coût de la remise en état les plus économiques. Or, par la nature aléatoire des sinistres automobiles, la valeur du coût moyen est influencée par des facteurs exogènes à l'expert. Les travaux de cette étude visent à construire une autre classification basée sur une variable de coût non affectée de ces facteurs exogènes. La variance de la variable d'intérêt est décomposée en partie exogène et endogène. L'objectif est de ne conserver que la variance liée aux facteurs endogènes, correspondant aux différents niveaux de performance entre les experts considérés. Un modèle GLM multiplicatif est utilisé pour évaluer l'impact des effets exogènes. La qualité d'information dans le modèle est évaluée à partir de la chute de variance. La modélisation réalisée permet aussi de donner une vue analytique en donnant les niveaux d'impacts des différentes variables et de leurs modalités sur le coût moyen de l'expert. Afin de réduire la dépendance entre variables endogènes et exogènes, de nouvelles modélisations sont réalisées avec l'ajout de pénalisation et d'interactions entre variables. La performance des modèles et les classifications qui en résultent sont étudiées et comparées. Ces travaux permettent d'obtenir une classification des experts plus pertinente, avec une influence minimisée des facteurs exogènes. On obtient ainsi une appréciation plus fine de la performance des experts.

Mots-clés : *automobile*, *expert*, *non-vie*, *pilotage*, *GLM*.

Abstract

In the management of motor claims, the insurer requests the intervention of a loss adjuster who acts as a cost control agent on behalf of the insurer. The insurer selects and manages its networks of loss adjuster to ensure that its forecasts are respected. To this end, it is necessary to have metrics allowing it to measure the performance of the loss adujster. To do this, we compare the ability of each of them to produce the most economical evaluations of the cost of rehabilitation. However, due to the random nature of motor claims, the value of the average cost is influenced by factors exogenous to the loss adjuster. This study aims to construct another classification based on a cost variable not affected by these exogenous factors. The variance of the variable of interest is broken down into exogenous and endogenous parts. The objective is to retain only the variance linked to endogenous factors, corresponding to the different levels of performance between the experts considered. A multiplicative GLM model is used to assess the impact of exogenous effects. The quality of information in the model is evaluated from the decrease of variance . The modeling carried out also makes it possible to provide an analytical view by giving the levels of impact of the various variables and their methods on the average cost of the expert. In order to reduce the dependence between endogenous and exogenous variables, new models are carried out with the addition of penalization and interactions between variables. The performance of the models and the resulting classifications are studied and compared. This work makes it possible to obtain a more relevant expert classification, with a minimized influence of exogenous factors. We thus obtain a more detailed assessment of the performance of the experts.

Keywords : *motor*; *loss-adjuster*; *non-life* ; *GLM.*

Note de Synthèse

En assurance non-vie, le marché des produits de couverture des risques de fréquence est concurrentiel et particulièrement en assurance automobile. De nombreux acteurs se disputent les parts de marché avec des dynamiques de croissance qui peuvent être très différentes. L'intensité de la concurrence est telle que les assureurs peinent à dégager des marges sur ces activités. La gestion des sinistres en application du principe indemnitaire est complexe et consommatrice de ressources. Le contexte de taux bas contraint les assureurs à améliorer leurs résultats techniques.

Dans la gestion des sinistres dommages l'assureur sollicite l'intervention d'un expert qui agit comme un agent de la maîtrise des coûts pour le compte de l'assureur. L'assureur sélectionne et anime ses réseaux d'experts pour s'assurer que ses prévisions seront respectées. Pour cela l'assureur doit disposer de métriques lui permettant de mesurer la performance des experts. Or l'expert intervenant sur des sinistres dont la nature et l'environnement sont par nature aléatoire, la mesure de la maîtrise du coût est une opération délicate.

De précédents travaux ont traité la question de la performance des réseaux d'experts en automobile. L'approche consistait à développer des modèles visant à comparer des réseaux d'experts, c'est à dire des ensembles d'expert. Sur la base des données des rapports d'expertise, de Lignaud de Lussac (2018) utilise un modèle de régression dans lequel est introduit une variable indiquant le réseau d'appartenance. L'influence de cette variable sur le coût de sinistre est mesurée à l'aide de la valeur du coefficient associée à la variable dans le cadre d'une modélisation de type GLM. Par la suite, Khougea (2019) et Wabo Foka (2020) ont traités la même problématique en comparant le pouvoir prédictif de différents modèles d'apprentissage automatique sur le même jeu de données.

L'approche menée au travers ces travaux revêt un double objectif :

- fournir des méthodes de classification des individus d'une population d'experts intervenant dans le cadre de sinistres automobiles : cela répond au besoin de sélection ;
- identifier les variables explicatives de la performance : cela répond au besoin d'animation.

L'approche retenue se différencie des travaux précédents dans le sens, où elle ne vise pas à comparer les résultats de différentes populations d'experts mais individuellement les experts entre eux.

Données

Le jeu de donnée utilisé comprend 506607 observations et 50 variables. Il s'agit de l'ensemble des rapports d'expertise effectués par 732 experts sur une année. Pour bon nombre de variables, des traitements ont été nécessaires : suppression ou imputation des valeurs manquantes, regroupement de modalités, écrêtage. Après l'ensemble des traitements et transformations sur les variables le jeu de données comporte 468363 observations, soit une perte de -7.5%, qui reste acceptable pour l'étude. Des

indicateurs de performance qui serviront à objectiver les experts dans la maîtrise de leur coût moyen ont été créés.

Formulation du problème

La classification des experts par niveau de performance qui semble la plus évidente est de comparer la capacité de chacun d'entre eux à produire des évaluations du coût de la remise en état les plus économiques. Soit calculer le rang de chaque expert à partir de la variable $\mathbb{E}[CM|Expert]$, la valeur du coût moyen évalué par chacun des n experts.

Par la nature des sinistres automobile, la valeur de $\mathbb{E}[CM|Expert]$ est influencée par des facteurs exogènes à l'expert. Dans l'optique de classification de performance basée sur le coût moyen, ces éléments extérieurs ne peuvent lui être imputés . Les travaux de cette étude visent à construire une autre classification basée sur une variable non affectée de ces facteurs exogènes. Ce qui conduit à vouloir supprimer l'effet des variables exogènes sur les résultats de l'expert afin de pouvoir les comparer entre eux.

On considère que CM se décompose en deux parties : $CM = Z + P$, avec Z la variable aléatoire correspondant aux effets exogènes et P la variable aléatoire correspondant au niveau de performance des experts $\mathbb{E}[CM|Z|Expert]$. On va chercher à évaluer $P = CM - Z$.

Modélisation

Dans le cas présent la distribution de la variable `CM` suit une loi `Gamma`. Un modèle de type GLM avec une fonction de lien $g = log$ est retenu pour modéliser l'influence des variables sur la variable d'intérêt. Il s'exprime sous la forme :

$$g(\mathbb{E}[CM]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k \tag{1}$$

$$\mathbb{E}[CM] = exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k). \tag{2}$$

Cette modélisation, donne, pour chaque modalité X_i, le coefficient β_i de la relation entre la modalité et la variable `CM`. La suppression des effets des différentes variables exogènes, est calculée pour chaque expert, par la différence $\Delta Z_i = Z_i - \overline{Z}$, et $\Delta Z_i \beta_i$, qui donne l'impact de la variable i sur le coût moyen. Le calcul d'une nouvelle variable de coût qui prend en compte l'ensemble des variables s'exprime ainsi :

$$\mathbb{E}[CM|Z] = \frac{\mathbb{E}[CM]}{\prod_{i=1}^{k} \exp(\Delta Z_i \beta_i)}. \tag{3}$$

Ce qui revient à chercher la valeur d'un intercept qui correspond à la valeur moyenne du niveau de performance des experts.

Quatre modèles GLM sont réalisés et leurs résultats comparés :

- version « standard » ;
- modification des coefficients β_i avec l'ajout d'une pénalisation de type $LASSO$;
- ajout d'interactions entre variables au modèle pénalisé ;
- réduction du nombre de variables du modèle précédent.

Réduction de variance

La suppression de l'effet des variables retenues dans le modèle sur le coût moyen de l'expert engendre une réduction de la variance de la variable `CM`. Afin de juger de la pertinence du modèle, la chute de variance est décomposée en parties exogène et endogène. A partir des relations suivantes :

$$\mathbb{V}(CM|Expert) = \mathbb{E}[\mathbb{V}(CM|Z|Expert)] + \mathbb{V}(\mathbb{E}[CM|Z|Expert]) \tag{4}$$
$$\mathbb{V}(CM|Expert) = \mathbb{V}(Z|Expert) + \mathbb{V}(\mathbb{E}[CM|Z|Expert]) + 2\,\mathrm{Cov}(\mathbb{E}[CM|Z|Expert], Z|Expert), \tag{5}$$

des indicateurs de la qualité d'information dans le modèle sont calculés. $\mathbb{V}(\mathbb{E}[CM|Z|Expert])$ représente la variance inter-classe, qui décrit la variance de la performance des experts.

La part de variance des impacts des effets exogènes (variance intra-classe) :

$$PV_{intra} = \frac{\mathbb{V}(Z|Expert) + 2\,\mathrm{Cov}(\mathbb{E}[CM|Z|Expert], Z|Expert))}{\mathbb{V}(CM|Expert)}, \tag{6}$$

et la part d'information pertinente dans la variance intra-classe :

$$PIP = \frac{\mathbb{V}(Z|Expert)}{\mathbb{V}(Z|Expert) + 2\,\mathrm{Cov}(\mathbb{E}[CM|Z|Expert], Z|Expert))}. \tag{7}$$

Le graphe de la figure 1 montre les densités des distributions de coûts moyen avant et après épuration des effets exogènes. La table 1 donne les éléments de comparaison entre les distribution : le modèle `interaction_reduit` est celui qui affiche les meilleurs résultats.

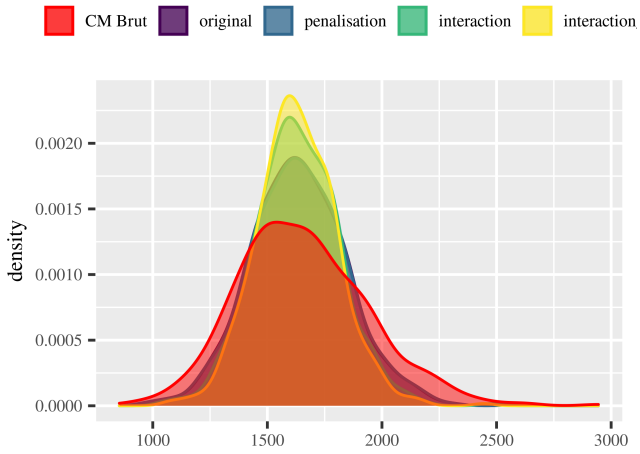


FIGURE 1 – Distribution du CM brut et des variables issues des différentes modèles

TABLE 1 – Comparaison de la qualité de l'information entre modèles

Indicateur	interaction	interaction_reduit	original	penalisation
Var_Inter	34738.06	33994.01	44832.37	45054.13
Var_Intra	47293.66	48037.72	37199.35	36977.59
PV_intra	0.58	0.59	0.45	0.45
PIP	0.32	0.32	0.25	0.25

Analyse d'impact

Quantifier l'impact des facteurs exogènes permet de donner des explications quant au niveau de classification d'un expert. Cela représente aussi des informations pertinentes pour l'assureur pour comprendre sa structure de coût.

Pour chaque modalité son impact se calcule de la sorte :

$$Impact_i = \sum_{m=1}^{n} \text{abs}(\frac{CM_m}{\exp(\Delta Z_{i,m}\beta_i)}), \tag{8}$$

avec $m_1...m_n$ l'indice de l'expert.

Le graphe de la figure 2 montre les niveaux de contribution des variables exogènes. On note que l'ajout d'interaction dans le modèle modifie significativement les niveaux d'impact des modalités

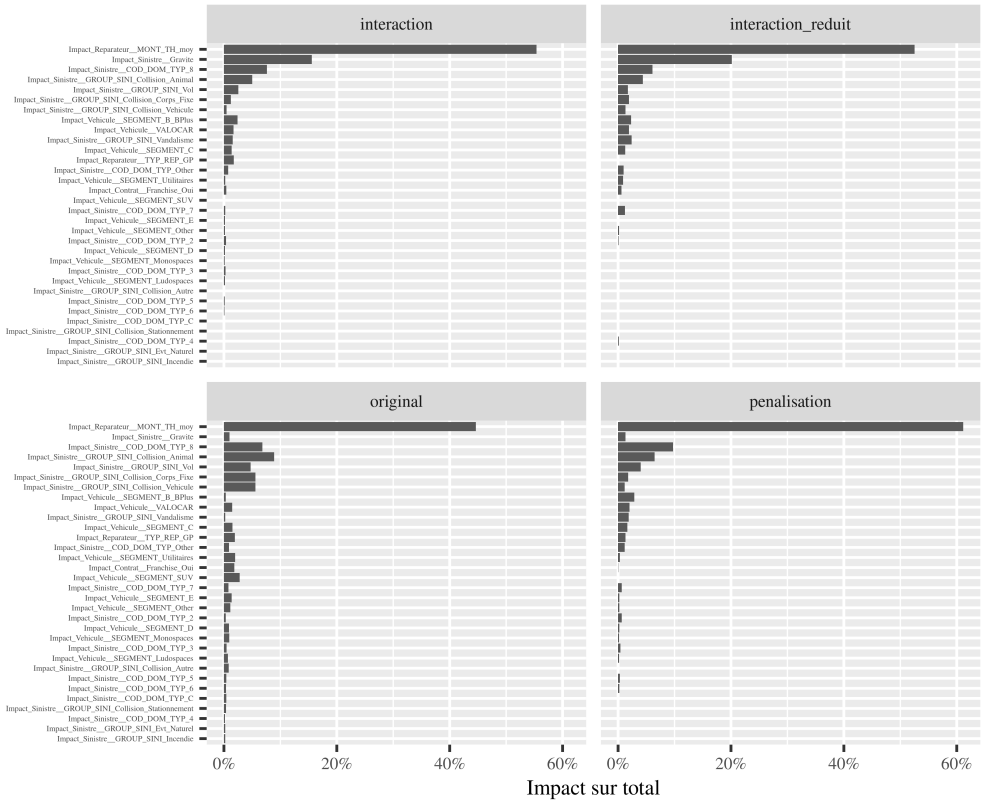


FIGURE 2 – Comparaison de l'impact des modalités entre les différents modèles

Classification des experts

Sur la base des comparaisons effectuées entre les modèles, le modèle `interaction_reduit` est le candidat le plus pertinent pour effectuer la classification des experts.

La classification des experts correspond à la statistique d'ordre de la variable $\mathbb{E}[CM|Expert]$. A partir des nouvelles variables de coûts, de nouvelles classifications sont effectuées. Les deux graphes de la

figure 3 comparent le classement initial fondé sur la valeur brute de `CM` avec les classements obtenus à partir des variables de coût issues des différents modèles.

La réduction de variance ne se résume pas à une contraction de la distribution, ce qui gage d'une classification plus pertinente. Sa forme est plus centrée, ce qui signifie que des situations extrêmes sont corrigées. Cela se traduit par des changements de position des experts dans le classement : dans le modèle `original` 45.3% des experts voient leur position augmenter en moyenne de 23.6 positions et 52.1% des experts leur position baisser en moyenne de 29.3 positions. Le modèle `interaction_reduit` est plus discriminant : les modifications apportées au classement brut sont plus nombreuses et d'une amplitude plus importante comme le montre la table 2.

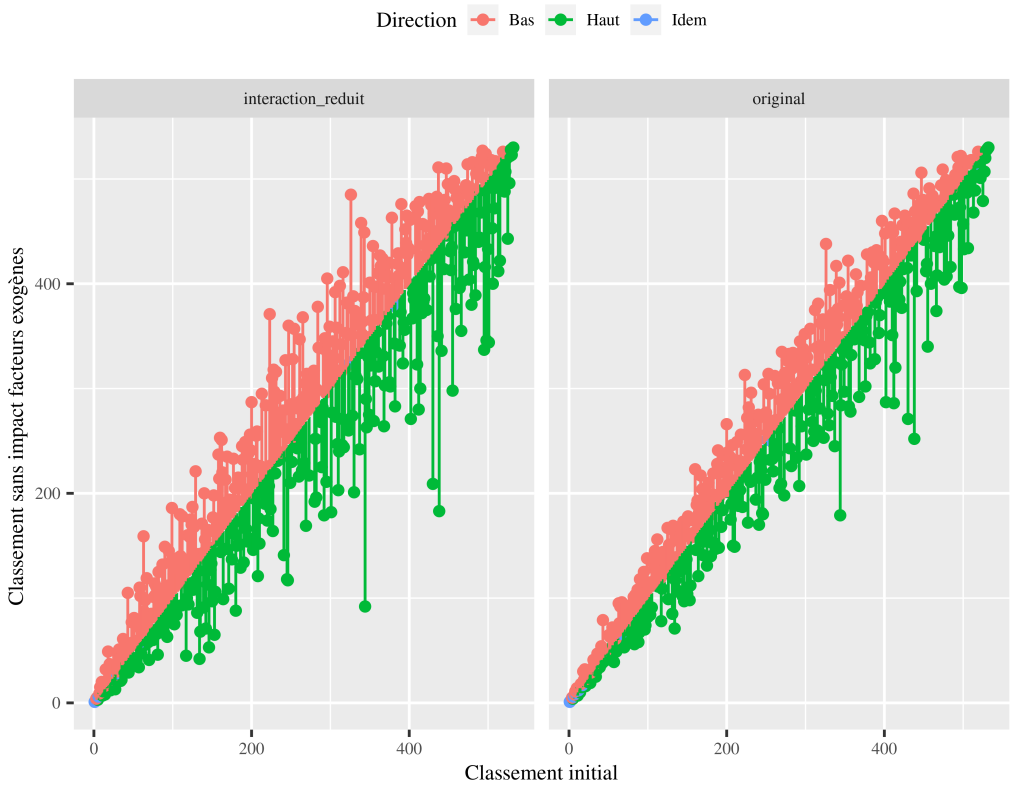


FIGURE 3 – Comparaison des classifications basée sur le CM brut et les classifications sans facteurs exogènes

Statistique	interaction_reduit	original
Part_experts_surclasses	0.46	0.45
Surclasse_max	151	111
Surclasse_moy	29.00	23.60
Part_experts_declasses	0.52	0.52
Declasse_max	-213	-187
Declasse_moy	-35.00	-29.30
Part_experts_stables	0.01	0.03

TABLE 2 – Comparaison des changement entre la classification basée sur le CM brut et des classifications sans facteurs exogènes

Conclusion

A partir des coefficients d'un modèle GLM, l'impact des variables exogènes sur le coût moyen de chaque expert est supprimé. La chute de variance constatée est utilisée comme indicateur de pertinence du modèle. La chute de variance obtenue est supérieure à 95% ce qui signifie que les variables utilisées possèdent un bon pouvoir explicatif. Les variables exogènes, quant à elles, créent une chute de variance de 57% ce qui gage d'une quantité significative d'information pertinente pour notre démarche.

Plusieurs modèles ont été réalisés. Face à la présence de dépendance entre certaines variables endogènes, une pénalisation de type Lasso a été appliquée afin de réduire le nombre de variables utilisées. Des interactions ont été ajoutées dans le modèle afin de modéliser la dépendance entre variables. La performance du modèle est largement améliorée dans le cadre de l'utilisation d'interaction et d'un nombre réduit de variables endogènes.

Ces travaux permettent d'obtenir une classification des experts plus pertinente, avec une influence minimisée des facteurs exogènes. On obtient ainsi une appréciation plus fine de la performance des experts. La modélisation choisie permet aussi de donner une vue analytique en donnant les niveaux d'impacts des différentes variables et de leurs modalités sur le coût moyen de l'expert.

Des améliorations peuvent être apportées à la modélisation. Sachant que le GLM inclut dans l'intercept la première des modalités d'une variable, son impact n'est pas calculé et n'est pas déduit du coût moyen. La prise en compte de ces modalités améliorerait la suppression des effets exogènes.

Étendre le domaine d'étude à plusieurs années apporterait plus d'information, et notamment lisserait l'impact de l'expert sur le coût moyen visant à réduire la dépendance entre facteurs exogènes et endogènes. La prise en compte de données d'environnement de l'expert, telles que la productivité, les secteurs géographiques, le niveau de formation et d'ancienneté d'exercice, donnerait des clefs de lecture intéressante pour les management des experts.

Enfin, ces travaux donnent une vue statique en étudiant la performance sur une période temporelle fixe. Une modélisation des facteurs d'inflation de coûts serait une suite logique des ces travaux et répondrait à une forte demande des assureurs et dirigeants amenés à piloter des experts en automobile de plus impliqués dans la maîtrise de la charge sinistre des compagnies d'assurance.

Synthesis note

In non-life insurance, the market for frequency risk coverage products is competitive, particularly in automobile insurance. Many players compete for market share with growth dynamics that can be very different. The intensity of competition is such, that insurers are struggling to generate margins on these activities. Claims management under the indemnity principle is complex and resource-intensive. The low interest rate environment is forcing insurers to improve their technical results.

In the management of non-life claims, the insurer seeks the intervention of a loss adjuster who acts as a cost control agent on behalf of the insurer. The insurer selects and animates its networks of loss adjusters to ensure that its forecasts will be respected. For this, the insurer must have metrics allowing it to measure the performance of loss adjusters. However, the loss adjuster intervening on claims whose severity is by nature random, the measurement of cost control is a delicate operation.

Previous works have addressed the issue of the performance of networks of automotive loss adjusters. The approach was to develop models to compare networks of loss adjusters, i.e. sets of loss adjusters. Based on the data from the loss adjuster reports, de Lignaud de Lussac (2018) uses a regression model in which a variable indicating the membership network is introduced. The influence of this variable on the claim cost is measured using the value of the coefficient associated with the variable in GLM modeling. Subsequently, Khougea (2019) and Wabo Foka (2020) addressed the same issue by comparing the predictive power of different machine learning models on the same dataset.

The approach carried out through this work has a twofold objective:

— provide methods for classifying individuals from a population of adjusters involved in automotive claims: this meets the need for selection;

— identify the explanatory variables of performance: this meets the need for animation.

The approach adopted differs from previous work in that it does not aim to compare the results of different populations of loss adjusters but individually loss adjusters with each other.

Data

The dataset used includes 506607 observations and 50 variables. This is the set of loss adjuster reports carried out by 732 loss adjusters over a year. For many of the variables of the treatments were necessary: deletion or imputation of missing values, grouping of modalities, clipping. After all the treatments and transformations on the variables the dataset includes 468363 observations, a loss of -7.5%, which remains acceptable for the study. Performance indicators that will be used to incentive loss adjusters in controlling their average cost have been created.

Problem formulation

The classification of loss adjusters by performance level that seems most obvious is to compare the ability of each of them to produce the most economical discount cost estimates. Thus, calculate the rank of each loss adjuster from the variable $\mathbb{E}[CM|Expert]$, the value of the average cost evaluated by each of the n loss adjusters.

By the nature of automobile claims, the value of $\mathbb{E}[CM|Expert]$ is influenced by factors exogenous to the loss adjuster. From the point of view of performance classification based on average cost, these external elements cannot be imputed to it. The work of this study aims to construct another classification based on an unaffected variable of these exogenous factors. This leads to wanting to remove the effect of exogenous variables on the loss adjuster's results in order to be able to compare them with each other.

We consider that CM is divided into two parts: $CM = Z+P$, with Z the random variable corresponding to the exogenous effects and P the random variable corresponding to the performance level of the loss adjusters $\mathbb{E}[CM|Z|Expert]$. We will try to evaluate $P = CM - Z$.

Modeling

In this case the distribution of the variable `CM` follows a law `Gamma`. A GLM model with a link function $g = log$ is used to model the influence of variables on the variable of interest. It is expressed in the as followed :

$$g(\mathbb{E}[CM]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k \tag{9}$$

$$\mathbb{E}[CM] = exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k). \tag{10}$$

This modeling gives, for each modality X_i, the coefficient β_i of the relationship between the modality and the variable `CM`. The removal of the effects of the different exogenous variables, is calculated for each loss adjuster, by the difference $\Delta Z_i = Z_i - \overline{Z}$, and $\Delta Z_i \beta_i$, which gives the impact of the variable i on the average cost. The calculation of a new cost variable that takes into account all the variables is expressed as follows:

$$\mathbb{E}[CM|Z] = \frac{\mathbb{E}[CM]}{\prod_{i=1}^{k} \exp(\Delta Z_i \beta_i)}. \tag{11}$$

This amounts to looking for the value of an intercept that corresponds to the average value of the performance level of the loss adjusters.

Four GLM models are produced and their results compared:
- version "standard";
- modification of the coefficients β_i with the addition of a penalty type $LASSO$;
- adding interactions between variables to the penalized model;
- reduction in the number of variables in the previous model.

Variance reduction

Removing the effect of the variables retained in the model on the average cost of the loss adjuster results in a reduction in the variance of the variable `CM`. In order to judge the relevance of the model, the

variance drop is broken down into exogenous and endogenous parts. From the following relationships:

$$\mathbb{V}(CM|Expert) = \mathbb{E}[\mathbb{V}(CM|Z|Expert)] + \mathbb{V}(\mathbb{E}[CM|Z|Expert]) \quad (12)$$
$$\mathbb{V}(CM|Expert) = \mathbb{V}(Z|Expert) + \mathbb{V}(\mathbb{E}[CM|Z|Expert]) + 2\,\mathrm{Cov}(\mathbb{E}[CM|Z|Expert], Z|Expert), \quad (13)$$

indicators of the quality of information in the model are calculated. $\mathbb{V}(\mathbb{E}[CM|Z|Expert])$ represents the inter-class variance, which describes the variance of the loss adjusters' performance.

The variance share of the impacts of exogenous effects (intra-class variance):

$$PV_{intra} = \frac{\mathbb{V}(Z|Expert) + 2\,\mathrm{Cov}(\mathbb{E}[CM|Z|Expert], Z|Expert))}{\mathbb{V}(CM|Expert)}, \quad (14)$$

and the share of relevant information in the intra-class variance:

$$PIP = \frac{\mathbb{V}(Z|Expert)}{\mathbb{V}(Z|Expert) + 2\,\mathrm{Cov}(\mathbb{E}[CM|Z|Expert], Z|Expert))}. \quad (15)$$

The graph in the figure 4 shows the densities of the average cost distributions before and after cleaning up exogenous effects. The table 3 gives the elements of comparison between the distributions: the model `interaction_reduit` is the one that displays the best results.

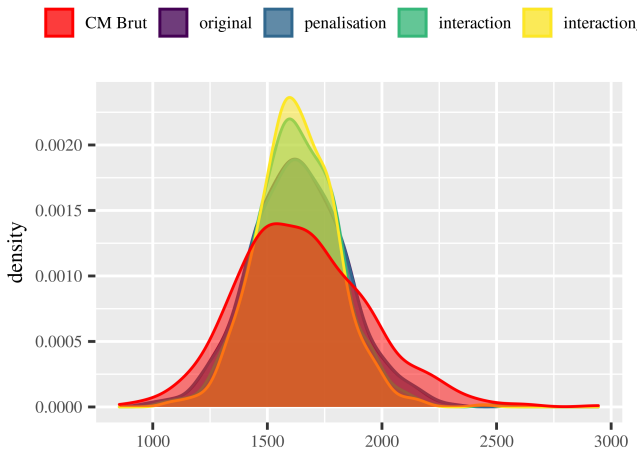


Figure 4 – Distribution of the raw CM and variables from the different models

Table 3 – Comparison of information quality between models

Indicator	interaction	interaction_reduit	original	penalization
Var_Inter	34738.06	33994.01	44832.37	45054.13
Var_Intra	47293.66	48037.72	37199.35	36977.59
PV_intra	0.58	0.59	0.45	0.45
PIP	0.32	0.32	0.25	0.25

Impact assessment

Quantifying the impact of exogenous factors makes it possible to give explanations when an loss adjuster's classification level. It also represents crucial information for the insurer in understanding its cost structure.

For each modality its impact is calculated as follows:

$$Impact_i = \sum_{m=1}^{n} \text{abs}(\frac{CM_m}{\exp(\Delta Z_{i,m}\beta_i)}), \tag{16}$$

with $m_1...m_n$ the loss adjuster's index.

The graph in figure 5 shows the contribution levels of exogenous variables. It is noted that the addition of interaction in the model significantly modifies the impact levels of the modalities

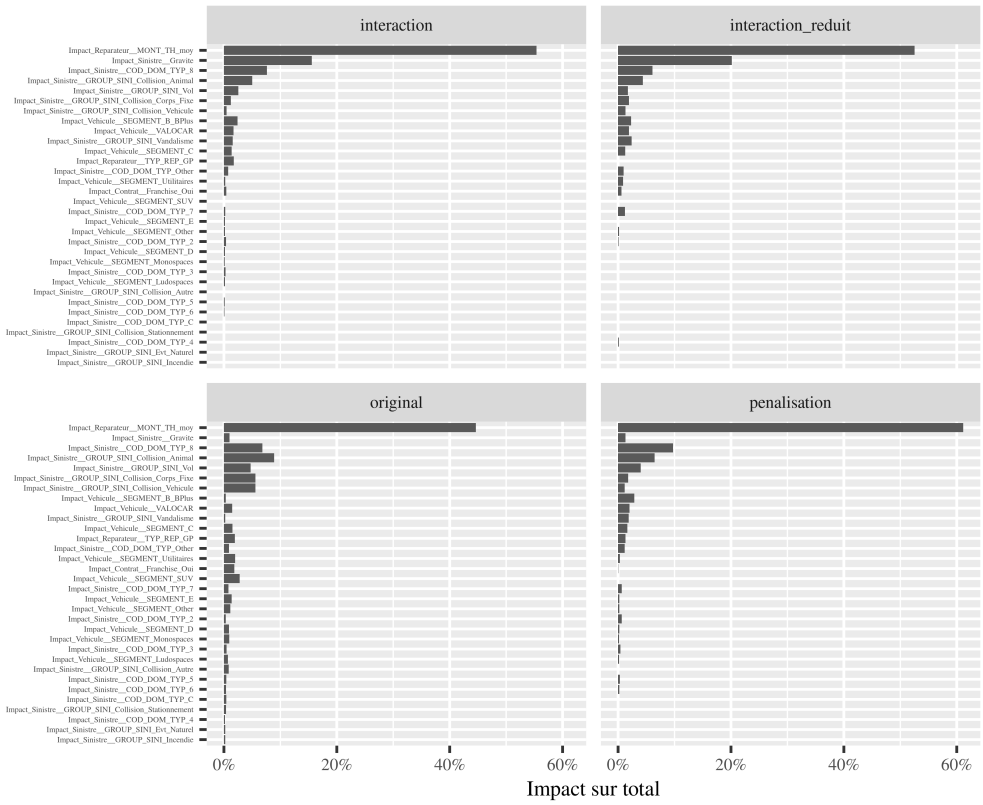


Figure 5 – Comparison of the impact of modalities between different models

Classification of loss adjusters

Based on the comparisons made between the models, the model `interaction_reduit` is the most relevant candidate for performing the classification of loss adjusters.

The classification of the loss adjusters corresponds to the order statistics of the variable $\mathbb{E}[CM|Expert]$. From the new cost variables, new classifications are made. The two graphs in figure 6 compare the initial ranking based on the gross value of `CM` with the rankings obtained from the cost variables from the different models.

The reduction in variance is not limited to a contraction of the distribution, which guarantees a more relevant classification. Its shape is more centered, which means that extreme situations are corrected. This results in changes in the position of loss adjusters in the ranking: in the model `original` 45.3% of loss adjusters see their position increase on average by 23.6 positions and 52.1% of loss adjusters their

position decrease on average by 29.3 positions. The `interaction_reduit` model is more discriminating: the changes made to the raw collation are more numerous and of a greater amplitude as shown in the table 4.

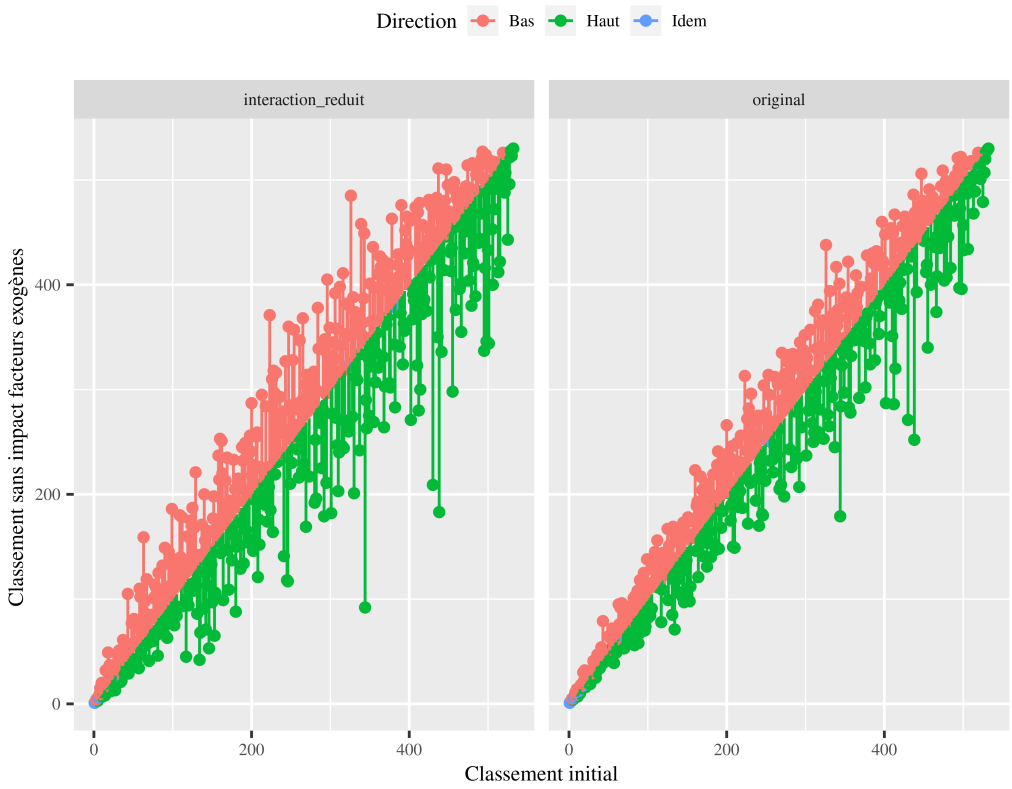


Figure 6 – Comparison of classifications based on raw CM and classifications without exogenous factors

Statistics	interaction_reduit	original
Share_loss adjusters_surclasses	0.46	0.45
Upgrade_max	151	111
Outperform_moy	29.00	23.60
Share_loss adjusters_declasses	0.52	0.52
Declass_max	-213	-187
Downgrade_moy	-35.00	-29.30
Share_loss adjusters_stables	0.01	0.03

Table 4 – Comparison of changes between raw CM-based classification and classifications without exogenous factors

Conclusion

From the coefficients of a GLM model, the impact of exogenous variables on the average cost of each loss adjuster is removed. The observed variance drop is used as an indicator of the relevance of the model. The variance drop obtained is greater than 95% which means that the variables used have a good explanatory power. Exogenous variables create a drop in variance of 57% which guarantees a significant amount of information relevant to our approach.

Several models have been made. Faced with the presence of dependence between certain endogenous

variables, a Lasso-type penalization was applied in order to reduce the number of variables used. Interactions were added to the model to model the dependency between variables. The performance of the model is greatly improved through the use of interaction and a reduced number of endogenous variables.

This work makes it possible to obtain a more relevant classification of loss adjusters, with a minimized influence of exogenous factors. This gives a finer assessment of the performance of the loss adjusters. The model chosen also makes it possible to give an analytical view by giving the levels of impact of the different variables and their modalities on the average cost of the loss adjuster.

Improvements can be made to modeling. Knowing that the GLM includes in the intercept the first of the modalities of a variable, its impact is not calculated and is not deducted from the average cost. Taking these modalities into account would improve the elimination of exogenous effects.

Extending the field of study to several years would provide more information, and in particular would smooth out the impact of the loss adjuster on the average cost of reducing the dependence between exogenous and endogenous factors. Taking into account the loss adjuster's environmental data, such as productivity, geographical sectors, level of training and seniority, will give interesting reading keys for the management of loss adjusters.

Finally, these works give a static view by studying performance over a fixed time period. A modeling of cost inflation factors would be a logical continuation of this work and will respond to a strong demand from insurers and executives who are required to lead automotive loss adjusters who are more involved in controlling the claims burden of insurance companies.

Remerciements

Je tiens à remercier sincèrement Frédéric Planchet pour avoir accepté d'encadrer ces travaux, la pertinence de ses remarques et sa grande disponibilité lors de la rédaction de ce mémoire.

Je ne me serai jamais lancé dans cette aventure sans mon passage à l'Ecole Nationale de l'Assurance, l'ENASS, où les professeurs et intervenants m'ont fait découvrir et donné goût à l'actuariat. Je pense notamment à Gilles Beneplanc, Philippe Picagne et Edith Bocquaire qui m'ont encouragé dans cette voie.

Je témoigne ma reconnaissance à Jean Prevost et Antoine Jove qui m'ont permis de suivre ce cursus dans de bonnes conditions et dans un cadre d'application opérationnel.

Un grand merci à mes collègues de promotion, et surout Gabriel, Sylvain, Kevin, Benjamin Yoann, Alice pour nos nombreuses séances de travail, acharnées, motivantes et sympathiques !

Je remercie l'ensemble du corps de professoral de Sorbonne Université et de l'Institut du Risk Management pour la qualité des cours et leur investissement dans l'enseignement qu'ils dispensent.

Toutes les personnes qui m'ont fait part de leur leurs avis et critiques sur mes travaux m'ont beaucoup apporté. Je les en remercie.

Je suis infiniment redevable à Camille pour son soutien et son aide précieuse à la rédaction de ce mémoire.

Je dédie ce mémoire à mes deux fils Lino et Maceo qui m'ont soutenu tout au long de cette aventure.

Table des matières

Introduction

Environnement

En assurance non-vie, le marché des produits de couverture des risques de fréquence est concurrentiel et particulièrement en assurance automobile. La figure 7 montre le positionnement des vingt premiers assureurs automobile du marché français. On constate que de nombreux acteurs se disputent les parts de marché avec des dynamiques de croissance qui peuvent être très différentes. L'indice de Herfindahl-Hirschmann *, calculé sur ces données, indique un marché à caractère concurrentiel.

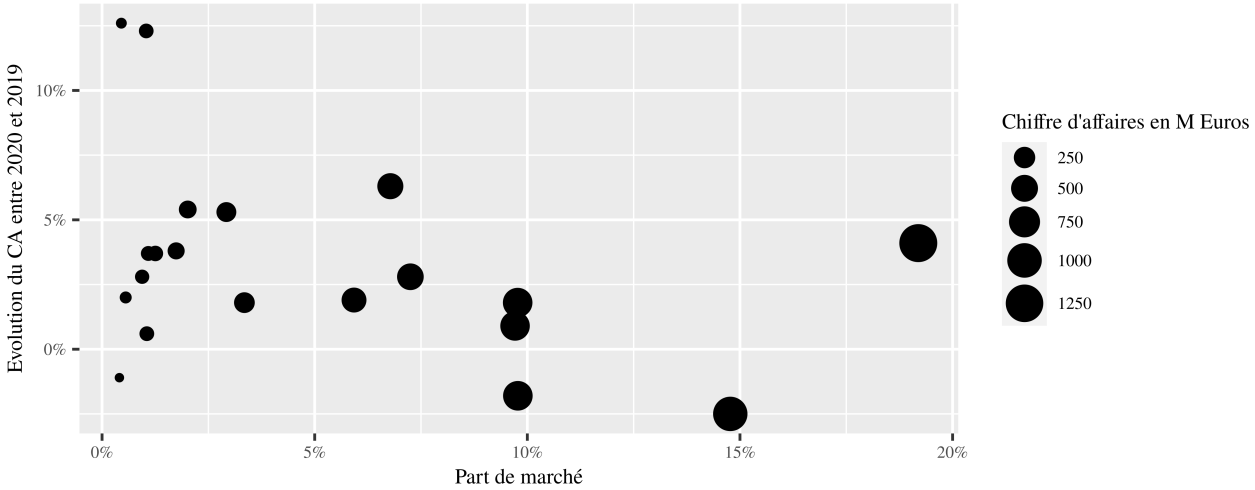


FIGURE 7 – Marché des assureurs auto en 2020

L'intensité de la concurrence est telle que les assureurs peinent à dégager des marges sur ces activités. Cette compétition entre assureurs s'est accrue avec l'arrivée de la loi Hamon en 2015, qui permet aux assurés de changer plus facilement d'assureur. De 2011 à 2019, le ratio combiné comptable de la branche automobile n'a jamais été en dessous de 100 †. Souvent considérés comme produits d'appels, ces contrats représentent une part importante de l'activité de l'assureur : près d'un tiers du chiffre d'affaires en 2021 d'après Abadie & Chabrier (2021) – détails en annexe.

Pour maintenir un niveau de prime concurrentiel, la maîtrise de la charge sinistre est essentielle. La gestion des sinistres en application du principe indemnitaire est complexe et consommatrice de ressources. Le contexte de taux bas contraint les assureurs à améliorer leurs résultats techniques. Dans le même temps, le développement de la digitalisation et de l'intelligence artificielle promet de nouvelles

∗. IHH normalisé vaut 0.056 et $\frac{1}{n} = 0.05$, https://fr.wikipedia.org/wiki/Indice_de_Herfindahl-Hirschmann

†. ratio après réassurance, source : France Assureurs https://www.franceassureurs.fr/nos-chiffres-cles/donnees-globales/les-donnees-cles-de-lassurance-francaise

façon de gérer les sinistres : automatisées, rapides et à faible coût.

Problématique métier

La gestion des sinistres dommages répond à une politique d'indemnisation. L'une des caractéristiques majeure de cette politique, est la stratégie d'expertise mise en place. L'expert mandaté par l'assureur est un agent de la maîtrise des coûts pour le compte de l'assureur. En 2014, l'assureur Allianz annonçait : *« A travers la réorganisation de son réseau [d'experts en automobiles], Allianz France entend continuer à réduire le coût moyen des sinistres »* Durand (2014). Autant le recours à un expert pour évaluer les dommages peut être imposé par la réglementation ou le cadre conventionnel, autant l'assureur possède une latitude pour y recourir ou non dans bon nombre de situations.

Ainsi la définition d'une politique d'indemnisation amène le décideur à se poser la question du recours ou non à l'expert pour des situations données. Dans ce cadre, la définition de la stratégie d'expertise optimale, dans le sens où elle minimise le couple charge sinistre – charge honoraires, consiste essentiellement à quantifier le niveau de maîtrise des coûts apporté par l'intervention de l'expert ; la charge honoraires, étant généralement connue de façon certaine.

Une fois la politique d'indemnisation établie, l'assureur sélectionne et anime ses réseaux d'experts pour s'assurer que ses prévisions seront respectées. Pour cela l'assureur doit disposer de métriques lui permettant de mesurer la performance des experts. Or l'expert intervenant sur des sinistres dont la gravité est par nature aléatoire, la mesure de la maîtrise du coût est une opération délicate.

Travaux antérieurs

La question de la définition de la politique d'indemnisation a été abordée dans le cadre de sinistres habitation, avec l'utilisation de méthodes stochastiques pour déterminer la méthode d'intervention : indemnisation en nature, gré à gré ou expertise Ebali (2011).

De récents travaux ont été menés sur la question de la performance des réseaux d'experts en automobile. L'approche a été de développer des modèles visant à comparer des réseaux d'experts, c'est à dire des ensembles d'expert. Sur la base des données des rapports d'expertise, de Lignaud de Lussac (2018) utilise un modèle de régression dans le lequel est introduit une variable indiquant le réseau d'appartenance. L'influence de cette variable sur le coût de sinistre est mesurée à l'aide de la valeur du coefficient associé à la variable dans le cadre d'une modélisation de type GLM. Par la suite, Khougea (2019) et Wabo Foka (2020) ont traité la même problématique en comparant le pouvoir prédictif de différents modèles d'apprentissage automatique sur le même jeu de données.

Objectifs du mémoire

Le périmètre d'étude sera celui des sinistres matériels automobiles. La population étudiée sera celle des experts en automobile, lorsqu'ils interviennent dans l'évaluation du coût de la remise en état de véhicules dans l'optique de leur réparation suite à un sinistre couvert par une garantie d'assurance.

A travers ce mémoire, nous proposons d'étudier la modélisation de la performance des experts. L'objectif est double :

- fournir des méthodes de classification des individus d'une population d'experts intervenant dans le cadre de sinistres automobiles : cela répond au besoin de sélection ;
- identifier les variables explicatives de la performance : cela répond au besoin d'animation.

L'approche retenue se différencie des travaux précédents dans le sens, où elle ne vise pas à comparer les résultats de différentes populations d'experts mais individuellement des experts entre eux. Nous comptons ainsi fournir des outils et des clefs de lecture quant à la mise en oeuvre d'une stratégie d'expertise efficiente.

Chapitre 1

Contexte et cadre d'étude

1.1 Dispositions contractuelles et principe indemnitaire

L'indemnisation des sinistres automobiles matériels répond au principe indemnitaire. Contrairement à d'autres typologies de risque, l'appréciation *à priori* du montant de l'indemnité n'est guère possible car il pèse de fortes inconnues à la fois sur la gravité des sinistres et le coût de la remise en état. Ainsi les assureurs recourent au service d'un expert en automobile pour évaluer le coût de la remise en état. Dans certaines situations, l'intervention de l'expert est même nécessaire afin de présenter un recours entre assureurs (notamment dans le cadre de la convention d'Indemnisation directe de l'assuré et de Recours entre Sociétés d'assurance Automobile - IRSA). Cette difficulté d'estimation du coût du sinistre conduit les assureurs à proposer une indemnisation des sinistres automobiles qui réponde au principe indemnitaire. L'expression contractuelle du montant de l'indemnité est la somme d'argent permettant de replacer l'assuré dans la même situation antérieure au sinistre.

On considère deux cas dans l'indemnisation de sinistres automobiles :
- le sinistre « total » ;
- le sinistre « partiel ».

Le sinistre est considéré « total »dès que le montant de la remise est supérieur à la valeur du véhicule avant le sinistre. Dans cette situation et en vertu du principe indemnitaire, le montant de l'indemnité ne peut excéder la valeur du bien assuré. Pour un sinistre dit « partiel », l'indemnité est égale au coût de la remise en état du véhicule dès lors que celui-ci est réparé. Ainsi, le principe indemnitaire définit *de facto* un plafond d'indemnisation aux sinistres automobiles.

On peut classer les garanties automobiles en deux grandes catégories :
- les assurance de responsabilité ;
- les assurances de dommages.

L'assurance de responsabilité civile automobile est obligatoire. Elle prévoit l'indemnisation des victimes d'accident de la circulation, dont les dommages matériels subis par le véhicule. En complément, les assureurs proposent des garanties « dommages tout accidents », qui couvre les dommages subis par le véhicule, quelque soit la faute commise par le conducteur. En cas de sinistre responsable, l'assureur applique généralement une franchise à la charge de l'assuré.

1.2 Le rôle de l'expert en automobile

La profession d'expert en automobile est réglementée Legifrance (2001). Son exercice requiert d'une part l'acquisition du diplôme d'expert en automobile délivré par le ministère de l'Éducation nationale,

et d'autre part être inscrit sur la liste nationale des experts en automobile.

L'expert est mandaté par un assureur susceptible de mettre en jeu une garantie d'un contrat d'assurance. Dans ce cadre, son rôle consiste a :

1. Identifier le véhicule ;
2. Procéder au relevé des dommages des et à leur imputation ;
3. Définir une méthodologie de réparation ;
4. Chiffrer le coût de la remise en état

L'exécution de ces opérations est réalisée systématiquement par l'expert : d'une part cela fait partie de la mission que l'assureur lui adresse, d'autre part le compte rendu de celles-ci est obligatoire dans les rapports standardisés Namin (2009). L'examen de ces opérations d'expertise permet de comprendre l'impact de l'expert sur le coût du sinistre.

Pour chacune d'entre elles, nous notons l'influence de la qualité du travail de l'expert :

1. une erreur d'identification peut conduire à l'application de tarifs de pièces de rechanges inadaptés ou à une erreur dans la détermination de la valeur du véhicule ;
2. un relevé erroné et/ou une mauvaise appréciation de l'imputation des dommages à la déclaration de sinistre peut conduire à faire prendre en charge des dommages qui ne sont pas couverts au titre de la garantie mise en jeu ;
3. quand plusieurs méthodologies de réparation conduisent à un résultat identique, une comparaison de leurs coûts doit déterminer le choix de la méthodologie retenue ;
4. sa capacité intrinsèque à chiffrer le plus justement le coût de la remise en état.

1.3 Le modèle de coût

L'expert consigne ses conclusions dans son rapport. Celles-ci se veulent techniques et indépendantes des dispositions contractuelles dont l'application relève de la gestion du sinistre par l'assureur. Le montant d'expertise inscrit sur le rapport correspond au coût de la remise en état des dommages imputables à la garantie mise en jeu au titre de la déclaration établie par l'assuré. En terme de coût de sinistre, il y a donc là une différence importante entre celui de l'assureur et celui de l'expert. L'expert évalue un montant qui correspond au débours pour replacer l'assuré dans la situation antérieure au sinistre : coût de la remise en état si le véhicule est réparé ; montant de la valeur de remplacement en cas de sinistre total. Le coût du sinistre pour l'assureur correspond au débours pour indemniser l'assuré au regard des dispositions contractuelles. Ainsi l'assureur, en appliquant le contrat d'assurance, peut être amené à appliquer un découvert obligatoire (ex : franchise) au montant déterminé par l'expert.

En appliquant le principe indemnitaire, avec R le coût de la remise en état et V_R la valeur de remplacement, le montant d'expertise X s'exprime de la sorte :

$$X = \begin{cases} R, \text{ si } R < V_R, \\ V_R, \text{ si } R \geq V_R, \end{cases} \tag{1.1}$$

$$= \min\{R, V_R\}. \tag{1.2}$$

L'indemnité Y versée à l'assuré peut être minorée d'une franchise * suivant les conditions contractuelles.

∗. Précisément nous considérons ici une franchise dite « déduite ».

Considérant le montant de la franchise F, le montant de l'indemnité vaut :

$$Y = \begin{cases} 0, \text{ si } X \leq F, \\ X - F, \text{ si } X > F, \end{cases} \tag{1.3}$$

$$= \max\{0, X - F\}. \tag{1.4}$$

La relation 1.4 a des conséquences importantes sur l'analyse des différentes variables. Nous pouvons considérer deux cas du point de vue de l'assureur :

- L'assuré estime par ses propres moyens que le montant X sera inférieur à la franchise F et il ne déclare pas le sinistre. Il s'agit d'une troncature des données ; l'assureur n'a pas eu connaissance de l'existence du sinistre.
- L'assuré a déclaré son sinistre et les conclusions de l'expert amènent l'assureur à constater que $X \leq F$; il ne délivrera pas d'indemnité. On parlera alors de censure des données.

Ainsi, la variable X est soumise au même phénomène de troncature décrit ci-dessus mais n'est pas censurée par l'application d'une franchise. Il en résulte une plus grande quantité d'information contenue dans X que dans Y.

Si l'on s'intéresse à la variable R, la relation 1.2 montre également que cette variable est censurée dès lors que nous sommes dans le cas du sinistre « total »(i.e. $R \geq V_R$). Car même si c'est l'expert qui détermine le coût de la remise en état, il y a une grande différence entre une estimation sur dégâts apparents R_{min} (qui peut sous-estimer la valeur réelle de R ayant $R_{min} < R$, mais suffit à classer le sinistre en perte totale en constatant que $R_{min} \geq V_R$) et un chiffrage R résultant de la remise en état effective du véhicule, incluant donc la totalité de l'information du coût.

1.4 Anatomie du coût de réparation

Le coût de remise en état R est la somme de plusieurs postes de coûts. Dans le cas de la réparation d'une automobile accidentée, on distingue trois postes principaux, ayant chacun des poids différents * :

- les pièces de rechange (50.9%), noté P_r ;
- la main d'oeuvre (38.8%), noté M_o ;
- les ingrédients de peinture (10.3%), noté I_p.

Cette répartition résulte de l'analyse des ensembles des coûts de réparation sur une année.

De façon détaillée, nous avons $R = P_r + M_o + I_p$, avec :

$$P_r = \text{Nb Pièces} \times \text{tarif pièces}, \tag{1.5}$$

$$M_o = \text{Nb Heures main d'oeuvre} \times \text{tarif horaire de main d'oeuvre} \tag{1.6}$$

$$I_p = \text{Nb Heures peinture} \times \text{tarif horaire de peinture} \tag{1.7}$$

Les données de ces composantes des postes de coût sont facilement accessibles car l'expert les consigne dans son rapport. Ce niveau de détail est nécessaire pour pouvoir apprécier les différents facteurs d'influence montrés en figure 1.1.

1.5 La relation avec le réparateur : intérêts divergents

L'expert et le réparateur poursuivent des objectifs qui peuvent diverger du point de vue économique selon l'assureur. L'expert cherchera établir le coût de la remise en fonction de critères concurrentiels.

∗. Statistique pour l'année 2020 fournie par SRA (Sécurité et Réparation Automobiles)Colas et al. (2021) , organisme professionnel dont la vocation est : *« promouvoir, au sein de la profession et avec les acteurs de l'automobile, toutes études et de mettre en œuvre tous moyens utiles à la réalisation des actions pouvant contribuer à la limitation du nombre et du coût des sinistres dans l'intérêt des assurés »*.

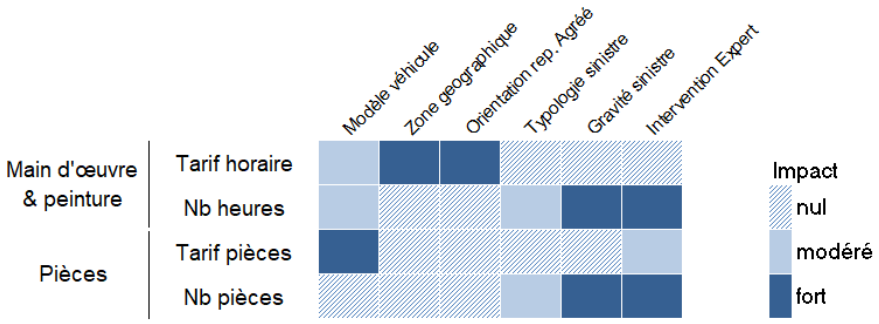


FIGURE 1.1 – Facteurs d'influence du coût de la remise en état.

Cela consiste à retenir, pour une prestation équivalente, celle étant la mieux disante d'un point de vue économique. Il remplit ainsi sa mission de régulateur des coûts, dictée par son devoir de conseil technique auprès de son mandant. Le réparateur recherche a tirer vers le haut des postes de facturation, répondant ainsi à une logique de maximisation de profit.

En pratique cette démarche de sélection d'un réparateur a rarement lieu, l'assuré ayant le libre de choix du réparateur. Cela se traduit par la réalisation de l'expertise contradictoirement entre l'expert et le réparateur. Dans cette négociation, l'expert argumente techniquement et peut faire valoir la concurrence locale pour asseoir ses arguments *. Le processus contradictoire débouche *in fine* sur une prise d'accord entre l'expert et le réparateur.

La maîtrise du coût de la remise en état dans ce cadre revêt un caractère asymétrique, tel que montré en figure 1.2. En effet, il y a plus à perdre qu'à gagner car le rapport entre intensité de la négociation et le niveau de facturation du réparateur n'est pas linéaire. En ce sens, on considère d'une part que le réparateur ne facture pas en dessous de son seuil de rentabilité ; que d'autre part, en l'absence totale de négociation – du moindre contrôle, le réparateur augmentera significativement le montant de sa facturation.

1.6 Le modèle de mesure de la performance

Depuis des années, les assureurs pilotent les experts en mesurant l'évolution de son coût moyen d'un année sur l'autre. Un expert est considéré « performant »lorsque l'évolution de son coût moyen est inférieure au niveau d'évolution fixé par l'assureur. L'expression de l'évolution du coût moyen est la suivante :

$$Evol = \frac{\frac{1}{N_v}\sum_{j=1}^{N_v} R_j}{\frac{1}{N_u}\sum_{i=1}^{N_u} R_i} - 1, \tag{1.8}$$

avec, N_u et R_i respectivement le nombre et le montant des expertises de l'année u, et N_v et R_j respectivement le nombre et le montant des expertises de l'année v, considérant que l'année u précède l'année v. Cet indicateur a l'avantage d'être aligné avec l'objectif de l'assureur dans sa recherche d'optimisation du ratio $\frac{Sinistre}{Prime}$. Considérant qu'il existe un certain nombre de facteurs d'influence du coût moyen, la question se pose quant au niveau d'impact de l'expert sur cette évolution. Il est donc

*. *« que si le réparateur fixe librement ses prix, il appartient à l'expert de se prononcer sur le tarif horaire applicable sans être tenu d'entériner les devis et factures présentés par le réparateur, et que, lorsque l'expertise a lieu dans un garage non agréé, il [l'expert] peut, pour faire jouer la concurrence, se baser sur les prix publics pratiqués par les professionnels voisins »*. Décision de la Cour de cassation du 2 février 2017, à propos des rapports entre les réparateurs et les experts en automobile. Namin (2017)

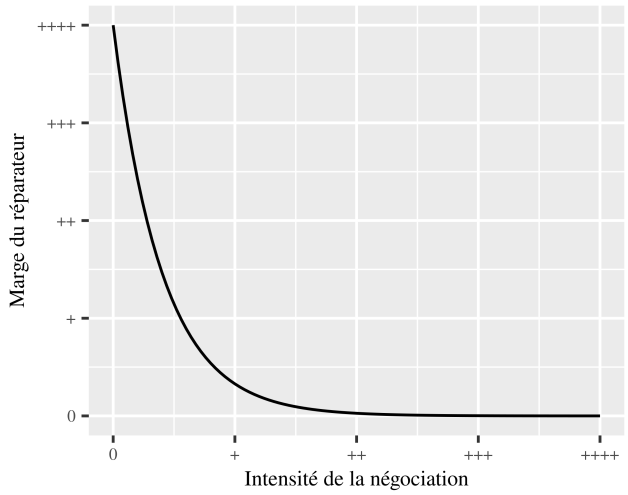


FIGURE 1.2 – Relation entre l'intensité de la négociation et le niveau de facturation du réparateur.

nécessaire de distinguer dans les composantes du coût moyen, celles qui relèvent de facteurs exogènes à l'expert et celles qui incombent à l'action propre de l'expert (i.e. endogènes). Celui-ci est aussi soumis aux évolutions dans le temps des tarifs de pièces de rechange et du tarif horaire de main d'oeuvre.

La fixation du niveau d'objectif *Obj* est une tâche délicate. A la fin de l'année u, l'assureur doit estimer l'impact des facteurs exogènes pour l'année v qui s'annonce. On définit une probabilité de réussite, comme suit :

$$\mathbb{P}(\text{Réussite}) = \mathbb{P}(\text{Evol} < \text{Obj}). \tag{1.9}$$

La fixation d'objectif permet d'aligner les intérêts entre l'assureur et l'expert. Le niveau de l'objectif détermine la quantité de ressources mobilisées pour l'atteindre. Il y a deux situations où l'expert ne mettra pas d'énergie pour réaliser l'objectif : dès lors qu'il a la certitude que l'objectif est inatteignable, ou alors qu'il est déjà acquis. Une caractéristique désirable d'un niveau d'objectif est d'éviter ces deux situations, que l'on peut formuler de la sorte :

$$\exists \text{ Obj t.q. } \mathbb{P}(\text{Réussite}) \ll 1 \wedge \mathbb{P}(\text{Réussite}) \gg 0. \tag{1.10}$$

La figure 1.3 illustre cette modélisation. Tout l'enjeu des travaux consistera à estimer l'impact sur le coût moyen des facteurs exogènes, afin que le risque porte uniquement sur l'expert et ainsi fixer le niveau d'objectif désiré. Aussi, l'évaluation de la performance d'un expert se fera sur son atteinte de l'objectif épuré des facteurs exogènes.

1.7 Synthèse

L'expert en automobile est un agent de la maîtrise des coûts. Il s'assure de l'imputabilité des dommages constatés à la déclaration de sinistre. En réalisant l'expertise contradictoirement avec le réparateur, il limite l'inflation des postes de coûts de la remise en état. La lecture directe du coût moyen de l'expert pose des difficultés car nous faisons face à données censurées avec des variables d'influence exogènes à l'expert. La quantité d'information contenue dans les données de l'expertise est supérieure à celle dont peut disposer l'assureur.

L'objectivation de cette performance est réalisée par l'observation de l'évolution de son coût moyen d'une année sur l'autre. Cet approche s'aligne avec les prévisions de sinistralité de l'assureur. Nous

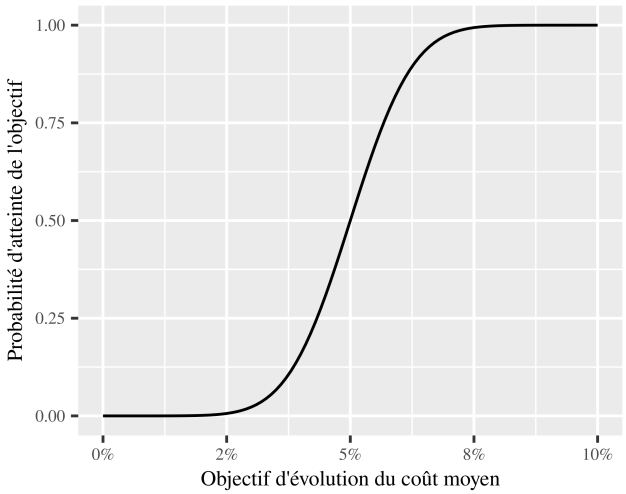


FIGURE 1.3 – Modèle de fixation de l'objectif d'évolution du coût. Valeurs à titre d'illustration.

proposons ici d'évaluer l'impact des facteurs exogènes sur le coût moyen, afin de pouvoir, *in fine*, obtenir une niveaux d'évolutions objective.

La finalité étant d'évoluer d'un univers d'incertitude vers un univers de risque. En ce sens, le passage de l'incertitude au risque, n'enlève pas le caractère aléatoire des phénomènes sous-jacents de l'évolution du coût moyen mais prétend à en connaître les lois qui les gouvernent.

Chapitre 2

Traitement des données

2.1 Description du jeu de données

Le jeu de données comprend 506607 observations et 50 variables [†]. Il s'agit de l'ensemble des expertises où le véhicule est réparé pour une année. Le choix de se limiter à une année pleine évite les problématiques liées à la saisonalité et les inflations de prix d'une année à l'autre.

Les variables sont classées en différentes thématiques :

- Variable dépendante
- Facteurs exogènes :
  - `Sinistre` : 7 variables
  - `Véhicule` : 10 variables
  - `Reparateur` : 6 variables
  - `Contrat` : 2 variables
- Facteurs endogènes :
  - `Pièces` : 4 variables
  - `Cout main d'oeuvre` : 2 variables
  - `Cout peinture` : 1 variable
  - `Cout pièces` : 1 variable
  - `Temps de main d'oeuvre` : 6 variables
  - `Chiffrage` : 3 variables

Dans les sections suivantes, nous allons décrire les différentes variables et les traitement qui leur seront appliqués. L'ensemble des traitements et calculs sont réalisés avec le logiciel open source R [‡] avec un usage intensif la librairie tidyverse développée par Wickham et al. (2019).

2.1.1 Variable dépendante : `CoutTotal__MONT_EXPERT_HT`

Le montant d'expertise est le résultat des investigations de l'expert. C'est ce montant qui est utilisé par l'assureur pour procéder au règlement du sinistre. Les statistiques d'expertise reposent sur cette variable, dont l'évolution est très observée car elle influe le coût des sinistres de l'assureur.

L'allure de la distribution cumulée représentée figure 2.1a montre que 99% des expertises ont un

†. Données issues de l'activité de Bca Expertise, www.bca.fr

‡. {R Core Team} (2022). R : A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

montant < 7500€. La figure 2.1b montre le nuage de points des observations où l'on constate que des valeurs extrêmes se détachant nettement au dessus d'un seuil de 15000 €. On relève que 1% des dossiers les plus chers pèsent 6.7% du total des coûts et 5% des dossiers pèsent 20.3% du total.

Il est intéressant de rechercher la loi approchant la distribution empirique de la variable, afin de définir la modélisation la plus adaptée.

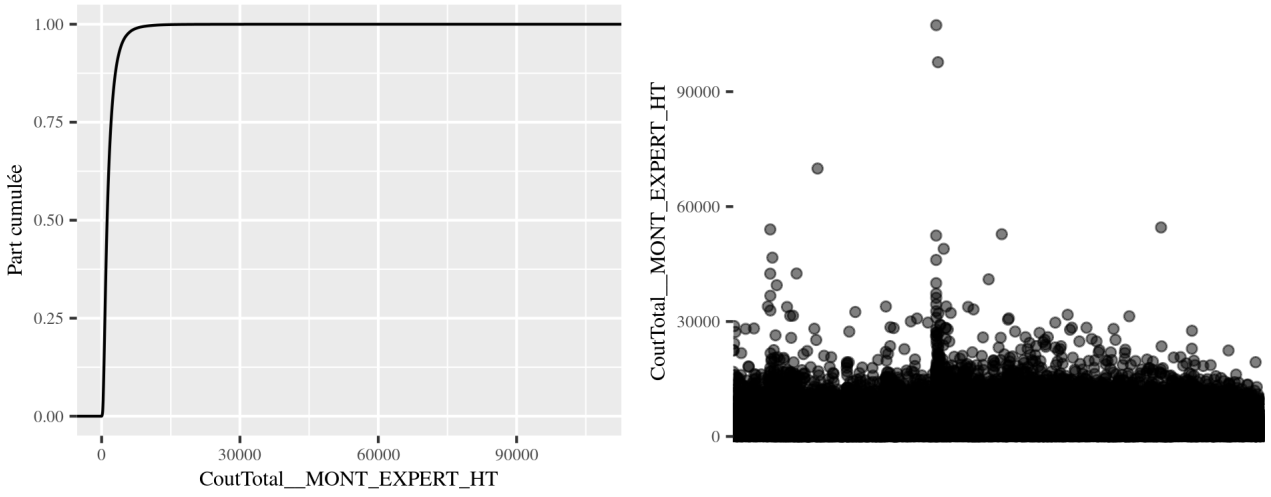
Nous traçons le graphe de Cullen et Frey sur la figure 2.2a. Celui ci positionne la variable sur un plan défini par le coefficient d'asymétrie (*skewness*) et le coefficient d'aplatissement (*kurtosis*). On peut ainsi voir les lois de distribution plausibles avec ce jeu de paramètres.

L'examen du graphe montre qu'il n'y a pas de loi de distribution candidate. Il apparaît alors nécessaire d'étudier l'influence de la queue de distribution. La démarche est d'exclure les montants d'expertises considérés comme extrêmes (ou importants). Pour cela nous utilisons un `mean excess plot`. A la lecture du graphe figure 2.3, nous choisissons un seuil d'écretage à 15000 €. Cette démarche amène à exclure 528 dossiers (0.1% du total), qui pèsent 1.15% du coût total des sinistres.

Avec le jeu de données réduit aux expertises inférieures au seuil de 15000€, le graphe de Cullen et Frey figure 2.2b nous indique que la distribution est proche d'une loi gamma. Ce qui est cohérent avec les modélisations utilisées en tarification automobile, où le coût des sinistres est souvent modélisé par une loi de distribution Gamma (Denuit & Charpentier (2005), Gourieroux (1999)).

D'un point de vue opérationnel, exclure les dossiers supérieurs à 15000€ n'est pas incompatible avec la démarche que nous menons :

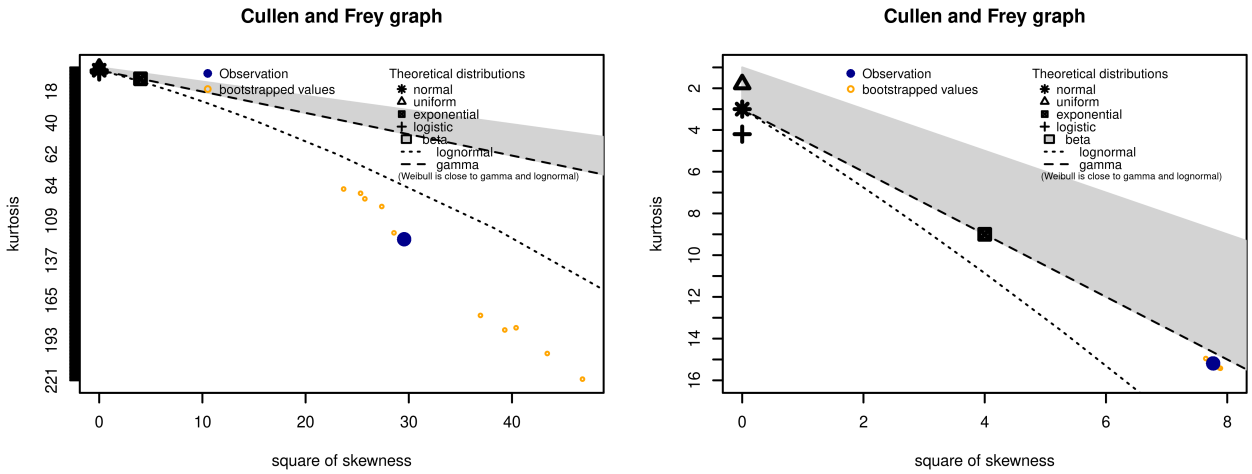
— Nous cherchons à caractériser le comportement de l'expert. Or la typologie de sinistres "graves" n'est pas représentative de l'activité courante de l'expert.
— Il existe des processus de contrôle et de validation pour les dossiers d'un montant élevé, les mettant ainsi sous contrôle.



(a) Fonction de répartition empirique

(b) Valeurs de la variable

FIGURE 2.1 – Variable `CoutTotal__MONT_EXPERT_HT`



(a) Sur l'ensemble des données (b) Après écretage à 15K€

FIGURE 2.2 – Graphe de Cullen et Frey de la variable `CoutTotal__MONT_EXPERT_HT`

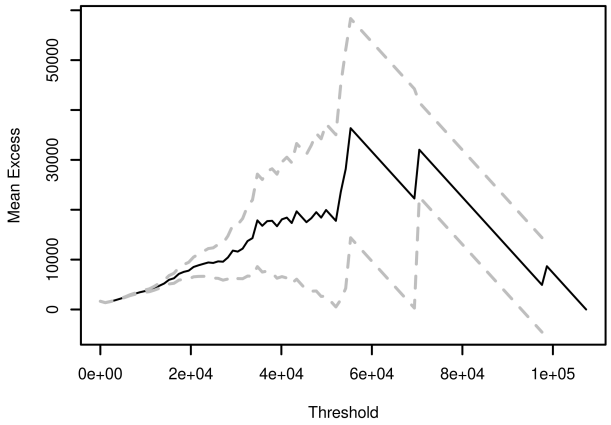


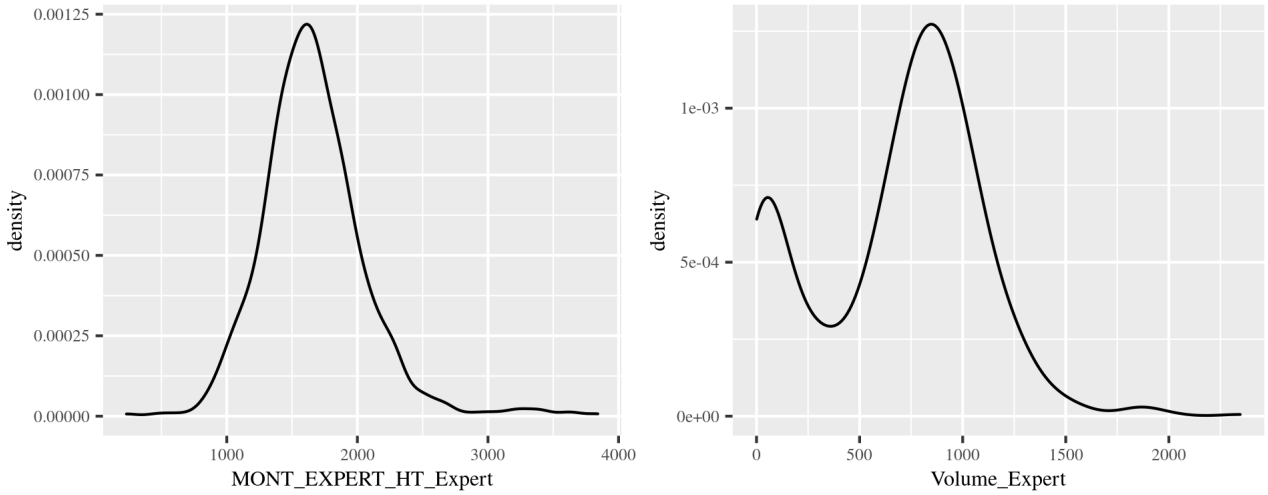
FIGURE 2.3 – Mean Excess Plot de la variable `CoutTotal__MONT_EXPERT_HT`

Agrégat par experts

Les résultats globaux sont l'agrégat des résultats de l'ensemble des experts. Pour autant que l'entreprise est jugée sur ces résultats globaux, le pilotage de l'activité s'effectue au niveau de l'expert. Cela pose des problème de stabilité dans les indicateurs, car le volume d'expertise par expert est faible par rapport à l'activité globale de l'entreprise, augmentant par là leur variance. La distribution des évolutions par experts représentée figure 2.4a a l'allure d'une distribution normale, avec une longue queue à droite. Ce qui n'est pas étonnant quand on connaît l'asymétrie de la maîtrise du coût : il y a beaucoup moins de freins à une augmentation qu'une réduction du coût d'expertise.

Il y a 733 experts différents dans le jeu de données. Le volume d'expertises réalisées annuellement par expert varie car il y a des variations dans les effectifs liées aux entrées et sorties de personnels. L'examen de la distribution des volumes par expert figure 2.4b montre une proportion significative d'experts avec de faibles volumes. Ceci s'explique par le turn-over conduisant à des experts arrivants ou partant en cours d'année, qui ne peut que réduire le volume d'expertise par expert.

Afin que les résultats soient exploitables, nous choisissons de ne retenir dans l'étude que les experts dont le volume annuel est supérieur à 500 expertises. Ce qui nous amène à retenir 532 experts.



(a) `CoutTotal__MONT_EXPERT_HT` par expert (b) Volume d'expertises par expert

FIGURE 2.4 – Distributions d'agrégats par expert

2.2 Facteurs exogènes

2.2.1 Variables liées au sinistre

`Sinistre__CP_MISSION`

Il s'agit du code postal où a lieu l'expertise du véhicule. La variable possède 5340 valeurs distinctes, ce qui est cohérent avec le référentiel des codes postaux.

`Sinistre__Gravite`

Il s'agit d'un score calculé pour chaque expertise qui détermine le niveau de gravité en se basant sur le nombre, la nature et les position des pièces de rechanges impactées. C'est une variable numérique continue corrélée positivement avec la variable `CoutTotal__MONT_EXPERT_HT` comme le montre la

figure 2.5.

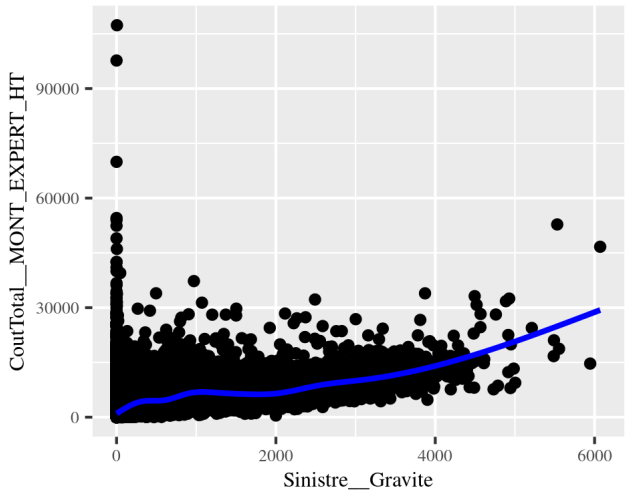


FIGURE 2.5 – Relation entre la gravité et le coût de l'expertise

38588 des valeurs (7.6 %) sont absentes. On constate dans la table 2.1, que cette proportion de valeurs absentes est répartie de façon quasi uniforme entre les différents points de chocs. Nous choisissons de remplacer les valeurs absentes par la valeur moyenne de la variable.

TABLE 2.1 – Valeurs manquantes de la variable `Sinistre__Gravite` par type de dommage

COD_DOM_TYP	Valeurs_Manquantes	mean(Gravite)	Volume	mean(MONT_EXPERT_HT)
1	0.0635870	72.17246	51976	1676.7360
2	0.0630877	93.35261	48916	1760.7163
3	0.0753221	50.99912	41993	1634.3978
4	0.0674226	53.42490	50354	1750.2724
5	0.0760245	39.62667	38093	1246.2062
6	0.0791311	37.42654	34525	1258.8239
7	0.0749677	54.39777	105152	1382.1811
8	0.0650416	96.42874	98783	2023.8529
9	0.1945032	42.49869	473	1938.8219
A	0.2143852	62.23390	2155	1840.7394
B	0.2335329	183.75938	3340	2005.0085
C	0.1416062	119.35707	30606	1886.1597
D	0.0454545	57.75661	198	997.7021
E	1.0000000	NaN	1	9102.4000
Z	0.6666667	88.00000	3	1517.5194
NA	0.5897436	40.00000	39	1667.6521

`Sinistre__GROUP_SINI`

Il s'agit du code de la nature de sinistre couverte par l'assureur. 5 modalités représentent 96.8% des observations. La table 2.2 montre qu'il existe des différences notables de montant d'expertise suivant la nature du sinistre.

`Sinistre__LIB_GAR`

Il s'agit de la nature de sinistre couverte par l'assureur. 20080 des valeurs (4 %) sont absentes. Elles sont assignées à une catégorie `unknown`. 5 modalités représentent 96.7% des observations. La pertinence

TABLE 2.2 – 10 premières natures de sinistres en terme de volume

Sinistre___GROUP_SINI	Volume	mean(CoutTotal___MONT_EXPERT_HT)
Collision_Vehicule	347763	1549.099
Collision_Corps_Fixe	91794	1947.593
Collision_Animal	24179	2015.987
Vandalisme	16842	1705.426
Vol	9846	1854.868
Collision_Autre	7567	1365.524
Autres	4010	1232.751
Collision_Stationnement	3000	1394.447
Evt_Naturel	830	1756.571
Incendie	776	2162.068

du regroupement de certaines modalités est à étudier.

`Sinistre__NUM_DEPART_SECT`

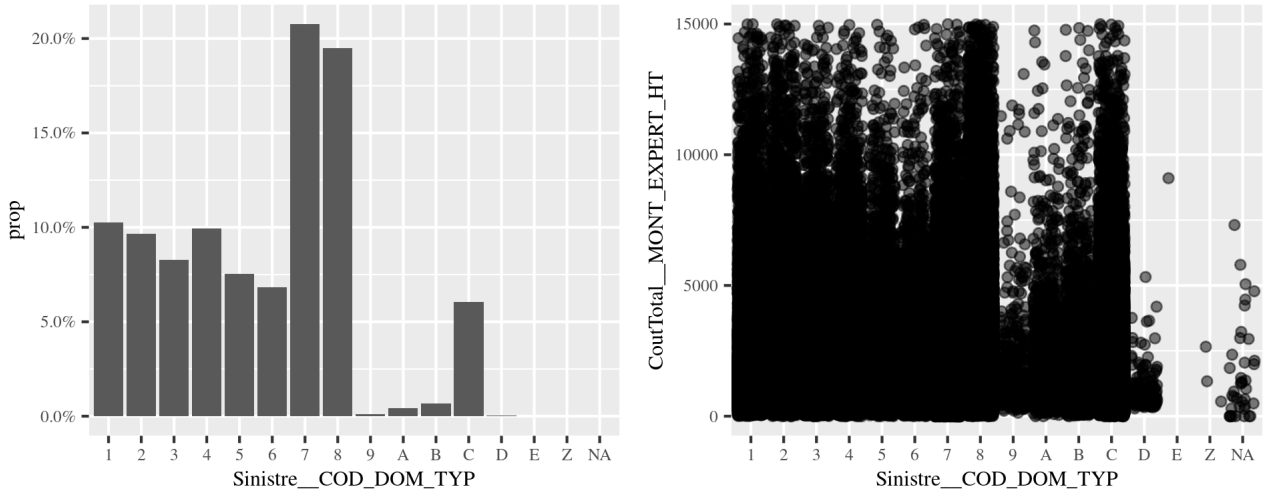
Il s'agit du département où a lieu l'expertise du véhicule. La variable possède 96 valeurs distinctes ce qui est cohérent avec le référentiel des départements. 13 valeurs sont absentes : elles sont supprimées.

`Sinistre__COD_DOM_TYP`

Il s'agit du code du point de choc sur le véhicule constaté par l'expert. La table 2.3 décrit les différentes modalités. Les sinistres avec un choc avant ou arrière sont les plus nombreux comme le montre la figure 2.6a.

Nous constatons également sur la figure 2.6b que le montant de la variable `CoutTotal__MONT_EXPERT_HT` varie fortement d'une modalité à l'autre. Ceci est cohérent, car suivant le point de choc, les conséquences sur le véhicule sont différentes. Typiquement un choc avant (modalité `8`) coûte plus cher à remettre en état à cause de la proximité du moteur et de ses accessoires.

Nous choisissons de regrouper les modalités les moins représentées : `9`, `A`, `B`, `D`, `E` et `Z`, et les valeurs absentes dans une seule modalité. Ce qui représente 6209 individus (1.2 % du total).



(a) Volume d'expertises

(b) Valeurs de la variable `CoutTotal__MONT_EXPERT_HT`

FIGURE 2.6 – Distribution par points de choc

TABLE 2.3 – Codification des points de choc

Cod Dom Typ	Lib Dom Typ
1	à l'avant gauche
2	à l'avant droit
3	sur le côté gauche
4	sur le côté droit
5	à l'arrière gauche
6	à l'arrière droit
7	à l'arrière
8	à l'avant
9	sur le pavillon
A	en partie supérieure
B	sous caisse
C	sur toute la carrosserie
D	sur le pare-brise
E	à l'intérieur
Z	Autre

`Sinistre__VAL_ORIENT_CHOC`

Cette variable indique l'angle du choc en degrés constaté par l'expert. Cela ajoute de l'information complémentaire car pour les modalités `1` à `8` de la variable `Sinistre__COD_DOM_TYP`, nous pouvons avoir plusieurs angles de choc. La figure 2.7 montre la distribution de l'orientation entre les différents points de chocs. On note que certaines orientations de chocs prévalent avec certains points de choc. 39 valeurs sont absentes : elles sont supprimées.

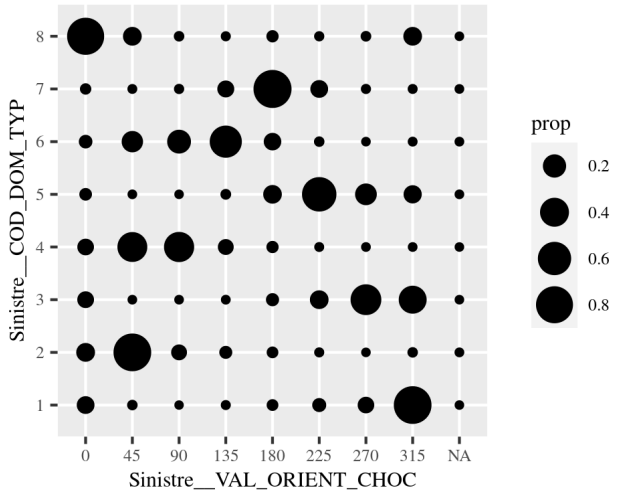


FIGURE 2.7 – Distribution de l'orientation pour chaque point de choc

2.2.2 Variables liées au véhicule

`Vehicule__AGE_MOIS_VEH`

Cette variable numérique représente l'âge du véhicule, calculé par la différence entre la date de première mise en circulation du véhicule et la date d'expertise. On constate sur la figure 2.8 des valeurs incohérentes : la valeur maximale est de 1430 mois, soit un véhicule âgé de 120 ans, ce qui n'est pas cohérent.

Dans bon nombre de cas, la date de première mise en circulation est saisie manuellement, dont les

erreurs peuvent produire des valeurs incohérentes. D'un autre coté on peut aussi rencontrer des véhicules ayant cent ans d'âge ; toute fois cela constitue des cas atypiques et non représentatif de l'activité que nous souhaitons modéliser. On choisit alors d'appliquer un seuil d'écretage à droite à la valeur de 400 mois afin d'exclure ces deux cas de figure. Ce qui représente 435 dossiers (0.1% du total). 247 valeurs sont absentes et sont supprimées.

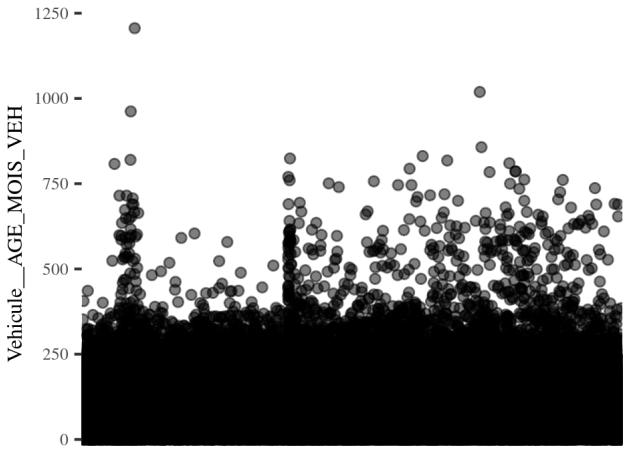


FIGURE 2.8 – Valeurs de la variable `Vehicule__AGE_MOIS_VEH`

`Vehicule__COD_COND_ACHAT`

Les trois modalités précisent les conditions d'achat du véhicule :

- usuelles
- location longue durée
- location avec option d'achat

77 valeurs sont absentes et sont supprimées.

`Vehicule__Cod_Ener`

Cette codification indique l'énergie de la motorisation du véhicule. 2 modalités représentent 97.2% des observations. Les autres modalités sont regroupées. La répartition est représentée sur le graphe figure 2.9. 76 valeurs sont absentes et sont supprimées.

`Vehicule__COD_MARQUE`

Cette codification indique la marque du véhicule. 10 modalités représentent 78.9% des observations.

`Vehicule__COD_MARQUE_MODELE`

Cette codification est une concaténation de la marque et du modèle du véhicule. 50 modalités représentent 45.5% des observations.

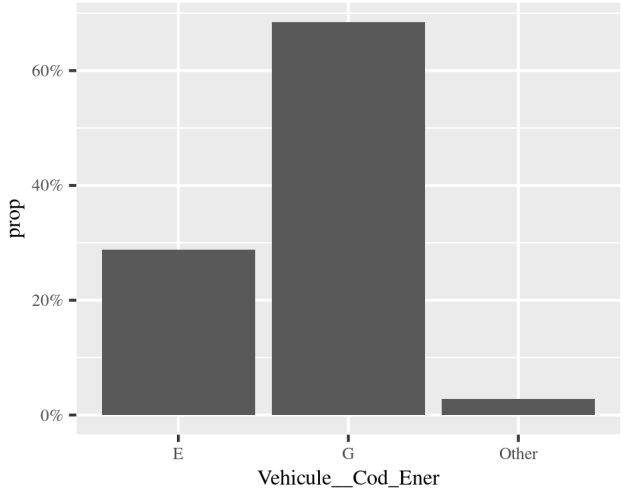


FIGURE 2.9 – Distribution du volume d'expertise par type de motorisation

`Vehicule__SEGMENT`

Cette variable indique le segment du véhicule au sens de la classification Européenne des véhicules *. La variable comporte 13 modalités avec des représentativités très différentes comme le montre le graphe figure 2.10a.

Nous constatons également sur la figure 2.10b que le montant de la variable `CoutTotal__MONT_EXPERT_HT` varie fortement d'une modalité à l'autre. Ceci est cohérent, car l'appartenance à un segment est déterminée par la taille du véhicule, dont la survenance du sinistre a des conséquences plus ou moins importantes en termes de coût de remise en état.

Nous choisissons de regrouper les modalités les moins représentées : `Autres`, `F`, `Coupe_Cabriolet`, `Sportive`, et les valeurs absentes dans une seule modalité. Ce qui représente 12389 individus (2.4 % du total).

`Vehicule__VAL_KM_HORO_COMPTEUR`

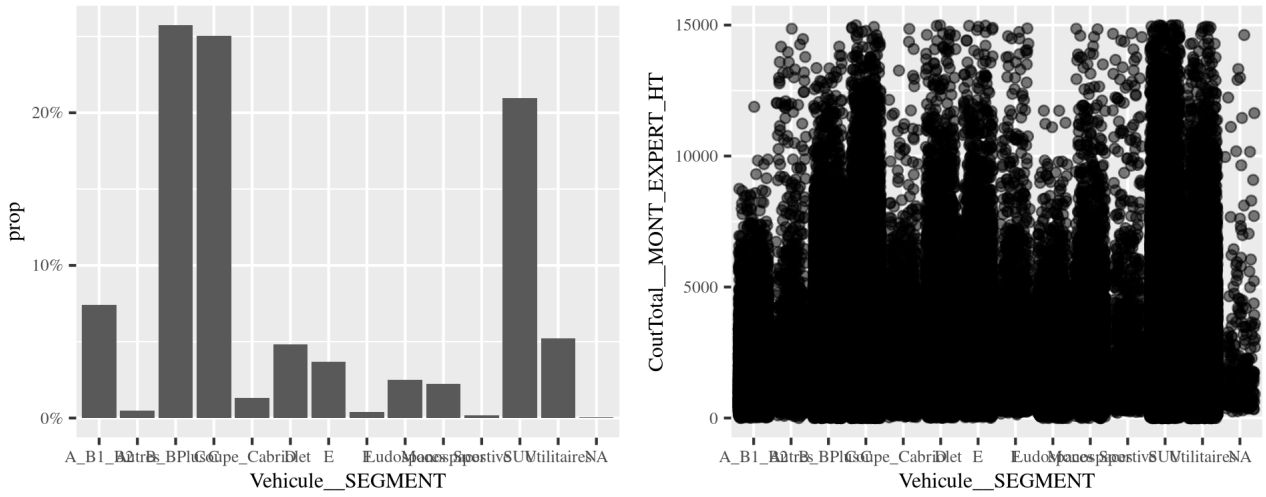
Il s'agit du kilométrage saisi par l'expert lors de l'examen du véhicule. La figure 2.11 montre que cette variable comporte des valeurs supérieures à un million de kilomètres, ce qui n'est sont pas compatible avec le type de véhicule. Il s'agit vraisemblablement des erreurs de saisie. Nous décidons d'exclure les valeurs les plus extrêmes de l'étude.

On choisit alors d'appliquer un seuil d'écretage à droite. Afin de déterminer la valeur de ce seuil, on retient les normes de kilométrage annuels fixés par l'argus automobile †. Pour les véhicules à motorisation diesel (`Vehicule__Cod_Ener` = "G") le kilométrage est de 25000 km par année. Pour les véhicules à motorisation essence (`Vehicule__Cod_Ener` = "E") le kilométrage est de 15000 km par année.

Au regard des proportions de ces types de motorisation dans le jeu de données, nous déterminons un kilométrage annuel moyen de $25000 \times 0.7 + 15000 \times 0.3 = 22000$. Avec le seuil maximum de 400 mois déterminé précédemment pour la variable `Vehicule__AGE_MOIS_VEH`, nous obtenons un seuil d'écretage de $22000 \times \frac{400}{12} = 733333$. Ce qui représente 395 dossiers (0.1% du total).

Il y a 7650 valeurs absentes (1.3 % du total) ; nous choisissons d'imputer la valeur du kilométrage

*. https://fr.wikipedia.org/wiki/Segment_automobile
†. https://www.largus.fr/faq/index.cfm?idTexte=4080



(a) Volume d'expertise

(b) Variable `CoutTotal__MONT_EXPERT_HT`

FIGURE 2.10 – Distribution par segments de véhicule

moyen dans ces cas-là.

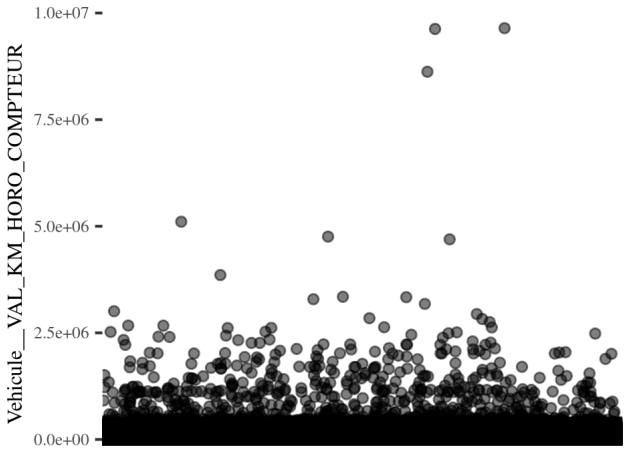


FIGURE 2.11 – Valeurs de la variable `Vehicule__VAL_KM_HORO_COMPTEUR`

`Vehicule__VAL_PUIS_DIN`

Cette variable représente la puissance réelle du véhicule, exprimée en chevaux din. Elle donne une indication de la valeur du véhicule. Elle est obtenue à partir des informations présentes sur la carte grise du véhicule. Il y a 5938 valeurs absentes ou nulles (1.2% du total). Ce qui peut être considéré comme des incohérences, la puissance d'un véhicule ne pouvant être nulle. Nous imputons la valeur moyenne dans ces cas-là.

Les valeurs maximales sont très élevées. Elles ne sont pas représentatives de l'ensemble des véhicules expertiser de l'étude. Nous choisissons d'exclure les valeurs au dessus du seuil de 350 chevaux din. Ce qui représente 1424 dossiers (0.3% du total).

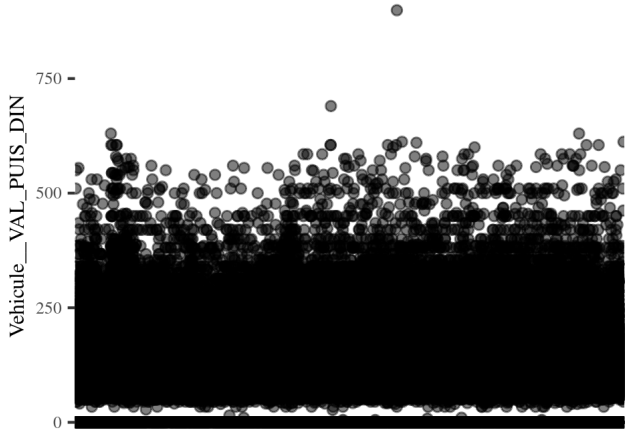


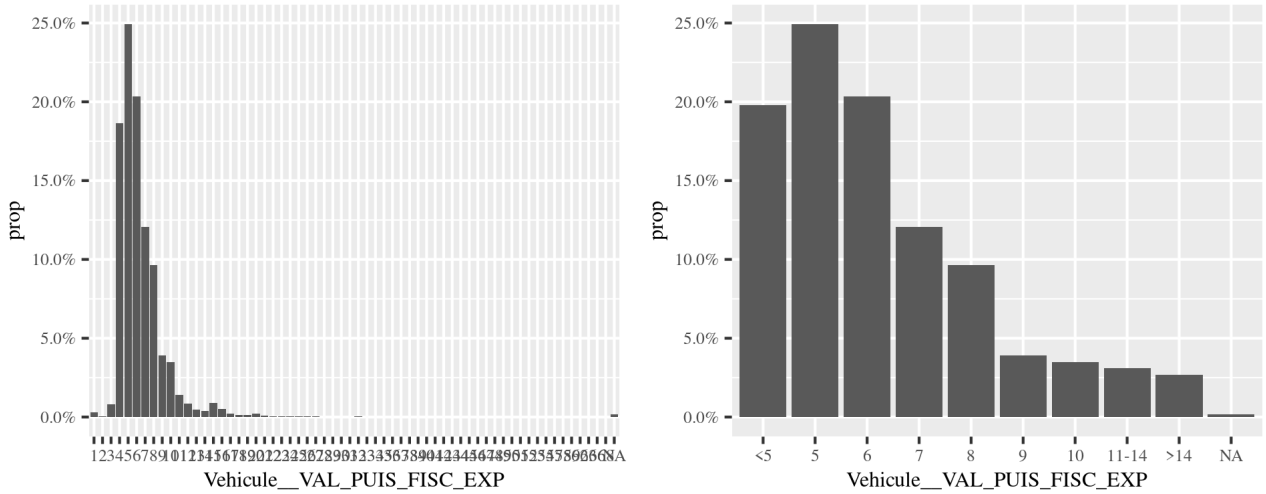
FIGURE 2.12 – Valeurs de la variable `Vehicule__VAL_PUIS_DIN`

`Vehicule__VAL_PUIS_FISC_EXP`

Cette variable représente la puissance fiscale indiquée sur la carte grise. Comme la puissance réelle, il s'agit là encore d'une indication sur la valeur du véhicule. Alors que la variable `Vehicule__VAL_PUIS_FISC_EX` est de type numérique, celle ci est qualitative ordinale.

Nous constatons sur la figure 2.13a qu'un grand nombre de modalités sont très peu représentées. Nous choisissons de regrouper certaines des modalités afin de créer 9 classes homogènes, représentées sur le graphe figure 2.13b.

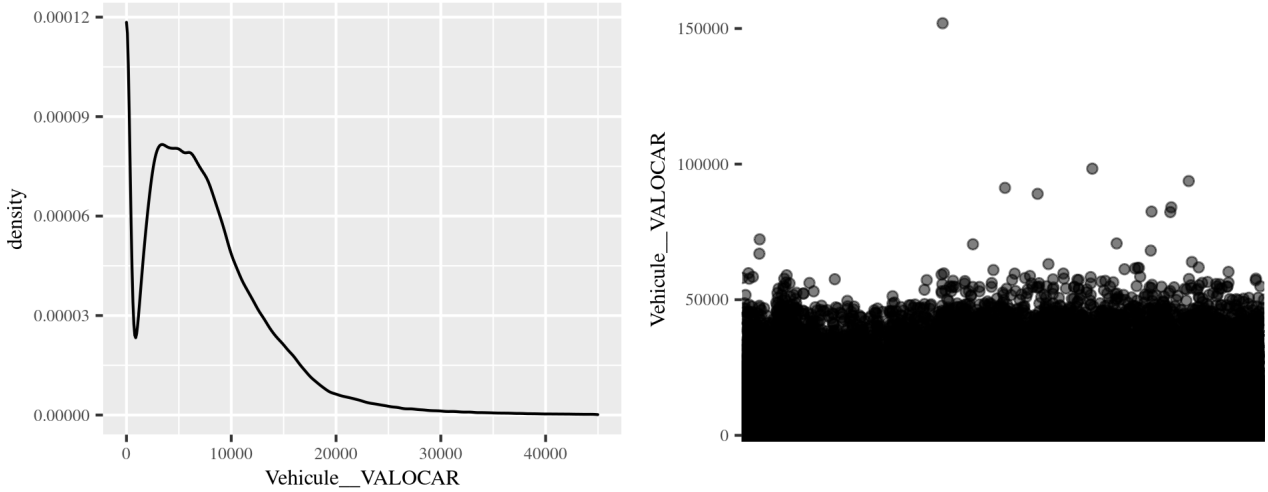
770 valeurs sont absentes (0.2% du total). Nous choisissons de les supprimer.



(a) Modalités originelles

(b) Après regroupement des modalités

FIGURE 2.13 – Distribution par puissance fiscale



(a) Densité de distribution

(b) Valeurs de la variable

FIGURE 2.14 – Variable `Vehicule__VALOCAR`

`Vehicule__VALOCAR`

Cette variable est calculée automatiquement pour chaque expertise. C'est une estimation en euros de la valeur du véhicule au regard de ses caractéristiques (âge, kilométrage, marque et modèle). La lecture du nuage de points figure 2.14b montre des valeurs extrêmes. Nous choisissons d'exclure les valeurs supérieures à 45000. Ce qui représente 494 dossiers (0.1% du total).

L'allure de la distribution figure 2.14a montre la présence de valeurs égales à zéro. 51885 valeurs sont égales à zéro (10.2% du total).

L'algorithme calculant cette variable prend en compte l'âge du véhicule. Sur la figure 2.15a, le nuage de point bi-dimensionnel entre les variables `Vehicule__VALOCAR` et `Vehicule__AGE_MOIS_VEH` montre que les valeurs à zéro sont essentiellement sur des véhicules d'un âge faible. Pour ces cas, l'âge moyen des véhicules est de 51.1 mois, contre un âge moyen sur l'ensemble du jeu de données de 72.7 mois.

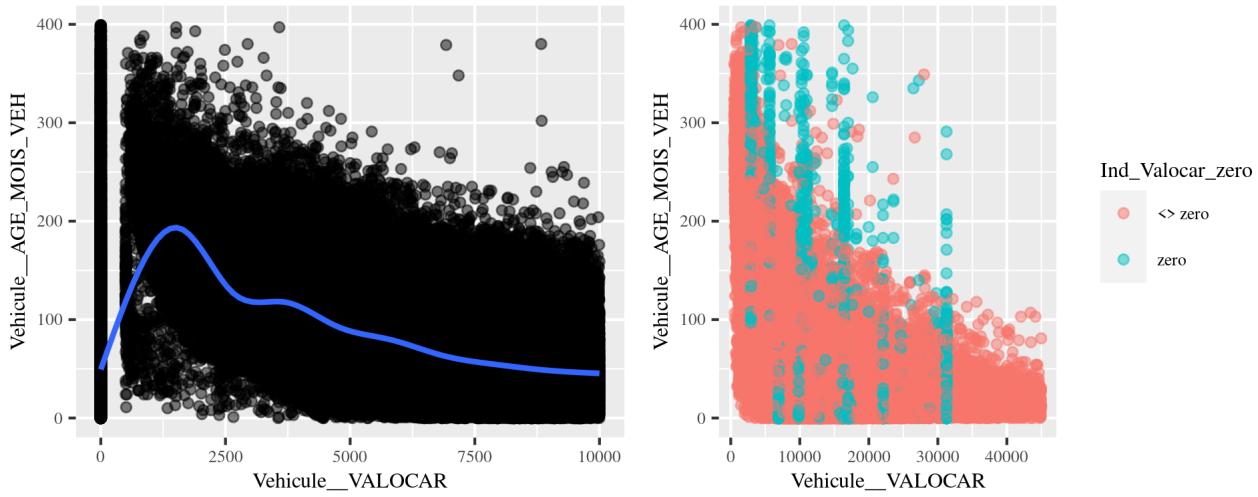
Au vu du volume de valeurs incohérentes, remplacer celles-ci par la valeur moyenne réaliserait une approximation trop importante, d'autant que les véhicules dans ce cas sont assez éloignés de l'individu moyen. Nous utilisons un modèle `bagged trees`* pour imputer les valeurs égales à zéro ou manquantes. Ce modèle utilise l'ensemble des autres variables pour déterminer la nouvelle valeur de `Vehicule__VALOCAR`. Le graphe figure 2.15a montre les valeurs obtenues. Même si des valeurs sont assez écartées du nuage de point, cela reste plus discriminant qu'une valeur constante comme zéro ou la moyenne.

2.2.3 Variables liées au réparateur

`Reparateur__TYP_REP`

Cette variable indique si le réparateur est agréé par l'assureur. Les réparateurs agréés (modalité `GP`) consentent des tarifs de main d'œuvre en contre partie du volume d'affaire apporté par l'assureur.

∗. Disponible dans le package *Recipes* (Kuhn & Wickham (2022)) et spécifiquement la fonction `step_impute_bag()` utilisant l'algorithme *bagged tree* basé sur les travaux de Kuhn & Johnson (2013)



(a) Données originelles (b) Après imputation des valeurs incohérentes

FIGURE 2.15 – Relation entre les variables `Vehicule__VALOCAR` et `Vehicule__AGE_MOIS_VEH`

Variables `Reparateur__MONT_TAR_*`

Les variables `Reparateur__MONT_TAR_IP`, `Reparateur__MONT_TAR_T1`, `Reparateur__MONT_TAR_T2`, `Reparateur__MONT_TAR_T3` et `Reparateur__MONT_TAR_TP` représentent les différentes tarifications du réparateur. En fonction de la nature et de la complexité des opérations de remise en état, le réparateur facture le tarif correspondant multiplié par le nombre d'heures de travail. Ainsi, le coût total de la main d'œuvre est la somme des poste de coûts.

On constate sur la figure 2.16 la présence de données aberrantes avec d'un coté des tarifs supérieurs à 200 euros et d'un autre coté des tarifs très faible voire égaux à zéro.

Nous choisissons d'exclure les cas où les valeurs de tarifs de main d'œuvre sont supérieures à 150 euros et inférieures à 35 euros et respectivement 100 et 20 euros pour les tarifs d'ingrédients. Ce qui représente 3356 dossiers, soit 0.7% du total.

La distribution des valeurs des différents tarifs sur la figure 2.17 montre une asymétrie à droite avec un comportement bi-modal pour le tarif `T3`. Le nombre d'heures facturé n'est pas égal entre les différents tarifs,comme le montre les statistiques de la table 20.

Nous choisissons de calculer un tarif moyen pondéré par le nombre d'heures moyen facturé par tarif. Soit :

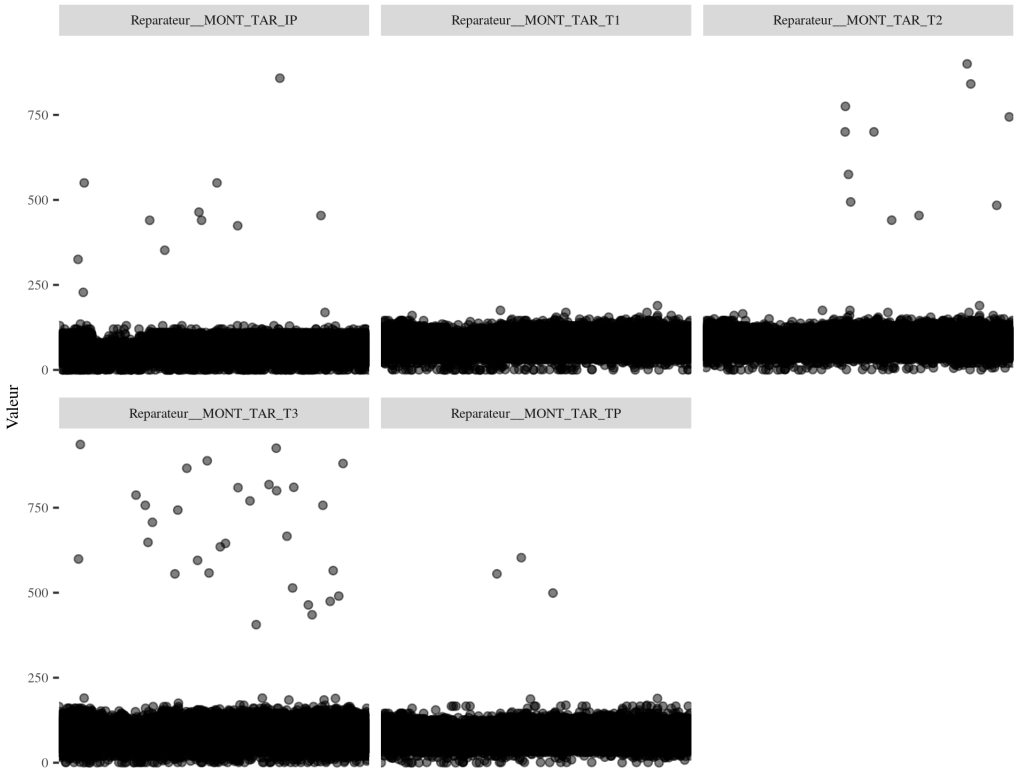
$$TH_{moy} = \frac{T1 \times 2.9 + T2 \times 3.9 + T3 \times 0.1 + (TP + IP) \times 4.5}{15.1}. \tag{2.1}$$

La distribution de la variable ainsi créée est représentée sur la figure 2.18. Les valeurs sont absentes, au nombre de 155, sont supprimées.

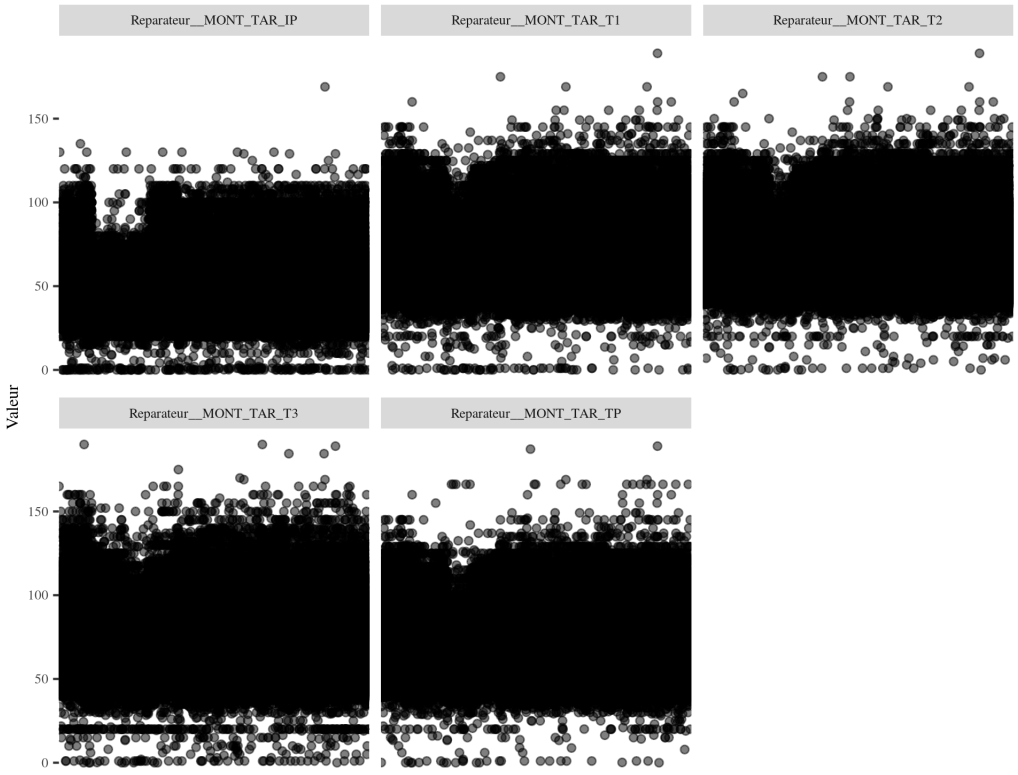
2.2.4 Variables liées au contrat

`Contrat__Franchise`

Cette variable indique si une franchise a été mise en jeu.



(a) Données originelles



(b) Valeurs après écrêtage

FIGURE 2.16 – Valeurs des variables `Reparateur__MONT_TAR_*`

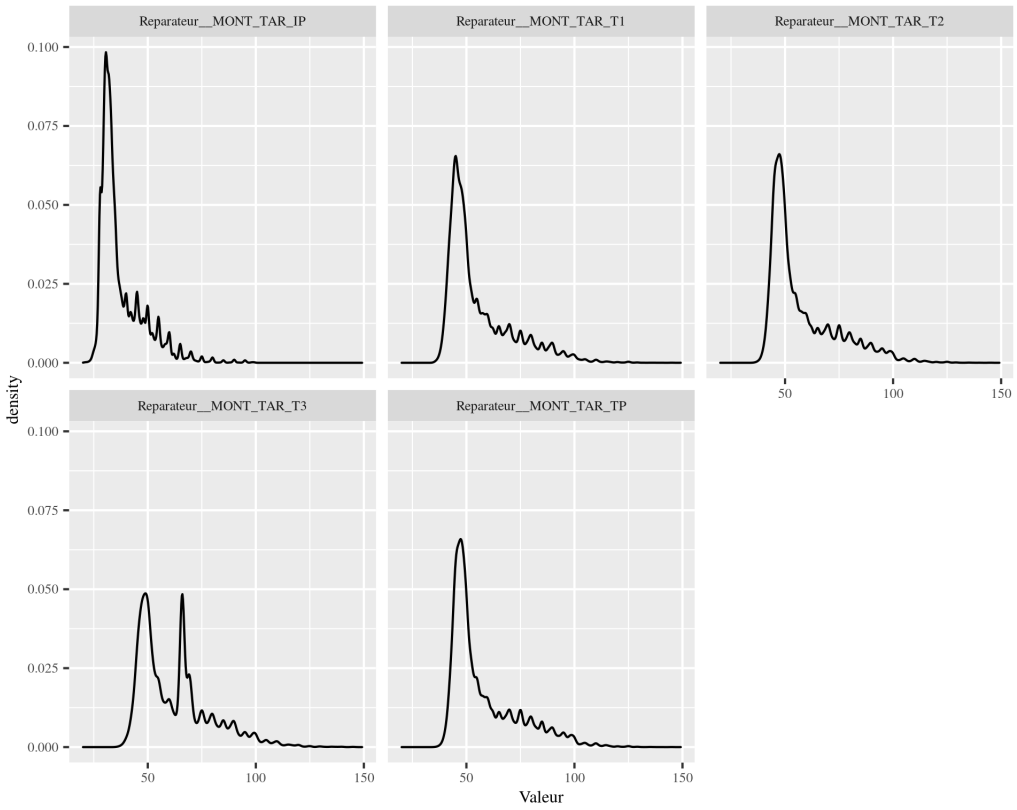


FIGURE 2.17 – Distribution des tarifs écrêtés

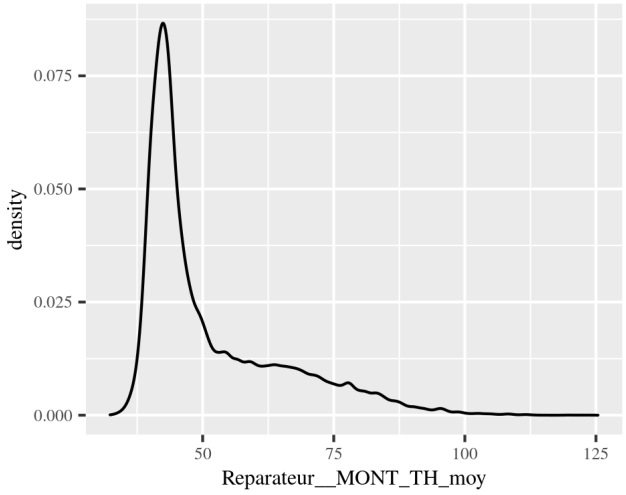


FIGURE 2.18 – Densité du tarif moyen pondéré

`Contrat__LIB_SOC`

Il s'agit du nom de la compagnie d'assurance.

2.2.5 Sélection des variables

Démarche générale

Le jeu de variables contient des variables de deux natures : qualitatives et quantitatives. Afin d'avoir une seule de mesure de dépendance sur l'ensemble des variables, nous utilisons le test V de Cramer. Basée sur le test du Khi2, cette statistique mesure l'intensité de la dépendance entre deux variables. Ce test est réalisé sur des tables de contingence entre variables qualitatives. Les variables quantitatives sont alors transformées en variables qualitatives en opérant une discrétisation sur la base des 15emes centiles. Ceci permet d'avoir des classes de taille homogènes.

Une sélection des variables est effectuée en calculant le V de Cramer entre la variable dépendante `CoutTotal__MONT_EXPERT_HT` et les variables retenues. Une fois le jeu de variables explicatives constitué, nous examinons les relations entre les variables elles mêmes toujours en utilisant le test V de Cramer.

V de Cramer

Soit une échantillon de taille n, où la table de contingence pour les variables A and B pour $i = 1, \ldots, r; j = 1, \ldots, k$ contient les fréquences n_{ij}, soit le nombre de fois où les valeurs (A_i, B_j) sont observées.

La statistique du χ^2 est donnée par :

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

Le V de Cramer est calculé en prenant la racine carrée de la statistique du χ^2 divisée par la taille de l'échantillon la plus petite dimension - 1 :

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

Avec :

- χ^2 est le test du χ^2 de Pearson,
- n est le total d'observations,
- k le nombre de colonnes,
- r le nombre de lignes.

2.2.6 Relation avec la variable dépendante

La table 2.4 donne les résultats du test V de Craner entre la variable dépendante `CoutTotal__MONT_EXPERT_HT` et les variables indépendantes. L'échelle d'interprétation de cette statistique décrite dans la table 2.5. En retenant un seuil à 0.05, on voit que 18 variables sont candidates.

2.2.7 Relation entre variables independantes

Nous allons maintenant étudier la relation entre les variables indépendantes. L'objectif est d'identifier les variables corrélées entre elles. La matrice figure 2.19 montre la valeur de la statistique V de Cramer pour chaque association de variables.

TABLE 2.4 – Dépendance entre la variable `CoutTotal__MONT_EXPERT_HT` et les variables indépendantes mesurée avec la statistique du V de cramer

Variable	V Cramer
Sinistre___Gravite	0.329
Contrat___Franchise	0.192
Reparateur___TYP_REP	0.181
Vehicule___COD_MARQUE_MODELE	0.114
Reparateur___MONT_TH_moy	0.108
Sinistre___CP_MISSION	0.100
Sinistre___COD_DOM_TYP	0.091
Sinistre___VAL_ORIENT_CHOC	0.082
Sinistre___GROUP_SINI	0.077
Vehicule___AGE_MOIS_VEH	0.073
Sinistre___NUM_DEPART_SECT	0.073
Vehicule___VALOCAR	0.070
Contrat___LIB_SOC	0.057
Vehicule___SEGMENT	0.057
Vehicule___VAL_PUIS_DIN	0.056
Vehicule___COD_COND_ACHAT	0.055
Vehicule___VAL_PUIS_FISC_EXP	0.050
Vehicule___Cod_Ener	0.050
Vehicule___COD_MARQUE	0.047
Sinistre___LIB_GAR	0.043
Vehicule___VAL_KM_HORO_COMPTEUR	0.042

TABLE 2.5 – Échelle d'interprétation de la statistique V de Cramer

V Cramer	Niveau d'association
< 0.5	Faible
0.05 - 0.10	Correct
0.10 - 0.20	Bon
> 0.20	Très bon

TABLE 2.7 – Variables exogènes retenues pour l'étude

Variables retenues
Sinistre___Gravite
Contrat___Franchise
Reparateur___TYP_REP
Sinistre___COD_DOM_TYP
Sinistre___GROUP_SINI
Vehicule___SEGMENT
Reparateur___MONT_TH_moy
Vehicule___VALOCAR

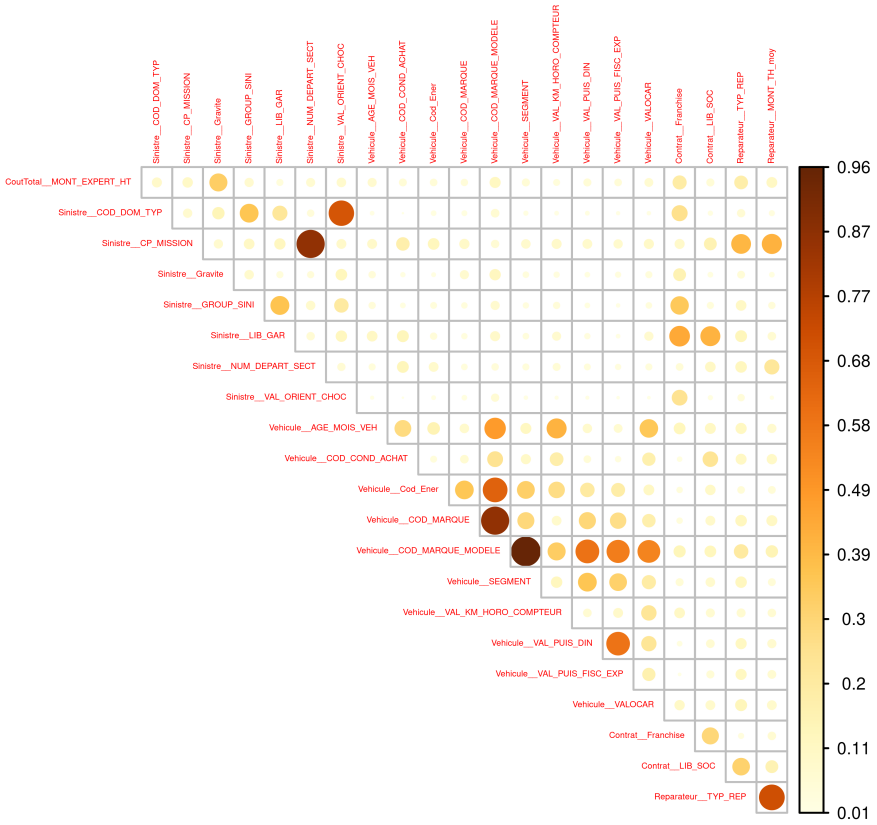
On peut constater que certaines variables sont fortement dépendantes :

- `Sinistre__CP_MISSION` avec `Sinistre__NUM_DEPART_SECT` : ce qui est tout à fait normal, le code postal étant un sous-ensemble du département ;
- `Sinistre__COD_DOM_TYP` avec `Sinistre__VAL_ORIENT_CHOC` : ce qui s'explique par le fait que suivant le point choc, l'orientation est souvent la même ;
- `Vehicule__Cod_Ener` avec `Vehicule__COD_MARQUE_MODELE` : expliqué par le fait que certains modèles sont associés à un type de motorisation ;
- `Sinistre__GROUP_SINI` avec `Sinistre__COD_DOM_TYP` et `Sinistre__LIB_GAR` : pour une nature de sinistre donnée, une garantie est mise en jeu qui correspond une typologie de dommages ;
- `Vehicule__COD_MARQUE_MODELE` avec `Vehicule__AGE_MOIS_VEH` : les modèles de véhicules sont générationnels, et correspondent donc à une tranche d'âge ;
- `Vehicule__AGE_MOIS_VEH` avec `Vehicule__COD_COND_ACHAT` : les véhicules en condition d'achat de type leasing sont souvent très récents ;
- `Vehicule__AGE_MOIS_VEH` avec `Vehicule__VAL_KM_HORO_COMPTEUR` : plus le véhicule est âgé, plus il peut parcourir de kilomètres ;
- `Vehicule__COD_MARQUE_MODELE` avec `Vehicule__COD_MARQUE` : les modèles appartiennent toujours à une marque donnée ;
- `Vehicule__COD_MARQUE_MODELE` avec `Vehicule__VAL_PUIS_FISC_EXP` et `Vehicule__VAL_PUIS_FISC_EXP` : ce sont les caractéristiques du modèle ;
- `Contrat__Franchise` avec `Sinistre__LIB_GAR` et `Sinistre__GROUP_SINI` : le contrat associe un niveau de franchise à des garanties et des typologies de sinistre ;
- `Contrat__LIB_SOC` avec `Sinistre__LIB_GAR` : chaque assureur propose plus moins de certaines garanties ;
- `Reparateur__TYP_REP` avec `Sinistre__CP_MISSION` : dépendance entre l'implémentation géographique des réparateurs agréés ;
- `Reparateur__MONT_TH_moy` avec `Sinistre__CP_MISSION` et `Reparateur__TYP_REP` : les niveaux de tarifs de main d'œuvre dépendent de la localisation géographique et des agréments des réparateurs.

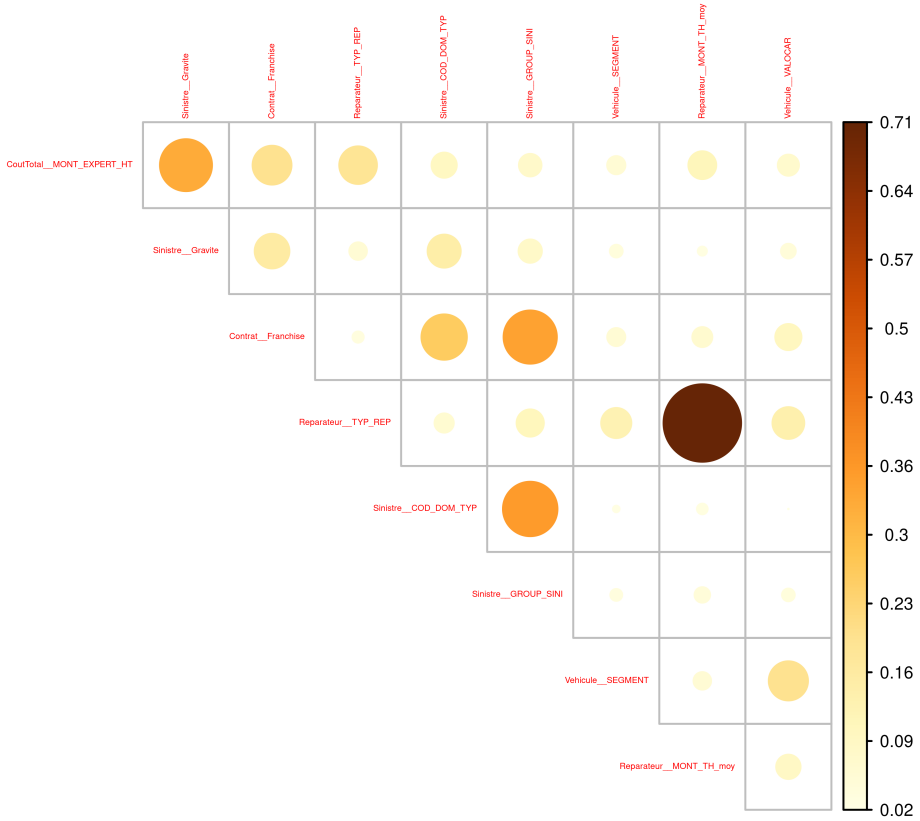
2.2.8 Variables retenues

Les variables `Sinistre__CP_MISSION` et `Vehicule__COD_MARQUE_MODELE` ne sont pas retenues au vue de leur nombre conséquent de modalités (respectivement 5228 et 1474). A la lumière des constats effectuées, nous choisissons de retenir les variables listées dans la table 2.7.

La matrice réduite aux seules variables retenues devient prend alors la forme affichée sur la figure 2.19b.



(a) Ensemble des variables



(b) Variables retenues

FIGURE 2.19 – Relation entres les variables exogènes via la statistique du V de Cramer

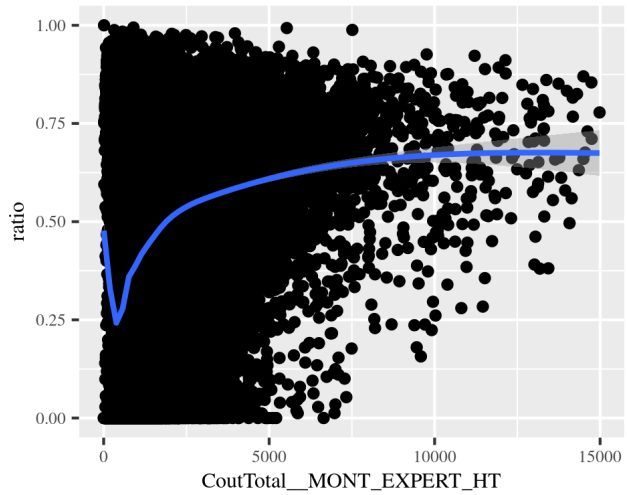


FIGURE 2.20 – Evolution du ratio $\frac{CoutPieces}{CoutTotal}$ en fonction de la variable `CoutTotal__MONT_EXPERT_HT`

2.3 Facteurs endogènes

Les facteurs endogènes sont les variables qui décrivent la performance de l'expert. Leur choix est essentiel car ils servent à piloter l'activité des experts et à leur assigner des objectifs. Ils doivent à la fois avoir du sens pour l'expert et être corrélés avec le coût moyen d'expertise. Afin de pouvoir assurer leur compréhension opérationnellement, des variables quantitatives sont retenues.

2.3.1 Analyse du coût

La variable `CoutTotal__MONT_EXPERT_HT` est la somme de quatre autres variables :

- `CoutPieces__MONT_PIECE_HT_REMI` ;
- `CoutMainOeuvre__MONT_MO_HT_REMI` ;
- `CoutPeinture__MONT_IP_HT_REMI` ;
- `CoutMainOeuvre__MONT_MOF_HT_REMI`.

Le poids de chacune est indiqué dans la table 2.8.

La variable `CoutPieces__MONT_PIECE_HT_REMI` est celle qui a le plus de poids. Celui-ci n'est pas constant suivant la valeur du montant d'expertise, comme le montre le graphe de la figure 2.20.

La relation entre les variables est représentée dans la matrice de corrélation figure 2.21. On voit que trois postes de coûts sont fortement corrélés avec la variable dépendante. La variable `CoutMainOeuvre__MONT_MOF_HT_REMI` est faiblement corrélée car d'une part, elle n'apparaît que dans 55.4% des observations, et d'autre part son montant varie peu. C'est pour cela qu'il est intéressant de distinguer les variables influant sur le montant pièces et celles sur le montant de la main d'œuvre.

Pour chacun des postes de coûts, des indicateurs de performance sont définis et présentés ci après.

TABLE 2.8 – Variables constitutives du coût moyen d'expertise : `CoutTotal__MONT_EXPERT_HT`

Variable	Moyenne	Part
CoutPieces___MONT_PIECE_HT_REMI	838	0.51
CoutMainOeuvre___MONT_MO_HT_REMI	618	0.37
CoutPeinture___MONT_IP_HT_REMI	174	0.11
CoutMainOeuvre___MONT_MOF_HT_REMI	23	0.01

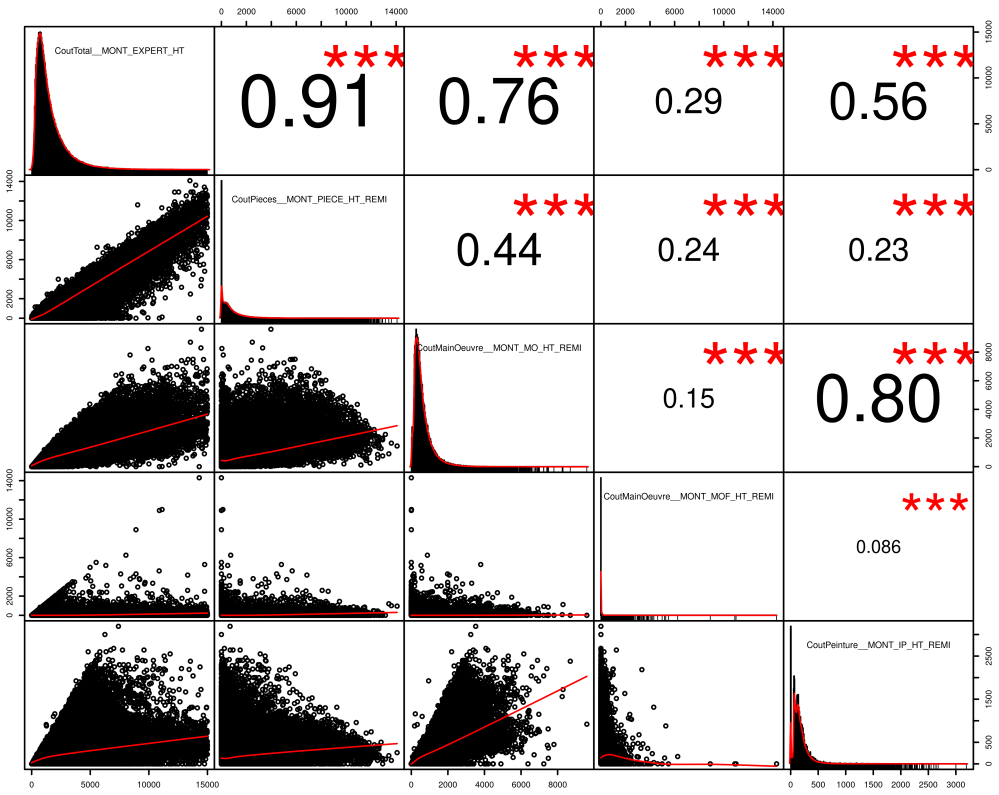


FIGURE 2.21 – Corrélation entres les différentes variables de coût

2.3.2 Sur le montant pièces

KPI_TAUX_REPAR

Le taux de réparation donne une indication sur la méthodologie prescrite par l'expert. Il est le rapport entre le nombre de pièces réparés et le nombre de pièces remplacées, calculé comme suit :

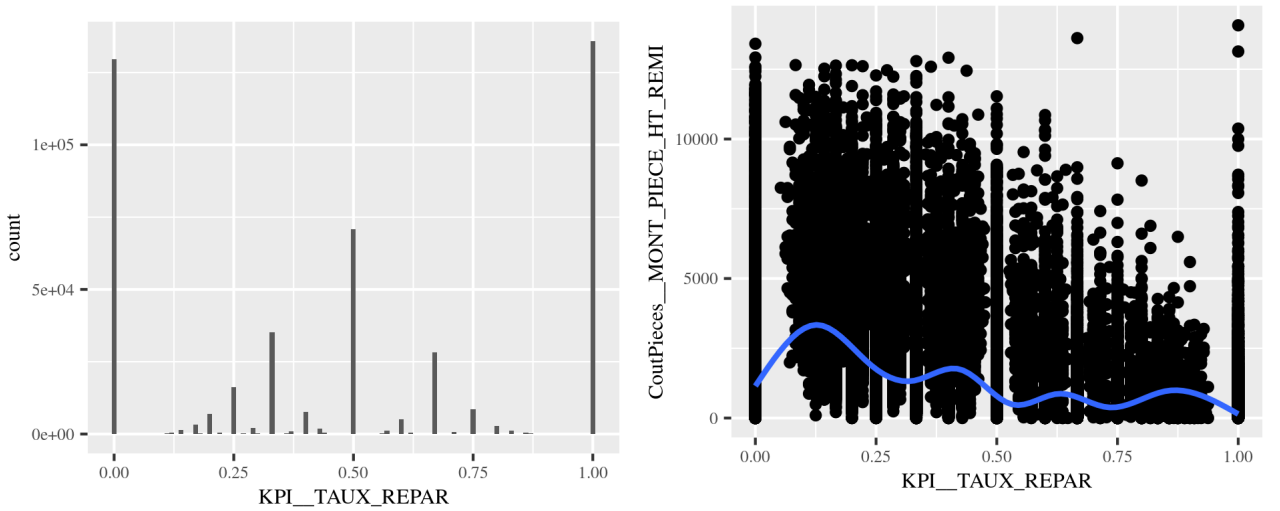
$$\texttt{KPI\_TAUX\_REPAR} = \frac{\sum \texttt{Pieces réparées}}{\sum \texttt{Pieces changées + réparées}} \tag{2.2}$$

Comme le montre la figure 2.22b, cette variable est liée au montant des pièces car la mise en œuvre de pièces du véhicule peut être une option préférable au remplacement d'un point de vue économique. Pour cela l'expert doit avoir les compétences techniques pour prescrire la réparation et effectuer le calcul d'opportunité.

Cet indicateur est calculé pour chaque expertise, dont la distribution des valeurs a la forme présentée sur le graphe figure 2.22a. La discontinuité s'explique par le fait que bon nombre d'expertises ne comportent que quelques pièces, générant ainsi une récurrence dans certaines valeurs.

KPI_TAUX_REDRESS

Le taux de redressage donne également une indication sur la méthodologie prescrite par l'expert. C'est le pendant du taux de réparation basé sur la nature de main d'œuvre. En effet, l'expert qualifie la nature de l'opération en deux types :



(a) Distribution de la variable

(b) Relation avec la variable `CoutPieces__MONT_PIECE_HT_REMI`

FIGURE 2.22 – Variable `KPI_TAUX_REPAR`

— DCC : opérations de dépose, changement ou contrôle,
— R : opération de redressage.

L'indicateur est calculé ainsi :

$$\texttt{TAUX\_REDRESS} = \frac{\texttt{TPS\_T2\_R}}{(\texttt{TPS\_T1 + TPS\_T2})}, \tag{2.3}$$

car les opérations de redressage ne se retrouvent que sur le niveau de main d'œuvre $T2$, mais la mise en œuvre d'un redressage diminue généralement le temps nécessaire à la dépose facturé soit en $T1$ ou en $T2$, sachant que `TPS_T2 = TPS_T2_R + TPS_T2_DCC`. Il y a donc un impact sur la variable `CoutPieces__MONT_PIECE_HT_REMI` comme le montre la figure 2.23b

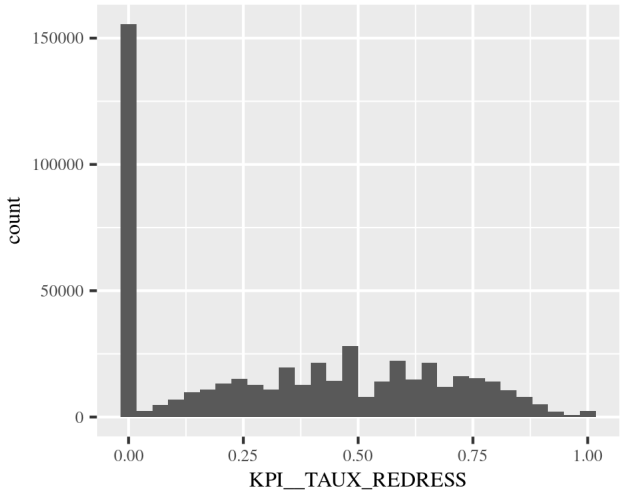
La distribution de la variable `KPI_TAUX_REDRESS` présente 30.6% de valeurs égales à zéro comme le montre la figure 2.23a. Il s'agit des dossiers où aucune opération de redressage n'a été prescrite. Les valeurs absentes sont imputées avec la valeur zéro.

`KPI__DNI`

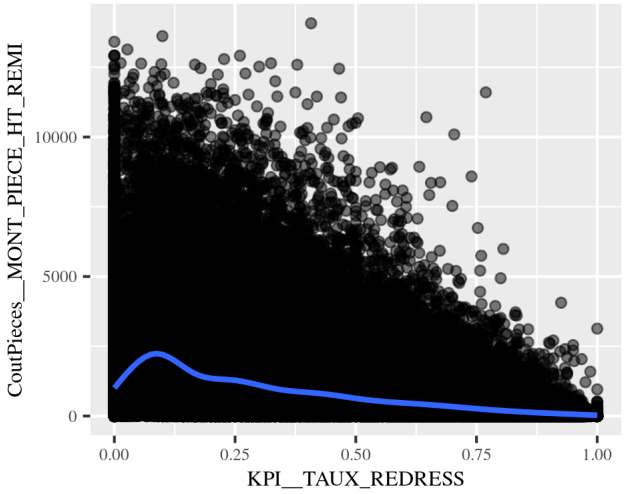
Cette variable qualitative prend la modalité `Oui` quand l'expert a exclu des dommages considérés comme non imputables au sinistre, c'est à dire non couverts par la garantie mise en jeu. L'indicateur, qui se veut quantitatif, consiste à retenir la proportion de modalités `Oui`. Ce cas de figure se produit dans 5.1% des expertises comme le montre la figure 2.24.

`KPI__NB_PIECES_C`

Il s'agit du nombre de pièces préconisées en remplacement par l'expert. La quantité de pièces détermine le coût total du poste pièces comme le montre la figure 2.25. 1.9% des valeurs sont absentes et imputées avec la valeur zéro.



(a) Distribution de la variable



(b) Relation avec la variable `CoutPieces__MONT_PIECE_HT_REMI`

FIGURE 2.23 – Variable `TAUX_REDRESS`

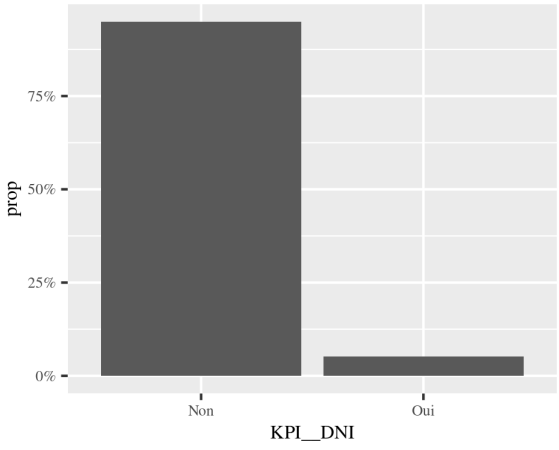


FIGURE 2.24 – Proportion de dossiers avec exclusion de dommages

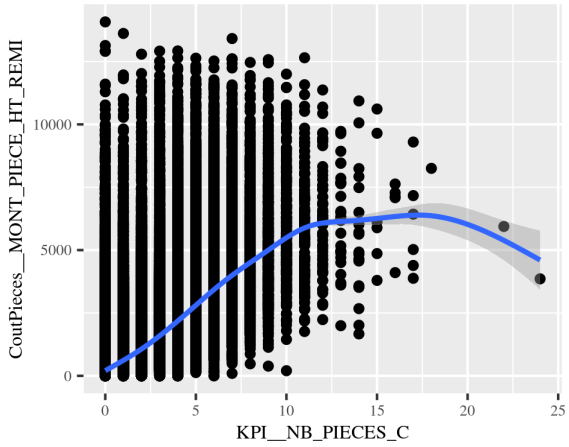


FIGURE 2.25 – Relation entre les variables `KPI__NB_PIECES_C` et`CoutPieces__MONT_PIECE_HT_REMI`

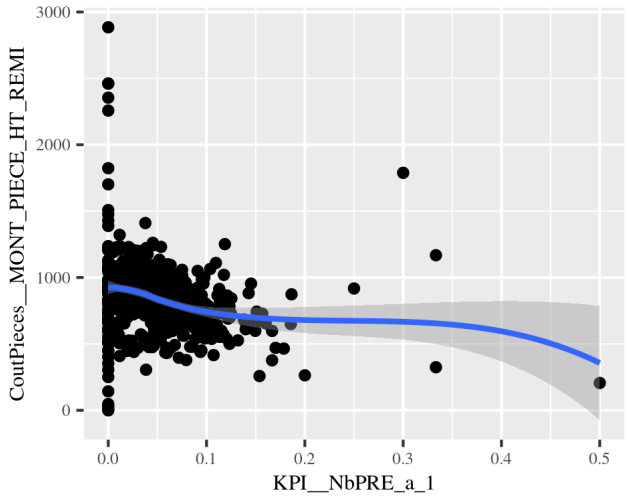


FIGURE 2.26 – Relation entre les variables `KPI__NbPRE_a_1` et `CoutPieces__MONT_PIECE_HT_REMI`

`KPI__NbPRE_a_1`

Il s'agit d'une indicatrice prenant la valeur `1` si l'expert a préconisé de la monte de pièces issues de l'économie circulaire. Ces pièces étant commercialisées à un tarif inférieur au prix de la pièce neuve, leur prescription a un impact sur le coût du poste pièces. 1.9% des valeurs sont absentes et imputés avec la valeur zéro. Le graphe de la figure 2.26 représente le taux de prescription par expert, comparé à son montant moyen de pièces.

2.3.3 Sur le montant de main d'oeuvre

Variables `Temps__TPS_T*`

Plusieurs variables décrivent le nombre d'heures prescrites par l'expert pour les postes de coûts correspondants. Il existe quatre tarifications de temps de main d'œuvre :

- T1 : opérations courantes ;
- T2 : opérations techniques ;
- T3 : haute technicité ;
- TP : peinture.

Ici le temps $T2$ est décomposé en `TPS_T2 = TPS_T2_R + TPS_T2_DCC`, deux natures d'opérations techniques.

La matrice de corrélation de la figure 2.27 montre la relation entre les différentes variables et la variable `CoutTotal__MONT_EXPERT_HT`. Au vu des niveaux de corrélation, nous décidons de retenir les variables `Temps__TPS_T1`, `Temps__TPS_T1` `Temps__TPS_TP`.

2.3.4 Sur le suivi des dossiers

Lors de ces opérations d'expertise, l'expert est amené à faire plusieurs chiffrages. En effet, depuis l'examen réalisé lors du chiffrage initial, l'expert peut modifier son chiffrage si des dommages complémentaires sont constatés, notamment après démontage du véhicule, ainsi de suite jusqu'au chiffrage final qui correspond à la facturation du réparateur et donc au montant du rapport d'expertise. Il est donc essentiel que l'expert apporte le plus grand soin dans la validation de ces compléments, quitte à effectuer un nouveau déplacement chez le réparateur. Dans le cas contraire, il peut s'exposer à une inflation de son coût moyen.

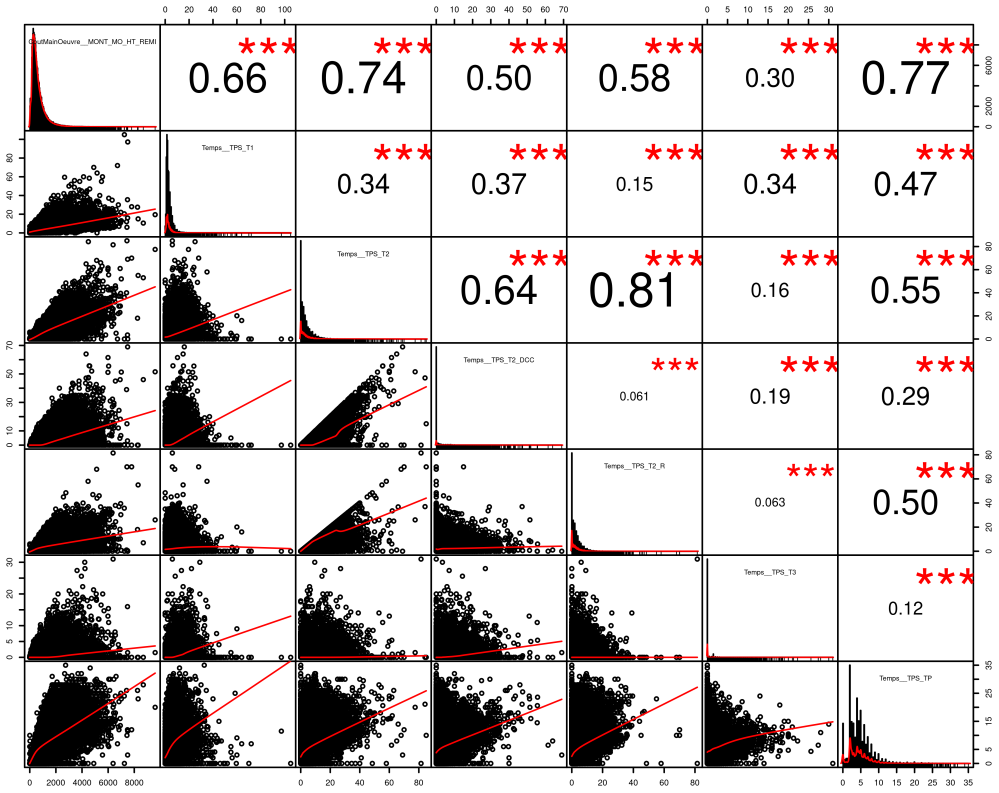


FIGURE 2.27 – Corrélation entre les variables de temps de main d'oeuvre

`KPI__ECART_CT`

L'indicateur est calculé de la sorte :

$$\frac{\text{Montant Rapport}}{\text{1er chiffrage}} - 1. \tag{2.4}$$

La figure 2.28 montre la relation avec la variable `CoutTotal__MONT_EXPERT_HT`. L'indicateur comporte des valeurs supérieures à 1 et inférieures à 0, comme indiqué dans la table 2.9. Dans ces cas, les valeurs sont assignées respectivement à 1 et 0.

TABLE 2.9 – Valeurs abbérantes de la variable `KPI__ECART_CT`

Situation	Part
KPI___ECART_CT > 1	0.0345528
KPI___ECART_CT < 0	0.0748271

2.3.5 Sélection des variables

La matrice de corrélation de la figure 2.29 montre deux variables "évidentes". Toutefois les autres variables sont conservées et leur pertinence sera jugée lors de la modélisation.

Comme réalisé précédemment avec les variables exogènes, nous utilisons la statistique du V de Cramer pur mesure l'intensité de la relation entre les variables. Les résultats figurent dans la table 2.10. Sur la

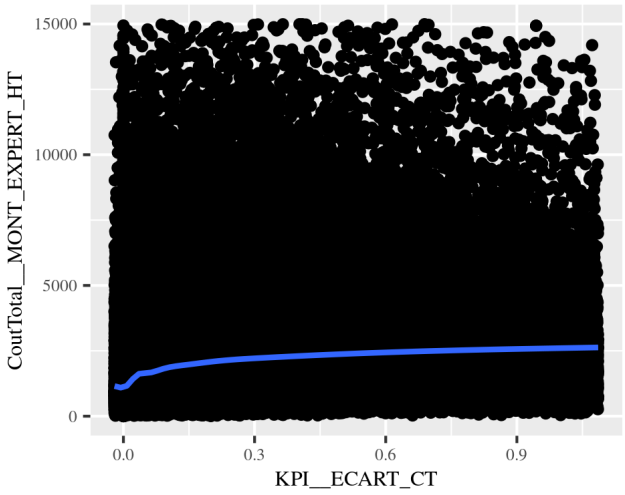


FIGURE 2.28 – Relation entre les variables `KPI__ECART_CT` et `CoutTotal__MONT_EXPERT_HT`

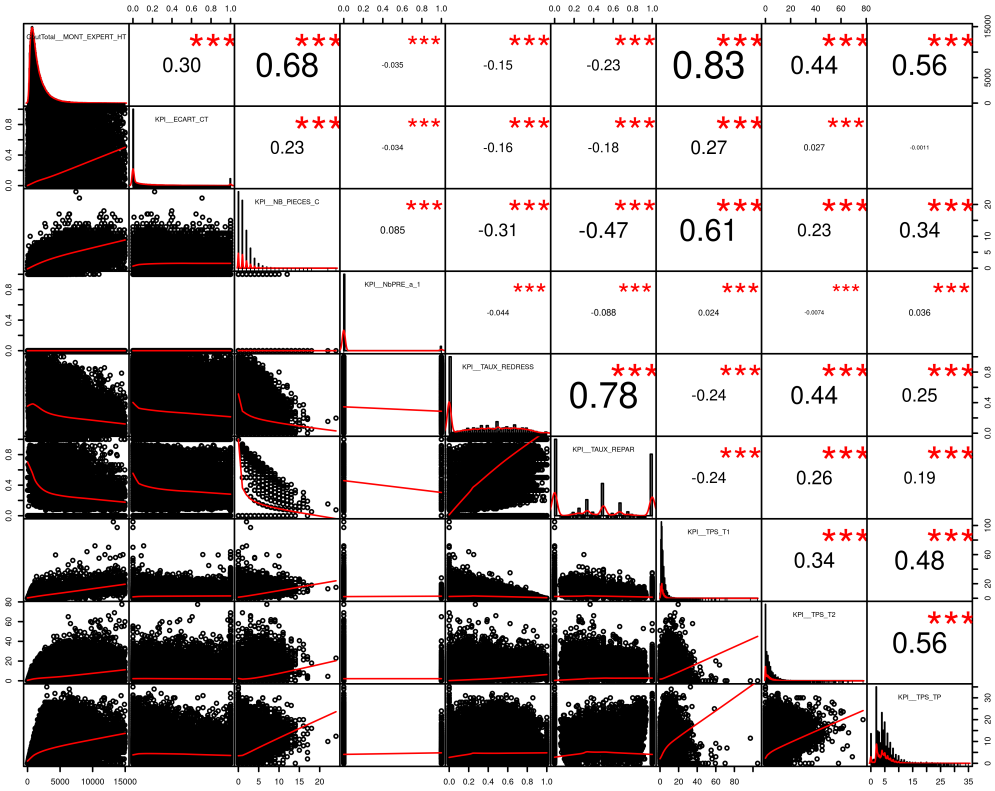


FIGURE 2.29 – Corellation entre les variables endogènes et la variable `CoutTotal__MONT_EXPERT_HT`

relation avec la variable dépendante, nous notons que sur les 8 variables, 7 sont candidates avec une valeur supérieure au seuil de 0.05.

Les relations entre les variables présentent des niveaux de dépendance significatifs tels que le montre la figure 2.30. Toutefois nous décidons de les conserver et de prendre en considération ce point lors de la modélisation.

Nous ajoutons les variables endogènes à la sélection précédente des variables exogènes. La table 2.11 liste ainsi le jeu de variables qui sera utilisé dans la modélisation.

TABLE 2.10 – Dépendance entre la variable `CoutTotal__MONT_EXPERT_HT` et les variables endogènes mesurée avec la statistique du V de cramer

Variable	V_Cramer
KPI___TPS_T1	0.434
KPI___NB_PIECES_C	0.409
KPI___TPS_TP	0.289
KPI___TAUX_REPAR	0.249
KPI___ECART_CT	0.187
KPI___TPS_T2	0.185
KPI___TAUX_REDRESS	0.151
KPI___NbPRE_a_1	0.085
KPI___DNI	0.012

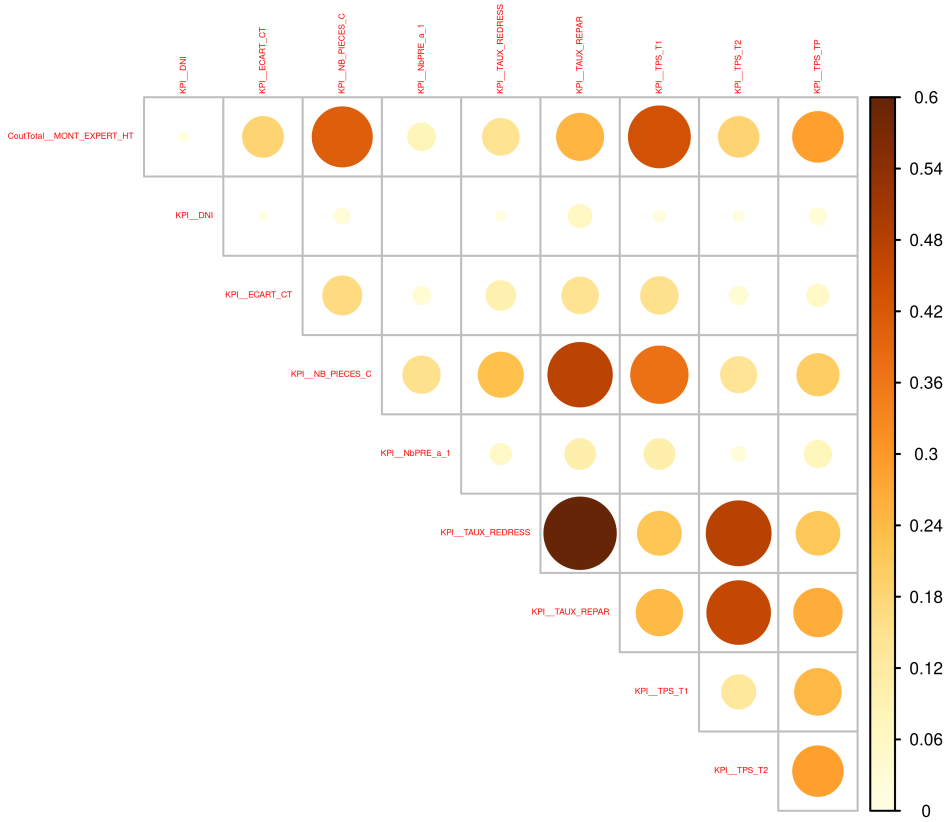


FIGURE 2.30 – Relation entres les variables endogènes via la statistique du V de Cramer

TABLE 2.11 – Variables retenues pour l'étude

Variables retenues
Sinistre___Gravite
Contrat___Franchise
Reparateur___TYP_REP
Sinistre___COD_DOM_TYP
Sinistre___GROUP_SINI
Vehicule___SEGMENT
Reparateur___MONT_TH_moy
Vehicule___VALOCAR
KPI___NB_PIECES_C
KPI___TAUX_REPAR
KPI___TAUX_REDRESS
KPI___ECART_CT
KPI___TPS_T1
KPI___TPS_T2
KPI___TPS_TP
KPI___NbPRE_a_1

2.4 Synthèse

Nous avons exploité un jeu de données complet et analysé les variables décrivant la variable d'intérêt. Pour bon nombre de variables des traitements ont été nécessaires : suppression ou imputation des valeurs manquantes, regroupement de modalités, écrêtage. Après l'ensemble des traitements et transformations sur les variables, le jeu de données comporte 468363 observations, soit une perte de -7.5%, ce qui reste acceptable. Nous avons créé des indicateurs de performance qui serviront à objectiver les experts dans la maîtrise de leur coût moyen. Le choix des variables a été effectué "à dire d'expert". Autant l'analyse de dépendance montre une relation significative avec la variable d'intérêt, autant l'examen des relations inter-variables montre des relations de dépendances assez marquées qui devront être prises en compte lors de la modélisation. Ce choix de conserver des variables dépendantes est motivé par le besoin de recueillir des variables qui aient du sens lors de leur utilisation opérationnelle.

Chapitre 3

Modélisation

3.1 Formulation du problème

La classification des experts par niveau de performance qui semble la plus évidente consiste à comparer le coût moyen de chacun d'entre eux. Pour cela on calcule l'espérance conditionnelle de la variable `CoutTotal__MONT_EXPERT_HT` (également nommée `CM` pour des raisons de commodité d'écriture), soit $\mathbb{E}[CM|Expert]$, la valeur du coût moyen pour chacun des n experts, dont la distribution est représentée dans le graphe de la figure 3.1.

Or il a été établi que des facteurs exogènes influencent la valeur de $\mathbb{E}[CM|Expert]$. Aussi, il a été choisi de construire une autre classification fondée sur une variable non affectée par ces facteurs exogènes, ce qui conduit à vouloir supprimer l'effet des variables exogènes sur les résultats de l'expert afin de pouvoir les comparer entre eux.

On considère que CM se décompose en deux parties : $CM = Z + P$, avec Z la variable aléatoire correspondant aux effets exogènes et P la variable aléatoire correspondant au niveau de performance des experts $\mathbb{E}[CM|Z|Expert]$. On va chercher à évaluer $P = CM - Z$.

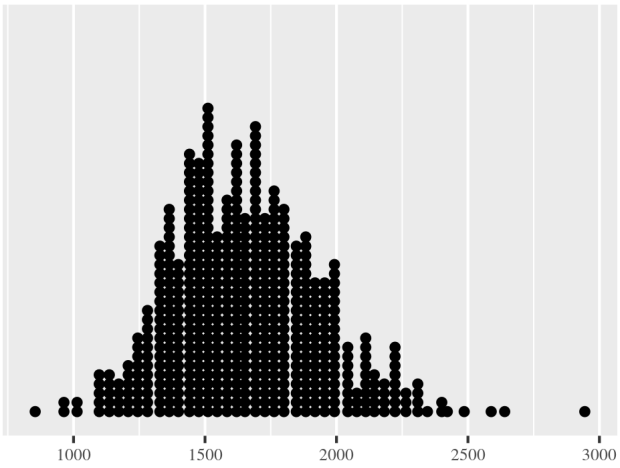


FIGURE 3.1 – Distribution de la variable `CoutTotal__MONT_EXPERT_HT` agrégée par expert

3.1.1 Critères de validité de la démarche

Quantité d'information

Considérons $CM = Y + X$, où X serait une variable de coût non affectée de facteur exogènes et où l'expert disposerait de l'exhaustivité des informations pour une évaluation parfaite du coût. Y est l'ensemble des facteurs (endogènes et exogènes). Ce qui revient à dire que X serait une constante, donc à variance nulle.

Évaluer la quantité d'information dans le modèle revient donc à mesurer la chute de variance obtenue avec les variables exogènes et endogènes. Un modèle contenant toute la quantité d'information vérifie

$$\mathbb{V}(CM) - \mathbb{V}(Y) = \mathbb{V}(X) = 0.$$

L'indicateur de quantité d'information est :

$$QI = 1 - \frac{\mathbb{V}(Y)}{\mathbb{V}(CM)},$$

avec une valeur de 1 pour un modèle contenant l'ensemble de l'information.

Variabilité de performance

Nous cherchons à identifier dans Y la part de variance correspondant au niveau de performance des experts, afin d'établir leur classification, soit $\mathbb{V}(P) = \mathbb{V}(\mathbb{E}[CM|Z])$.

La loi de la variance totale :

$$\mathbb{V}(CM|Expert) = \mathbb{E}[\mathbb{V}(CM|Z|Expert)] + \mathbb{V}(\mathbb{E}[CM|Z|Expert]), \tag{3.1}$$

décompose la variance en, respectivement, variance intra-classe et variance-inter classe. CM étant conditionné aux facteurs exogènes Z, la variance inter-classe correspond à la variance entre les coûts moyen des experts non impactés par les variables exogènes, autrement dit la variabilité de la performance entre les experts.

Nous le verrons plus tard, mais la construction du modèle revenant à soustraire à la moyenne l'impact des écarts des facteurs exogènes, il n'y a que $\mathbb{E}[CM|Z|Expert]$ qui est connu. La variance intra-classe ne peut donc être mesurée directement. Afin de l'estimer, la décomposition de la variance est utilisée :

$$\mathbb{V}(CM|Expert) = \mathbb{V}(Z|Expert) + \mathbb{V}(\mathbb{E}[CM|Z|Expert]) + 2\,\mathrm{Cov}(\mathbb{E}[CM|Z|Expert], Z|Expert), \tag{3.2}$$

où $\mathbb{V}(Z|Expert)$ est la variance de l'impact des effets exogènes, et $\mathrm{Cov}(\mathbb{E}[CM|Z|Expert], Z|Expert)$ la covariance entre les impact des facteurs exogènes et le coût moyen par expert.

En posant :

$$\begin{aligned} \mathbb{E}[\mathbb{V}(CM|Z|Expert)] + \mathbb{V}(\mathbb{E}[CM|Z|Expert]) &= \mathbb{V}(Z|Expert) + \mathbb{V}(\mathbb{E}[CM|Z|Expert]) \\ &\quad + 2\,\mathrm{Cov}(\mathbb{E}[CM|Z|Expert], Z|Expert) \end{aligned} \tag{3.3}$$

$$\begin{aligned} \mathbb{E}[\mathbb{V}(CM|Z|Expert)] &= \mathbb{V}(Z|Expert) \\ &\quad + 2\,\mathrm{Cov}(\mathbb{E}[CM|Z|Expert], Z|Expert), \end{aligned} \tag{3.4}$$

nous obtenons une autre expression de la variance intra-classe. La covariance entre variables signifie une dépendance linéaire entre elles et donc n'apportent pas plus d'informations dans le modèle. L'information pertinente contenue dans la variance intra-classe est donc $\mathbb{V}(Z|Expert)$.

Nous avons alors deux indicateurs de la qualité d'information dans le modèle. La part de variance des impacts des effets exogènes, soit la part de la variance intra-classe dans la variance du coût moyen * :

$$PV_{intra} = \frac{\mathbb{V}(Z|Expert) + 2\,\mathrm{Cov}(\mathbb{E}[CM|Z|Expert], Z|Expert))}{\mathbb{V}(CM|Expert)}, \tag{3.5}$$

et la part d'information pertinente dans la variance intra-classe :

$$PIP = \frac{\mathbb{V}(Z|Expert)}{\mathbb{V}(Z|Expert) + 2\,\mathrm{Cov}(\mathbb{E}[CM|Z|Expert], Z|Expert))}. \tag{3.6}$$

Nous sommes dans un cas multifactoriel, nous avons alors $Z_1 ... Z_k$ variables exogènes et $P_1 ... P_k$ variables endogènes, soit $\mathbb{V}(Z) = \sum_{i=1}^{k} \mathbb{V}(Z_i) + 2\sum_{1\leq i<j\leq k} \mathrm{Cov}(Z_i, Z_j)$ et $\mathbb{V}(P) = \sum_{i=1}^{k} \mathbb{V}(P_i) + 2\sum_{1\leq i<j\leq k} \mathrm{Cov}(P_i, P_j)$.

Notons que si les variables évaluant le niveau de performance de l'expert sont indépendantes des effets exogènes, on a $\mathrm{Cov}(P, Z) = 0$, donnant $\mathbb{V}(P + Z) = \mathbb{V}(P) + \mathbb{V}(Z)$.

3.1.2 Choix de la modelisation

Après avoir sélectionné les variables qui nous semblent être les prédicteurs les plus pertinents de la variable `CoutTotal__MONT_EXPERT_HT`, nous passons à l'étape de la modélisation. Il s'agit de trouver un modèle qui exploite au mieux les informations des prédicteurs pour expliquer la variable dépendante.

Au regard du sujet traité, nous pouvons dégager deux grandes caractéristiques attendues pour ce modèle. Les implications étant opérationnelles, le modèle se doit d'être intelligible par le plus grand nombre et à des fins d'analyse, il doit être à même de donner l'impact de chaque variable sur le coût d'expertise.

Un modèle de type GLM † est retenu pour répondre à ces attentes. Le modèle est de la forme $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k$. Ce qui répond au besoin d'intelligibilité. Dans le cas présent la distribution de la variable dépendante est `Gamma` (comme vu au précédent chapitre) et la fonction de lien $g = log$. Le modèle est alors de la forme $g(\mathbb{E}[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k$. Pour exprimer les coefficients dans les mêmes unités que la variable dépendante, il suffit d'effectuer une transformation inverse $exp(g(\mathbb{E}[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k)$, soit $\mathbb{E}[Y] = exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k)$.

Le GLM donne ainsi, pour chaque modalité i des variables X_k qualitatives, un coefficient β_i qui exprime l'accroissement de la valeur de la variable dépendante quand cette modalité est rencontrée. Dans le cas d'une variable X_i quantitative, le coefficient β_i exprime l'accroissement de la valeur de la variable dépendante pour une unité de variation de la variable X_i. Il y a autant de coefficient β que de modalités et de variables quantitatives. Ce qui correspond au besoin d'expliciter les impacts des différentes variables.

3.2 Modèle de base

Les variables retenues dans le modèle sont les suivantes :

```
## CoutTotal__MONT_EXPERT_HT ~ Sinistre__Gravite + Contrat__Franchise +
```

∗. Cette grandeur est aussi connue sous le nom de *part de variance résiduelle*

†. « Modèle Linéaire Généralisé » introduit en 1972 par Nelder & Wedderburn (1972)

```
##     Reparateur__TYP_REP + Sinistre__COD_DOM_TYP + Sinistre__GROUP_SINI +
##     Vehicule__SEGMENT + Reparateur__MONT_TH_moy + Vehicule__VALOCAR +
##     KPI__NB_PIECES_C + KPI__TAUX_REPAR + KPI__TAUX_REDRESS +
##     KPI__ECART_CT + KPI__TPS_T1 + KPI__TPS_T2 + KPI__TPS_TP +
##     KPI__NbPRE_a_1
```

La plate-forme d'apprentissage automatique H2O est utilisée *. Ce choix est effectué, car elle permet d'appliquer nativement une pénalisation Lasso. L'algorithme détermine les coefficients du modèle via le principe du maximum de vraisemblance. La plate-forme possède une interface dans R, ce qui la rend facilement intégrable dans le code utilisé sur R.

3.2.1 Détermination des coefficients β

Le modèle est entraîné sur le jeu de données comprenant 468363 observations. La sortie du GLM donne pour chaque modalité sa P-valeur associée. Celle-ci aide à déterminer si les relations observées dans l'échantillon utilisé existent également dans une population plus large.

La P-valeur de chaque variable indépendante teste l'hypothèse nulle selon laquelle la variable n'a aucune corrélation avec la variable dépendante. S'il n'y a pas de corrélation, alors il n'y a pas d'association entre les changements de la variable indépendante et ceux de la variable dépendante. En d'autres termes, il n'y a pas suffisamment de preuves pour conclure qu'il y a un effet au niveau de la population.

Si la P-valeur d'une variable est inférieure au seuil fixé, les données d'échantillon fournissent alors suffisamment de preuves pour rejeter l'hypothèse nulle pour l'ensemble de la population. Dans ce cas, les données favorisent l'hypothèse d'une corrélation non nulle. Les changements de la variable indépendante sont associés aux changements de la variable dépendante au niveau de la population. Cette variable est statistiquement significative et constitue probablement un ajout intéressant au modèle.

La lecture des résultats dans la table 24 montre que 3 modalités sur 47 ont une P-valeur > 0.1. Elles ne sont donc ne sont pas considérées comme représentatives. Elles ne seront pas retenues dans le modèle.

Pour les variables qualitatives à x modalités, la modélisation GLM comporte x-1 coefficients, la première modalité de la variable étant comprise dans le coefficient `intercept`. Le choix de l'ordre des modalités est alors important, car nous le verrons par la suite, le coefficient `intercept` n'étant pas utilisé pas le calcul d'impacts.

3.2.2 Analyse des résidus

La différence entre la valeur observée de la variable dépendante y_i et la valeur prédite μ_i est appelée *résidu*. L'ensemble des résidus peuvent être vus comme une partie de la variance non-attribuable aux variables explicatives du modèle. Dans une optique de modélisation, un modèle avec un fort pouvoir explicatif correspond à des résidus sans structures particulière, donc effectivement des « restes » sans intérêt quelconque ; dans le cas contraire, l'étude des résidus permet de voir les défauts dans les restes, car une structure, une information intéressante dans les résidus signifie que le modèle n'en a pas tenu compte.

L'examen de la distribution des résidus fourni donc une indication sur la qualité du modèle. Nous utilisons les résidus de Pearson calculés de la sorte :

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{V}(\hat{\mu}_i)}},$$

*. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/glm.html

où μ_i la valeur prédite pour y_i et $\widehat{V}(\hat{\mu}_i)$ l'estimation de la variance pour μ_i. Les résidus doivent suivre une distribution normale comprise dans l'intervalle $[-2, 2]$ pour que le modèle soit considéré de qualité.

A la lecture du graphe de la figure 3.2 nous constatons la normalité de la distribution, avec une queue à droite. Il y a 0.1% des résidus avec une valeur supérieure à 2, ce qui est acceptable.

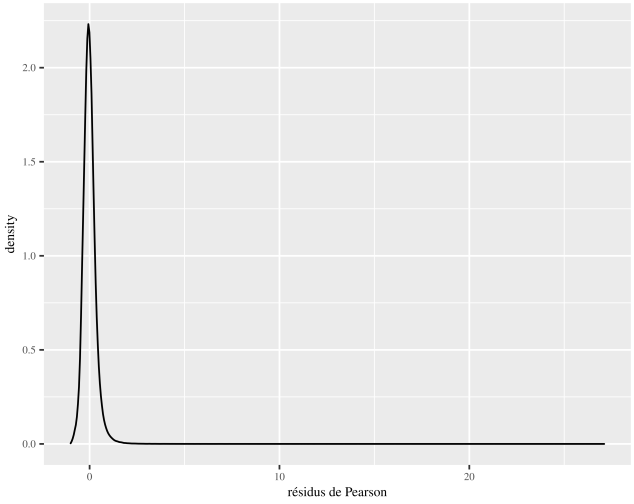


FIGURE 3.2 – Résidus de pearson du modèle

3.3 Suppression des effets exogènes

3.3.1 Démarche générale

Le modèle GLM nous donne, pour chaque modalité X_i, le coefficient β_i de la relation entre la modalité et la variable `CoutTotal__MONT_EXPERT_HT`. Afin de supprimer les effets des différentes variables, nous calculons pour chaque expert, la différence $\Delta Z_i = Z_i - \overline{Z}$, et donc $\Delta Z_i \beta_i$, ici pour le cas des variables exogènes. Ceci nous donne la valeur de l'impact en euros à partir de laquelle nous pouvons ajuster la valeur de la variable `CoutTotal__MONT_EXPERT_HT`. Soit

$$\mathbb{E}[CM|Z] = \mathbb{E}[CM] - \sum_{i=1}^{k} \Delta Z_i \beta_i.$$

Dans le cadre du modèle GLM multiplicatif retenu ici, la décomposition entre effets endogènes et exogènes du coût moyen est de la forme $CM = Z \times P$ et $\mathbb{E}[CM] = exp(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + ... + \beta_k Z_k)$ si nous ne retenons que les coefficients des variables exogènes. Avec $Z = \prod_{i=1}^{k} \exp(\Delta Z_i \beta_i)$, nous avons $\mathbb{E}[CM] = exp(\beta_0) \times \prod_{i=1}^{k} \exp(\Delta Z_i \beta_i)$.

Par identification, $P = exp(\beta_0)$, soit :

$$\mathbb{E}[P] = \frac{\mathbb{E}[CM]}{\prod_{i=1}^{k} \exp(\Delta Z_i \beta_i)}.$$

Ce qui revient à chercher la valeur d'un intercept qui correspond à la valeur moyenne du niveau de performance des experts.

La démarche est identique si l'on souhaite supprimer les impacts de la variance de performance des experts $\mathbb{V}(P)$, en utilisant la variable P, soit :

$$\mathbb{E}[Z] = \frac{\mathbb{E}[CM]}{\prod_{i=1}^{k} \exp(\Delta P_i \beta_i)}.$$

3.3.2 Réduction de variance

En supprimant les impacts des variables exogènes à la variable CM, nous obtenons une estimation de la variable par $P = CM - Z$, soit $\mathbb{E}[CM|Expert|Z]$. La variance de P correspond au coût moyen de chaque expert expurgé des effets exogènes. La table 3.1 donne les informations de la comparaison entre les deux distributions. Nous constatons que la variance a ainsi chuté de 45.3 %. Le graphe de la figure 3.3a montre les densités des deux distributions de coûts moyen avant et après épuration.

La valeur de l'espérance du coût moyen $\mathbb{E}[CM|Expert]$ varie de -1.1%. L'asymétrie à droite de la distribution est réduite, la distribution de $\mathbb{E}[CM|Expert|Z]$ étant plus proche d'une distribution normale.

TABLE 3.1 – Impact de la suppression des effets exogènes sur la variable `CoutTotal__MONT_EXPERT_HT`

Indicateur	Valeur
Quantite_Information	0.4534752
mean(CM_dec_Exo)/mean(CoutTotal___MONT_EXPERT_HT) - 1	-0.0109630
skewness(CM_dec_Exo)/skewness(CoutTotal___MONT_EXPERT_HT) - 1	-0.5178191

Les variables endogènes ayant été intégrées dans le modèle GLM, nous effectuons la même démarche avec les facteurs endogènes. Il s'agit là d'épurer la variable $\mathbb{E}[CM|Expert|Z]$ des variations de performances entre experts, soit

$$\mathbb{E}[CM|Expert|Z|P] = \frac{\mathbb{E}[CM]}{\prod_{i=1}^{k} \exp(\Delta Z_i \beta_i) \prod_{j=1}^{l} \exp(\Delta P_j \beta_j)}. \tag{3.7}$$

On s'attend à trouver, *idéalement*, une variance nulle pour la variable ainsi obtenue. Les résultats dans la table 3.2 montrent que la réduction de variance constatée est de 95.7%. Finalement cette variance peut être interprétée comme celle des résidus dans un modèle où l'on tente d'expliquer l'ensemble des facteurs influençant le coût d'un sinistre, l'ensemble de l'information disponible ayant été utilisée.

La valeur de l'espérance du coût moyen $\mathbb{E}[CM|Expert]$ varie de -2.6%. L'asymétrie de la distribution est quasi nulle et avec une allure normale comme constatée sur le graphe figure 3.3b.

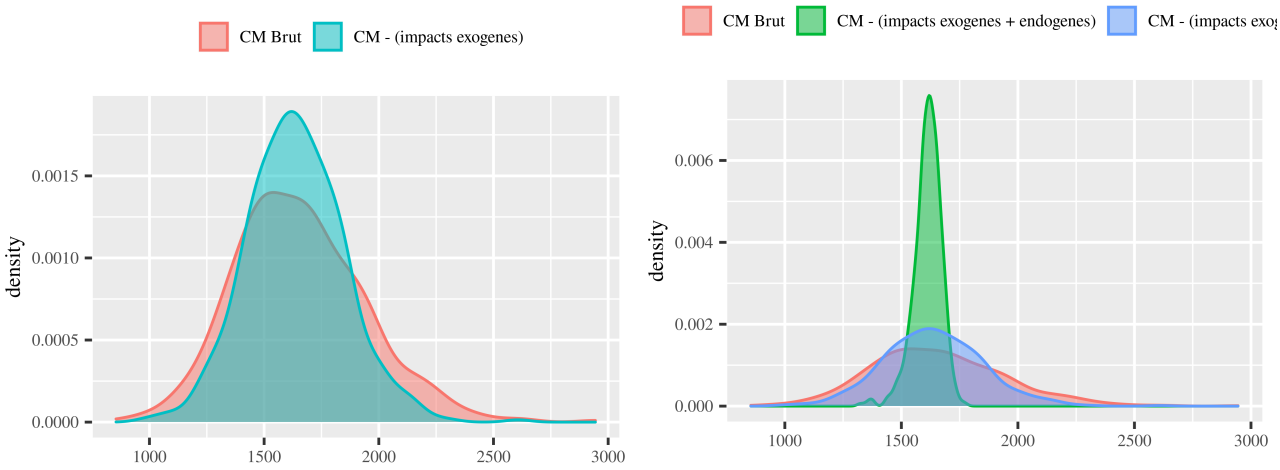
TABLE 3.2 – Impact de la suppression des effets exogènes et endogènes sur la variable `CoutTotal__MONT_EXPERT_HT`

Indicateur	Valeur
Quantite_Information	0.9572774
mean(CM_dec_ALL)/mean(CoutTotal___MONT_EXPERT_HT) - 1	-0.0261014
skewness(CM_dec_ALL)/skewness(CoutTotal___MONT_EXPERT_HT) - 1	-2.9171158

Les résultats dans la table 3.4 correspondent aux termes de l'équation 3.2. La table 3.3 donne les correspondances entre les termes employés dans les équations et la dénomination utilisée lors des calculs.

L'égalité obtenue entre $\mathbb{V}(CM|Expert)$ et $\mathbb{V}(Z|Expert) + \mathbb{V}(\mathbb{E}[CM|Z|Expert])$ $+ 2\,\mathrm{Cov}(\mathbb{E}[CM|Z|Expert], Z|Expert)$ permet de s'assurer de la validité des calculs effectués.

Partant de là, il est possible de calculer les indicateurs définis à travers les relation 3.5 et 3.6. La table 3.5 montre les résultats obtenus pour ce modèle. Ces indicateurs seront utilisés pour comparer les modèles entre eux.



(a) CM originel et CM épuré des facteurs exogènes

(b) CM originel et CM épuré des facteurs exogènes et endogènes

FIGURE 3.3 – Densités des différents coûts moyen

TABLE 3.3 – Correspondance termes - dénomination

Dénomination	Terme
var(CoutTotal___MONT__EXPERT__HT)	$\mathbb{V}(CM\|Expert)$
var(delta__Exo)	$\mathbb{V}(Z\|Expert)$
var(CM__dec__Exo)	$\mathbb{V}(\mathbb{E}[CM\|Z\|Expert])$
2 * cov(delta__Exo, CM__dec__Exo)	$2\,\mathrm{Cov}(\mathbb{E}[CM\|Z\|Expert], Z\|Expert))$

TABLE 3.4 – Décomposition de la variance du coût moyen

Statistique	original
var(CoutTotal___MONT__EXPERT__HT)	82032
-var(delta__Exo)	-9358
-var(CM__dec__Exo)	-44832
-2 * cov(delta__Exo, CM__dec__Exo)	-27841
Total	0

3.3.3 Orthogonalité

On recherche toujours à obtenir un modèle orthogonal, c'est-à-dire avec des variables toutes indépendantes entre elles. C'est un cas idéal, dans lequel, les estimations des coefficients du modèle réduit seront les mêmes que celles du modèle complet. Autrement dit, on obtient les mêmes effets estimés pour les variables indépendantes, qu'elles soient testées individuellement ou simultanément. Il est possible d'ajouter ou soustraire des variables orthogonales sans affecter les coefficients des autres variables. Il en va de même pour l'inclusion ou l'exclusion des effets d'interaction.

TABLE 3.5 – Indicateurs de la qualité d'information dans le modèle

Indicateur	original
Var_Inter	44832.37
Var_Intra	37199.35
PV_intra	0.45
PIP	0.25

Cela assure une meilleure compréhension des phénomènes, les variables n'étant pas influencées par la variation d'autres. L'orthogonalité est vérifiée par l'absence de covariance entre les variables. Dans ce cas nous avons $\text{Cov}(P, Z) = 0$, donnant $\mathbb{V}(P + Z) = \mathbb{V}(P) + \mathbb{V}(Z)$ car la covariance de variables indépendantes est nulle.

D'un autre coté, lorsque les variables ne sont pas orthogonales, les coefficients peuvent changer lors de l'ajustement des variables dans le modèle. Les effets dépendent dans une certaine mesure des variables du modèle.

Pour le cas étudié, la covariance entre les variables P et Z est non nulle comme indiqué dans la table 3.4, donc le modèle n'est pas orthogonal. Il faut alors s'attendre à des dépendances entre certaines variables du modèle.

3.3.4 Analyse d'impact

Dans le cadre de l'application opérationnelle de l'étude, il est important de pouvoir expliciter les résultats, à des fins de justification et également de compréhension des phénomènes. Le modèle GLM permet de calculer l'impact de chaque modalité sur la valeur de la variable dépendante, $\mathbb{E}[CM]$. Les facteurs exogènes ou endogènes vont tirer à la hausse ou à la baisse la valeur de $\mathbb{E}[CM|Expert]$. L'objectif est de quantifier quel est le poids de chaque modalité dans cette variation.

Impact des facteurs exogènes

L'intérêt de quantifier l'impact des facteurs exogènes est manifeste : il permet de de donner des explications quant à tel ou tel niveau de classification d'un expert. Cela représente aussi des informations intéressantes pour l'assureur dans la compréhension de sa structure de coût.

Pour chaque modalité nous calculons son impact tel que :

$$Impact_i = \sum_{m=1}^{n} \text{abs}(\frac{CM_m}{\exp(\Delta Z_{i,m}\beta_i)}), \tag{3.8}$$

avec $m_1...m_n$ l'indice de l'expert. Cette approche est indicative, s'agissant d'une évaluation individuelle des impacts, leur somme est différente des impacts totaux réels, le modèle étant multiplicatif.

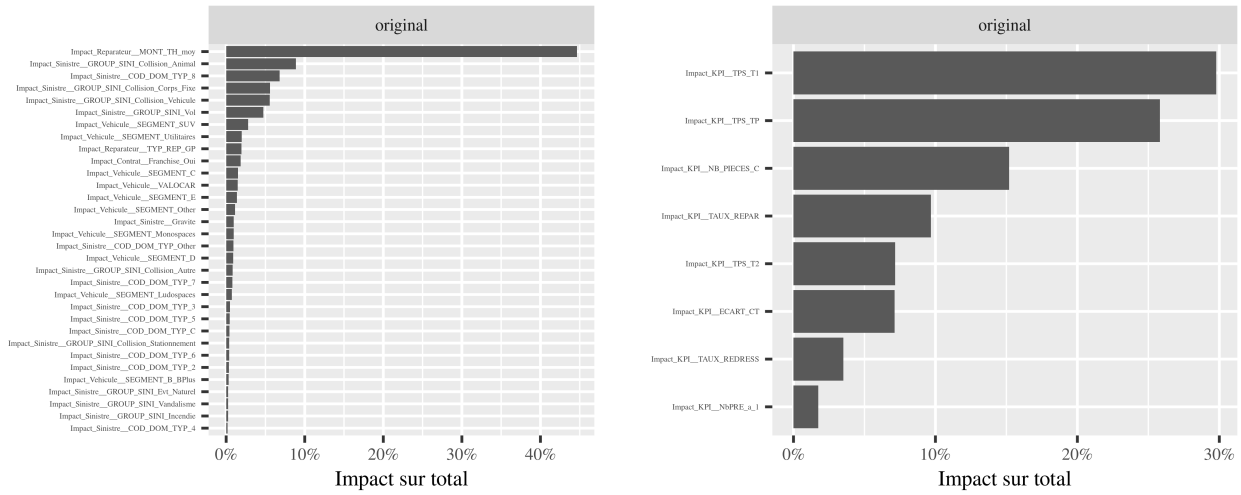
Le graphe de la figure 3.4a montre la contribution de chaque modalité. On note que le premier facteur exogène en termes d'impact est le tarif horaire du réparateur. Ce qui est cohérent, puisqu'il entre directement dans la détermination du montant de la main d'oeuvre d'une part et d'autre part il y a une forte variance sur cette variable en raison des disparités dans les niveaux de facturation des réparateurs. Vient ensuite les modalités liées à la nature du sinistre et du point de choc. Encore des éléments cohérents dans le sens où l'on observe de fortes variation de coût suivant le point de choc ou la nature du sinistre (figure 2.7 et table 2.1).

Impact des indicateurs de performance

De la même manière que pour les facteurs exogènes, nous calculons l'impact des variables décrivant la performance des experts. La formulation est similaire à celle utilisée pour les facteurs exogènes :

$$Impact_i = \sum_{m=1}^{n} \text{abs}(\frac{CM_m}{\exp(\Delta P_{i,m}\beta_i)}). \tag{3.9}$$

De même, la démarche revêt une finalité opérationnelle. Dans ce cas, l'approche permet de comparer l'impact des différentes modalités des facteurs endogènes. Ces derniers sont utilisés comme indicateurs de performance et servent à décliner un plan d'objectivation des experts. Le choix se portera naturellement sur les indicateurs de performance ayant le plus d'impact sur le coût moyen que l'expert cherche à maîtriser. Le graphe de la figure 3.4b montre la contribution de chaque modalité calculée par la valeur absolue des impacts.



(a) Facteurs exogènes (b) Facteurs endogènes

FIGURE 3.4 – Comparaison du niveau d'impact des modalités sur le coût moyen

3.4 Classification des experts

Connaissant désormais $\mathbb{E}[CM|Expert|Z]$, nous pouvons effectuer une classification des experts qui ne souffre pas des facteurs exogènes. La distribution de cette variable présente une variabilité moindre représentée sur le graphe de la figure 3.3a. Cela signifie que les coûts moyens des experts sont plus proches car moins influencés par la variance des facteurs exogènes. Afin de juger de la pertinence de la classification, nous comparons les deux classifications, initiale et après suppression des facteurs exogènes.

La table 3.6 et le graphe de la figure 3.5 montrent les changements de position des experts. On constate donc que la réduction de variance ne se résume pas à une contraction de la distribution, ce qui gage d'une classification plus pertinente. Sa forme est plus centrée (le skewness diminue de 51% comme le montre la table 3.1), ce qui signifie que des situations extrêmes sont corrigées. Cela se traduit par des changements de position des experts dans le classement : 45.3% des experts voient leur position augmenter en moyenne de 23.6 positions et 52.1% des experts voient leur position baisser en moyenne de 29.3 positions.

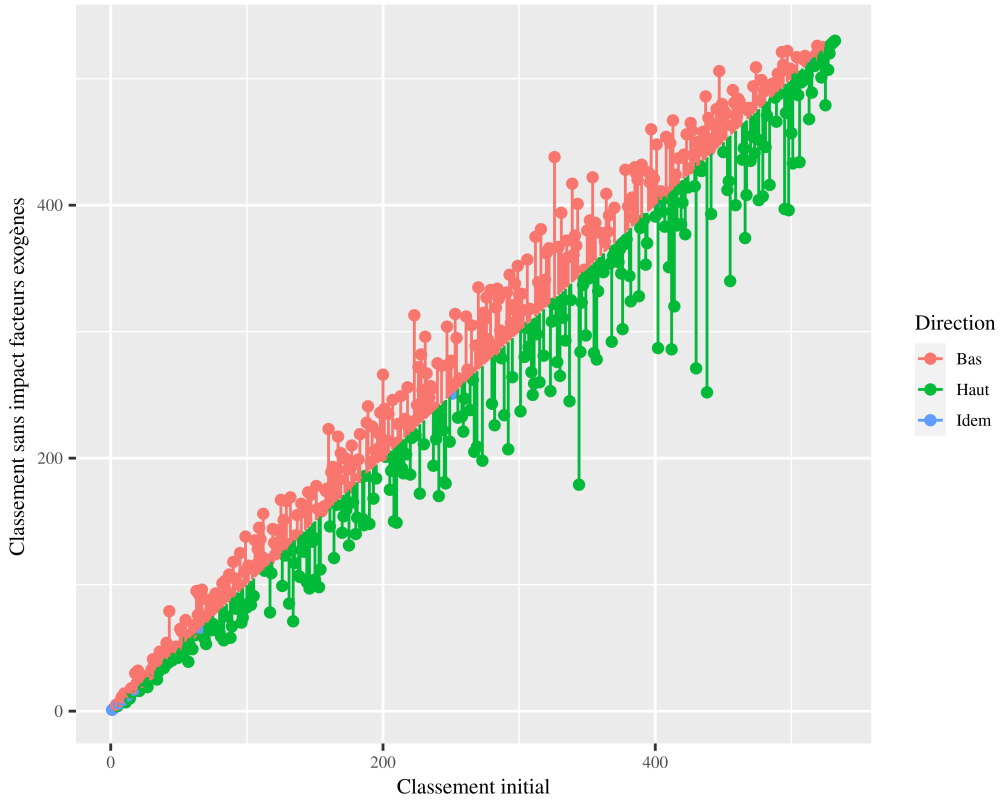


FIGURE 3.5 – Changement de position des experts après classification avec le CM épuré des facteurs exogènes

TABLE 3.6 – Statistiques sur le changement de classification des experts

Statistique	Valeur
Part_experts_surclasses	0.45
Surclasse_max	111.00
Surclasse_moy	23.60
Part_experts_declasses	0.52
Declasse_max	-187.00
Declasse_moy	-29.30
Part_experts_stables	0.03

3.5 Modèle avec pénalisation

La sélection de modèles statistiques consiste à estimer les performances prédictives de différents modèles et à choisir le meilleur modèle parmi les candidats. Le modèle idéal est simple et ayant une bonne précision de prédiction, malheureusement il demeure peu rencontré dans la pratique : le meilleur modèle est souvent la meilleure approximation.

Les méthodes de sélection traditionnelles (telles que *forward*, *backward*, ou *stepwise selection*) identifient d'abord un sous-ensemble de variables prédictives en ajoutant ou en supprimant successivement des variables (ou les deux), puis elles ajustent un modèle sur l'ensemble réduit de variables. Les algorithmes de ces méthodes sont gourmands en ressources de calcul car ils procèdent de manière itérative pour chercher un optimum. Non seulement ces méthodes sont gourmandes en temps de calcul mais peuvent aussi échouer à trouver un optimum. Plus encore, les modèles sélectionnés peuvent également être extrêmement variables, en ce sens qu'un petit changement dans les données peut entraîner un ensemble très différent de variables et de prédictions (Harrell (2001)).

La régression pénalisée résout cette instabilité en diminuant la variance impliquée dans l'estimation des coefficients. Parce qu'elle produit des résultats plus stables pour les données corrélées ou les cas où le nombre de prédicteurs est beaucoup plus grand que la taille de l'échantillon, la régression pénalisée est souvent préférée aux méthodes de sélection traditionnelles. Contrairement aux méthodes de sélection de sous-ensembles, les méthodes de régression pénalisées ne sélectionnent pas explicitement les variables ; au lieu de cela, elles rétrécissent (*« shrinkage »*) les coefficients en introduisant une contrainte qui va pénaliser la log-vraisemblance du modèle.

Cette pénalité fait que les coefficients de régression vont tendre vers zéro. Avec une pénalisation de type `Lasso` (*Least Absolute Shrinkage and Selection Operator*) certains coefficients de régression pourront être mis à zéro, ce qui rend les autres coefficients plus significatifs. Ainsi, les méthodes de régression pénalisées effectuent simultanément la sélection des variables et l'estimation des coefficients.

3.5.1 Régression avec régularisation Lasso

La régularisation Lasso est implémentée sur le modèle original. Dans un premier temps, les résultats obtenus sont comparés avec ce premier modèle. La pénalisation sera conservée par la suite pour les autres modélisations.

Concrètement, la régularisation s'effectue via un paramètre de pénalisation λ qui contrôle la quantité de régularisation appliquée. Pour obtenir le meilleur modèle possible, il faut trouver la valeur optimale de λ. Le modèle GLM h2o permet d'effectuer une recherche (« grid search ») pour déterminer le paramètre optimal. La démarche consiste à calculer des modèles pour différentes valeurs de λ, en partant de la valeur la plus élevée qui assigne la valeur zéro à tous les coefficients, puis en diminuant le niveau de régularisation.

Résidus

Avec la pénalisation Lasso, il y a 0.1% des résidus avec une valeur supérieure à 2, ce qui est semblable aux résultats obtenus avec le modèle original. La distribution est identique, on ne distingue pas les deux courbes sur la figure 3.6.

Réduction de variance

Les tables 3.7 et 3.8 comparent la chute de variance entre les deux modèles, respectivement après suppression de l'impact des facteurs exogènes et impacts des facteurs exogènes et endogènes. Les

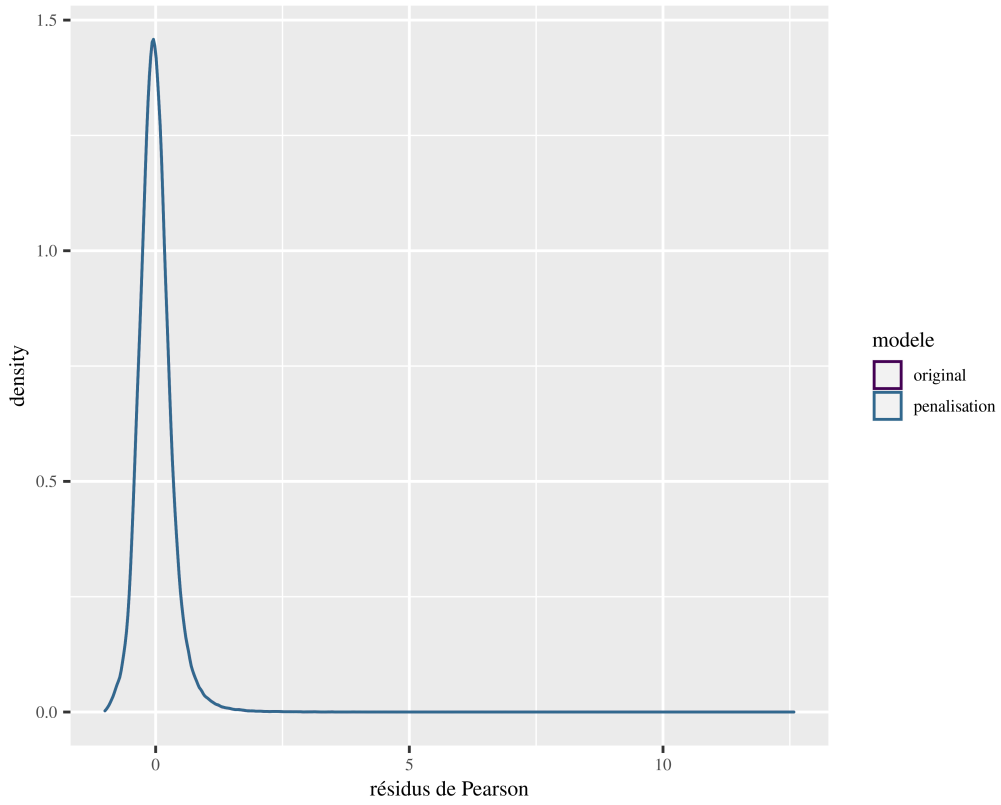


FIGURE 3.6 – Comparaison des résidus de pearson

résultats sont sensiblement identiques au modèle original. La quantité d'information captée par le modèle reste identique avec une réduction de variance similaire au modèle original.

La table 3.9 indique la chute de variance avec le modèle pénalisé. Là encore, les résultats sont très proches du modèle original : la mise en place de la pénalisation n'a pas d'influence sur la réduction de variance. Conséquence : les indicateurs de qualité d'information sont également très proches du modèle original telle que le montre la table 3.10.

TABLE 3.7 – Impact de la suppression des effets exogènes du modèle avec pénalisation

Indicateur	original	penalisation
Quantite_Information	0.4534752	0.4507719
mean(CM_dec_Exo)/mean(CoutTotal___MONT_EXPERT_HT) - 1	-0.0109630	-0.0108208
skewness(CM_dec_Exo)/skewness(CoutTotal___MONT_EXPERT_HT) - 1	-0.5178191	-0.5259607

Analyse d'impact

L'impact de chaque modalité est comparée pour chacun des modèles. Pour les variables exogènes, contrairement aux précédents constats, des différences plus marquées sont constatées. La pénalisation Lasso modifie l'impact des modalités. On peut voir sur le graphe de la figure 3.7 que certaines modalités ont un impact qui devient nul et que d'autres voient leur poids augmenter.

A contrario, comme le montre la figure 3.8, la contribution de chaque variable endogène varie faiblement. Il y a peu de changement dans les contributions d'impact entre les deux modèles.

TABLE 3.8 – Impact de la suppression des effets exogènes et endogènes du modèle avec pénalisation

Indicateur	original	penalisation
Quantite_Information	0.9572774	0.9567947
mean(CM_dec_ALL)/mean(CoutTotal___MONT_EXPERT_HT) - 1	-0.0261014	-0.0260054
skewness(CM_dec_ALL)/skewness(CoutTotal___MONT_EXPERT_HT) - 1	-2.9171158	-2.9548264

TABLE 3.9 – Décomposition de la variance du CM du modèle avec pénalisation

Statistique	original	penalisation
var(CoutTotal___MONT_EXPERT_HT)	82032	82032
-var(CM_dec_Exo)	-44832	-45054
-var(delta_Exo)	-9358	-9275
-2 * cov(delta_Exo, CM_dec_Exo)	-27841	-27702
Total	0	0

TABLE 3.10 – Indicateurs de la qualité d'information dans le modèle avec pénalisation

Indicateur	original	penalisation
PV_intra	0.4535	0.4508
PIP	0.2516	0.2508

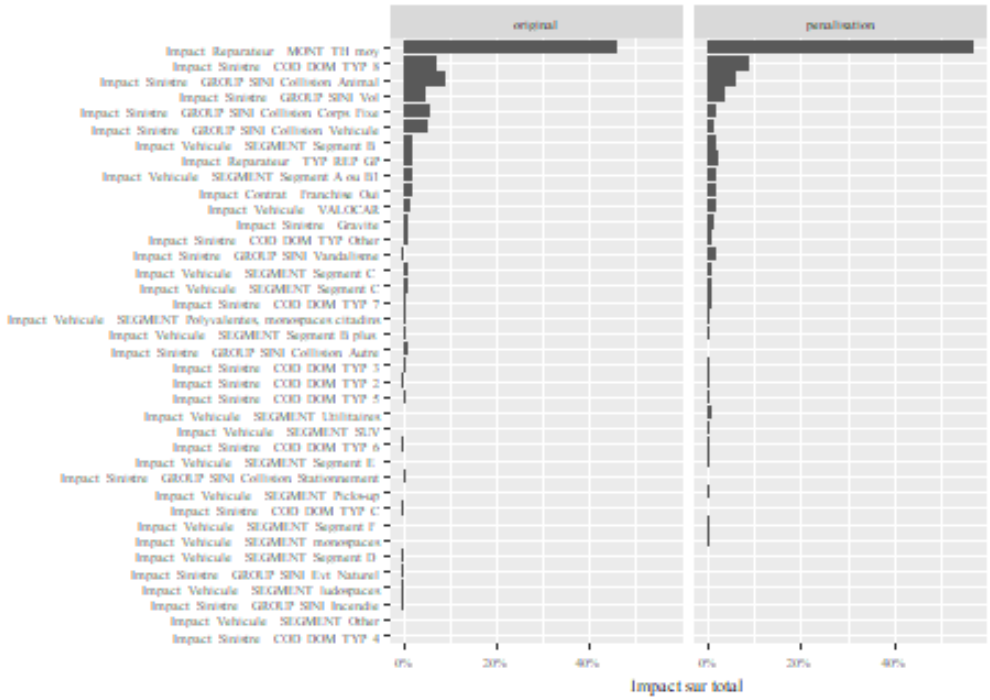


FIGURE 3.7 – Comparaison du niveau d'impacts des facteurs exogènes, par modalités pour le modèle avec pénalisation

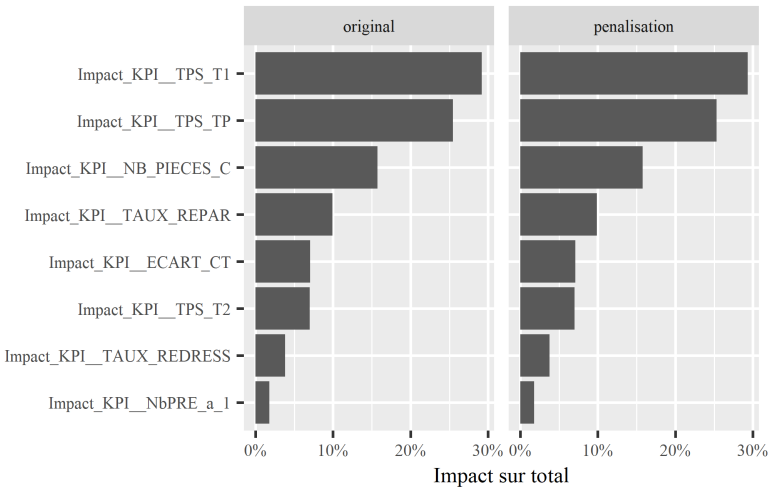


FIGURE 3.8 – Comparaison du niveau d'impacts des facteurs endogènes, par modalités pour le modèle avec pénalisation

3.5.2 Modalités éliminées

La contribution d'impact de certaines modalités varie sensiblement. La régularisation Lasso pénalise certaines et augmente par conséquent la contribution d'autres. La table 3.11 donne le niveau de contribution des différentes modalités des variables exogènes pour les deux modèles. Pour 7 modalités, la contribution est nulle et leur niveau contribution est reporté sur d'autres modalités. Néanmoins aucune variable ne voit l'ensemble de ses modalités éliminées.

3.5.3 Sélection des indicateurs

La régularisation Lasso permet d'obtenir un modèle ayant le même pouvoir explicatif avec un nombre inférieur de modalités. Toutefois les variables endogènes ne sont pas affectées par la pénalisation. Ces variables vont servir à fixer les objectifs aux experts. Il est souhaitable d'éviter qu'elles soient corrélées entre elles afin d'obtenir un schéma d'animation opérationnelle le plus pertinent possible. La lecture de la matrice de la figure 3.9 montre la dépendance des variables en termes d'impact mesurée par covariance entre variables.

Il est nécessaire d'opérer un choix « manuellement » dans les variables exogènes. Le choix se porte sur un jeu de variables ayant le moins de dépendance. En ce sens, sont retenues les variables suivants qui serviront d'indicateur de performance :

```
## [1] "KPI__ECART_CT"    "KPI__TAUX_REPAR"  "KPI__NB_PIECES_C" "KPI__TPS_TP"
## [5] "KPI__NbPRE_a_1"
```

3.6 Modèle avec interactions

3.6.1 Ajout d'interactions dans le modèle

Un effet d'interaction existe lorsque l'effet d'une variable indépendante sur une variable dépendante change, en fonction de la ou des valeurs d'une ou plusieurs autres variables indépendantes.

TABLE 3.11 – Niveau d'impact des modalités avant et après pénalisation

term	original	Part cumulée (Original)	pénalisation	Part cumulée (Pénalisation)
Impact_Reparateur___MONT_TH_moy	38068	0.46	38128	0.57
Impact_Sinistre___COD_DOM_TYP_8	5769	0.53	5866	0.66
Impact_Sinistre___GROUP_SINI_Collision_Anim	7495	0.62	4013	0.72
Impact_Sinistre___GROUP_SINI_Vol	3965	0.67	2478	0.76
Impact_Reparateur___TYP_REP_GP	1600	0.69	1569	0.78
Impact_Vehicule___SEGMENT_Segment B	1847	0.71	1367	0.80
Impact_Contrat___Franchise_Oui	1557	0.73	1334	0.82
Impact_Vehicule___SEGMENT_Segment A ou B1	1738	0.75	1330	0.84
Impact_Vehicule___VALOCAR	1296	0.77	1302	0.86
Impact_Sinistre___GROUP_SINI_Vandalisme	251	0.77	1139	0.88
Impact_Sinistre___GROUP_SINI_Collision_Corps_F	4674	0.83	1083	0.89
Impact_Sinistre___Gravite	870	0.84	855	0.91
Impact_Sinistre___GROUP_SINI_Collision_Vehic	4589	0.90	794	0.92
Impact_Sinistre___COD_DOM_TYP_Other	812	0.91	744	0.93
Impact_Sinistre___COD_DOM_TYP_7	699	0.91	567	0.94
Impact_Vehicule___SEGMENT_Segment C	887	0.92	469	0.94
Impact_Vehicule___SEGMENT_Segment C	850	0.93	454	0.95
Impact_Vehicule___SEGMENT_Utilitaires	126	0.94	454	0.96
Impact_Vehicule___SEGMENT_Polyvalentes	535	0.94	390	0.96
Impact_Sinistre___COD_DOM_TYP_2	283	0.95	326	0.97
Impact_Vehicule___SEGMENT_SUV	192	0.95	317	0.97
Impact_Vehicule___SEGMENT_Segment E	91	0.95	299	0.98
Impact_Vehicule___SEGMENT_Segment B plus	448	0.96	273	0.98
Impact_Sinistre___COD_DOM_TYP_3	352	0.96	266	0.99
Impact_Sinistre___COD_DOM_TYP_5	342	0.96	239	0.99
Impact_Vehicule___SEGMENT_monospaces	64	0.96	194	0.99
Impact_Vehicule___SEGMENT_Picks-up	146	0.97	180	1.00
Impact_Sinistre___COD_DOM_TYP_6	276	0.97	174	1.00
Impact_Vehicule___SEGMENT_Segment F	140	0.97	124	1.00
Impact_Sinistre___GROUP_SINI_Collision_Autre	710	0.98	5	1.00
Impact_Vehicule___SEGMENT_Other	82	0.98	0	1.00
Impact_Vehicule___SEGMENT_Segment D	242	0.98	0	1.00
Impact_Vehicule___SEGMENT_ludospaces	212	0.99	0	1.00
Impact_Sinistre___GROUP_SINI_Collision_Stat	335	0.99	0	1.00
Impact_Sinistre___GROUP_SINI_Evt_Naturel	218	0.99	0	1.00
Impact_Sinistre___GROUP_SINI_Incendie	200	1.00	0	1.00
Impact_Sinistre___COD_DOM_TYP_4	82	1.00	0	1.00
Impact_Sinistre___COD_DOM_TYP_C	293	1.00	0	1.00

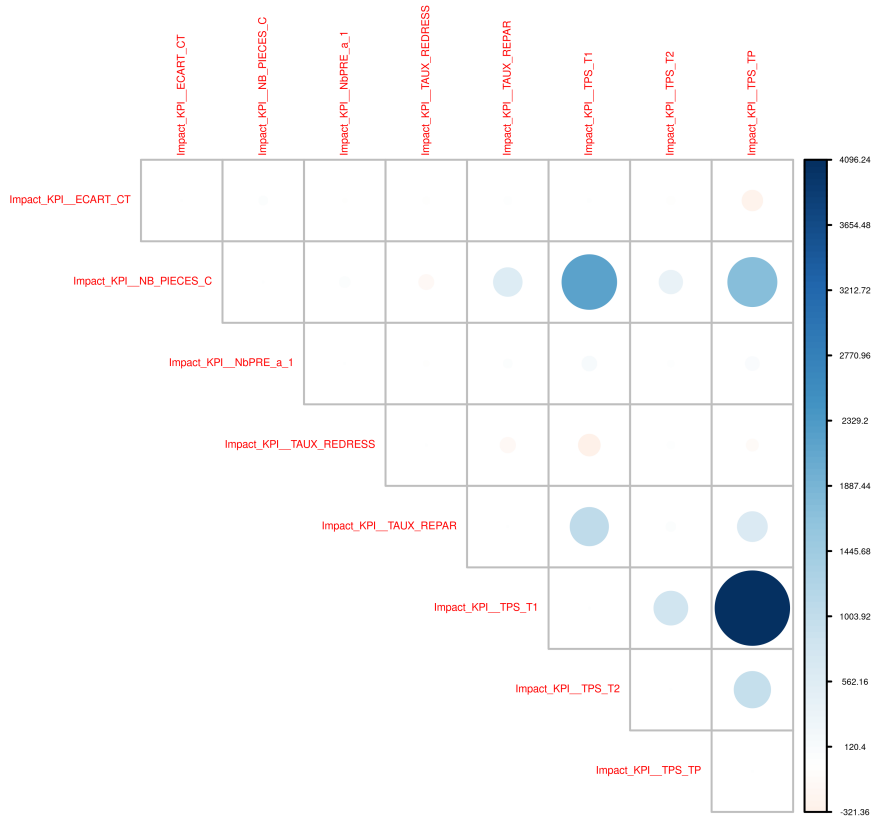


FIGURE 3.9 – Covariance des variables endogènes

Dans une équation de régression, un effet d'interaction est représenté comme le produit de deux variables indépendantes ou plus. La forme du modèle actuel est une équation de régression sans interaction :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k, \tag{3.10}$$

avec Y la variable dépendante, X_k les variables indépendantes et β_k les coefficients de régression. En modélisant un effet d'interaction entres les variables X_1 et X_2, l'équation devient :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + ... + \beta_k X_k, \tag{3.11}$$

où β_3 est un coefficient de régression et $X_1 X_2$ est l'interaction. Il s'agit ici d'interactions du premier ordre. Il est possible d'intégrer des interactions d'ordre supérieur telle que $\beta_4 X_1 X_2 X_3$ par exemple pour une interaction du second ordre, mais cela complique considérablement le modèle et sa compréhension.

Afin d'intégrer dans le modèle dans la notion de dépendance entres les variables, nous modifions le modèle en y intégrant les paramètres d'interaction entre les variables. Pour chaque interaction, le modèle va estimer un coefficient, qui indique l'accroissement de la valeur d'une variable relativement à la variation d'une autre.

Nous choisissons d'intégrer dans le modèle les interactions entre les indicateurs de performance et les variables exogènes ayant le plus d'impact, soit les variables `Reparateur__MONT_TH_moy` et `Sinistre__Gravite`. Ce qui nous amène à ajouter 10 paramètres supplémentaires au modèle (identifiés par le signe : entre le nom des modalités). La pénalisation Lasso est conservée et s'applique à ces nouvelles variables.

Après l'introduction des interactions, nous créons un autre modèle qui restreint les variables endogènes

aux 5 variables correspondantes aux indicateurs de performance retenus. Nous allons alors pouvoir faire une comparaison entres les quatre modèles ainsi créés.

Pour les modèles `original` et `penalisation`, la forme est :

```
## CoutTotal__MONT_EXPERT_HT ~ Sinistre__Gravite + Contrat__Franchise +
##     Reparateur__TYP_REP + Sinistre__COD_DOM_TYP + Sinistre__GROUP_SINI +
##     Vehicule__SEGMENT + Reparateur__MONT_TH_moy + Vehicule__VALOCAR +
##     KPI__NB_PIECES_C + KPI__TAUX_REPAR + KPI__TAUX_REDRESS +
##     KPI__ECART_CT + KPI__TPS_T1 + KPI__TPS_T2 + KPI__TPS_TP +
##     KPI__NbPRE_a_1
```

Pour le modèle `interaction` la forme est :

```
## CoutTotal__MONT_EXPERT_HT ~ Sinistre__Gravite + Contrat__Franchise +
##     Reparateur__TYP_REP + Sinistre__COD_DOM_TYP + Sinistre__GROUP_SINI +
##     Vehicule__SEGMENT + Reparateur__MONT_TH_moy + Vehicule__VALOCAR +
##     KPI__NB_PIECES_C + KPI__TAUX_REPAR + KPI__TAUX_REDRESS +
##     KPI__ECART_CT + KPI__TPS_T1 + KPI__TPS_T2 + KPI__TPS_TP +
##     KPI__NbPRE_a_1 + Reparateur__MONT_TH_moy:KPI__ECART_CT +
##     Reparateur__MONT_TH_moy:KPI__TAUX_REPAR + Reparateur__MONT_TH_moy:KPI__NB_PIECES_C +
##     Reparateur__MONT_TH_moy:KPI__TPS_TP + Reparateur__MONT_TH_moy:KPI__NbPRE_a_1 +
##     Sinistre__Gravite:KPI__ECART_CT + Sinistre__Gravite:KPI__TAUX_REPAR +
##     Sinistre__Gravite:KPI__NB_PIECES_C + Sinistre__Gravite:KPI__TPS_TP +
##     Sinistre__Gravite:KPI__NbPRE_a_1
```

Pour le modèle `interaction_reduit` la forme est :

```
## CoutTotal__MONT_EXPERT_HT ~ Sinistre__Gravite + Contrat__Franchise +
##     Reparateur__TYP_REP + Sinistre__COD_DOM_TYP + Sinistre__GROUP_SINI +
##     Vehicule__SEGMENT + Reparateur__MONT_TH_moy + Vehicule__VALOCAR +
##     KPI__ECART_CT + KPI__TAUX_REPAR + KPI__NB_PIECES_C + KPI__TPS_TP +
##     KPI__NbPRE_a_1 + Reparateur__MONT_TH_moy:KPI__ECART_CT +
##     Reparateur__MONT_TH_moy:KPI__TAUX_REPAR + Reparateur__MONT_TH_moy:KPI__NB_PIECES_C +
##     Reparateur__MONT_TH_moy:KPI__TPS_TP + Reparateur__MONT_TH_moy:KPI__NbPRE_a_1 +
##     Sinistre__Gravite:KPI__ECART_CT + Sinistre__Gravite:KPI__TAUX_REPAR +
##     Sinistre__Gravite:KPI__NB_PIECES_C + Sinistre__Gravite:KPI__TPS_TP +
##     Sinistre__Gravite:KPI__NbPRE_a_1
```

3.6.2 Analyse des interactions

Possédant désormais des modèles intégrant l'interaction entre les variables exogènes et endogènes, la valeur des coefficients peut être observée afin de mesurer le degré de dépendance. La table 3.12 donne les coefficient pour les modèles avec interaction.

La valeur `estimate` correspond à $\exp(\beta_i) - 1$, soit le coefficient de variation entre les deux variables. On constate que de nombreuses valeurs des coefficients sont très proches de zéro, traduisant ainsi une faible dépendance entre les variables.

3.6.3 Comparaison des résidus entre les modèles

L'examen de la densité des résidus sur la figure 3.10 montre que la dispersion est minimale pour le modèle `interaction`. Le modèle `interaction_reduit` présente une dispersion supérieure à celle

TABLE 3.12 – Niveaux d'interaction entre les variables

term	estimate
Reparateur___MONT_TH_moy_KPI___ECART_CT	0.0000000
Reparateur___MONT_TH_moy_KPI___TAUX_REPAR	0.0000000
Reparateur___MONT_TH_moy_KPI___NB_PIECES_C	0.0000000
Reparateur___MONT_TH_moy_KPI___TPS_TP	0.0000000
Reparateur___MONT_TH_moy_KPI___NbPRE_a_1	-0.0000002
Sinistre___Gravite_KPI___ECART_CT	-0.0001231
Sinistre___Gravite_KPI___TAUX_REPAR	0.0002383
Sinistre___Gravite_KPI___NB_PIECES_C	-0.0000952
Sinistre___Gravite_KPI___TPS_TP	-0.0000795
Sinistre___Gravite_KPI___NbPRE_a_1	0.0000000

du modèle `original`. Toutefois, il y a 0.4% des résidus avec une valeur supérieure à 2, ce qui reste acceptable.

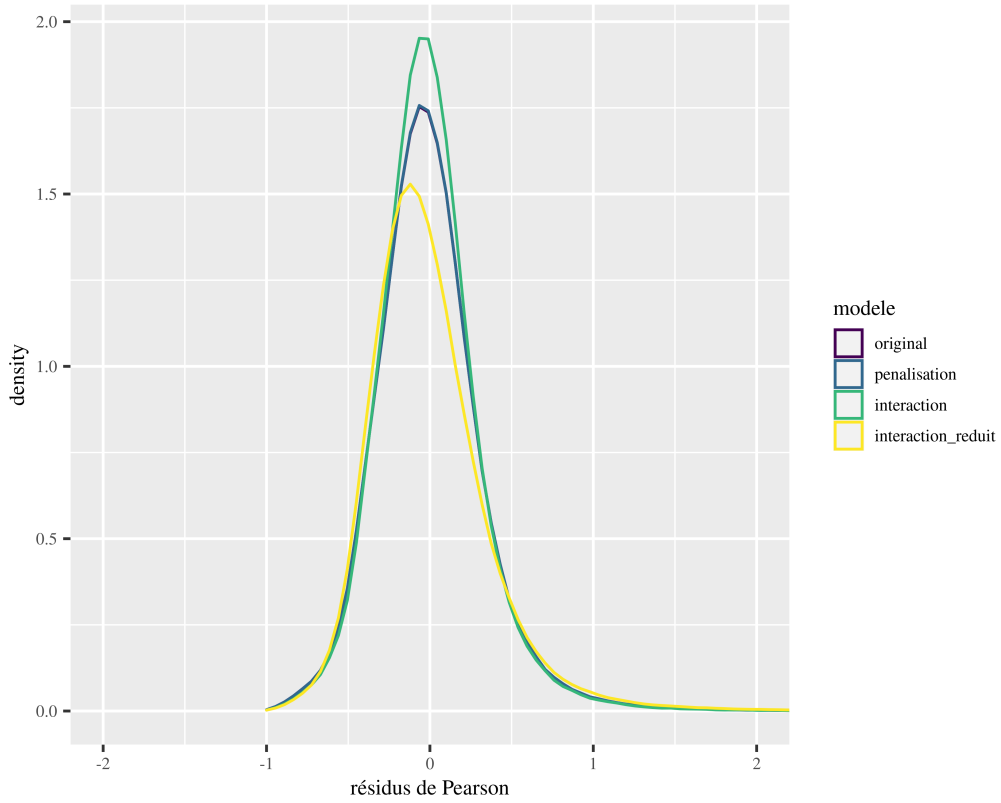


FIGURE 3.10 – Comparaison des résidus entre les différents modèles

3.6.4 Comparaison de performance des modèles

Afin de comparer les modèles entre eux, la métrique RMSE * est utilisée. Elle mesure la moyenne des carrés des erreurs ou des écarts. Elle se calcule de la sorte :

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\mu}_i)^2}, \tag{3.12}$$

∗. Root Mean Square Error

avec N le nombre d'observations, $\hat{\mu}_i$ la valeur prédite pour y_i. Plus les valeurs sont faibles, meilleure est la performance du modèle.

La table 3.13 montre les niveaux de performance mesurés. L'ajout de la pénalisation dans le modèle `original` n'améliore pas la performance – *si ce n'est de façon relative en ayant 7 modalités en moins.* On note une amélioration sensible par l'ajout d'interaction et la suppression des variables endogènes non utilisées. Le modèle `interaction_reduit` est le plus performant sur ce critère.

La table 3.14 compare la qualité de l'information dans le modèle. On voit que l'ajout d'interaction augmente la part de la variance intra-classe dans le modèle. La part d'information pertinente dans la variance intra-classe augmente.

TABLE 3.13 – Comparaison de la performance des modèles

modele	rmse	lambda
interaction_reduit	1275	0.0051
interaction	132321	0.0001
original	189012	0.0000
penalisation	196805	0.0001

TABLE 3.14 – Comparaison de qualité de l'information entre modèles

Indicateur	interaction	interaction_reduit	original	penalisation
Var_Inter	34738.06	33994.01	44832.37	45054.13
Var_Intra	47293.66	48037.72	37199.35	36977.59
PV_intra	0.58	0.59	0.45	0.45
PIP	0.32	0.32	0.25	0.25

TABLE 3.15 – Décomposition de la variance des modèles

Statistique	interaction	interaction_reduit	original	penalisation
var(CoutTotal___MONT_EXPERT_HT)	82032	82032	82032	82032
-var(delta_Exo)	-15150	-15380	-9358	-9275
-var(CM_dec_Exo)	-34738	-33994	-44832	-45054
-2 * cov(delta_Exo, CM_dec_Exo)	-32143	-32658	-27841	-27702
Total	0	-0	0	0

3.6.5 Impact sur la variance

On constate que le modèle `interaction_reduit` est celui qui présente la plus faible variance du coût moyen épuré des facteurs exogènes (`CM_dec_Exo`, ou la variance inter-classe). Cette réduction de la variance inter-classe s'obtient au prix d'une augmentation de la covariance, mais largement inférieure à l'augmentation de la variance des impacts des effets exogènes `delta_Exo`, ce qui signifie que le modèle contient plus d'information pertinente : 32% de la variance intra-classe (indicateur `PIP`). La quantité d'information totale reste à des niveaux identiques pour les quatre modèles.

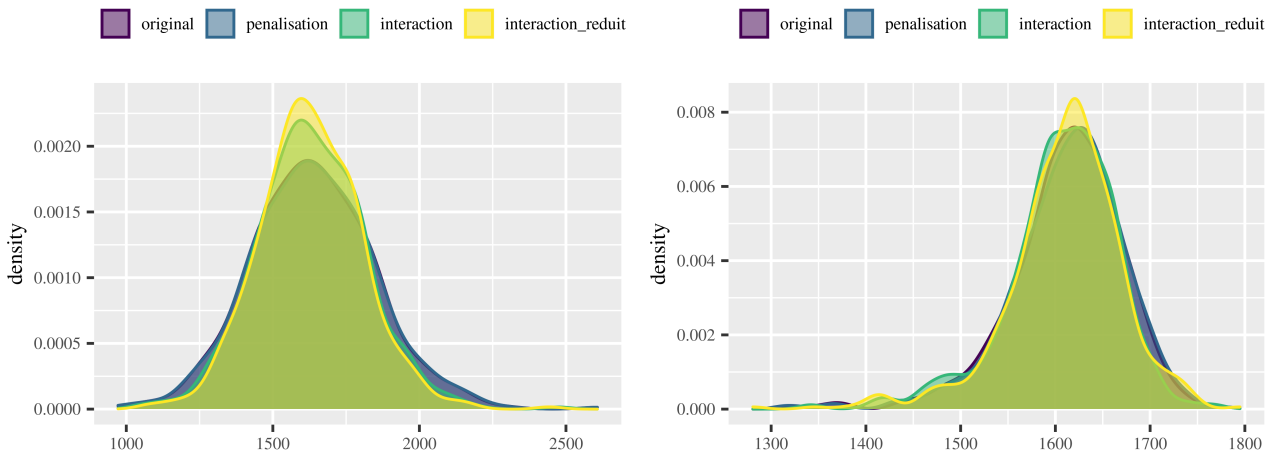
On peut conclure que le modèle `interaction_reduit` capte plus d'information pertinente pour notre démarche : réduire la variance entre les coûts moyen des experts en supprimant l'effet des facteurs exogènes.

3.6.6 Vue d'ensemble des modalités

Le graphe de la figure 3.12 représente l'impact de chacune des modalités des variables exogènes. On note que le modèle `interaction_reduit` est celui qui compte le moins de modalités, tout en leur octroyant une contribution plus importante.

TABLE 3.16 – Quantité d'information totale dans les modèles

Indicateur	interaction	interaction_reduit	original	penalisation
Quantite_Information	0.9600	0.9535	0.9573	0.9568
mean(CM_dec_ALL)/mean(CoutTotal___MONT_EXPERT_HT) - 1	-0.0293	-0.0261	-0.0261	-0.0260
skewness(CM_dec_ALL)/skewness(CoutTotal___MONT_EXPERT_HT) - 1	-2.7331	-2.5951	-2.9171	-2.9548



(a) CM épuré des facteurs exogènes

(b) CM épuré des facteurs exogènes et endogènes

FIGURE 3.11 – Comparaison des densités des variables de coûts entre les différents modèles

L'ajout d'interactions augmente l'impact des variables endogènes. Dans le modèle `interaction_reduit`, on constate que les cinq variables retenues reprennent la contribution des variables supprimées.

3.7 Comparaison des classifications

Avec le modèle `interaction_reduit` une nouvelle classification est réalisée. Elle est comparée avec la classification établie précédemment avec le modèle `original`. L'examen de la figure 3.14 montre que le second modèle modifie la classification initiale, essentiellement en terme d'amplitude, mais ne modifie très peu l'ordre du classement. Les chiffres de la table 3.17 montrent que peu de changement dans le nombre de sur-classement et de déclassements interviennent. Toutefois cette nouvelle classification peut être considérée comme plus pertinente au vu de l'amélioration de la qualité du modèle.

Statistique	interaction_reduit	original
Part_experts_surclasses	0.46	0.45
Surclasse_max	151	111
Surclasse_moy	29.00	23.60
Part_experts_declasses	0.52	0.52
Declasse_max	-213	-187
Declasse_moy	-35.00	-29.30
Part_experts_stables	0.01	0.03

TABLE 3.17 – Comparaison des changement dans la classification

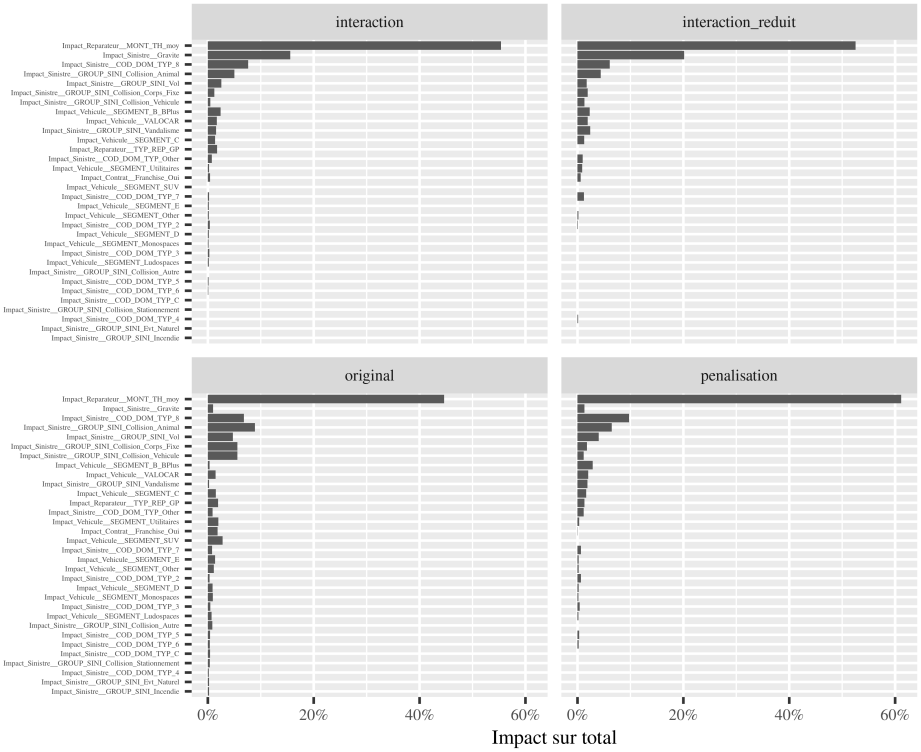


FIGURE 3.12 – Comparaison de l'impact des modalités entre les différents modèles

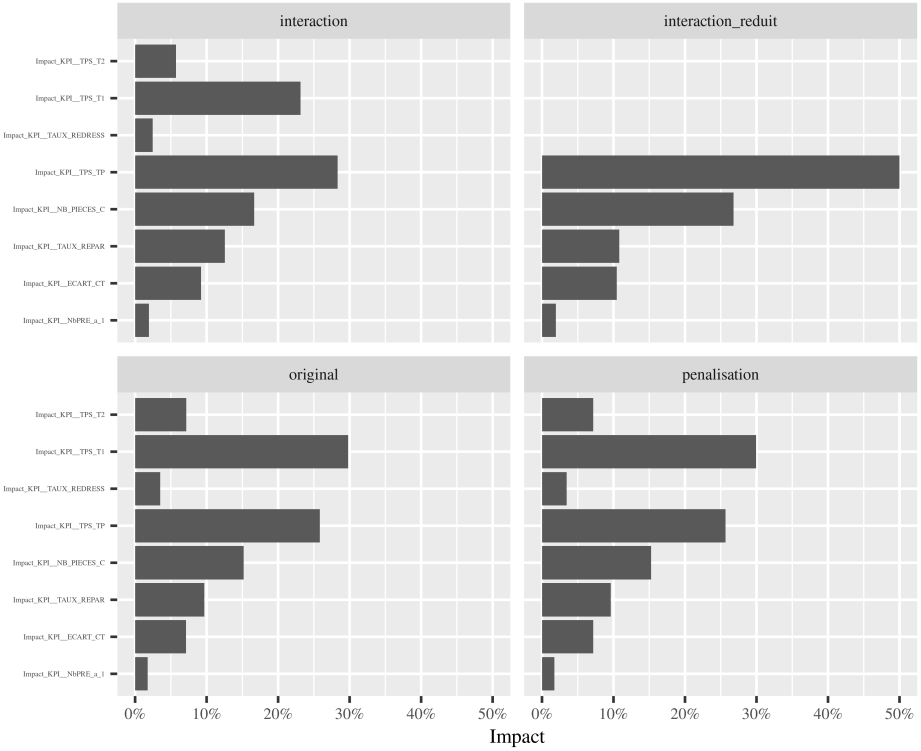


FIGURE 3.13 – Comparaison de l'impact des indicateurs entre les différents modèles

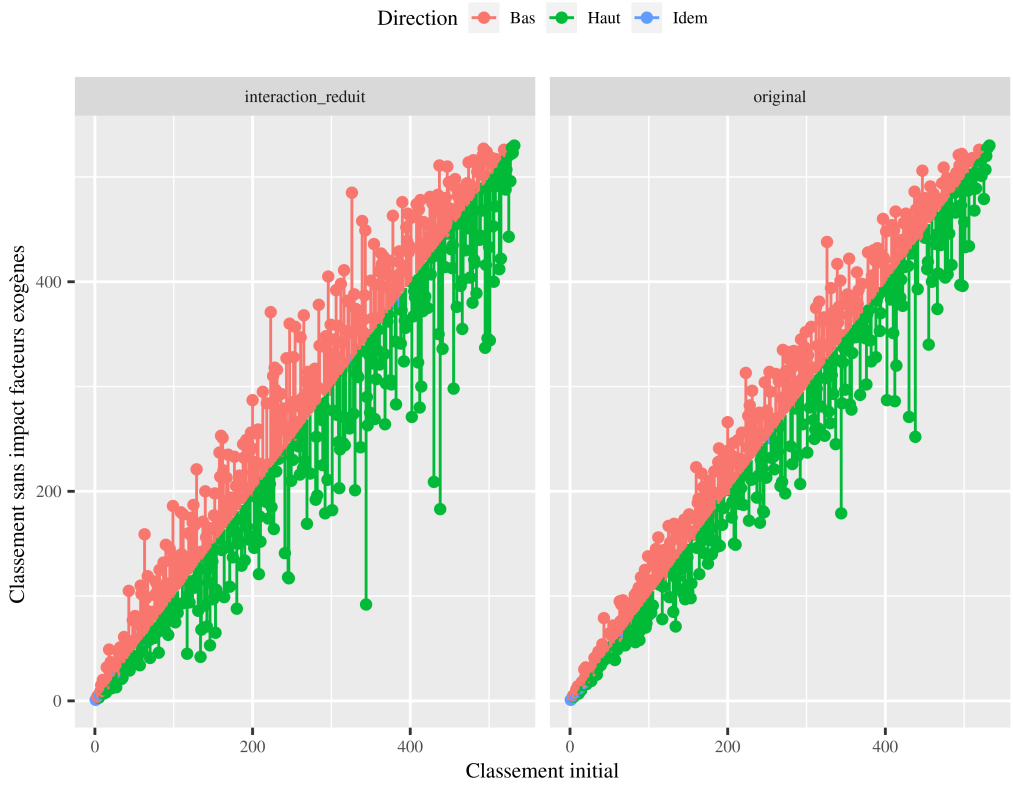


FIGURE 3.14 – Comparaison de la classification des experts

3.8 Synthèse

Nous avons décrit une modélisation via un GLM afin de réduire l'effet des facteurs exogènes sur la variable `CoutTotal__MONT_EXPERT_HT`. Pour cela, nous soustrayons, pour chaque expert, l'impact des facteurs exogènes sur la variable dépendante. Cela nous permet d'obtenir une nouvelle classification des experts qui reflète de façon plus pertinente leur performance. Afin de contrôler la pertinence du modèle, nous essayons de tendre vers une variable à variance nulle en soustrayant l'impact de l'ensemble des variables, exogènes et endogènes. La réduction de variance s'élève à 95%, mais avec une valeur de covariance non nulle. Cette colinéarité peut perturber les indicateurs de performance qui servent au pilotage des experts.

Nous avons recherché des pistes d'amélioration du modèle. D'abord nous avons appliqué une pénalisation Lasso qui réduit la valeur des coefficients β, jusqu'à leur assigner la valeur zéro pour les variables les moins contributrices. Nous obtenons ainsi un modèle avec le même pouvoir explicatif, mais avec 7 modalités en moins.

Souhaitant modéliser la dépendance entre les variables exogènes et endogènes, nous avons ajouté des interactions dans le modèle. Dans le nouveau modèle ainsi obtenu, les coefficients d'interactions entre ces deux familles de variables, montrent qu'il y a peu de dépendance entre elles. La prise en compte de ces interactions améliore la performance du modèle mesurée par la statistique du $RMSE$.

Enfin, nous avons sélectionné un nombre limité de variables endogènes pour servir d'objectifs aux experts. Leur choix a été dicté par le souhait d'éviter de la redondance avec des indicateurs colinéaires. La réduction du nombre de variables améliore significativement la performance du modèle.

Conclusion

Dans le cadre du pilotage des experts en automobile, il est apparu nécessaire de décomposer le coût de sinistre en deux parties : d'un côté l'impact de facteurs exogènes - qui sont indépendants de l'action de l'expert - et d'un autre côté les facteurs endogènes, qui traduisent la performance de l'expert dans la maîtrise des coûts. L'étude a été menée sur un jeu de données de plus de 500 000 expertises réalisées en 2019 par 532 experts. L'objectif de la démarche vise à obtenir une classification des experts sur la base du montant des expertises réalisées qui soit épurée des facteurs exogènes.

S'agissant d'une problématique de pilotage opérationnel, la classification obtenue servant à objectiver des individus, il était nécessaire de disposer d'un modèle compréhensible par le plus grand nombre et offrant un niveau de détail suffisant pour expliciter les résultats. Le choix de la modélisation s'est porté sur un modèle linéaire généralisé (GLM), car, dans son principe, il est intelligible, et permet de quantifier l'impact de chaque modalité sur la variable d'intérêt.

La démarche a d'abord été de préparer le jeu de données, en procédant au traitement des valeurs manquantes et aberrantes et des modalités pour les variables catégorielles. Ces opérations induisent une perte de 7% du volume de données totales, ce qui est raisonnable et assure un représentativité de l'échantillon utilisé. Cette étape a permis de classifier les variables en deux catégories : celles considérées comme exogènes et celles dépendantes de l'action de l'expert, dites endogènes.

A partir des coefficients du modèle GLM, nous supprimons l'impact des variables exogènes sur le coût moyen de chaque expert. Nous obtenons ainsi une nouvelle classification. Nous utilisons la chute de variance constatée comme indicateur de pertinence du modèle. La chute de variance obtenue est supérieure à 95% ce qui signifie que les variables utilisées possèdent un bon pouvoir explicatif. Les variables exogènes quant à elles créent une chute de variance de 57% ce qui gage d'une quantité significative d'information pertinente pour notre démarche.

Face à la présence de dépendance entre certaines variables endogènes, on utilise une sélection de type Lasso afin de réduire le nombre de variables utilisées. Sept variables sont ainsi supprimées tout en conservant le même niveau explicatif du modèle. Nous avons ajouté des interactions dans le modèle afin de modéliser la dépendance entre variables. Nous avons pu comparer ainsi le modèle à trois autres modèles. La performance du modèle est largement améliorée dans le cadre de l'utilisation d'interaction et d'un nombre réduit de variables endogènes.

Ces travaux permettent d'obtenir une classification des experts plus pertinente, avec une influence minimisée des facteurs exogènes. On obtient ainsi une appréciation plus fine de la performance des experts. La modélisation choisie permet aussi de donner une vue analytique en donnant les niveaux d'impacts de des différentes variables et de leurs modalités sur le coût moyen de l'expert.

Des améliorations peuvent être apportées à la modélisation. Sachant que le GLM inclut dans l'intercept la première des modalités d'une variable, son impact n'est pas calculé et n'est pas déduit du coût moyen. La prise en compte de ces modalités améliorerait la suppression des effets exogènes. Le problème ne se

pose pas pour les variables endogènes qui sont quantitatives.

Les données utilisées sont issues de l'ensemble des expertises d'une année civile. Étendre le domaine d'étude à plusieurs années apporterait plus d'information, et notamment lisserait l'impact de l'expert sur le coût moyen visant à réduire la dépendance entre facteurs exogènes et endogènes. Cela nécessiterait toutefois de prendre en compte les variations de saisonalité d'une année sur l'autre ainsi que de redresser l'inflation des coûts annuelle.

La prise en compte de données d'environnement de l'expert, telles que la productivité, les secteurs géographiques, le niveau de formation et d'ancienneté d'exercice, donnerait des clefs de lecture intéressantes pour le management des experts.

Enfin ces travaux donnent une vue statique en étudiant la performance sur une période temporelle fixe. Une modélisation des facteurs d'inflation de coûts serait une suite logique des ces travaux et répondrait à une forte demande des assureurs et dirigeants amenés à piloter des experts en automobile de plus en plus impliqués dans la maîtrise de la charge sinistre des compagnies d'assurance.

Bibliographie

Abadie, A. & Chabrier, B. (2021), 'Classement auto-MRH 2021 : entre solidarité et conquête commerciale, les assureurs résistent'.
URL: *https://www.argusdelassurance.com/assurance-dommages/habitation/classement-auto-mrh-2021-entre-solidarite-et-conquete-commerciale-les-assureurs-resistent.182174*

Breiman, L. (1996), 'Bagging predictors', *Machine Learning 1996 24 :2* **24**(2), 123–140.
URL: *https ://link.springer.com/article/10.1007/BF00058655*

Colas, M., Loisel, A. & Deng, B. (2021), 'Communication statistique SRA janvier 2021'.
URL: *https://www.sra.asso.fr/statistiques/Communication*

de Lignaud de Lussac, M. (2018), Comparaison de modèles prédictifs pour l'évaluation des coûts matériels automobiles, Mémoire d'actuariat, Université Paris Dauphine.

Denuit, M. & Charpentier, A. (2005), *Mathematiques de l'Assurance Non-Vie. Tome II : Tarification et Provisionnement*, Vol. II, Economica.

Durand, E. (2014), 'Gestion de sinistres : Allianz France va travailler avec un réseau d'experts en automobile resserré'.
URL: *https://www.argusdelassurance.com/acteurs/gestion-de-sinistres-allianz-france-va-travailler-avec-un-reseau-d-experts-en-automobile-resserre.85855*

Ebali, H. (2011), Impact du mode de gestion des sinistres sur l'optimisation des coûts IRD, Mémoire d'actuariat, ISUP.

Gourieroux, C. (1999), *Statistique de l'assurance*, Economica.

Harrell, F. E. (2001), 'Resampling, Validating, Describing, and Simplifying the Model', *Regression Modeling Strategies* pp. 87–103.

Khougea, D. (2019), Tarification IARD avec des modèles de régression avancés, Mémoire d'actuariat, Université de strasbourg.

Kuhn, M. & Johnson, K. (2013), *Applied predictive modeling*, springer edn, Springer New York.

Kuhn, M. & Wickham, H. (2022), 'recipes : Preprocessing and Feature Engineering Steps for Modeling'.
URL: *https://recipes.tidymodels.org/*

Legifrance (2001), 'Article 1 - Loi n°72-1097 du 11 décembre 1972 relative a l'organisation de la profession d'expert en automobile - Légifrance'.
URL: *https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000006840069/1990-07-01*

Namin, L. (2009), *L'expertise automobile : Aspects juridiques et pratiques de la profession.*

Namin, L. (2017), 'Réparateurs contre experts en automobile'.
URL: *https://pro.largus.fr/actualites/reparateur-vs-experts-en-automobile-8467387.html*

Nelder, J. A. & Wedderburn, R. W. M. (1972), 'Generalized Linear Models', *Journal of the Royal Statistical Society : Series A (General)* **135**(3), 370–384.

{R Core Team} (2022), 'R : A Language and Environment for Statistical Computing'.
URL: https://www.R-project.org/

Wabo Foka, A. (2020), Mesure de l'écart de performance entre deux réseaux d'experts en assurance automobile, Mémoire d'actuariat, Université Paris Dauphine.

Wickham, H., Averick, M., Bryan, J. & Chang, W. (2019), 'Welcome to the {tidyverse}', *Journal of Open Source Software* .

Annexes

A Marché de l'assurance automobile en 2021

Chiffres issus du classement auto-MRH 2021 de l'argus de l'assurance (Abadie & Chabrier (2021)). Quand la part auto est absente, le chiffre d'affaire est calculé avec une part de 34%.

Assureur	CA_2020	CA_2019	evol_CA	Part_Auto	Nb_Contrats	evol_Contrats	CA_Auto_2020
Covea	3822.0	3918.0	-0.025	0.259	9856838	0.004	989.9
Axa	2438.0	2394.0	0.018	0.092	4145244	-0.001	224.3
Groupe Macif	2105.0	2022.0	0.041	0.611	6138352	0.015	1286.2
Groupama	1926.1	1891.2	0.018		4061617	0.001	654.9
Allianz	1914.0	1896.0	0.009		4208445	0.013	650.8
Groupe Maif	1394.0	1420.0	-0.018	0.470	3497761	0.010	655.2
Credit Agricole Assurances	1336.0	1257.0	0.063		3185180	0.027	454.2
Groupe Ass. du Credit mutuel	1125.0	1067.0	0.054	0.120	2869901	0.032	135.0
Generali	1073.0	1053.0	0.019	0.370	2542000	0.022	397.0
Groupe Matmut	1053.9	1025.2	0.028	0.461	2814175	0.015	485.8
BPCE	569.4	540.8	0.053	0.344	1272229	0.067	195.9
Aviva	343.3	330.8	0.038		741222	0.019	116.7
Suravenir	187.1	180.5	0.037	0.449	573898	0.032	84.0
Areas Assurances	180.4	179.3	0.006	0.391	330007	0.025	70.5
Smacl Assurances	174.0	155.0	0.123	0.400	478000	0.115	69.6
Thelem	162.4	156.6	0.037	0.449	453640	0.030	72.9
Mutuelle de Poitiers	161.5	157.1	0.028	0.391	435736	0.026	63.1
Groupe Macsf	156.5	153.4	0.020	0.238	338512	0.015	37.2
Société Generale	152.0	135.0	0.126	0.200	372916	0.349	30.4
La Banque Postale Assurances	100.5	101.6	-0.011	0.272	263862	-0.041	27.3

B Statistiques descriptives de la variable `CoutTotal__MONT_EXPERT_HT`

Avant écretage :

```
## summary statistics
## ------
## min:  0   max:  107347.1
## median:  1191.667
## mean:  1649.844
## estimated sd:  1587.246
## estimated skewness:  5.436015
## estimated kurtosis:  121.804
```

Après écretage à 15K€

```
## summary statistics
## ------
## min:  0   max:  14990.96
## median:  1190.65
## mean:  1632.518
## estimated sd:  1456.35
## estimated skewness:  2.787091
## estimated kurtosis:  15.1878
```

C Statistiques descriptives des données avant traitement

C.1 Variables liées au sinistre

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
1	Sinistre___COD_DOM_TYP	1. 7	105152 (20.8%)	39
	[factor]	2. 8	98783 (19.5%)	(0.0%)
		3. 1	51976 (10.3%)	
		4. 4	50354 (9.9%)	
		5. 2	48916 (9.7%)	
		6. 3	41993 (8.3%)	
		7. 5	38093 (7.5%)	
		8. 6	34525 (6.8%)	
		9. C	30606 (6.0%)	
		10. B	3340 (0.7%)	
		[5 others]	2830 (0.6%)	
2	Sinistre___CP_MISSION	1. 51100	2156 (0.4%)	0
	[factor]	2. 91420	1750 (0.3%)	(0.0%)
		3. 75015	1562 (0.3%)	
		4. 25000	1499 (0.3%)	
		5. 93400	1499 (0.3%)	
		6. 76000	1493 (0.3%)	
		7. 66000	1441 (0.3%)	
		8. 30900	1411 (0.3%)	
		9. 82000	1403 (0.3%)	
		10. 69400	1388 (0.3%)	
		[5218 others]	491005 (96.9%)	
3	Sinistre___Gravite	Mean (sd) : 70.1 (197.9)	2713 distinct values	38588
	[numeric]	min < med < max :		(7.6%)
		0 < 26 < 6070		
		Q1 - Q3 : 6 - 64		
4	Sinistre___GROUP_SINI	1. Collision_Vehicule	347763 (68.6%)	0
	[factor]	2. Collision_Corps_Fixe	91794 (18.1%)	(0.0%)
		3. Collision_Animal	24179 (4.8%)	
		4. Vandalisme	16842 (3.3%)	
		5. Vol	9846 (1.9%)	
		6. Collision_Autre	7567 (1.5%)	
		7. Autres	4010 (0.8%)	
		8. Collision_Stationnement	3000 (0.6%)	
		9. Evt_Naturel	830 (0.2%)	
		10. Incendie	776 (0.2%)	
5	Sinistre___LIB_GAR	1. Tous risques	318683 (65.5%)	20080
	[factor]	2. Défense recours	67246 (13.8%)	(4.0%)
		3. Autre	52502 (10.8%)	
		4. Responsabilité civile	16724 (3.4%)	
		5. Avance sur recours	15145 (3.1%)	
		6. Vol	14184 (2.9%)	
		7. Incendie	732 (0.2%)	
		8. Tierce collision	342 (0.1%)	
		9. Bris de glace	332 (0.1%)	
		10. Force de la nature	326 (0.1%)	
		[9 others]	311 (0.1%)	
6	Sinistre___NUM_DEPART_SECT	1. 13	16995 (3.4%)	13
	[factor]	2. 59	15014 (3.0%)	(0.0%)
		3. 33	14946 (3.0%)	
		4. 69	14558 (2.9%)	
		5. 77	12242 (2.4%)	
		6. 91	11806 (2.3%)	
		7. 44	11396 (2.2%)	
		8. 6	11117 (2.2%)	
		9. 95	10328 (2.0%)	
		10. 31	10099 (2.0%)	
		[86 others]	378093 (74.6%)	

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
7	Sinistre___VAL_ORIENT_CHOC [factor]	1. 0	131295 (25.9%)	75 (0.0%)
		2. 180	98857 (19.5%)	
		3. 45	77805 (15.4%)	
		4. 315	71037 (14.0%)	
		5. 225	37796 (7.5%)	
		6. 90	32565 (6.4%)	
		7. 270	29204 (5.8%)	
		8. 135	27973 (5.5%)	

C.2 Variables liées au véhicule

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
1	Vehicule___AGE_MOIS_VEH [integer]	Mean (sd) : 73.1 (56.7) min < med < max : 0 < 62 < 1206 Q1 - Q3 : 28 - 106	638 distinct values	247 (0.0%)
2	Vehicule___COD_COND_ACHAT [factor]	1. 1 2. 3 3. 2	449246 (88.7%) 41485 (8.2%) 15799 (3.1%)	77 (0.0%)
3	Vehicule___Cod_Ener [factor]	1. G 2. E 3. C 4. I 5. Q 6. K 7. M 8. R 9. L 10. D 11. Z	346545 (68.4%) 145997 (28.8%) 8943 (1.8%) 2476 (0.5%) 1006 (0.2%) 1003 (0.2%) 316 (0.1%) 147 (0.0%) 71 (0.0%) 24 (0.0%) 3 (0.0%)	76 (0.0%)
4	Vehicule___COD_MARQUE [factor]	1. 6 2. 5 3. 1 4. 27 5. 24 6. 23 7. 22 8. 21 9. 77 10. 90 [82 others]	93919 (18.5%) 89894 (17.7%) 59553 (11.8%) 40135 (7.9%) 22734 (4.5%) 21532 (4.3%) 20330 (4.0%) 18825 (3.7%) 16646 (3.3%) 16382 (3.2%) 106657 (21.1%)	0 (0.0%)
5	Vehicule___COD_MARQUE_MODELE [factor]	1. 6_48 2. 5_78 3. 27_16 4. 5_48 5. 6_68 6. 77_8 7. 27_18 8. 6_49 9. 1_27 10. 5_98 [1464 others]	16534 (3.3%) 16178 (3.2%) 9082 (1.8%) 7593 (1.5%) 7428 (1.5%) 7064 (1.4%) 6989 (1.4%) 6984 (1.4%) 6787 (1.3%) 6424 (1.3%) 415544 (82.0%)	0 (0.0%)
6	Vehicule___SEGMENT [factor]	1. B_BPlus 2. C 3. SUV 4. A_B1_B2 5. Utilitaires 6. D 7. E 8. Ludospaces 9. Monospaces 10. Coupe_Cabriolet [3 others]	130345 (25.7%) 126870 (25.1%) 106285 (21.0%) 37470 (7.4%) 26415 (5.2%) 24355 (4.8%) 18585 (3.7%) 12613 (2.5%) 11280 (2.2%) 6597 (1.3%) 5486 (1.1%)	306 (0.1%)
7	Vehicule___VAL_KM_HORO_COMPTEUR [integer]	Mean (sd) : 94248.3 (84496.8) min < med < max : 0 < 79889.5 < 9644175 Q1 - Q3 : 35915 - 136810	205636 distinct values	5553 (1.1%)
8	Vehicule___VAL_PUIS_DIN [integer]	Mean (sd) : 115.2 (45.2) min < med < max : 0 < 110 < 899 Q1 - Q3 : 90 - 131	349 distinct values	245 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
9	Vehicule___VAL_PUIS_FISC_EXP [factor]	1. 5 2. 6 3. 4 4. 7 5. 8 6. 9 7. 10 8. 11 9. 15 10. 12 [51 others]	126238 (25.0%) 103025 (20.4%) 94421 (18.7%) 61031 (12.1%) 48921 (9.7%) 19702 (3.9%) 17597 (3.5%) 7098 (1.4%) 4529 (0.9%) 4254 (0.8%) 19021 (3.8%)	770 (0.2%)
10	Vehicule___VALOCAR [integer]	Mean (sd) : 7539.1 (6170.9) min < med < max : 0 < 6413 < 151931 Q1 - Q3 : 3295 - 10332	30519 distinct values	0 (0.0%)

C.3 Variables de temps de main d'oeuvre

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
1	Temps___TPS_T1 [numeric]	Mean (sd) : 2.9 (2.6) min < med < max : 0 < 2.2 < 320 Q1 - Q3 : 1.5 - 3.5	392 distinct values	0 (0.0%)
2	Temps___TPS_T2 [numeric]	Mean (sd) : 3.1 (3.9) min < med < max : 0 < 2 < 205 Q1 - Q3 : 0 - 4	435 distinct values	0 (0.0%)
3	Temps___TPS_T3 [numeric]	Mean (sd) : 0.1 (0.7) min < med < max : 0 < 0 < 50 Q1 - Q3 : 0 - 0	178 distinct values	0 (0.0%)
4	Temps___TPS_TP [numeric]	Mean (sd) : 4.5 (3) min < med < max : 0 < 4 < 40 Q1 - Q3 : 2.4 - 6	272 distinct values	0 (0.0%)

C.4 Variables de tarifs de main d'oeuvre

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
1	Reparateur___MONT_TAR_IP [numeric]	Mean (sd) : 37.8 (11.3) min < med < max : 0 < 33.4 < 858 Q1 - Q3 : 30.7 - 42.4	3506 distinct values	155 (0.0%)
2	Reparateur___MONT_TAR_T1 [numeric]	Mean (sd) : 56.1 (15.6) min < med < max : 0 < 50 < 189 Q1 - Q3 : 45 - 63	3903 distinct values	0 (0.0%)
3	Reparateur___MONT_TAR_T2 [numeric]	Mean (sd) : 58.1 (16.2) min < med < max : 0 < 51 < 900 Q1 - Q3 : 47 - 66	4097 distinct values	0 (0.0%)
4	Reparateur___MONT_TAR_T3 [numeric]	Mean (sd) : 62.5 (17.3) min < med < max : 0 < 58.7 < 936 Q1 - Q3 : 49 - 70	4608 distinct values	0 (0.0%)
5	Reparateur___MONT_TAR_TP [numeric]	Mean (sd) : 58.2 (16.1) min < med < max : 0 < 51 < 603 Q1 - Q3 : 47 - 66.2	4067 distinct values	0 (0.0%)
6	Reparateur___TYP_REP [factor]	1. GP 2. GNP	283527 (56.0%) 223080 (44.0%)	0 (0.0%)

C.5 Variables liées au contrat

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
1	Contrat___Franchise [factor]	1. Non 2. Oui	302755 (59.8%) 203852 (40.2%)	0 (0.0%)

C.6 Variables endogènes

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
1	KPI___DNI [factor]	1. Non 2. Oui	480658 (94.9%) 25949 (5.1%)	0 (0.0%)
2	KPI___ECART_CT [numeric]	Mean (sd) : Inf (NaN) min < med < max : -1 < 0 < Inf Q1 - Q3 : 0 - 0.2	359870 distinct values	516 (0.1%)
3	KPI___NB_PIECES_C [integer]	Mean (sd) : 1.4 (1.5) min < med < max : 0 < 1 < 32 Q1 - Q3 : 0 - 2	22 distinct values	9685 (1.9%)
4	KPI___NbPRE_a_1 [integer]	Min : 0 Mean : 0.1 Max : 1	0 : 469839 (94.5%) 1 : 27083 (5.5%)	9685 (1.9%)
5	KPI___TAUX_REDRESS [numeric]	Mean (sd) : 0.3 (0.3) min < med < max : 0 < 0.3 < 1 Q1 - Q3 : 0 - 0.6	8222 distinct values	751 (0.1%)
6	KPI___TAUX_REPAR [numeric]	Mean (sd) : 0.5 (0.4) min < med < max : 0 < 0.5 < 1 Q1 - Q3 : 0 - 1	120 distinct values	40843 (8.1%)
7	Temps___TPS_T1 [numeric]	Mean (sd) : 2.9 (2.6) min < med < max : 0 < 2.2 < 320 Q1 - Q3 : 1.5 - 3.5	392 distinct values	0 (0.0%)
8	Temps___TPS_T2 [numeric]	Mean (sd) : 3.1 (3.9) min < med < max : 0 < 2 < 205 Q1 - Q3 : 0 - 4	435 distinct values	0 (0.0%)
9	Temps___TPS_TP [numeric]	Mean (sd) : 4.5 (3) min < med < max : 0 < 4 < 40 Q1 - Q3 : 2.4 - 6	272 distinct values	0 (0.0%)

D P-valeur du modèle GLM original

TABLE 24 – Coefficients et P-Valeurs

	modalite	coefficients	P_valeur
1	Intercept	5.67	0.00
2	Vehicule___SEGMENT.Other	-0.03	0.00
3	Vehicule___SEGMENT.Picks-up	0.04	0.00
4	Vehicule___SEGMENT.Polyvalentes, monospaces citadins	-0.17	0.00
5	Vehicule___SEGMENT.SUV	-0.01	0.01
6	Vehicule___SEGMENT.Segment A ou B1	-0.15	0.00
7	Vehicule___SEGMENT.Segment B	-0.13	0.00
8	Vehicule___SEGMENT.Segment B plus	-0.10	0.00
9	Vehicule___SEGMENT.Segment C	-0.07	0.00
10	Vehicule___SEGMENT.Segment C	-0.07	0.00
11	Vehicule___SEGMENT.Segment D	-0.04	0.00
12	Vehicule___SEGMENT.Segment E	0.01	0.01
13	Vehicule___SEGMENT.Segment F	0.08	0.00
14	Vehicule___SEGMENT.Utilitaires	0.01	0.02
15	Vehicule___SEGMENT.ludospaces	-0.04	0.00
16	Vehicule___SEGMENT.monospaces	0.01	0.02
17	Sinistre___GROUP_SINI.Collision_Animal	0.28	0.00
18	Sinistre___GROUP_SINI.Collision_Autre	0.11	0.00
19	Sinistre___GROUP_SINI.Collision_Corps_Fixe	0.17	0.00
20	Sinistre___GROUP_SINI.Collision_Stationnement	0.10	0.00
21	Sinistre___GROUP_SINI.Collision_Vehicule	0.11	0.00
22	Sinistre___GROUP_SINI.Evt_Naturel	0.17	0.00
23	Sinistre___GROUP_SINI.Incendie	0.19	0.00
24	Sinistre___GROUP_SINI.Vandalisme	0.02	0.00
25	Sinistre___GROUP_SINI.Vol	0.40	0.00
26	Sinistre___COD_DOM_TYP.2	0.01	0.00
27	Sinistre___COD_DOM_TYP.3	-0.03	0.00
28	Sinistre___COD_DOM_TYP.4	-0.01	0.02
29	Sinistre___COD_DOM_TYP.5	-0.02	0.00
30	Sinistre___COD_DOM_TYP.6	-0.02	0.00
31	Sinistre___COD_DOM_TYP.7	-0.03	0.00
32	Sinistre___COD_DOM_TYP.8	0.16	0.00
33	Sinistre___COD_DOM_TYP.C	-0.02	0.01
34	Sinistre___COD_DOM_TYP.Other	0.17	0.00
35	Reparateur___TYP_REP.GP	-0.02	0.00
36	Contrat___Franchise.Oui	0.07	0.00
37	KPI___ECART_CT	0.35	0.00
38	KPI___NB_PIECES_C	0.11	0.00
39	KPI___NbPRE_a_1	-0.08	0.00
40	KPI___TAUX_REDRESS	0.18	0.00
41	KPI___TAUX_REPAR	-0.34	0.00
42	Sinistre___Gravite	0.00	0.00
43	KPI___TPS_T1	0.11	0.00
44	KPI___TPS_T2	0.03	0.00
45	KPI___TPS_TP	0.08	0.00
46	Vehicule___VALOCAR	0.00	0.00
47	Reparateur___MONT_TH_moy	0.01	0.00

E Imputation bagged tree

Les techniques dites d'ensemble sont apparues au début des années 1990. Le concept est d'utiliser plusieurs algorithmes d'apprentissage pour obtenir de meilleures prédictions. L'une des premières techniques d'ensemble a été le *Bagging* (contraction de *Bootstrap* et de *aggregation*) par Breiman (1996). On appelle bagged tree la technique ensembliste qui conjugue le boostrap avec un modèle d'arbre de régression. Plusieurs modèles sont entraînés sur des échantillons du jeu de données. Au final, le meta-modèle obtenu peut se voir comme la moyenne des modèles.

Cette technique est utilisée pour imputer les valeurs manquantes de la variable `Vehicule__VALOCAR` via la fonction `step_impute_bag`* du package *recipes*. Les arguments de la fonction sont :

```
step_impute_bag(Vehicule__VALOCAR,
                impute_with = imp_vars(Vehicule__AGE_MOIS_VEH, Vehicule__VAL_PUIS_DIN,
                Vehicule__VAL_PUIS_FISC_EXP, Vehicule__Cod_Ener, Vehicule__COD_MARQUE_MODELE,
                Vehicule__VAL_KM_HORO_COMPTEUR, Vehicule__COD_COND_ACHAT),
                trees = 25)
```

*. https://recipes.tidymodels.org/reference/step_impute_bag.html