Aix Marseille Université

Faculté d'Économie et de Gestion

Quels sont les déterminants de la gravité des accidents corporels en France ?

Michaela SORHO

Antoine PINTO

SOMMAIRE

Partie 1 : Modèle Linéaire

1. Introduction

Les décès causés par les accidents de la route avoisinent 1,25 millions de décès par année dans le monde. Pour réduire significativement ces accidents, la décennie 2011-2020 est proclamée « Décennie d'action pour la Sécurité Routière » par l'Organisation des Nations Unies (ONU).[1]

En France, les accidents corporels de la circulation sont à l'origine de 3 500 à 3 600 décès chaque année. Ces accidents font d'autres victimes notamment 75 000 blessés par an et 2 500 autres personnes qui en ressortent avec des lésions graves et irréversibles. En vue de ces chiffres, l'État français met en place chaque année des mesures qui visent à la réduction des accidents de la route donc à posteriori du nombre de victimes.

Pour veiller à l'amélioration de la sécurité routière, en 2012, les sanctions contre l'usage d'un téléphone ou d'un appareil à écran en conduisant deviennent plus lourdes. En Novembre 2017, un décret est mis en place autorisant les préfets départementaux à interdire la conduite lors de négligence ou refus du contrôle médical d'aptitude à la conduite dans le délai prescrit. On peut également citer en 2018 la limitation de vitesse à 80Km/h sur les routes hors agglomération, hors autoroute, sans séparation centrale. La même année, la France instaure une sanction pour le transport d'occupants en surnombre dans un véhicule.

Cependant il faut noter la constance dans les décès, liés aux accidents de la route, chaque année depuis 2012. Ceci attire notre attention sur les déterminants de la gravité des accidents routiers.

On caractérise comme étant un « accident corporel de la route » tout accident impliquant au moins un véhicule routier en mouvement, survenant sur une voie ouverte à la circulation publique, et **dans lequel au moins une personne est blessée ou tuée**[2].

Nous définissons ici la gravité comme étant le niveau de blessure des individus concernés par un accident. En effet les usagers impliqués dans ces accidents peuvent s'en sortir indemne alors que d'autres peuvent être tués ou blessés.

Définition des différents niveaux de gravité[3] :

[1] "Décennie d'action pour la sécurité routière 2011-2020"
[2] Source : Institut Nationale de la Statistique et des Études Économiques
[3] Source : Observatoire national interministériel de la sécurité routière

- Les indemnes sont les individus impliqués non décédés et qui n'ont besoin d'aucun soin médical.
- Dans la catégorie « blessé » il existe deux sous-catégories notamment « les blessés légers » : blessés dont l'état nécessite entre 0 et six jours d'hospitalisation ou un soin médical et « les blessés graves » : blessés dont l'état nécessite plus de six jours d'hospitalisation.
- Les personnes tuées sont les personnes qui sont décédées sur le coup lors de l'accident ou qui décèdent durant les six jours suivant l'accident.

Les issues différentes des accidents pour les victimes témoignent de l'existence de différents facteurs pouvant expliquer la gravité de l'accident.

Dans notre étude nous essayerons de trouver les déterminants de la gravité des accidents de la route en France sur la période 2012-2017. Pour ce faire nous mettrons en œuvre un modèle économétrique en utilisant les données du Bordereau d'Analyse des Accidents Corporels (BAAC).

a. Revue de littérature

La fréquence des accidents de la route a donné suite à une diversité d'étude dont l'objectif était d'expliquer leurs causes mais rarement leur gravité.

Selon le bilan de l'accidentalité routière de l'année 2018[4] (réalisé par l'ONISR), deux individus sur trois tués, lors d'un accident de la route en 2018, le sont en zone hors agglomération. Durant l'été (ainsi que le printemps), on a en moyenne 183 personnes tuées par mois contre 153 les autres mois dans ces zones. Les accidents mortels de la route, selon ce bilan, sont causés majoritairement par la vitesse, ensuite viennent l'alcool et la consommation de stupéfiant. Le refus de priorité des usagers ainsi que les malaises pouvant découler de la fatigue sont aussi à l'origine d'accidents. La probabilité d'être tué est 22 fois plus important pour un motocycliste (conducteur de véhicule deux roues motorisés) que pour un conducteur de voiture (à distance parcourue équivalente).

Le rapport de la sécurité routière de l'OMS en 2015[5], révèle que la gravité des conséquences d'un accident ainsi que la probabilité que celui-ci ait lieu augmente grâce à différents facteurs en plus de ceux cités plus haut. Ce sont notamment :

- Le fait de ne pas utiliser les équipements de protection individuelle (casque pour les cyclistes, la ceinture de sécurité et autres dispositifs de sécurité)
- La non-application rigoureuse du code de la route

[4] "La sécurité routière en France - Bilan de l'accidentalité de l'année 2018" - ONISR - 2018
[5] "Rapport de situation sur la sécurité routière dans le monde" - OMS - 2015

- La distraction au volant.

Ce rapport souligne bien que la décision de la limitation de vitesse dans les différents pays est un enjeu majeur pour privilégier la sécurité routière.

Le rapport de 2018 présente les piétons, motocyclistes et cyclistes comme les plus vulnérables en cas d'accidents routiers. Ils représentent plus de 50% des décès d'accident dans le monde.

Dans une analyse comparative de procédures d'accidents mortels et non mortels, l'INRETS tente de donner des réponses à la mortalité issue des accidents de la route[6]. Cette étude a pour but d'estimer si la probabilité d'être impliqué dans un accident mortel est plus forte pour un conducteur de 2 roues motrices à l'aide d'une régression logistique.

Ce modèle intègre des variables telles que : Type de jour, Horaire de l'accident, Luminosité au moment de l'accident, Localisation, Type d'intersection, Sexe du conducteur, Age du conducteur, Port du casque etc.

Au seuil de 10%, les résultats de l'étude témoignent que la probabilité d'être impliquée dans un accident mortel dépend du sexe, de l'âge du conducteur. La localisation, le port du casque sont également des facteurs à risque. Les accidents en campagne sont les plus graves et les accidents impliquants des conducteurs n'ayant pas porté de casque entrainent très souvent la mort.

b. La base de données

La base que nous exploitons pour notre analyse est le Bordereau d'Analyse des Accidents Corporels (BAAC) agrégée des années 2012 à 2017[7] : accidents_de_la_route_2012_2018.csv

Le fichier BAAC est un outil qui regroupe l'ensemble des accidents corporels de la circulation ayant eu lieu en France métropolitaine ainsi que dans les départements d'Outre-mer, durant une année précise.

Ces informations sont collectées sur le lieu de l'accident par l'unité des forces de l'ordre sur place. Pour avoir le fichier final recouvrant tous les accidents de l'année, il faut récupérer les fichiers de différentes autorités. Les données du BAAC notamment proviennent du ministère de la défense, des brigades locales et de gendarmeries, et du centre de traitement informatique du Ministère de l'Intérieur (chargé du traitement des accidents de la sécurité publique et du Plan Académique de Formation (PAF)). Ces données sont ensuite rassemblées et corrigées au Ministère de l'Équipement, par le SETRA (service d'études techniques des routes et autoroutes). Ensuite est réalisée une correction au sein de l'ONISR avant la diffusion des données aux organismes tels que l'INRETS (institut national de recherche sur les

[6] « Analyse comparative de procédures d'accidents mortels et non mortels » - INRETS (2008)
[7] "Accident corporels de la circulation millésimé" - BAAC

transports et leur sécurité), le CERTU (centre d'études sur les réseaux, les transports, l'urbanisme et les constructions publiques).

Dans notre base nous disposons de quatre rubriques : les caractéristiques de l'usager, celles de la route, celles de la voiture et celle des zones concernées. La base de données concerne les accidents sur le territoire français entre 2012 et 2017 identifiés de façon unique par la variable « identifiant de l'accident », soient 359 288 observations au total. Pour chaque accident est indiqué la catégorie de chaque usager, leur état de blessure, leurs sexes, leurs années de naissance. Il est aussi indiqué la collision à l'origine de l'accident, le lieu, la date, l'heure ainsi que les conditions atmosphériques. Il y a également d'autres données sur le type de voie, l'éclairage, le type de circulation, le type d'infrastructure etc.

Dans le fichier d'origine certaines données étaient codées d'une façon qui ne nous permettait pas de les traiter aisément. Par exemple, s'il y avait 3 individus dans un accident, la variable $année\_de\_naissance$ était codée « 1986, 1982, 2001 ». Ainsi, nous avons dû agir directement sur le fichier Excel en créant de nouvelles variables telles que :

- Le nombre de tués, de blessés et d'indemnes.
- Le nombre d'individus et de piétons accidentés.
- Le nombre de véhicules par type (nombre de voitures, de camions, de tracteurs…).
- L'âge moyen des individus accidentés.
- Le nombre d'individus n'ayant pas mis leurs dispositifs de sécurité (nombre de ceintures non mises, nombre de casques non mis…)
- Nombre de véhicules ayant reçu un choc par l'avant, par l'arrière ou par le côté.

Avec ces modifications, nous obtenons un nouveau fichier de données : accidents.csv.

Variable dépendante : le coefficient de gravité.

Nous souhaitons déterminer les facteurs qui expliquent la gravité d'un accident de la route. Nous nous inspirons des chiffres du bilan de l'accidentalité routière de L'ONISR afin de déterminer notre variable dépendante :

En 2017 (les chiffres pour les années antérieures sont assez similaires), le coût de l'insécurité routière s'élève aux montants suivants :

- 3 331 000 pour une personne tuée.
- 416 403 euros pour un blessé hospitalisé plus de 24 heures.
- 16 656 euros pour un blessé léger.
- 5 108 euros pour les dégâts matériels.

Les travaux de l'IFSTARR ont montré que le fichier BAAC sous-estime le nombre réel de « blessé légers » sur les routes puisque les policiers ne sont pas systématiquement appelés pour les petits accidents[8]. Notre base de données ne fait pas la différence entre les blessés légers et les blessés hospitalisés. Ainsi, nous nous basons que le coût d'un tué comparé au coût d'un blessé hospitalisé plus de 24 heures afin d'établir qu'un tué coûte en moyenne 8 fois plus cher qu'un blessé. Nous codons donc notre variable dépendante, le coefficient de gravité, de la manière suivante :

$$coef\_grav = 1 \times nombre\ de\ blessés + 8 \times nombre\ de\ tués$$

Remarque : multipliez le coefficient de gravité par 416 000, ajoutez 5 108 et vous obtiendrez le coût de l'accident de la route.

c. Statistiques descriptives

Présentons d'abord les variables essentielles de notre modèle (les détails donnés dans cette partie seront repris dans les tableaux) :

- $\boldsymbol{coef\_grav}$: En moyenne les accidents accusent une gravité de 1,76.
- $\boldsymbol{hors\_agglo}$: La variable $Localisation$ est marquée par deux modalités : agglomération et hors-agglomération. Nous créons une variable factorielle prenant pour valeur 1 lorsque l'accident a eu lieu hors agglomération. Les accidents sont relativement plus graves hors agglomération.
- $\boldsymbol{nuit\_sans\_eclairage}$ **:** Selon notre base de données, les accidents peuvent avoir lieu au crépuscule, en plein jour ou encore la nuit sans éclairage public. Nous utiliserons donc cette variable factorielle prenant pour valeur 1 lorsque l'accident a eu lieu la nuit et que la route n'était pas éclairée. Seulement 8,4% des accidents corporels se rapportant à notre base ont eu lieu en pleine nuit sans éclairage public, cependant ces accidents sont relativement les plus graves avec un niveau moyen de gravité de 2,07.
- $\boldsymbol{nb\_indiv}$: Variable numérique représentant le nombre d'individus concernés par l'accident. En moyenne un accident implique 2 personnes. Dans notre base on a au maximum 63 personnes ; cet accident concernait un bus.
- $\boldsymbol{nb\_secu\_non\_mis}$: variable numérique comptabilisant le nombre d'individus n'ayant pas mis leur équipement de sécurité (ceinture, casque…). Les accidents présentant au moins un usager sans équipement de sécurité ne représentent que 5,1% de nos données mais sont relativement plus graves que ceux avec équipement de sécurité mis.

[8] "Bilan de l'accidentalité de l'année 2018" - ONISR - 2018 - Page 23

- $\boldsymbol{age\_moyen}$ **:** Variable numérique représentant l'âge moyen des individus présent lors de l'accident.
- $\boldsymbol{nb\_choc\_arriere}$ **:** Variable numérique représentant le nombre de véhicules concernés par l'accident pour lesquels le choc a été fait par l'arrière. Il n'est pas absurde de croire que les accidents dans lesquels les véhicules se heurtent par l'arrière sont moins graves que les autres, pour lesquels le point de choc peut être en face à face ou bien par le côté. Le nombre maximum de chocs arrière enregistrés lors d'un accident est de 16.

Dans le tableau suivant nous présentons les statistiques concernant plusieurs variables numériques.

Tableau 1 : Statistiques sur les variables explicatives numériques

Libellés variables	Moyenne	Écart type	Minimum	Maximum
Coefficient gravité	1,76	2,17	1	286
Nombre individus	2,25	1,16	1	63
Age moyen	38,99	14,38	0	102
Nombre de chocs arrière	0,26	0,54	0	16
Nombre de sécurités non mises	0.061	0.31	0	24

Les accidents recensés dans notre base de données comptent toujours au moins un individu blessé ce qui justifie le minimum du coefficient de gravité. L'accident qui enregistre le plus grand nombre d'individus est un accident impliquant un bus.

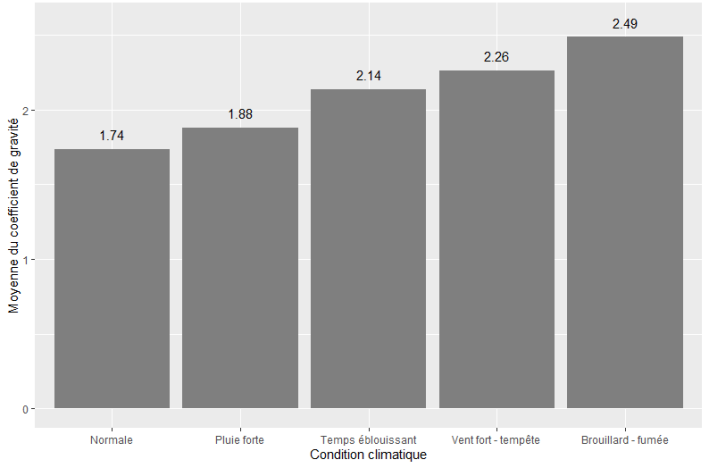
Le tableau suivant traduit quelques statistiques sur plusieurs variables indicatrices.

Tableau 2 : Statistiques sur les variables explicatives dummies

Variables indicatrices	Part des accidents pour lesquels la variable prend 1	Moyenne de gravité
Nuit sans éclairage public	8,4%	2,08
Hors agglomération	34,3%	1,91

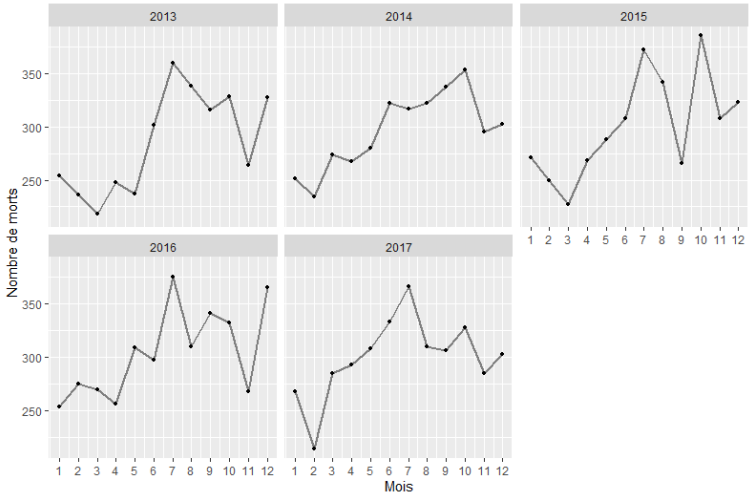
Étudions maintenant les caractéristiques de nos données grâce à des graphiques.

Figure 1 : Niveau de gravité de l'accident en fonction des conditions climatiques



Le graphique présente la gravité moyenne des accidents selon les conditions atmosphériques. Les accidents sont relativement plus graves lorsque la météo est mauvaise : par exemple lorsqu'il y a du brouillard la gravité moyenne des accidents est de 2,49 ; quand il y a une tempête la gravité moyenne est de 2,26. Il apparait selon nos données que 80,51% des accidents ont eu lieu dans des conditions normales de températures, cependant on note que plus la météo est dégradée plus les accidents sont graves. Sous de mauvaises conditions le conducteur peut voir sa visibilité baissée, notamment sa capacité à voir les autres usagers de la route ou encore à mieux voir la route elle-même. La visibilité des panneaux de signalisation peut devenir impossible, le conducteur peut par exemple ignorer la présence d'un virage imminent.

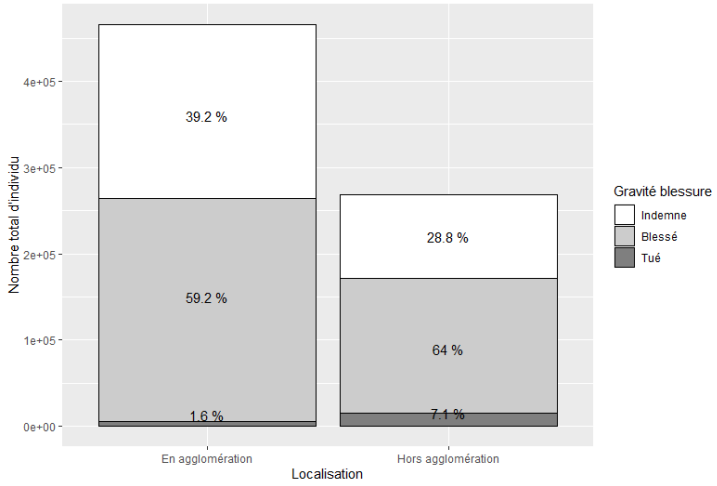
Figure 2 : Évolution du nombre de tués sur les routes pour différentes années



On observe une augmentation du nombre de décès entre le premier mois de chaque année et le dernier. Le nombre d'individus tués atteint un pic chaque année en été (période de vacances) et notamment en période de fin d'année. Ces mouvements de départs et de retours de vacances densifient la circulation. Les usagers sont souvent moins vigilants durant ces périodes, trop pressés ou

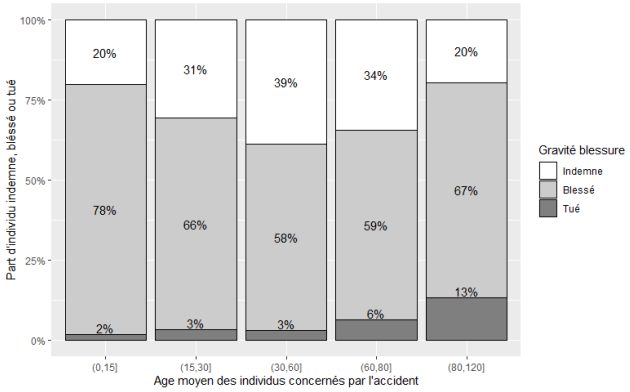
fatigués ce qui réduit leur prudence dans la conduite. L'alcool, la vitesse sont alors privilégiés durant cette période, d'où la hausse du nombre d'accidents corporels sur ces mois de l'année.

Figure 3 : Nombre d'indemnes, de blessés et de tués en agglomération et hors agglomération



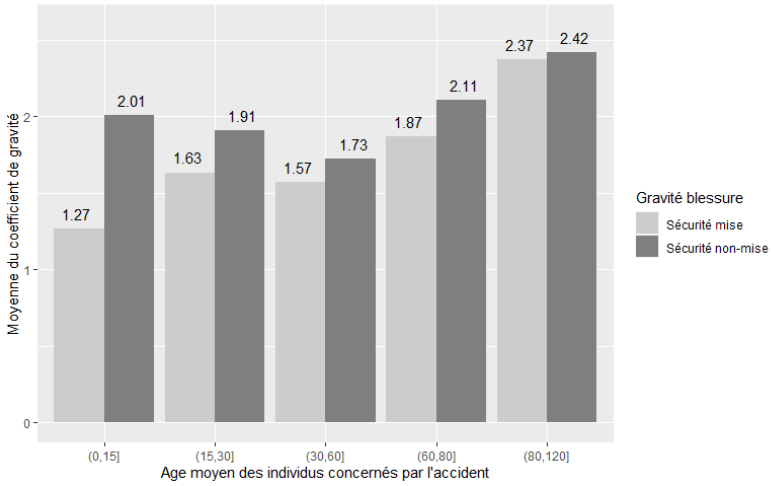
Ce graphique présente selon la gravité de la blessure, le nombre d'individus total accidentés en fonction de la localisation. Les accidents sont plus fréquents en agglomération. Les blessés et les tués représentent plus de 60% des individus impliqués dans les accidents dans chacune des localisations. Cependant, les accidents qui ont eu lieu hors agglomération sont plus meurtriers que ceux qui se sont déroulés en agglomération. On a 7,1% des individus accidentés en zone hors agglomération qui sont tués tandis que seulement 1 ,6% le sont en agglomération. De même on observe une différence de 10,4% entre le nombre total d'individus indemnes en agglomération et celui des individus hors agglomération. Les accidents hors agglomération engendrent plus de victimes que les autres. La limitation de vitesse en agglomération est de 50km/h tandis qu'elle est de 80km/h hors agglomération. Les excès de vitesse sur ces routes situées hors agglomération sont favorables à un niveau de gravité élevé des accidents dans cette zone.

Figure 4 : Part d'indemne, de blessé et de tué en fonction de l'âge moyen des individus concernés par l'accident



Le document présente la part des individus indemnes, blessés et tués selon les différentes classes d'âge. Il ressort que quel que soit la classe d'âge, les individus ont plus tendance à sortir blessés de l'accident. Les classes les plus vulnérables sont celles qui ont le plus de séquelles graves. En effet, la plus grande part de tué est enregistrée dans la classe [85,120] ; les plus jeunes [0,10] eux sont majoritairement blessés. C'est la classe d'âge [30 ;60] qui est majoritaire dans la population de notre base ; elle enregistre la plus grande proportion d'indemne. Quant aux individus de la classe [80,120], le bilan de l'accident est plus lourd pour eux car ils sont à environ 60% blessés ou ont environ 18% de chance d'être tués.

Figure 5 : Niveau de gravité de l'accident en fonction de l'âge moyen des individus et selon l'utilisation, ou non, des équipements de sécurité



Les usagers de la route qui n'utilisent pas d'équipement de sécurité sont plus vulnérables en cas d'accident. Selon le graphique, les accidents avec au moins un équipement de sécurité non mis sont relativement les plus graves car les usagers sont plus exposés.

2. Modélisation

a. Modèle de référence

Variables utilisées & effets attendus

1) $\boldsymbol{hors\_agglo}$: Nous nous attendons à ce que les accidents soient plus graves lorsqu'ils sont localisés en dehors de l'agglomération. En effet, la vitesse des véhicules étant plus élevée hors agglomération, les chocs entre plusieurs véhicules dans cette zone sont plus violents et peuvent causer davantage de dégâts.

2) $\boldsymbol{nuit\_sans\_eclairage}$: La gravité de l'accident devrait être plus grande lorsqu'il n'y a pas d'éclairage sur la route. En effet, une diminution de la visibilité des conducteurs va impacter leurs

réflexes et ils seront moins aptes à, par exemple, éviter un véhicule ou un piéton au dernier moment.

3) $\boldsymbol{nb\_indiv}$: Il est presque trivial de constater que plus il y a d'individu dans un accident, plus la probabilité qu'il y ait un blessé ou un mort augmente et donc, plus l'accident sera grave.

4) $\boldsymbol{nb\_secu\_non\_mis}$: En théorie, les équipements de sécurité routière sont importants et existent explicitement pour diminuer la probabilité d'être blessé ou tué lors d'un accident. Nous nous attendons donc à une augmentation de la gravité de l'accident lorsque les usagers ne mettent pas leur équipement de sécurité.

5) $\boldsymbol{age\_moyen}$: Il se peut que les personnes âgées, moins résistantes physiquement, aient plus de chance d'être blessées lors d'un accident ; de même pour les jeunes enfants. Ainsi, nous incluons l'âge moyen ainsi que le carré de l'âge moyen dans notre modèle, afin de prendre en compte cet effet non linéaire. Étant donné que nous nous attendons à un coefficient de gravité plus élevé pour les jeunes et les personnes âgés, alors nous nous attendons à ce que le coefficient estimant l'impact de age_moyen soit négatif et que celui de age_moyen^2 soit positif, ce qui permettra d'obtenir cette forme en U.

6) $\boldsymbol{nb\_choc\_arriere}$: Enfin, nous prévoyons un effet négatif du nombre de chocs par l'arrière sur le coefficient de gravité. En effet, les chocs par le côté ou par l'avant semblent plus graves que les chocs par l'arrière.

Tableau 3 : Effet attendu pour chaque variable explicative

Variable	Effet attendu sur la gravité de l'accident
$\boldsymbol{hors\_agglo}$	POSITIF
$\boldsymbol{nuit\_sans\_eclairage}$	POSITIF
$\boldsymbol{nb\_indiv}$	POSITIF
$\boldsymbol{nb\_secu\_non\_mis}$	POSITIF
$\boldsymbol{age\_moyen}$	NÉGATIF
$\boldsymbol{age\_moyen}^2$	POSITIF
$\boldsymbol{nb\_choc\_arriere}$	NÉGATIF

Étude des valeurs aberrantes :

Nous décidons de supprimer les accidents dans lesquels il y a plus de 10 individus, ils ne représentent que 0.1% de notre échantillon.

Corrélation entre les différentes variables :

Comme le montre le tableau suivant, il n'y a pas de corrélations trop importantes entre les différentes variables explicatives, nous pouvons donc poursuivre et passer à l'étape de modélisation.

Tableau 4 : Coefficients de corrélation de Pearson entre chaque variable

	Coefficients de corrélation de Pearson						
	coef_grav	***hors_agglo***	***nuit_sans_eclairage***	***nb_indiv***	***nb_secu_non_mis***	***age_moyen***	***nb_choc_arriere***
coef_grav		0,23	0,15	0,26	0,15	0,02	-0,03
hors_agglo	0,23		0,32	0,09	0,027	0,014	0,053
nuit_sans_eclair	0,15	0,32		0,0063	0,06	-0,082	-0,022
nb_indiv	0,26	0,09	0,0063		0,08	-0,065	0,31
nb_secu_non_mis	0,15	0,027	0,06	0,08		-0,054	-0,033
age_moyen	0,02	0,014	-0,082	-0,065	-0,054		0,026
nb_choc_arriere	-0,03	0,053	-0,022	0,31	-0,033	0,026	

b. **Application du modèle et interprétations :**

Ainsi, notre premier modèle s'écrit :

$$\begin{aligned} coef\_grav_i = \beta_0 \\ &+\beta_1 \times hors\_agglo_i \\ &+\beta_2 \times nuit\_sans\_eclairage_i \\ &+\beta_3 \times nb\_indiv_i \\ &+\beta_4 \times nb\_secu\_non\_mis_i \\ &+\beta_5 \times age\_moyen_i \\ &+\beta_6 \times age\_moyen_i^2 \\ &+\beta_7 \times nb\_choc\_arrière_i + \varepsilon_i \end{aligned}$$

Les hypothèses du modèle sont les suivants :

$$H_1 : E(\varepsilon_i) = 0$$

$$H_2 : Var(\varepsilon_i) = \sigma^2$$

$$H_3 : Cov(\varepsilon_i, \varepsilon_j) = 0 \ \ \forall i \neq j$$

$$H_4 : E(\varepsilon_i | X) = 0 \ \ \forall i$$

$$H_5 : \ Rang(X) = K + 1$$

Avec K le nombre de variables explicatives et ε_i le terme d'erreur pour l'observation i.

Le modèle peut se réécrire sous la forme matricielle suivante :

$$\begin{pmatrix} coef\_grav_1 \\ coef\_grav_2 \\ \vdots \\ coef\_grav_n \end{pmatrix} = \begin{pmatrix} 1 & hors\_agglo_1 & \cdots & \cdots & nb\_choc\_arrière_1 \\ 1 & hors\_agglo_2 & & & nb\_choc\_arrière_2 \\ \vdots & \vdots & & & \vdots \\ 1 & hors\_agglo_n & \cdots & \cdots & nb\_choc\_arrière_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_7 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{pmatrix}$$

Nous effectuons une première régression par la méthode des moindres carrés ordinaires.

Tableau 5 : Estimations sous le modèle des moindres carrés ordinaires (MCO)

Résultats estimés des paramètres				
Variable	**Valeur estimée des paramètres**	**Erreur type**	**Valeur du test t**	**P.value**
$\boldsymbol{Constante}$	0.73	0.024	31	<.0001
$\boldsymbol{hors\_agglo}$	0.81	0.0075	107	<.0001
$\boldsymbol{nuit\_sans\_eclairage}$	0.62	0.013	48	<.0001
$\boldsymbol{nb\_indiv}$	0.55	0.0034	162	<.0001
$\boldsymbol{nb\_secu\_non\_mis}$	0.83	0.012	70	<.0001
$\boldsymbol{age\_moyen}$	-0.031	0.00111	-28	<.0001
$\boldsymbol{age\_moyen}^2$	0.00045	0.000012	36	<.0001
$\boldsymbol{nb\_choc\_arriere}$	-0.438	0.00667	-66	<.0001

Tous les coefficients estimés sont significatifs au seuil de 5% et ont le signe attendu. Cependant, notre modèle pourrait être exposé à l'hétéroscédasticité et à l'endogénéité qui biaisent nos interprétations. Nous allons donc procéder pas-à-pas afin de tester leur présence, d'y trouver une solution et d'obtenir une estimation finale que nous interpréterons.

Étude de la multicolinéarité :

Afin de tester la présence de multi-colinéarité, nous nous appuyons sur l'indice du VIF (Variance inflation factor). Si cet indice est trop élevé pour une variable, alors cela indique que cette variable est fortement corrélée avec une autre variable. Le tableau ci-dessous présente les résultats :

Tableau 6 : Estimations sous MCO et calcul du Variance inflation factor (VIF)

Résultats estimés des paramètres - MCO					
Variable	**Estimation**	**Erreur type**	**Valeur du test t**	**Pr > |t|**	**Inflation de variance**
$Constante$	0.73	0.024	31	<.0001	0
$hors\_agglo$	0.81	0.0075	107	<.0001	1.13
$nuit\_sans\_eclairage$	0.62	0.013	48	<.0001	1.13
nb_indiv	0.55	0.0034	162	<.0001	1.15
$nb\_secu\_non\_mis$	0.83	0.012	70	<.0001	1.02
age_moyen	-0.031	0.00111	-28	<.0001	22.75
age_moyen^2	0.00045	0.000012	36	<.0001	22.74
$nb\_choc\_arriere$	-0.44	0.0067	-66	<.0001	1.13

L'indice est très élevé pour les variables age_moyen et age_moyen^2 mais cela est normal puisque l'une d'entre elle est le carré de l'autre : il n'y a pas de problème ici.

C. Étude de l'hétéroscédasticité

Test d'hétéroscédasticité :

Malgré la présence d'hétéroscédasticité, les coefficients estimés restent non-biaisés mais le problème se trouve dans l'estimation de la variance des coefficient estimés. En effet, ces estimations ne sont plus valables, ce qui rend les tests de significativité et les intervalles de confiance invalides. Nous appliquons ainsi un test de White dont les hypothèses sont les suivantes :

$$\begin{cases} H_0 : V(\varepsilon_i) = V(\varepsilon_j) \ \forall i,j \\ H_1 : \exists i,j \ s.t. \ V(\varepsilon_i) \neq V(\varepsilon_j) \end{cases}$$

Le résultat du test de White est le suivant :

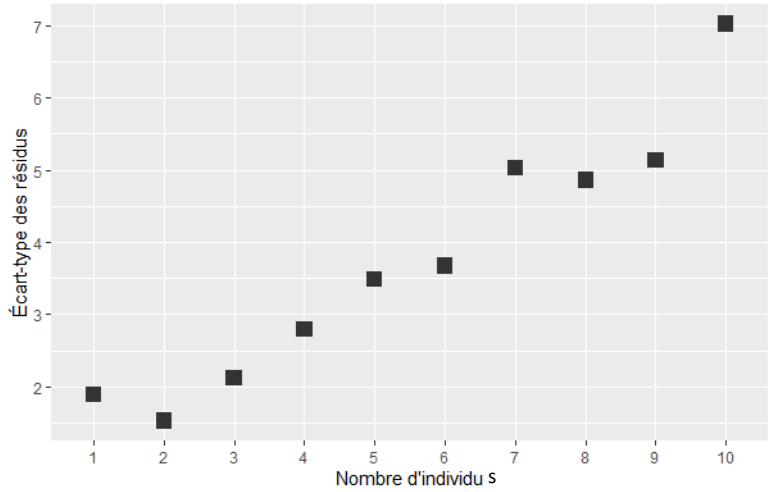
Tableau 7 : Test de White

Test de spécification du premier et du deuxième moment		
DDL	**Khi-2**	**Pr > Khi-2**
28	8747	<.0001

Ainsi, on constate clairement la présence d'hétéroscédasticité dans ce modèle. Au seuil de 95%, nous rejetons l'hypothèse nulle d'un modèle homoscédastique.

Pour constater la présence d'hétéroscédasticité, nous avons regroupé le nombre d'individus par classe et nous avons calculé l'écart-type des résidus pour chacune de ces classes. Bien que les intervalles de confiance de chacun des résultats ne se voient pas sur le graphique suivant, ce dernier permet de comprendre la présence d'hétéroscédasticité dans ce modèle.

Figure 6 : Écart-type des résidus en fonction du nombre d'individus



On peut justifier cela en effectuant un test de Breusch-Pagan sur la variable nb_indiv :

Tableau 8 : Test de Breusch-Pagan

Test d'hétéroscédasticité					
Equation	**Test**	**Statistique**	**DDL**	**Pr > khi-2**	**Variables**
$\boldsymbol{coef\_grav}$	Breusch-Pagan	4693	1	<.0001	nb_indiv, 1

Par ailleurs, il est logique de constater que la variation du coefficient de gravité augmente à mesure que le nombre d'individus augmente. En effet, par construction nous avons $coef\_grav = nb\_blessé + 8 \times nb\_tue$. Ainsi, la valeur minimale que peut prendre le coefficient est toujours 1 (puisqu'il y a toujours au moins un blessé ou tué dans les accidents recensés) alors que la valeur maximale qu'il peut prendre est de $8 \times nb\_individu$ (si tous les individus sont tués). L'intervalle de valeur que peut prendre le coefficient de gravité augmente à mesure que le nombre d'individus augmente, ce qui explique que la variance du coefficient de gravité soit croissante avec le nombre d'individus.

<u>Traitement de l'hétéroscédasticité :</u>

Afin de traiter le problème d'hétéroscédasticité, nous allons réévaluer les écarts-types estimés avec la matrice des écarts-types robustes.

Méthode des écarts-type robustes :

Tableau 9 : Estimations sous MCO avec écarts-types robustes

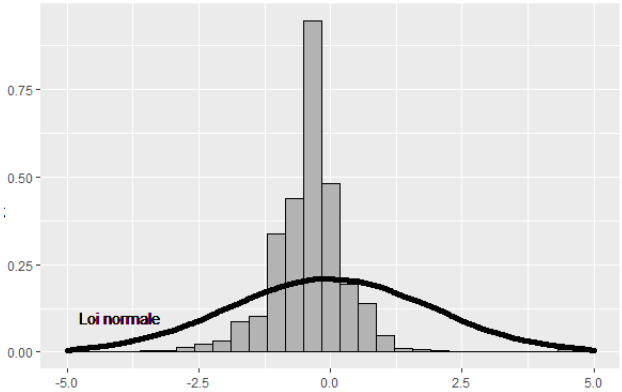
Résultats estimés des paramètres - MCO							
Variable	Estimations	Erreur type	t_{OBS}	Pr > |t|	Cohérent avec l'hétéroscédasticité		
					Erreur type	t_{OBS}	Pr > |t|
$Constante$	0.73	0.024	31	<.0001	0.026	29	<.0001
$hors\_agglo$	0.81	0.0075	107	<.0001	0.0088	84	<.0001
$nuit\_sans\_eclairage$	0.62	0.013	48	<.0001	0.020	31	<.0001
nb_indiv	0.55	**0.0034**	162	<.0001	**0.0062**	89	<.0001
$nb\_secu\_non\_mis$	0.83	0.012	70	<.0001	0.027	19	<.0001
age_moyen	-0.031	0.00111	-28	<.0001	0.0013	-25	<.0001
age_moyen^2	0.00045	0.000012	36	<.0001	0.000015	30	<.0001
$nb\_choc\_arriere$	-0.44	0.0067	-66	<.0001	0.0068	-64	<.0001

Le tableau ci-dessus présente les résultats en ayant tenu compte de l'hétéroscédasticité grâce à la méthode des écarts-type robustes. On observe que l'écart-type estimé augmente pour chacun des coefficients et, pour la variable nb_indiv, il est presque multiplié par 2. Cependant, cela n'affecte pas la conclusion des tests de significativité pour chacune des variables : les coefficients estimés restent significativement différents de 0 au seuil de 5%.

d. Étude de la normalité des résidus

Sur le graphique suivant, nous pouvons observer que les résidus ne sont clairement pas distribués selon une loi normale mais cela n'est pas un problème car nous avons un très grand nombre d'observations (plus de 300 000).

Figure 7 : Distribution des résidus



e. Étude de l'endogénéité

Selon l'étude des baromètres vacances des européens et des américains réalisé par Europ assistance et IPSOS (entreprise de sondages française et une société internationale de marketing d'opinion)[9], 56 % des vacanciers français décident de partir en vacances dans une autre région de France. Ainsi, ils sortent souvent de l'agglomération pendant la période des vacances, on peut donc s'attendre à ce que notre variables $hors\_agglo$ soit endogène et s'explique par la période de l'accident.

Afin d'utiliser les IV et GMM, nous utilisons les deux variables instrumentales suivantes :

- $nb\_tracteur$: Variable numérique représentant le nombre de tracteurs présent dans l'accident.
- $juillet\_aout$: Variable indicatrice prenant pour valeur 1 si l'accident a eu lieu en juillet ou en août et zéro sinon.

En effet, nous pensons que ces deux variables instrumentales ont un impact sur le coefficient de gravité de l'accident uniquement à travers leur effet sur la variable $hors\_agglo$. Autrement dit, les accidents mettant en cause un tracteur et les accidents ayant lieu en juillet ou en août sont particulièrement graves parce qu'ils ont lieu hors agglomération.

Nous suspectons également la variable $nuit\_sans\_eclairage$ d'être endogène. En effet, cette dernière est corrélée avec $hors\_agglo$ puisque la plupart des accidents ayant lieu la nuit sans éclairage extérieure sont des accidents qui ont lieu hors agglomération, étant donné que la plupart des agglomérations sont éclairées la nuit par l'éclairage public. Les variables $hors\_agglo$ et $nuit\_sans\_eclairage$ sont par ailleurs celles qui admettent le plus grand coefficient de corrélation de Pearson, s'élevant à 0.32, comme évoqué précédemment (Voir page **13**). Ainsi, nous utilisons une troisième variable instrumentale :

- $heure\_nuit$: Variable indicatrice prenant pour valeur 1 si l'accident a eu lieu entre 22 heures et 8 heures du matin.

Le tableau suivant présente donc les résultats de la méthode des moments généralisés, qui considérant l'existence d'hétéroscédasticité, nous offre des estimations plus précises que l'estimation standard des IV.

[9] "Baromètre vacances des Européens et des Américains" - ÉTUDE IPSOS/EUROP ASSISTANCE - 2019

Tableau 10 : Estimations sous la méthode des MCO ; IV et GMM

Variable	OLS Cohérent avec l'hétéroscédasticité			IV Cohérent avec l'hétéroscédasticité			GMM		
	Estim.	Erreur type	Pr > |t|	Estim.	Erreur type	Pr > |t|	Estim.	Erreur type	Pr > |t|
$Constante$	0.73	0.026	<.0001	0.32	0.03	<.0001	0.31	0.03	<.0001
$hors\_agglo$	0.81	0.0088	<.0001	2.43	0.11	<.0001	2.48	0.10	<.0001
$nuit\_sans\_eclairage$	0.62	0.020	<.0001	0.98	0.08	<.0001	0.96	0.088	<.0001
nb_indiv	0.55	0.0062	<.0001	0.49	0.007	<.0001	0.49	0.0071	<.0001
$nb\_secu\_non\_mis$	0.83	0.027	<.0001	0.75	0.026	<.0001	0.75	0.026	<.0001
age_moyen	-0.031	0.0013	<.0001	-0.032	0.0013	<.0001	-0.032	0.0013	<.0001
age_moyen^2	0.00045	0.000015	<.0001	0.00045	0.000015	<.0001	0.00045	0.000015	<.0001
$nb\_choc\_arriere$	-0.44	0.0068	<.0001	-0.47	0.008	<.0001	-0.48	0.008	<.0001

L'observation de ces résultats nous donne d'abord deux points positifs : premièrement, nous observons de grosses différences par rapport aux estimations par la méthode des moindres carrés ordinaires, ce qui atteste un peu plus de la possible endogénéité des variables $hors\_agglo$ et $nuit\_sans\_eclairage$. Deuxièmement, nous observons que les coefficients estimés ne diffèrent pas trop entre les méthodes IV et GMM, ce qui est positif.

Afin de nous assurer que les instruments utilisés sont exogènes, nous effectuons un test de Hansen. Comme le montre le résultat, nous acceptons, au seuil de 5%, l'hypothèse selon laquelle les variables instrumentales sont exogènes.

Tableau 11 : Test de Hansen

Statistique de test GMM			
Test	**DDL**	**Statistique**	**P-value**
Restrictions de sur-identification	1	1.46	0.227

Afin de nous assurer que les variables $hors\_agglo$ et $nuit\_sans\_eclairage$ sont endogènes, nous pouvons utiliser l'approche de la régression augmentée. Étant donné que nous avons de l'hétéroscédasticité, l'objectif est d'effectuer les moindres carrés généralisés faisables (FGLS). Ainsi,

- Nous prenons le résidu de la régression sous IV effectuée précédemment, nous l'élevons au carré et calculons son logarithme.
- Nous régressons ce logarithme du résidu au carré sur quelques-unes de nos variables explicatives et en tirons la valeur prédite.
- Soit Ω l'exponentielle de la prédite obtenue, nous calculons le poids $weight = \frac{1}{\Omega^{1/2}}$

- Nous effectuons la régression de base, en ajoutant comme variable explicative les résidus de la régression IV et en pondérant toutes les observations, y compris celles de la variable dépendante par le poids calculé $\frac{1}{\Omega^{1/2}}$.
- Nous effectuons un test de Student sur la significativité de l'effet des résidus de la régression IV sur la variable dépendante. S'ils sont significatifs, alors on rejette l'hypothèse d'exogénéité de la variable $hors\_agglo$.

Le résultat de la dernière régression est donné sur le tableau suivant :

Tableau 12 : Approche de la régression augmentée

Résultats estimés des paramètres				
Variable	**Valeur estimée des paramètres**	**Erreur type**	**Valeur du test t**	**P.value**
$Constante$	0.45	0.022	20	<.0001
$hors\_agglo$	2.61	0.0064	41	<.0001
$nuit\_sans\_eclairage$	0.97	0.060	16	<.0001
nb_indiv	0.41	0.0038	108	<.0001
$nb\_secu\_non\_mis$	0.47	0.010	46	<.0001
age_moyen	-0.031	0.0008	-36	<.0001
age_moyen^2	0.00040	0.000009	43	<.0001
$nb\_choc\_arriere$	-0.30	0.0057	-53	<.0001
$hors\_agglo\_residuals$	-2.1	0.066	-32	<.0001
$nuit\_sans\_eclairage\_residuals$	-1.56	0.053	29	<.0001

Ainsi, les résidus influencent significativement la variable dépendante : au seuil de 5%, nous rejetons l'hypothèse nulle d'exogénéité des variables $hors\_agglo$ et $nuit\_sans\_eclairage$.

En vue des résultats et sachant que l'estimateur des GMM est un meilleur estimateur que celui des IV, nous allons pouvoir conclure sur les effets des différentes variables en utilisant les résultats obtenus à partir des GMM.

3. Conclusion

a. Résumé de notre démarche

Cette première partie était dévouée à la mise en place d'une régression linéaire. Nous avons choisi d'expliquer la gravité d'un accident de la route en fonction de plusieurs variables en notre possession. Après avoir supprimer les valeurs « aberrantes » et vérifier que nos variables explicatives n'étaient pas

trop corrélées, nous avons effectué une première régression via la méthode des moindres carrés ordinaires.

Le test de White et celui de Breush-Pagan nous ont permis de constater la présence d'hétéroscédasticité dans notre modèle. Si toutes les autres hypothèses des MCO sont justes, alors les estimations de la valeur de nos coefficients sont valides mais leurs écarts-types ne le sont plus. Ainsi, nous avons dû tenir compte de cette hétéroscédasticité en réévaluant les écarts-types estimés grâce à la matrice des écarts-types robustes. Cela nous donne les mêmes coefficients estimés, mais des écarts-types estimés différents. Malgré ce changement, les coefficients estimés sont restés significativement différents de 0 au seuil de 5%.

Ensuite, nous avons émis l'hypothèse d'endogénéité des variables $hors\_agglo$ et $nuit\_sans\_eclairage$. Afin de traiter cela, nous avons utilisé trois instruments : $juillet\_aout$; $nb\_tracteur$ et $heure\_nuit$. Nous avons ainsi appliqué la méthode des variables instrumentales (en tenant compte de l'hétéroscédasticité) et la méthode des moments généralisés. Le test de sur-identification de Hansen nous a permis de conclure que les variables instrumentales sont exogènes alors que l'approche de la régression augmentée nous a assuré de l'endogénéité des variables $hors\_agglo$ et $nuit\_sans\_eclairage$.

Ainsi, nous avons choisi d'utiliser la méthode des moments généralisés afin d'obtenir nos résultats finaux. Nous allons maintenant les interpréter.

b. Interprétation de nos résultats

Notre régression finale donne les résultats suivants :

Tableau 13 : Estimations avec la méthode des moments généralisés (GMM)

Résultats estimés des paramètres par la méthode des GMM				
Variable	**Valeur estimée des paramètres**	**Erreur type**	**Valeur du test t**	**P.value**
$Constante$	0.31	0.03	8	<.0001
$hors\_agglo$	2.48	0.10	24	<.0001
$nuit\_sans\_eclairage$	0.96	0.088	11	<.0001
nb_indiv	0.49	0.0071	69	<.0001
$nb\_secu\_non\_mis$	0.75	0.026	29	<.0001
age_moyen	-0.032	0.0013	-24	<.0001
age_moyen^2	0.00045	0.000015	29	<.0001
$nb\_choc\_arriere$	-0.48	0.008	-59	<.0001

Au seuil de 5%, toutes nos variables explicatives sont significatives. Les variables $hors\_agglo$; $nuit\_sans\_eclairage$; nb_indiv et $secu\_non\_mis$ influent positivement sur le coefficient de gravité alors que la variable $nb\_choc\_arriere$ influe négativement. La variable age_moyen a un effet non linéaire et significatif sur le coefficient de gravité.

Parmi les variables indicatrice, la variable $hors\_agglo$ est celle qui influence le plus le coefficient de gravité dans notre modèle. Un accident qui a lieu dans une zone hors agglomération voit sa gravité augmenter de 2,48 par rapport à un accident en agglomération.

La visibilité de la route par le conducteur est essentielle à la bonne conduite de celui-ci afin d'éviter des obstacles source d'accidents. C'est certainement cela qui justifie que, selon notre modèle, la gravité d'un accident augmente de 0,96 quand celui-ci a lieu dans la nuit et dans une zone sans éclairage public allumé.

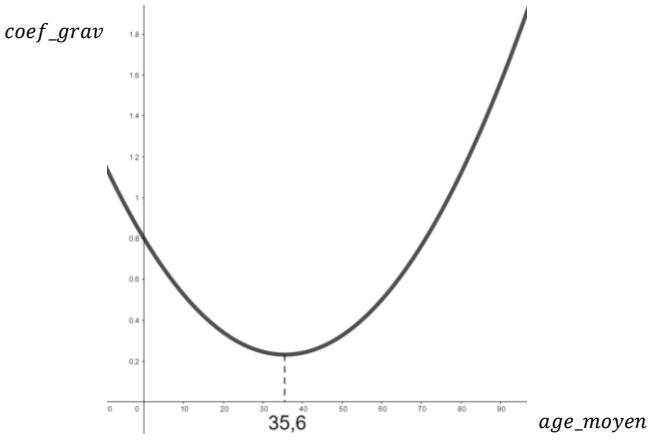
L'impact estimé du nombre d'individus sur le coefficient de gravité de l'accident est de $0{,}49$. Ainsi, cela signifie que lorsque le nombre d'individus concernés par l'accident augmente de 1, le coefficient de gravité augmente de 0,49. Nous pourrions interpréter ce résultat ainsi : étant donné le calcul du coefficient de gravité ($coef\_grav = 1 \times nb\_blesse + 8 \times nb\_tue$), on peut dire que « lorsque le nombre d'individus concernés par l'accident augmente de 2, cela équivaut à un blessé en plus ».

Étudions maintenant l'impact de l'âge moyen des individus concernés par l'accident sur le coefficient de gravité de l'accident. La variable age_moyen apparaît également au carré. Ainsi, l'effet de l'âge moyen sur la gravité de l'accident se calcule de la façon suivante. Gardons simplement la variable age_moyen et mettons toutes les autres variables à 0 sauf $nb\_individu$ qu'on égalise à 1 pour rester dans la logique de notre étude. On a ainsi l'équation estimée suivante :

$$\widehat{coef\_grav} = 0.31 + 0.49 \times 1 - 0.032 \times age\_moyen + 0.00045 \times age\_moyen^2$$

$$=> \widehat{coef\_grav} = 0.80 - 0.032 \times age\_moyen + 0.00045 \times age\_moyen^2$$

Il s'agit donc d'un polynôme du second degré représenté ci-dessous sur l'intervalle $[0{,}100]$:



Le minimum d'un polynôme du second degré $ax^2 + bx + c$ est définit par $min = -\frac{b}{2a}$. Ainsi, nous trouvons que l'âge moyen pour lequel le niveau de gravité de l'accident est le plus bas est de 35,6 ans. Plus on s'éloigne de cette valeur, plus la gravité de l'accident augmente, et cela peu importe la valeur que prenne les autres variables. En effet, le fait de les faire varier n'influe pas sur le minimum de cette fonction polynomiale mais seulement sur sa la façon dont elle est courbée.

Il est également possible de construire des intervalles de confiance. Prenons par exemple la variable $nb\_choc\_arriere$. Le coefficient estimé s'élève à -0.48 et semble très significativement différent de zéro selon le test de Student. Voici son intervalle de confiance au seuil de 95% :

$$IC_{\beta_8}^{5\%} = \left[\hat{\beta}_7 \pm 1.96 \times \widehat{\sigma_{\beta_7}}\right] = [-0.495; -0.464]$$

Selon ce modèle et selon les hypothèses évoquées précédemment, il y a 95% de chance pour que, lorsque le nombre de points de choc arrière augmente d'une unité, l'impact sur le coefficient de gravité soit compris dans l'intervalle de confiance $IC_{\beta_8}^{5\%}$.

La gravité des accidents augmente de 0,75 lorsque au moins un des usagers impliqués dans l'accidents ne mets pas son équipement de sécurité. En effet, les dispositifs de sécurité jouant un rôle important dans la prévention des blessures ou de la mortalité lors d'accidents de la route, leur non-utilisation expose les usagers à des séquelles graves, si ce n'est la mort.

c. Ouverture

Afin de savoir quels sont les facteurs qui influencent le plus l'insécurité sur la route et donc le coût de l'insécurité en France (qui s'élève à près de 2% du PIB comme nous l'avons vu précédemment), nous allons effectuer la somme de toute nos valeurs prédites et nous allons constater quelles sont les variables qui ont le plus influencées le coût total de l'insécurité durant ces dernières années.

Pour des soucis de lisibilité, nous définissons les variables suivantes :

$$X_{1i} = hors\_agglo_i$$
$$X_{2i} = nuit\_sans\_eclairage_i$$
$$X_{3i} = nb\_indiv_i$$
$$X_{4i} = nb\_secu\_non\_mis_i$$
$$X_{5i} = age\_moyen_i$$
$$X_{7i} = nb\_choc\_arrière_i$$

On fait la somme de toutes nos valeurs prédites :

$$\sum_{i=1}^{n} \widehat{coef\_grav_i} = \sum_{i=1}^{n} \left[0.31 + 2.48X_{1i} + 0.96X_{2i} + 0.49X_{3i} + 0.75X_{4i} - 0.032X_{5i} + 0.0045X_{5i}^2 - 0.48X_{7i}\right]$$

$$= \sum_{i=1}^{n} 0.31 + 2.48\sum_{i=1}^{n} X_{1i} + 0.96\sum_{i=1}^{n} X_{2i} + 0.49\sum_{i=1}^{n} X_{3i} + 0.75\sum_{i=1}^{n} X_{4i} - 0.032\sum_{i=1}^{n} X_{5i} + 0.0045\sum_{i=1}^{n} X_{5i}^2 - 0.48\sum_{i=1}^{n} X_{7i}$$

Le nombre d'accident pour lesquels nous avons une prédiction est de $n = 329\,883$, parmi eux :

- 112 859 ont lieu hors agglomération : $\sum_{i=1}^{n} X_{1i} = 112\,859$.
- 30 627 ont lieu la nuit sans éclairage. $\sum_{i=1}^{n} X_{2i} = 30\,627$
- 729 820 individus sont concernés. $\sum_{i=1}^{n} X_{3i} = 729\,820$
- 19 769 équipements de sécurité n'ont pas été mis. $\sum_{i=1}^{n} X_{4i} = 19\,769$
- 87 487 chocs arrière ont eu lieu. $\sum_{i=1}^{n} X_{7i} = 87\,487$

Ainsi on a :

$$\sum_{i=1}^{n} \widehat{coef\_grav_i} = 102\,263 + 279\,890 + 29\,402 + 357\,612 + 14\,827 - 155\,187 - 41\,994$$

Poids de la constante ; Poids de $hors\_agglo$; Poids de $nuit\_sans\_eclairage$; Poids de nb_indiv ; Poids de $nb\_secu\_non\_mis$; Poids de age_moyen ; Poids de $nb\_choc\_arriere$

Nous pouvons dresser le tableau suivant dont les explications se trouvent en-dessous.

	Impact estimé	**Nombre d'occurrence dans les données**	**Poids dans la gravité totale des accidents**	**Contribution au poids total**
Constante	0.31	$n = 329\,883$	102 263	17.4%
hors_agglo	2.48	112 859	279 890	47.7%
nuit_sans_eclairage	0.96	30 627	29 402	5.01%
nb_indiv	0.49	729 820	357 612	60.9%
nb_secu_non_mis	0.75	19 769	14 827	2.53%
age_moyen	−0.032	12 864 386	−411 660	−70%
age_moyen^2	0.00045	569 941 917	256 473	44%
nb_choc_arrière	−0.48	87 487	−41 994	−7.16%
		Total	**586 814**	**100%**

Tout d'abord, rappelons-nous que le coût d'un accident n'est qu'une combinaison linéaire du coefficient de gravité. En effet, lorsque nous avons défini notre variable dépendante, nous avons fait remarquer que $cout\_accident \approx 5108 + 416\,000 \times coef\_grav$. Ainsi, le coût total des accidents présent dans la base de données est la somme du coût de chacun des accidents et s'établit à 245 milliards d'euros sur toute la durée présente dans la base.

On observe dans notre tableau que toutes les variables ont un poids important sur le total de la gravité des accidents et donc sur le coût total des accidents. Le gouvernement ne contrôle pas toutes ces variables mais peut en influencer certaines, notamment $hors\_agglo$ et $nb\_secu\_non\_mis$.

- Le faits qu'un accident de la route se trouve hors agglomération semble prendre une place importante dans l'explication du coût total des accidents. Ainsi, le gouvernement pourrait investir afin de réduire la circulation hors agglomération, pourquoi pas en créant de nouveaux

réseaux de transports publics ou bien en réduisant les prix des transports publics inter-régionaux afin d'inciter les individus à privilégier les transports publics au lieu de leur véhicule personnel.

- Nous n'avons pas trouvé de données relatives au coût de la prévention routière sur les équipements de sécurité mais, en vue de nos résultats, le fait d'oublier son équipement de sécurité coute cher à la société et le gouvernement pourrait investir davantage dans ce type de prévention.

Partie 2 : Modèle logistique

1. Introduction

À la suite d'une longue phase de réarrangement des données, nous sommes parvenus à obtenir une nouvelle base de données dans laquelle chaque observation ne correspond plus à chaque accident, mais à chaque individu accidenté. Ainsi, la nouvelle base de données contient 798 425 observations et nous allons pouvoir faire une prédiction sur la gravité de blessure d'un individu présent dans un accident en fonction des caractéristiques de l'accident mais également de ses caractéristiques personnelles telles que son âge, son sexe ou encore selon le fait qu'il portait ou non la ceinture de sécurité.

Puisque cela ne se voit pas dans le rapport final de notre projet, nous tenons à insister sur le fait que le réarrangement des données a été une phase difficile. En effet, comme vous pouvez le voir sur la capture d'écran ci-dessous, chaque ligne de la base initiale correspondait à un accident dans lequel plusieurs individus étaient concernés et leurs caractéristiques étaient séparées par des virgules. De plus, le nombre de virgules de certaines cases ne correspondaient pas toujours, nous avons donc dû comprendre dans quel cas cette distinction apparaissait et trouver une solution dans le réarrangement des données.

Nous avons également estimé le véhicule de chacun des individus (cela n'était pas renseigné dans la base de données initiale) à partir des véhicules présents dans l'accident et du type d'équipement de sécurité que possédait l'individu. Par exemple si un individu est passager, que son équipement de sécurité est une ceinture, et que l'accident s'est fait entre une voiture et un scooteur, alors on en déduit que l'individu était dans la voiture (puisqu'il n'y a pas de ceintures sur les scooteurs).

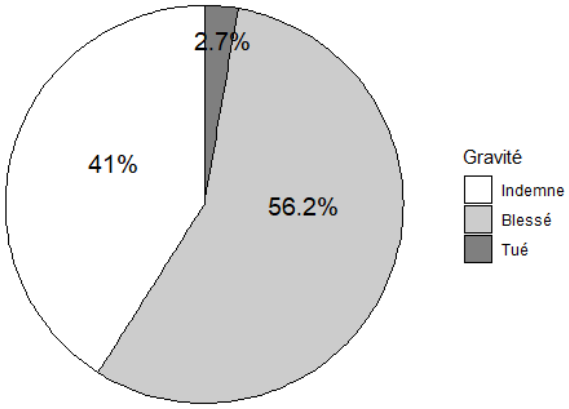
Année de naissance	Sexe	Action piéton	Gravité	Existence équipement
1972,1949,2010,1949	Féminin,Féminin,Féminin,Mascu	Se déplaçant,Se déplaçant,Se dép	Indemne,Indemne,Ind	Ceinture,Casque,Ceint
2015,1990,1976,2008,2010,1981	Masculin,Masculin,Masculin,Mas	Se déplaçant,Se déplaçant,Se dép	Blessé,Blessé,Blessé,Bl	Dispositif enfants,Cein
1941,1934	Masculin,Féminin	Se déplaçant,Se déplaçant	Blessé,Indemne	Ceinture,Ceinture
1993	Masculin	Se déplaçant	Blessé	Ceinture
1995,1944,1995	Masculin,Masculin,Masculin	Se déplaçant,Se déplaçant,Se dép	Blessé,Blessé,Blessé	Ceinture,Ceinture,Cein
1994	Masculin	Se déplaçant	Blessé	Ceinture
1992,1955	Féminin,Masculin	Se déplaçant,Se déplaçant	Indemne,Blessé	Ceinture,Ceinture
1969,199	Masculin,Masculin	Se déplaçant,Se déplaçant	Blessé,Indemne	Casque,Ceinture
1954,1934	Masculin,Masculin	Se déplaçant,Autre	Indemne,Blessé	Ceinture
1963	Masculin	Se déplaçant	Blessé	Ceinture
1972	Masculin	Se déplaçant	Blessé	Ceinture
1980,1982,2013,1955	Masculin,Féminin,Masculin,Masc	Se déplaçant,Se déplaçant,Se dép	Blessé,Blessé,Blessé,Bl	Ceinture,Ceinture,Cein

Ainsi, nous possédons une nouvelle base de données qui porte un intérêt différent de la première et grâce à laquelle nous allons pouvoir calculer la probabilité qu'un individu ressorte blessé ou tué sachant qu'il a eu un accident de la route ayant fait au moins un blessé.

a. Statistiques descriptives

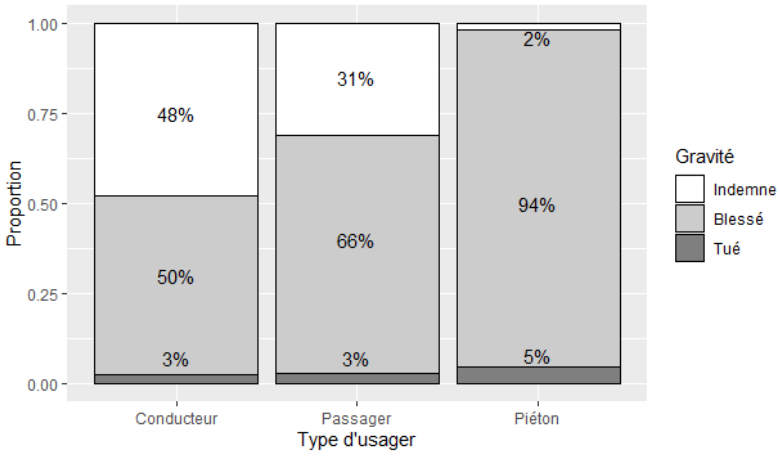
Cette nouvelle base de données nous permet d'effectuer plusieurs statistiques descriptives auxquelles nous n'avions pas accès avec la première base de données. En plus, d'être intéressantes, ces statistiques nous aideront dans le choix de notre modèle : visualiser pour mieux prédire.

Figure 8 : Part d'indemne, de blessé et de tué des accidents de la route



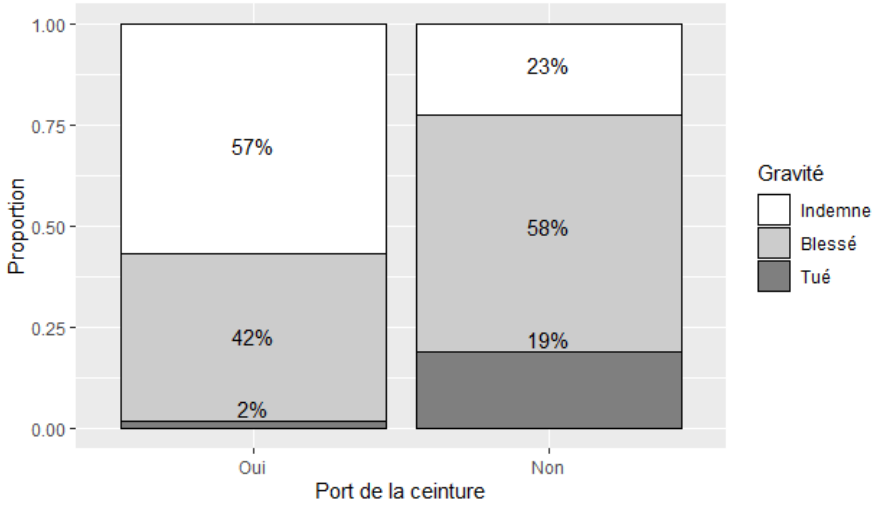
Le graphique précédent nous indique qu'il y a une majorité de blessé dans notre base de données (56.2%) alors que les tués représentent 2.7% de nos individus. Ainsi, notre variable dépendante $blesse\_ou\_tue$ prend la valeur 1 dans 58.9% des cas.

Figure 9 : Part d'indemne, de blessé et de tué en fonction du type d'usager



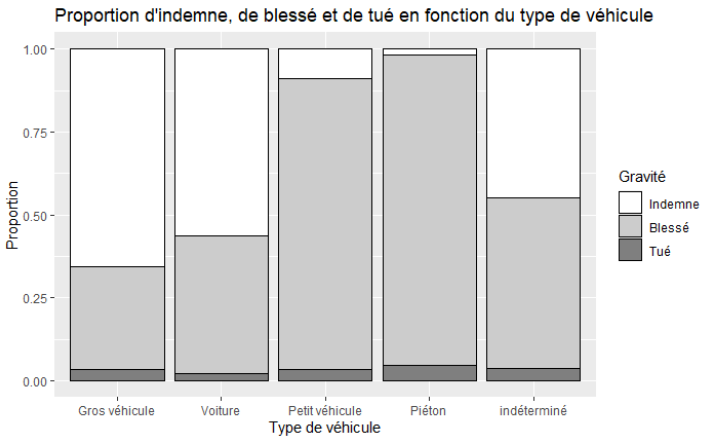
Dans nos données, la part de blessé est plus élevée lorsque l'individu est passager que lorsqu'il est conducteur. On remarque également que lorsque l'individu est un piéton, il a très peu de chance de ressortir indemne de l'accident (seulement 1.7% de chance que le piéton ressorte indemne).

Figure 10 : Part d'indemne, de blessé et de tué en fonction du port, ou non, de la ceinture



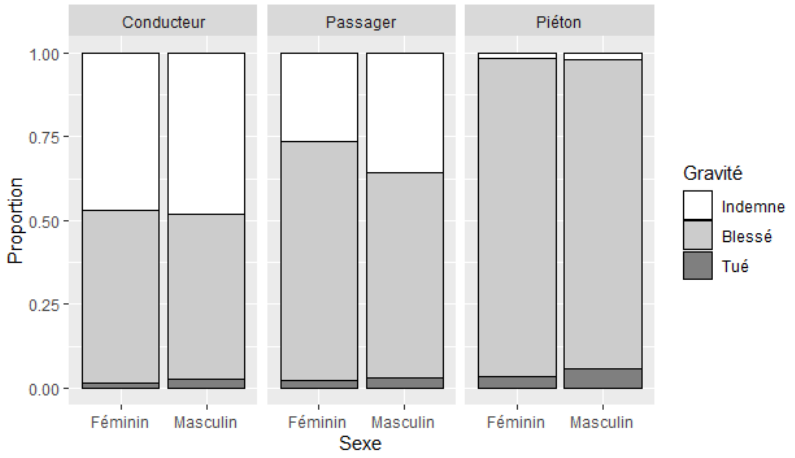
Le graphique précédent représente les individus accidentés qui étaient dans une voiture. Comme le montre les chiffres, la variable $secu\_mise$ (qui représente le fait que l'individu a, ou n'a pas, utilisé son équipement de sécurité) semble pertinente pour notre étude car comme vous pouvez le constater, la part des blessés et des tués est bien plus grande lorsque l'individu n'a pas mis sa ceinture de sécurité.

Figure 11 : Part d'indemne, de blessé et de tué en fonction du type de véhicule de l'usager



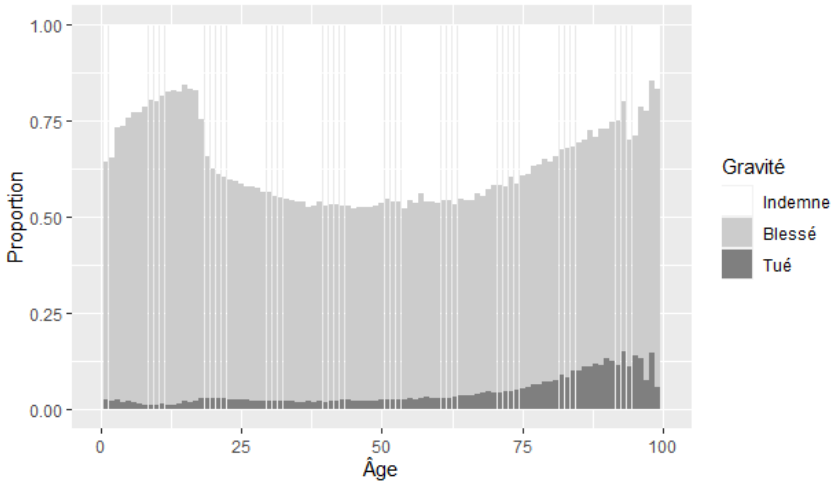
Une autre de nos variables concerne le type de véhicule de l'usager accidenté. Sur le graphique précédent, nous pouvons constater que plus le véhicule est gros, plus vous avez de chance de ressortir indemne de l'accident. Remarque : les gros véhicules sont les bus, les tracteurs et les camions alors que les petits véhicules sont les deux roues motorisées, les vélos et les quads.

Figure 12 : Part d'indemne, de blessé et de tué en fonction du sexe et du type d'usager



Dans notre base de données, les femmes subissent en moyenne des accidents plus graves que les hommes, mais cela peut s'expliquer par le fait que les femmes sont sur-représentées en tant que passagère et que, comme nous l'avons vu précédemment, les passagers subissent des accidents plus graves que les conducteurs. Sur le graphique suivant, nous observons la proportion d'indemne, de blessé et de tué en fonction de la catégorie d'usager et en fonction du sexe. On observe ainsi peu de différence entre les sexes lorsque l'individu est conducteur, cependant les femmes passagères subissent des accidents plus graves que les hommes passagers.

Figure 13 : Part d'indemne, de blessé et de tué en fonction de l'âge



Il est plutôt intéressant d'observer la part d'indemne, de blessé et de tué en fonction de l'âge. Comme le montre le graphique précédent, ceux qui subissent les accidents les plus graves sont les adolescents et les plus âgés. La part cumulée de blessé et de tué est en croissance de 0 à 16 ans puis décroit jusqu'à environ 45 ans, et entame ensuite une nouvelle phase de croissance. Cela peut s'expliquer par le fait

que les adolescents sont essentiellement des piétons ou des passagers, et que les personnes âgées, étant plus fragiles physiquement, ont probablement moins de chance de ressortir indemne d'un accident.

2. Modèle logistique

La régression logistique s'applique lorsque la variable dépendante est binaire : elle prend la valeur 1 en cas d'occurrence de l'évènement et 0 sinon. Dans notre base de données, cette variable est $blesse\_tue$. Elle prend la valeur 1 lorsque l'individu est blessé ou tué lors de l'accident et 0 si l'individu est indemne. On cherche donc à calculer la probabilité qu'un individu soit blessé ou tué, alors qu'il est victime d'un accident faisant au moins un blessé, en fonction des différentes variables explicatives que nous possédons.

Pour que notre variable dépendante ne puisse prendre que des valeurs entre 0 et 1 (puisque c'est une probabilité), nous utilisons la fonction de répartition de la loi logistique :

$$f(x) = \frac{e^x}{1 + e^x}$$

La fraction $\frac{P(y_i = 1|X)}{1-P(y_i = 1|X)}$ représente le rapport de cote. Il représente la probabilité d'occurrence d'un évènement par rapport à un autre : par exemple, si le numérateur représente le succès et le dénominateur représente l'échec et que cette fraction est égale à 3, alors on dit que la probabilité du succès est de 3 contre 1 (la probabilité de l'échec est alors de 1 contre 3).

Ainsi, nous appliquons la transformation LOGIT afin de pouvoir travailler sur des valeurs $\in [-\infty; +\infty]$ pour nos variables explicatives :

$$=> \ln\left(\frac{P(y_i = 1|X)}{1 - P(y_i = 1|X)}\right) = \ln\left(\frac{\frac{e^{Xb}}{1 + e^{Xb}}}{1 - \frac{e^{Xb}}{1 + e^{Xb}}}\right) = \ln\left(\frac{\frac{e^{Xb}}{1 + e^{Xb}}}{\frac{1}{1 + e^{Xb}}}\right)$$

$$=> \ln\left(\frac{P(y_i = 1|X)}{1 - P(y_i = 1|X)}\right) = Xb$$

a. Modèle de référence

Dans une analyse microéconométrique portant sur les accidents routiers en Tunisie[10], les auteurs ont tenté d'estimer la probabilité que l'individu soit indemne, blessé ou tué à la suite d'un accident. Ils ont pour cela utilisé certaines variables dont nous pouvons nous inspirer comme le sexe de l'individu, son

[10] Analyse microéconométrique des accidents routiers en Tunisie - Aloulou, Naouar - 2016

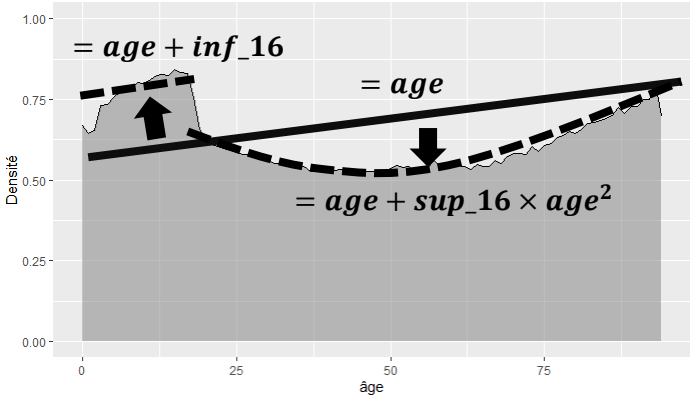
âge, le port de la ceinture, le moment de l'accident (jour / nuit) ou encore le lieu (en agglomération et hors agglomération).

Ainsi, nous décidons d'inclure les variables suivantes dans notre modèle :

- $passager$: indique 1 si l'individu est un passager et 0 sinon.
- $pieton$: indique 1 si l'individu est un piéton et 0 sinon.
- $securite\_oublie$: indique 1 si l'individu N'A PAS utilisé son dispositif de sécurité et 0 sinon.
- $homme$: indique 1 si l'individu est un homme et 0 sinon.
- $hors\_agglo$: indique 1 si l'accident a eu lieu hors agglomération et 0 sinon.
- $heure\_nuit$: indique 1 si l'accident a eu lieu entre 21h00 et 7h00 et 0 sinon.
- age : variable numérique représentant l'âge de l'individu.
- $usager\_petit\_vehiv$: indique 1 si le véhicule de l'individu est un petit véhicule (vélo, quad ou deux roues motorisés) et 0 sinon.
- $usager\_voiture$: indique 1 si le véhicule de l'individu est une voiture et 0 sinon.
- $presence\_pieton$: indique 1 si un piéton (autre que l'individu observé) est présent dans l'accident et 0 sinon.
- $presence\_petit\_vehic$: indique 1 si un petit véhicule (autre que celui de l'individu observé) est présent est 0 sinon.
- sup_16 : indique 1 si l'individu a plus de 16 ans et 0 sinon.
- inf_16 : indique 1 si l'individu a moins de 16 ans et 0 sinon.

Comme nous l'avons vu sur le graphique (répété ci-dessous) représentant la part d'indemne, de blessé et de tué en fonction de l'âge, l'évolution de la proportion de blessé ou de tué semble assez linéaire pour les individus âgés de moins de 16 ans alors qu'elle semble non-linéaire pour les individus âgés de plus de 16 ans. Ainsi, nous créons les variables indicatrices inf_16 et sup_16 qui indiquent respectivement si l'individu a plus ou moins de 16 ans. Nous allons ainsi ajouter inf_16 ainsi que l'interaction entre sup_16 et l'âge au carré afin de prendre en compte les effets non linéaires décrits ci-dessous :

Figure 14 : Proportion de blessé ou tué en fonction de l'âge



Ainsi, notre modèle est le suivant

$$\ln\left(\frac{P(blessé\ ou\ tué|X)_i}{1-P(blessé\ ou\ tué|X)_i}\right) = \beta_0 + \beta_1 \times secu\_oublie_i + \beta_2 \times usager\_petit\_vehic_i + \beta_3 \times nb\_indiv_i + \beta_4 \times passager_i + \beta_5 \times pieton_i + \beta_6 \times homme_i + \beta_7 \times hors\_agglo_i + \beta_8 \times heure\_nuit_i + \beta_9 \times presence\_pieton_i + \beta_{10} \times usager\_voiture_i \times presence\_petit\_vehic_i + \beta_{11} \times age_i + \beta_{12} \times inf\_16_i + \beta_{13} \times sup\_16 \times age_i^2$$

Effets attendus

Nous nous attendons à ce que les estimations de notre régression logistique nous apportent les effets suivants sur la probabilité d'être blessé ou tué :

- $\boldsymbol{securite\_oublie}$: les dispositifs de sécurité existent justement pour diminuer la probabilité d'être blessé ou tué, nous nous attendons donc à un effet positif de l'oublie du dispositif de sécurité.
- $\boldsymbol{usager\_petit\_vehiv}$: les petits véhicules sont les moins sécurisés. Ainsi, étant donné que l'effet des piétons est déjà pris en compte dans le modèle, nous nous attendons à un signe positif pour cette variable.
- $\boldsymbol{nb\_indiv}$: les accidents recensés comportent au moins un blessé ou tué. Ainsi, si le nombre d'individus est de 1, alors l'individu est forcément blessé ou tué, et plus le nombre d'individus augmente, plus il est probable que l'individu soit parmi les indemnes. Nous nous attendons donc à un effet négatif du nombre d'individus.
- $\boldsymbol{passager}$: nous avons vu qu'être passager semble plus risqué que d'être conducteur mais moins risqué que d'être piéton. Cependant, $pieton$ est également présent dans le modèle donc $conducteur$ est la modalité de référence. Ainsi, nous nous attendons à un effet positif de la variable $passager$ sur la probabilité d'être blessé ou tué.
- $\boldsymbol{pieton}$: les piétons ne sortent indemnes que dans 1.7% des cas, nous nous attendons donc à un effet positif élevé.
- $\boldsymbol{homme}$: les hommes sont-ils significativement plus résistants aux accidents que les femmes ? Difficile à dire. Nous optons pour un effet négatif et nous verrons ce que disent les données.

- $\boldsymbol{hors\_agglo}$: comme nous l'avons vu dans la première partie, les accidents sont bien plus graves hors agglomération, nous nous attendons donc à un effet positif sur la probabilité d'être blessé ou tué.
- $\boldsymbol{heure\_nuit}$: comme nous l'avons également vu dans la partie précédente, les accidents semblent plus graves la nuit, nous nous attendons donc à un effet positif.
- $\boldsymbol{presence\_pieton}$: si l'individu n'est pas un piéton et qu'il y a un piéton dans l'accident, alors il y a de forte chance pour que ce soit celui-ci qui soit blessé ou tué et pas l'individu, l'influence de cette variable est donc attendue négative.
- $\boldsymbol{usager\_voiture} \times \boldsymbol{presence\_petit\_vehic}$: lorsque l'usager est dans une voiture et qu'il heurt un petit véhicule, alors il est probable qu'il s'en sorte indemne, nous nous attendons donc à un signe négatif.
- $\boldsymbol{age}$: nous prévoyons une forme en U de l'âge sur la part de blessé ou tué, donc l'influence de cette variable devrait être négative.
- $\boldsymbol{inf\_16}$: la part de blessé ou tué chez les moins de 16 ans est relativement grande, nous nous attendons donc à un signe positif.
- $\boldsymbol{sup\_16} \times \boldsymbol{age^2}$: nous nous attendons à un signe positif.

Tableau 14 : Effet attendu de chaque variable

Variable	Effet attendu sur la gravité de l'accident
$\boldsymbol{secu\_oublie}$	POSITIF
$\boldsymbol{usager\_petit\_vehic}$	POSITIF
$\boldsymbol{nb\_indiv}$	NÉGATIF
$\boldsymbol{passager}$	POSITIF
$\boldsymbol{pieton\_}$	POSITIF
$\boldsymbol{homme}$	NÉGATIF
$\boldsymbol{hors\_agglo}$	POSITIF
$\boldsymbol{heure\_nuit}$	POSITIF
$\boldsymbol{presence\_pieton}$	NÉGATIF
$\boldsymbol{usag\_voiture} \times \boldsymbol{presence\_petit\_vehic}$	NÉGATIF
$\boldsymbol{age}$	NÉGATIF
$\boldsymbol{inf\_16}$	POSITIF
$\boldsymbol{sup\_16} \times \boldsymbol{age^2}$	POSITIF

Corrélation entre les différentes variables

Pour des soucis d'affichage, nous définissons les variables suivantes dont nous présentons les corrélations croisées ci-dessous.

$Y = blesse\_ou\_tue$
$X_1 = secu\_oublie_i$
$X_2 = usager\_petit\_vehic_i$
$X_3 = nb\_indiv_i$
$X_4 = passager_i$
$X_5 = pieton_i$
$X_6 = homme_i$
$X_7 = hors\_agglo_i$
$X_8 = heure\_nuit_i$
$X_9 = presence\_pieton_i$
$X_{10} = usager\_voiture_i \times presence\_petit\_vehic_i$
$X_{11} = age_i$

Tableau 15 : Coefficients de corrélation de Pearson entre les variables

	Coefficient de corrélation de Pearson											
	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
Y		0,097	0,34	-0,12	0,095	0,24	-0,081	0,072	0,065	-0,28	-0,44	-0,073
X_1	0,097		0,068	-0,02	0,042	-0,022	0,033	0,0021	0,053	-0,027	-0,06	-0,036
X_2	0,34	0,068		-0,23	-0,099	-0,16	0,19	-0,11	-0,028	-0,069	-0,22	-0,12
X_3	-0,17	-0,02	-0,23		0,45	-0,16	-0,084	0,22	0,029	-0,075	-0,075	-0,01
X_4	0,095	0,042	-0,099	0,45		-0,13	-0,18	0,15	0,11	-0,1	-0,092	-0,19
X_5	0,24	-0,022	-0,16	-0,12	-0,13		-0,12	-0,18	-0,059	0,0053	-0,13	0,06
X_6	-0,081	0,033	0,19	-0,084	-0,18	-0,12		0,011	0,07	0,0086	-0,008	-0,042
X_7	0,072	0,0021	-0,11	0,22	0,15	-0,18	0,011		0,042	-0,19	-0,089	0,0057
X_8	0,065	0,053	-0,028	0,029	0,11	-0,059	0,07	0,042		-0,059	-0,065	-0,13
X_9	-0,28	-0,027	-0,069	-0,075	-0,1	0,0053	0,0086	-0,19	-0,059		-0,13	0,054
X_{10}	-0,44	-0,06	-0,22	-0,075	-0,092	-0,129	-0,008	-0,089	-0,065	-0,13		0,075
X_{11}	-0,073	-0,039	-0,12	-0,01	-0,19	0,06	-0,042	0,0057	-0,13	0,054	0,075	

Aucune de nos variables ne semble en trop forte corrélation avec une autre variable, nous pouvons donc poursuivre.

b. Application du modèle et interprétations

Nous utilisons donc le modèle logistique pour obtenir les résultats suivants :

Tableau 16 : Estimations du modèle Logistique

Analyse des valeurs estimées du maximum de vraisemblance				
Paramètre	**Estimation**	**Erreur type**	**Valeur du test t**	**P-value**
$Intercept$	2.04	0.022	93	<.0001
$secu\_oublie$	1.10	0.022	51	<.0001
$usager\_petit\_vehic$	2.54	0.012	219	<.0001
nb_indiv	-0.33	0.0024	-139	<.0001
$passager$	0.90	0.0098	92	<.0001
$pieton\_$	5.26	0.034	157	<.0001
$homme$	-0.79	0.0071	-111	<.0001
$hors\_agglo$	0.45	0.0067	68	<.0001
$heure\_nuit$	0.38	0.0091	42	<.0001
$presence\_pieton$	-3.55	0.016	-225	<.0001
$usag\_voiture \times presence\_petit\_vehic$	-3.00	0.012	-256	<.0001
age	-0.030	0.00093	-32	<.0001
inf_16	-0.066	0.019	-3.46	0.0005
$sup\_16 \times age^2$	0.00036	0.0000099	36	<.0001

Tous les coefficients sont significatifs au seuil de 5% et tous les coefficients estimés ont le signe attendu sauf l'un d'entre eux : inf_16. Cela peut être dû à la présence d'hétéroscédasticité dans notre modèle qui, contrairement au modèle linéaire, peut ici biaiser la valeur des coefficients estimés.

c. Étude de l'hétéroscédasticité

Étant donné cette différence entre l'effet attendu et l'effet estimé d'une variable en relation avec la variable age, nous soupçonnons de l'hétéroscédasticité dans notre modèle. Nous supposons que les résidus ne sont pas constants et ont la forme fonctionnelle suivantes :

$$\sigma_i^2 = \sigma^2 \times e^{\lambda_1 \times age_i + \lambda_2 \times sup\_16 \times age_i^2}$$

En prenant en compte cette spécificité, nous obtenons les résultats suivants :

Tableau 17 : Estimations du modèle Logistique en prenant en compte l'hétéroscédasticité

Analyse des valeurs estimées du maximum de vraisemblance						
	Modèle LOGIT			Modèle LOGIT avec hétéroscédasticité		
Paramètre	Estimation	Erreur type	P-value	Estimation	Erreur type	P-value
$Intercept$	2.04	0.022	<.0001	1.54	0.025	<.0001
$secu\_oublie$	1.10	0.022	<.0001	0.76	0.018	<.0001
$usager\_petit\_vehic$	2.54	0.012	<.0001	1.74	0.024	<.0001
nb_indiv	-0.33	0.0024	<.0001	-0.22	0.0034	<.0001
$passager$	0.90	0.0098	<.0001	0.62	0.010	<.0001
$pieton\_$	5.26	0.034	<.0001	3.60	0.051	<.0001
$homme$	-0.79	0.0071	<.0001	-0.54	0.0085	<.0001
$hors\_agglo$	0.45	0.0067	<.0001	0.31	0.0061	<.0001
$heure\_nuit$	0.38	0.0091	<.0001	0.26	0.0071	<.0001
$presence\_pieton$	-3.55	0.016	<.0001	-2.40	0.034	<.0001
$usag\_voiture \times presence\_petit\_vehic$	-3.00	0.012	<.0001	-2.01	0.028	<.0001
age	-0.030	0.00093	<.0001	-0.027	0.00074	<.0001
inf_16	-0.066	0.019	0.0005	0.035	0.015	0.0224
$sup\_16 \times age^2$	0.00036	0.0000099	<.0001	0.00030	0.000014	<.0001
$Hetero - age$				-0.34	0.0013	<.0001
$Hetero - sup\_16 \times age^2$				0.00030	0.000014	<.0001

Afin de tester la présence d'hétéroscédasticité, nous faisons un test du ratio des vraisemblance. La statistique du test est la suivante :

$$QLR = -2 \times \ln\left(\frac{L_0}{L_1}\right) \sim \mathcal{X}^2(p)$$

La procédure nous renvoie les données suivantes : le logarithme de vraisemblance du modèle logistique sans prendre en compte l'hétéroscédasticité est de -315136 alors que le logarithme de vraisemblance du modèle en prenant en compte l'hétéroscédasticité est de -313915.

Ainsi, nous calculons la statistique du test :

$$QLR = -2 \times \ln\left(\frac{L_0}{L_1}\right) = -2 \times (\ln(L_0) - \ln(L_1)) = -2 \times (-315136 + 313915)$$

$$=> QLR = 2442$$

Cette statistique est bien au-dessus du fractile du Khi-deux à 1 degrés de liberté qui, au seuil de 5%, vaut 3.84. Ainsi, au seuil de 5%, nous rejetons l'hypothèse H_0 d'homoscédasticité du modèle.

d. Étude de l'endogénéité

Comme dans la première partie de notre étude, nous soupçonnons la variable $hors\_agglo$ d'être endogène. Ainsi, nous utilisons les mêmes instruments que précédemment :

- $nb\_tracteur$: Variable numérique représentant le nombre de tracteurs présent dans l'accident.
- $juillet\_aout$: Variable indicatrice prenant pour valeur 1 si l'accident a eu lieu en juillet ou en août et zéro sinon.

Afin de tester l'endogénéité de la variable $hors\_agglo$, nous voulons appliquer une commande SAS qui teste la corrélation entre notre variable suspectée endogène et le terme d'erreur. Malheureusement, la commande SAS permettant cela admet un temps de compilation beaucoup trop long, nous l'avons abandonné au bout de plusieurs heures de compilation. Nous souhaitions également effectuer un test de sur-identification des instruments $juillet\_aout$ et $nb\_tracteur$ mais pour les mêmes raisons, cela n'a pas été possible. Cependant nous avons laissé, en fichier joint à notre rapport, les commandes SAS utiles à effectuer ce calcul.

Ainsi, nous sommes en incapacité de tester l'endogénéité de $hors\_agglo$ et la validité des instruments. Nous décidons donc de supposer que toutes nos variables explicatives sont exogènes et nous allons conclure à partir des résultats issus du modèle logistique en prenant en compte l'hétéroscédasticité.

3. Conclusion

a. Interprétations du résultat final

La régression finale de notre modèle nous donne les résultats suivants :

Tableau 18 : Estimations du modèle Logistique en prenant en compte l'hétéroscédasticité

Analyse des valeurs estimées du maximum de vraisemblance en tenant compte de l'hétéroscédasticité				
Paramètre	**Estimation**	**Erreur type**	**Valeur du test t**	**P-value**
$Intercept$	1.54	0.025	61	<.0001
$secu\_oublie$	0.76	0.018	42	<.0001
$usager\_petit\_vehic$	1.74	0.024	72	<.0001
nb_indiv	-0.22	0.0034	-66	<.0001
$passager$	0.62	0.010	61	<.0001
$pieton\_$	3.60	0.051	70	<.0001
$homme$	-0.54	0.0085	-64	<.0001
$hors\_agglo$	0.31	0.0061	51	<.0001
$heure\_nuit$	0.26	0.0071	37	<.0001
$presence\_pieton$	-2.40	0.034	-72	<.0001
$usag\_voiture \times presence\_petit\_vehic$	-2.01	0.028	-71	<.0001
age	-0.027	0.00074	-36	<.0001
inf_16	0.035	0.015	2.28	0.022
$sup\_16 \times age^2$	0.00030	0.000014	21	<.0001
$Hetero - age$	-0.34	0.0013	-26	<.0001
$Hetero - sup\_16 \times age^2$	0.00030	0.000014	21	<.0001

Dans les résultats, nous pouvons observer que toutes les variables ont les effets attendus, y compris inf_16.

Prenons l'exemple d'une jeune femme de 20 ans qui rentre d'une soirée à 3h00 du matin. Elle passe ainsi par une route départementale avec son scooter et n'oublie pas de mettre son casque. Malheureusement, elle est victime d'un accident faisant au moins une victime avec une voiture dans laquelle il n'y a qu'une personne.

Ainsi, on calcule le logarithme du rapport des côtes grâce auquel on en tire le rapport des côtes.

$$\ln\left(\frac{P(blessé\ ou\ tué|X)}{1-P(blessé\ ou\ tué|X)}\right) = 4.264$$

$$=> \frac{P(blessé\ ou\ tué|X)}{1-P(blessé\ ou\ tué|X)} = \exp(4.264) = 71.09$$

La jeune femme a 71 fois plus de chances d'être blessée ou tuée que de ne pas l'être.

$$=> P(blessé\ ou\ tué|X) = \big(1 - P(blessé\ ou\ tué|X)\big) \times 71.09$$

$$=> P(blessé\ ou\ tué|X) = 0.986$$

La probabilité que la jeune femme soit blessée ou tuée lors de l'accident est de 98.6%.

Au lieu d'interpréter directement les coefficients estimés, nous pouvons plutôt nous attarder sur les effets marginaux et les rapports de côtes. Les effets marginaux nous indiquent pour chaque observation, quelle était l'impact d'une variable sur la probabilité estimée. Nous pouvons ainsi faire un résumer de la moyenne des effets marginaux pour chacune des variables.

Voici un tableau décrivant la moyenne des effets marginaux et le rapport des côtes :

Tableau 19 : Effets marginaux et rapport des côtes pour chaque variable explicative

		Effet marginal moyen	Rapport des côtes		
Variable	**Estimation**	**Impact moyen**	**Rapport des côtes**	**Borne inf 5%**	**Borne sup 5%**
$secu\_oublie$	0.76	0.14	2,138	2,064	2,215
$usager\_petit\_vehic$	1.74	0.32	5,697	5,436	5,972
nb_indiv	-0.22	-0.042	0,803	0,797	0,808
$passager$	0.62	0.12	1,859	1,823	1,896
$pieton\_$	3.60	0.67	36,598	33,117	40,446
$homme$	-0.54	-0.10	0,583	0,573	0,593
$hors\_agglo$	0.31	0.06	1,363	1,347	1,38
$heure\_nuit$	0.26	0.05	1,297	1,279	1,315
$presence\_pieton$	-2.40	-0.45	0,091	0,085	0,097
$usag\_voiture \times presence\_petit\_vehic$	-2.01	-0.38	0,134	0,127	0,142
age	-0.027	-0.0050	0,973	0,972	0,975
inf_16	0.035	0.007	1,036	1,006	1,067
$sup\_16 \times age^2$	0.00030	0.000057	1	1	1

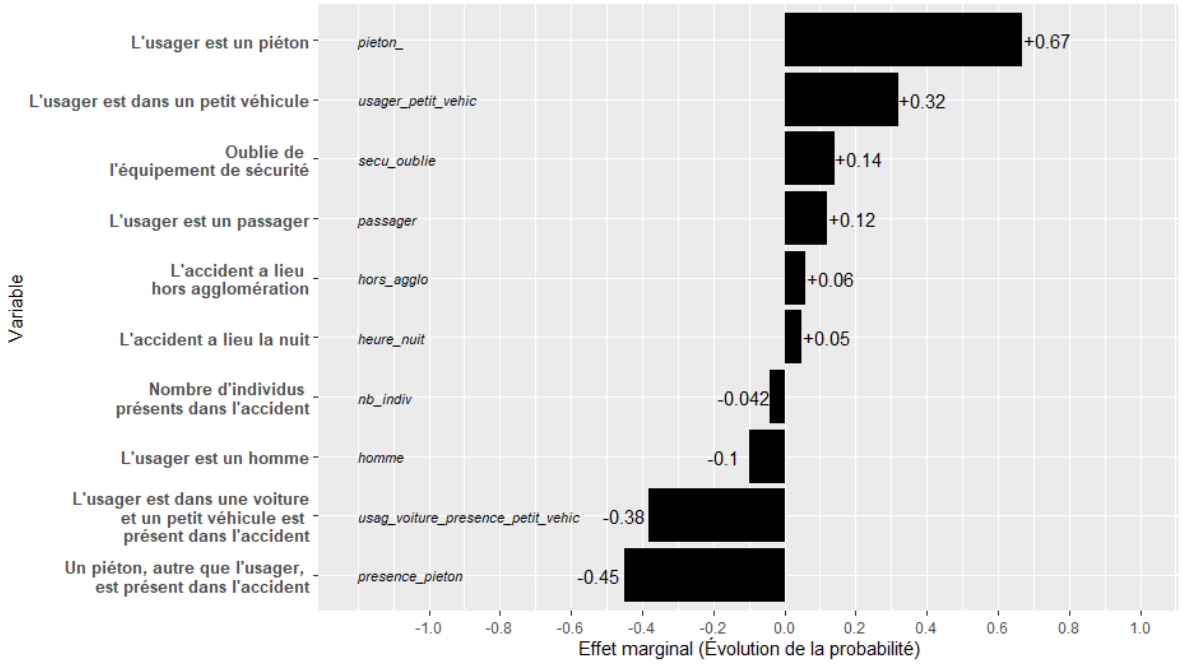
Nous pouvons interpréter quelques coefficients :

- En moyenne, le fait de ne pas mettre son dispositif de sécurité (ceinture, casque etc…) augmente de 14% la probabilité d'être blessé ou tué lors d'un accident de la route.
- Le fait d'être un homme diminue en moyenne de 10% la probabilité d'être blessé ou tué pendant un accident de la route.

- À chaque fois que le nombre d'individus présent dans l'accident augmente de 1, la probabilité d'être blessé ou tué diminue en moyenne de 4.2%.
- Être un piéton dans un accident de la route augmente en moyenne de 67% la probabilité d'être blessé ou tué.

Nous représentons les effets marginaux moyens sur le graphique suivant :

Figure 15 : Effet marginal moyen pour plusieurs variables



b. Évaluation du modèle – Matrice de confusion

Afin d'évaluer si notre modèle estime bien la probabilité d'être blessé ou tué, nous lui faisons faire une prédiction en fonction de la probabilité estimée. Si la probabilité estimée d'être blessé ou tué est supérieure à 0.5, alors le modèle prédit que l'individu sera blessé ou tué, sinon il prédit qu'il sera indemne. Afin de constater nos résultats, nous calculons la matrice de confusion qui résume l'échec ou la réussite de la prédiction.

Tableau 20 : Matrice de confusion (seuil de décision 50%)

		Matrice de confusion		
		blesse_ou_tue		
		0	**1**	**Total**
Prédiction	**0**	28.37 %	6.05 %	34.41 %
	1	12.56 %	53.03 %	65.59 %
	Total	40.92 %	59.08 %	100.00 %

Ainsi, notre modèle donne une *accuracy* de 81.4%.

- La sensibilité représente la capacité du modèle à retrouver les positifs. Ici. $SE = \frac{53.03}{59.08} = 89.8\%$.
- La spécificité représente la capacité à retrouver les négatifs. Ici. $SP = \frac{28.37}{40.92} = 69.3\%$

Pour mesurer si notre modèle fait de bonnes estimations, nous nous appuyons sur la valeur de la log-vraisemblance maximisée. Ainsi, nous calculons le pseudo R^2 de McFadden's défini ainsi :

$$McFadden's\ R^2 = 1 - \frac{\ln(L)}{\ln(L_0)}$$

Où L_0 représente la valeur de la vraisemblance lorsque le modèle est estimé avec seulement une constante.

On trouve $McFadden's\ R^2 = 0.4185$.

En fonction des différents gains lorsque vous faites une bonne prédiction sur cette variable, vous pourriez être plus ou moins averse au risque et décider de changer le seuil de décision (précédemment fixé à 0.5). Fixons désormais le seuil de décision à 20%. Ainsi, si la probabilité estimée est supérieure à 80%, le modèle prédit « Blessé ou tué » ; si la probabilité estimée est inférieure à 20%, le modèle prédit « Indemne » ; si la probabilité est entre 20% et 80%, le modèle ne se prononce pas. Voyons les résultats grâce à la matrice de confusion.

Tableau 21 : Matrice de confusion (seuil de décision 20%)

		Matrice de confusion		
		blesse_ou_tue		
		0	**1**	**Total**
Prédiction	**0**	20.86%	1.54%	22.39%
	1	2.1%	33.55%	35.65%
	NSP	17.96%	23.99%	41.96%
	Total	40.92%	59.08%	100.00%

Le modèle fait une prédiction pour 58% des observations

Dans ce cas, la prédiction n'a eu lieu que pour 458 066 observations (soit 58% de notre base de données) et, lorsque le modèle fait une prédiction, ne se trompe que dans 6.27% des cas, soit une *accuracy* de 93.73%.

c. Ouverture

Nos estimations nous amènent à des résultats plus ou moins évidents :

- Dans nos résultats estimés, il est assez logique de constater que les usagers de petits véhicules, les piétons influent positivement et significativement la probabilité d'être blessé ou tué lors d'un accident : il est plus probable d'être blessé ou tué lorsque nous sommes dans un scooter que lorsque nous sommes dans un camion, toutes choses égales par ailleurs.
- Notre analyse confirme également le fait qu'il est plus risqué d'être passager que d'être conducteur : la probabilité d'être blessé ou tué en étant passager augmente en moyenne de 12%.
- Lorsque vous êtes usagers d'une voiture, si un petit véhicule est présent dans l'accident, alors vous avez en moyenne 38% de chance de plus de vous en sortir indemne par rapport au cas où un véhicule plus gros est présent dans l'accident.
- Enfin, le fait d'avoir moins de 16 ans influe positivement et significativement la probabilité d'être blessé ou tué lors d'un accident, de 7% en plus en moyenne.
- Cependant, on observe que le sexe de l'usager influe significativement sur la probabilité qu'il soit blessé ou tué lors d'un accident. En effet, les femmes de la base de données ont en moyenne 10% de chance de plus que les hommes de ne pas sortir indemne d'un accident. Nous avons vu que les femmes de la base de données étaient sur-représentées en tant que passagère mais cela ne peut pas être une explication puisque la variable $passager$ est prise en compte dans le modèle donc son effet est déjà intégré. Ainsi, soit la variable $homme$ est endogène, soit les hommes ont effectivement plus de chance de s'en sortir que les femmes, nous pouvons soupçonner des différences en termes de résistances physiques aux accidents entre les hommes et les femmes.

Références

- Page 4 : "Décennie d'action pour la sécurité routière 2011-2020"
- Page 4 : Source : Institut Nationale de la Statistique et des Études Économiques
- Page 4 : Source : Observatoire national interministériel de la sécurité routière
- Page 5 : "La sécurité routière en France - Bilan de l'accidentalité de l'année 2018" - ONISR - 2018
- Page 5 : "Rapport de situation sur la sécurité routière dans le monde" - OMS – 2015
- Page 6 : « Analyse comparative de procédures d'accidents mortels et non mortels » - INRETS (2008)
- Page 6 : "Accident corporels de la circulation millésimé" - BAAC
- Page 7 : "Bilan de l'accidentalité de l'année 2018" - ONISR - 2018 - Page 23
- Page 19 : "Baromètre vacances des Européens et des Américains" - ÉTUDE IPSOS/EUROP ASSISTANCE - 2019
- Page 31 : Analyse microéconométrique des accidents routiers en Tunisie - Aloulou, Naouar - 2016
- Sécurité des agents et des usagers Etude des accidents corporels, CEREMA
- Description des bases de données annuelles des accidents corporels de la circulation routière : https://www.data.gouv.fr/fr/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2019/#_

Sigles et abréviations

BAAC : Bordereau d'Analyse des Accidents Corporels

INRETS : Institut national de recherche sur les transports et leur sécurité

INSEE : Institut national de la statistique et des études économiques

OMS : Organisation Mondiale de la Santé

ONISR : Observatoire national interministériel de la sécurité routière

ONU : Organisation des Nations Unies

PAF : Plan Académique de Formation

SETRA : service d'études techniques des routes et autoroutes

Figures et tableaux