Université Lille 1 – Sciences et Technologies

THÈSE

pour obtenir le grade de
Docteur de l'Université Lille 1

Discipline : Statistiques
École doctorale : Sciences Pour l'Ingénieur

Estimation par approximation de Laplace dans les modèles GLM Mixtes : application à la gravité corporelle maximale des accidents de la route

par

Fatima MEGUELLATI

soutenue publiquement le 07 février 2014

Jury

Nicolas WICKER	Pr. Lille 1	Président
Gilbert SAPORTA	Pr. CNAM Paris	Rapporteur
Sylvain LASSARE	DR. INRETS	Rapporteur
Assi N'GUESSAN	MdC, HDR, Lille 1	Directeur de thèse
Thierry Hermitte	LAB Renault	Co-encadrant
Joseph NGTCHOU-WANDJI	Pr. Nancy	Examinateur
Anne GUILLAUME	Directrice LAB Renault	Invitée
Yves PAGES	Renault, TechnoCentre	Invité

Remerciements

Je tiens à exprimer ma reconnaissance à mon directeur de thèse Assi N'Guessan et à mon encadrant Thierry Hermite pour leur encadrement, pour leur travail et pour leurs conseils qui m'ont aidée à approfondir mes réflexions et à améliorer et guider mon travail.

J'adresse mes remerciements aux rapporteurs de ce travail de recherche Gilbert Saporta et Sylvain Lassare qui ont su apporter un regard critique et constructif sur ce mémoire. Je remercie également Joseph Ngtchou-Wandji, Anne Guillaume et Yves Pages pour avoir accepté de faire parti du jury et Nicolas Wicker pour avoir accepté de présider le jury

J'adresse mes remerciements à Anne la directrice du LAB et Nicolas le directeur adjoint pour m'avoir acceuillie au sein de leur laboratoire ainsi que pour le suivi de ma thèse.

J'adresse ma gratitude à l'ensemble du personnel du LAB et du CEESAR pour son soutien et sa bonne humeur. Je remercie en particulier Carole, Mariem, Zayed, Martine, Yary, Vuthy, Romain, Sophie, Véronique, Konthea, Reeka, Franck, Maxime, Filipe, Lionel, Jean-Marc, Audrey et Caterine

Je remercie également le Laboratoire Paul Painlevé pour son accueil.

Je remercie également mon mari pour m'avoir soutenue et toujours remotivée, ma famille pour avoir été présente pendant ces années lors desquelles elle a toujours été là pour me soutenir et enfin ma belle-soeur Naima pour sa relecture assidue du mémoire.

Enfin je remercie mes amis qui, de près ou de loin, m'ont soutenue.

Résumé

Cette thèse est une contribution à la construction de méthodes statistiques applicables à l'évaluation (modélisation et estimation) de certains indices utilisés pour analyser la gravité des accidents de la route. On se focalise sur quatre points lors du développement de la méthodologie : la sélection des variables à effets aléatoires, la construction de modèles logistique-normaux mixtes, l'estimation des paramètres et la comparaison des méthodes d'estimation. Dans une première contribution, on construit un modèle logistique-Normal avec « Type de collision » comme variable à effet aléatoire. Des méthodes d'estimation fondées sur l'approximation de Laplace de la log-vraisemblance sont proposées pour estimer et analyser la contribution des variables présentes dans le modèle. On compare, par simulation, cette approximation Laplacienne à celle basée sur l'adaptation des polynômes de Gauss-Hermite (AGH) ainsi qu'à la quasi-vraisemblance pénalisée (PQL). On montre que les trois approches sont équivalentes par rapport à la précision de l'estimation bien qu'AGH soit légèrement supérieure. Une seconde contribution consiste à la construction d'un second modèle logistique-Normal avec « EES » (Energie de déformation) comme variable à effets aléatoires. On montre pour ce cas que les trois méthodes d'estimation sont équivalentes seulement pour l'estimation des effets fixes. Concernant l'estimation de l'écart-type de la variable à effets aléatoires, la méthode AGH est la plus précise. La différence entre les deux modèles se trouvant principalement au niveau de la valeur de l'écart-type de l'effet aléatoire, l'une inférieure à un et l'autre supérieure à deux. Dans une troisième contribution, on identifie plusieurs modèles logistique-normaux mixtes avec plus d'une variable à effets aléatoires. La convergence des algorithmes (Laplace, AGH, PQL) ainsi que la proximité des estimations sont étudiées. Une base de données d'accidents est utilisée pour analyser la performance des modèles à détecter des véhicules contenant au moins un blessé grave. L'analyse montre que le modèle logistique-normal mixte avec « EES » comme variable à effets aléatoires a une plus grande capacité à détecter les accidents graves comparativement au modèle classique sans effets aléatoires. Une programmation orientée R accompagne l'ensemble des résultats obtenus. La thèse se termine sur des perspectives relatives aux critères de sélection des modèles GLM Mixtes et à l'extension de nos modèles à la famille multinomiale.

Abstract

This thesis is a contribution to the construction of statistical methods for the evaluation (modeling and estimation) of some indices used to analyze the injury severity of road crashes. We focus on four points during the development of the methodology: the random variables selection, the construction of mixed logistic-Normal model, the parameters estimation and the comparison of the estimation methods. In a first contribution , a logistic-Normal model is constructed with "'collision type"' as random variable to analyze the injury severity observed in a sample of crashed vehicles. Estimation methods based on the Laplace approximation of the log-likelihood are proposed to estimate and analyze the contribution of variables in the model. We compare, by simulation, the Laplacian approximation to those based on the adaptation of Gauss-Hermite polynomials (AGH) and to Penalized Quasi-Likelihood method (PQL). We show that the three approaches are equivalent with respect to the accuracy of the estimate although AGH is slighty superior. A second contribution is the construction of a second logistic-Normal model with "'EES"' (Energy Equivalent Speed) as a random effect variable. We show that in this case the three estimation methods are equivalent only for the estimates of the fixed effects. As regards the standard deviation of the random effects variable, the AGH method is the most accurate. The difference between the two models is mainly located at the value of the standard deviation of the random effects variable. As a matter of fact the first is smaller than one and the second is higher than two. Two examples of simulated data illustrate the results. In a third contribution, we identify several mixed logistic-normal models with more than one random effect. The convergence of the algorithms (Laplace, AGH, PQL) and the closeness of the estimates are investigated. Simulations as well as a database of crash data are used to analyze the models performance to detect vehicles containing users with maximum injury severity. The analysis shows that the logistic-Normal model with "'EES"' as a random effect variable has a greater ability to detect serious accidents compared to the classical model without random effects. Programming oriented R accompany all results. The thesis concludes with perspectives on GLM Mixed models selection criteria and the extension of these models to the multinomial family.

Contents

List of Tables

List of Figures

Introduction

Problème de transport au premier abord, les accidents de la route constituent également un problème de santé publique : de santé par ses conséquences humaines, de santé publique car tout le monde utilise la route et risque sa santé et sa vie quotidiennement.

Selon le rapport de l'OMS (2013) environ 1,24 million de personnes meurent chaque année dans des accidents de la route dans le monde et entre 20 et 50 millions seraient blessées. En 2012, les accidents de la circulation représentaient la 8ème cause de mortalité dans le monde et la 1ère cause de mortalité chez les jeunes âgés de 15 à 29 ans (OMS, 2013). 92% des décès par accident de la circulation surviennent dans des pays à revenu faible ou intermédiaire bien que ces pays ne comptent pourtant que 53% des véhicules immatriculés dans le monde.

L'Europe en 2011 dénombrait 1 016 521 accidents corporels faisant 26 002 morts. Pour rappel, la commission européenne s'était fixée pour objectif de réduire de moitié le nombre de tués sur les routes en Europe entre 2001 et 2010. Si cet objectif n'a pas été atteint, sa mise en place a cependant permis une prise de conscience des différents états, mais aussi de l'ensemble des acteurs (constructeurs automobiles, équipementiers, constructeurs de route, citoyens, etc.) ce qui a permis une nette amélioration de la mortalité malgré un élargissement de l'Union Européenne dans la même période. Fort de ces efforts encourageants, la Commission Européenne a décidé en 2011 de se fixer un nouveau challenge pour 2020, en conservant comme objectif une diminution du nombre de morts par deux par rapport aux résultats de 2011 (figure 1).

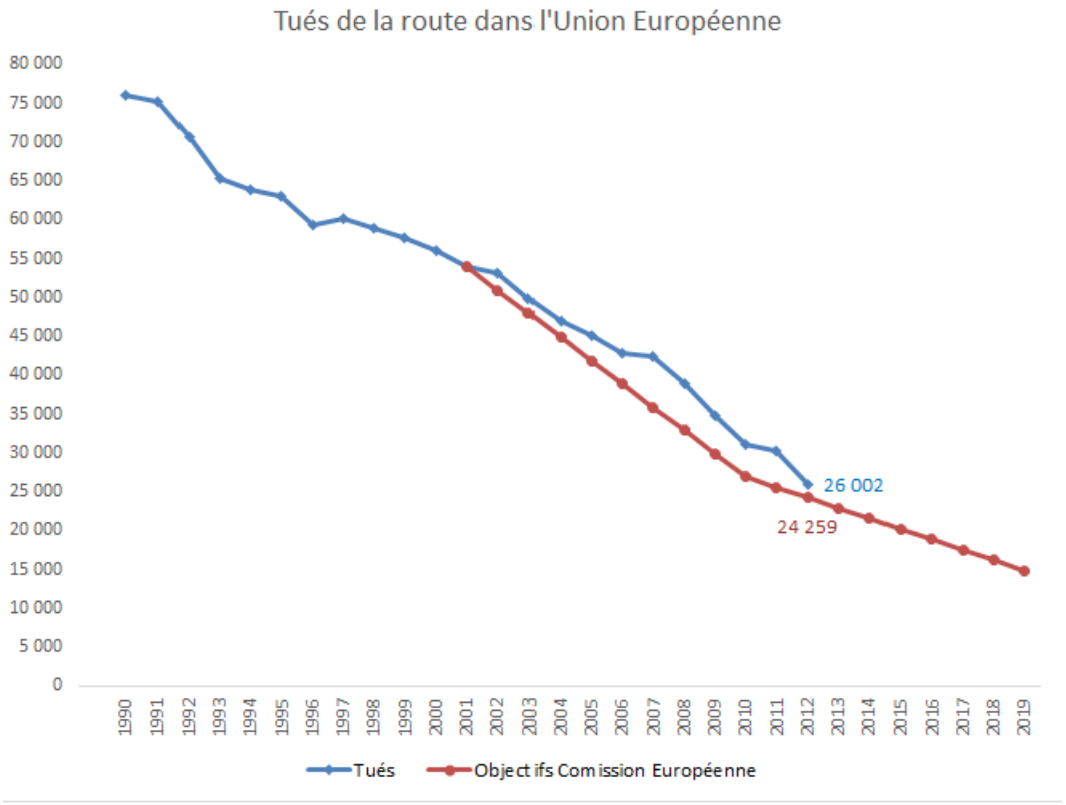

Figure 1: Evolution du nombre de tués de la route dans l'Union Européenne

Les principales initiatives proposées par la Commission Européennes sont :

- Inciter les usagers à un meilleur comportement en harmonisant les sanctions à l'échelle européenne.
- Tirer parti des progrès techniques, rendre les véhicules plus sûrs par l'harmonisation des mesures de sécurité passive.
- Encourager l'amélioration des infrastructures routières avec l'élimination des points noirs.
- La sécurité du transport professionnel de marchandises et de passagers en diminuant le nombre d'accidents de poids lourds, en réglementant la formation des conducteurs professionnels et en faisant respecter les temps de conduite et de repos.
- Les secours et soins aux accidentés de la route en étudiant les meilleurs pratiques dans le domaine des soins médicaux post accidents.

- La collecte, l'analyse et la diffusion des données sur les accidents, pour pouvoir les éviter et atténuer leur gravité.

En France , l'insécurité routière est un problème public depuis plus de 30 ans. En effet, c'est en 1972 qu'a eu lieu une prise de conscience de la part du gouvernement face au pic du nombre de tués atteignant environ 17 000 tués sur les routes françaises. Par la suite en 1973, les premières mesures pour pallier ce problème, apparaissent, ce qui engendre une baisse immédiate du nombre de tués. En 40 ans, le nombre de tués en France a été divisé par 3, tandis que le parc automobile a doublé et le nombre de kilomètres parcourus multiplié par 3. Le bilan provisoire de 2012 fait état de 60556 accidents corporels. Ces accidents ont causé la mort de 3645 personnes (tués à 30 jours) et en ont blessé 75636 dont 27337 hospitalisés. Comparativement à l'année 2011 les chiffres sont en baisse, le nombre de blessés a diminué de 6.9% et celui de la mortalité de 8% soit plus de 300 vies sauvées en 2012. Cependant, malgré les progrès réalisés, la France reste en deçà de l'objectif de 3 000 tués en 2012, objectif fixé en 2007 par le gouvernement Français mais aussi en deçà des résultats obtenus dans la plupart des autres pays de même niveau de vie, notamment en Europe.

Ce bilan fait de la sécurité routière un enjeu social, économique, technologique et politique. L'importance de ces enjeux s'est traduite sous forme de plusieurs types de contre-mesures qui se manifestent à travers des lois, des réglementations, des actions de prévention et des actions de recherche.

Les objectifs fixés ne sont pas simples à atteindre et l'amélioration de la sécurité routière doit s'effectuer de manière quotidienne par l'intervention des pouvoirs publics, des équipementiers, des usagers et des constructeurs automobiles.

En tant qu'organisme de recherche rattaché aux constructeurs automobiles, l'une des missions principales du LAB (Laboratoire d'Accidentologie, de Biomécanique et du comportement humain) est de pouvoir fournir aux constructeurs les éléments nécessaires afin de les aider à connaître les grands enjeux de sécurité actuels et à anticiper les enjeux de demain.

Pour remplir cette mission, Le LAB s'appuie sur les travaux de ses trois pôles : pôle d'accidentologie, de biomécanique et du facteur humain.
L'accidentologie a pour but de comprendre les mécanismes des accidents et d'améliorer la sécurité des véhicules afin de protéger les occupants et réduire les traumatismes subis lors d'un accident de la route. C'est un champs pluridisciplinaire qui nécessite entre autre des compétences en épidémiologie, en statistique, en dynamique

du véhicule ou encore en médecine (pour la compréhension des bilans lésionnels). La Biomécanique des chocs a pour objectifs la compréhension du comportement du corps humain en choc et de spécifier le fonctionnement des stratégies en interaction avec le corps humain.
L'étude du comportement humain a pour but de comprendre et d'évaluer principalement le comportement du conducteur mais aussi des autres occupants lorsqu'ils utilisent les véhicules automobiles.

Les activités menées en accidentologie s'articulent autour des 3 principaux axes suivants :

- Le système d'information sur les accidents de la route (les bases de données).
- Le diagnostic et l'évaluation.
- Les recherches pour développer les méthodes et les outils.

C'est dans ce dernier axe que s'inclut ce travail de thèse. En effet, le diagnostic et l'évaluation requiert sans cesse l'amélioration des méthodes et outils existants ainsi que le développement de nouveaux procédés. De plus, la majorité des travaux et études menés au LAB, quelqu'en soit le domaine d'expertise, s'appuient essentiellement sur l'exploitation et l'analyse des bases de données. Il est donc nécessaire de pouvoir mettre en place des méthodes qui soient capables d'extraire de ces bases des informations nouvelles et utiles pour le détenteur de ces données. Cette thèse entre donc dans cet objectif d'amélioration des méthodes et outils. Elle consiste à développer des méthodes statistiques applicables à l'accidentologie.

Au LAB, l'accidentologie qui est l'étude des accidents de la route, traite des trois domaines : l'accidentologie primaire, l'accidentologie secondaire et l'accidentologie tertiaire.

- L'accidentologie primaire (avant accident) tente d'analyser les causes d'un accident afin de proposer des contre-mesures innovantes en matière de sécurité automobile pour éviter l'accident.

- L'accidentologie secondaire (une fois que l'accident a eu lieu) étudie les conséquences des accidents de la route afin de proposer des contre-mesures pour protéger les occupants en cas de choc.

- L'accidentologie tertiaire est l'ensemble des recherches ayant pour objectif d'améliorer les secours et la prise en charge des usagers ayant été impliqués dans l'accident afin d'éviter l'aggravation de leurs blessures.

C'est dans ce dernier domaine qu'intervient notre travail d'amélioration des méthodes bien que ces dernières soient transposables aux autres domaines. En effet, un des besoins afin d'améliorer la prise en charge des blessés est de pouvoir identifier rapidement le niveau de gravité de l'accident. Cette identification rapide permet aux secours et aux services hospitaliers de traiter et trier plus efficacement les blessés. Un blessé léger ne nécessitera pas de prise en charge immédiate tandis qu'un blessé grave devra être plus urgemment traité.

L'objectif principal de la thèse est donc de développer un modèle statistique qui puisse prédire le niveau de gravité d'un accident.

Ce rapport s'articule en cinq chapitres :

Le premier chapitre intitulé « Enjeux et problématique » a pour but d'expliciter le cadre ainsi que le contexte dans lequel se place ce travail. Ce chapitre définit également la problématique du travail. Tout d'abord, nous définissons ce qu'est la sécurité routière puis nous exposons ses enjeux. Ensuite, nous présentons les différents acteurs intervenant dans la sécurité routière et exposons une revue sur l'évolution des différentes mesures de sécurité mises en place au niveau national. Le LAB est ensuite présenté ainsi que la problématique liée à ce travail. Ce premier chapitre se termine par une introduction aux outils permettant de répondre à la problématique.

Le second chapitre quant à lui traite de l'outil de travail principal, les modèles linéaires généralisés mixtes (GL2M), qui est la famille de modèles que nous utiliserons tout au long du travail. Pour cela nous présentons tout d'abord les modèles linéaires généralisés (GLM) ainsi que les modèles linéaires mixtes (L2M) qui sont les modèles à la base des GL2M et donc nécessaires afin de bien comprendre les GL2M. Les méthodes d'estimation pour ces deux familles de modèles seront également exposées car elles sont à la base de certaines méthodes d'estimation utilisées pour les GL2M. La famille des GL2M est alors ensuite présentée ainsi que la problématique d'estimation qui lui est liée.

Le chapitre 3 traite des méthodes d'estimation des GL2M par approximation de la vraisemblance. En effet, la fonction de vraisemblance des GL2M fait intervenir une intégrale qui n'est pas calculable formellement sauf pour quelques cas

particuliers comme les L2M, les modèles béta-binomial ou les modèles de Poisson-gamma (Martinez, 2006). De ce fait, contrairement aux GLM ou L2M il n'existe pas de méthodes standards d'estimation pour les GL2M. Plusieurs approches ont ainsi été développées afin de soit résoudre, soit contourner les difficultés liées au calcul. Parmi ces méthodes figurent l'approximation numérique de la vraisemblance. Ainsi, après une introduction et une revue sur l'ensemble des méthodes permettant d'estimer les paramètres d'un GL2M, les méthodes d'estimation par approximation numérique de la vraisemblance sont exposées et développées. Nous commençons par présenter les approximations de Laplace du second ordre mais aussi à des ordres supérieurs puis continuons par les approximations de Gauss-Hermite (GH) et de Gauss-Hermite Adaptative (AGH). Une partie sera également dédiée à l'utilisation de ces méthodes dans le logiciel R qui est le logiciel que nous utiliserons pour estimer les paramètres de nos modèles. Enfin, la dernière partie de ce chapitre est consacrée à la comparaison des méthodes de Laplace et AGH par simulations.

Le chapitre 4, quant à lui traite d'un autre type de méthode d'estimation des GL2M, la méthode de quasi-vraisemblance pénalisée (PQL). Cette fois, ce n'est plus la fonction de vraisemblance qui est considérée, comme cela était le cas pour le chapitre 3, mais la fonction de quasi-vraisemblance. Néanmoins tout comme pour la fonction de vraisemblance l'intégrale ne peut être calculée directement et doit être approximée. Ce chapitre expose donc la méthode d'estimation par quasi-vraisemblance pénalisée pour les GL2M. Une partie sera également consacrée à l'utilisation de cette méthode dans le logiciel R et tout comme pour le chapitre précédent une autre partie sera dédiée à la comparaison de cette méthode avec les méthodes de Laplace et AGH du chapitre 4 par simulations. Enfin la dernière partie de ce chapitre traite des avantages et inconvénients de chaque méthode recensés dans la littérature.

Le chapitre 5 qui constitue le dernier chapitre de cette thèse traite de la modélisation de la gravité corporelle par régression logistique mixte. Tout d'abord nous traitons des données utilisées. Pour cela, nous présentons la base de données utilisée, la variable que l'on cherche à modéliser, les variables utilisées ainsi que leur sélection et enfin la constitution de l'échantillon. Ensuite, nous exposons le concept de modélisation mixte ainsi que de la sélection des variables à effets aléatoires. A l'issue de cette sélection de variables à effets aléatoires sept modèles logistiques mixtes sont obtenus, trois à un effet aléatoire, trois à deux effets aléatoires et un à trois effets aléatoires. Le modèle logistique mixte à un effet aléatoire est alors présenté ainsi que l'expression des différentes techniques d'estimation dans ce cas précis. Ensuite les modèles 1 et 2 à un effet aléatoire chacun sont étudiés sur des

jeux de données simulées de différentes tailles mais aussi sur les données réelles. L'objectif étant de pouvoir effectuer une comparaison entre les différentes méthodes et voir si les conclusions établies aux chapitres 3 et 4 demeurent inchangées. Egalement l'étude sur les données réelles permet de voir quels facteurs ont une influence sur la gravité corporelle et comment se quantifie cette influence. Pour finir, les autres modèles mixtes sont présentés ainsi que le modèle logistique classique (sans effet aléatoire) puis comparés en termes d'estimation et de performance prédictive.

Chapter 1

Enjeux et problématique

1.1 La sécurité routière

1.1.1 Définition

La sécurité routière a pour finalité spécifique d'assurer les déplacements routiers sans effets externes indésirables (e.g. sentiment d'insécurité, accidents de la route, blessures, etc.). Elle doit être assurée tant dans l'organisation des déplacements qu'à l'occasion de chaque déplacement (Le Coz and Page, 2003).
Il existe au moins deux visions de la sécurité routière. La première considère trois points de vue : la sécurité primaire (réduire le nombre d'accidents par leur évitement et leur prévention), la sécurité secondaire (augmenter la protection des occupants) et la sécurité tertiaire (améliorer les secours après accident). La seconde fait la distinction entre la sécurité active (référence aux contre-mesures nécessitant l'intervention du conducteur) et la sécurité passive (référence aux contre-mesures indépendantes du conducteur) (Perron and Bocquet, 1997; Ben Ahmed, 2004).

1.1.2 Enjeux

Le bilan annuel des victimes et des dégâts matériels fait de la sécurité routière en France (et dans le monde en général) un enjeu social, économique, technologique et politique :

- Un enjeu humain : les accidents de la route causent des souffrances pour les victimes mais aussi pour la famille, les amis et les proches. Les rescapés et leurs familles subissent également les conséquences souvent longues et douloureuses des traumatismes, des incapacités et de la réadaptation. De plus, il n'est pas rare que les coûts des réparations, des soins ou des obsèques

mais aussi la perte d'un salaire plongent la famille dans la pauvreté (OMS, 2004). Ce sont les familles pauvres ou défavorisées qui sont le plus affectées par ces conséquences financières.

- Un enjeu social : l'insécurité routière affecte la société sur plusieurs plans. En effet, elle se manifeste comme un problème de santé publique puisqu'elle cause la mort et des dégâts dramatiques sur les personnes. L'inconfort, le bruit et la pollution sont considérés aussi comme des défis à surmonter pour assurer la sécurité du public.

- Un enjeu économique : les accidents de la route engendrent aussi des dégâts matériels et humains importants. Dans la plupart des pays, le coût des accidents de la route s'élève à 1% ou 2% du PNB (OMS, 2004). Le coût tutélaire de la sécurité routière en 2002 est de 1 054 949 euros pour un tué, 158 243 euros pour un blessé grave, 23 209 euros pour un blessé léger et 5802 euros pour un accident matériel. Le coût global s'élève à 24,1 milliards d'euros.

- Un enjeu technique : l'enjeu social et l'enjeu économique imposent des contraintes contradictoires. Le compromis recherché est d'assurer la sécurité tout en respectant les contraintes économiques mais en minimisant les coûts. Le développement de nouvelles techniques et technologies paraît l'un des éléments essentiels pour assurer ce compromis en intégrant les objectifs de sécurité dès les phases de conception des véhicules et infrastructures. D'où l'apparition de nouveaux matériaux, de nouvelles structures et de nouveaux dispositifs de sécurité.

- Un enjeu politique : face à l'enjeu socio-économique, la sécurité routière est devenue un enjeu politique : en 2002, le président de la République décide de faire de la sécurité routière un des trois chantiers de son quinquennat. En 2003, un projet de loi « à finalité préventive » a été voté au parlement pour lutter contre l'insécurité routière. De nouvelles mesures de prévention et de répression ont été prises (e.g. la suppression des « permis blancs », la création de nouvelles peines complémentaires, la mise en place d'une chaîne automatisée contrôle-sanction, la mise en place d'un permis probatoire). En 2010, des gouvernements du monde entier ont proclamé une Décennie d'action pour la sécurité routière. L'objectif de cette initiative (2011-2020) est de stabiliser puis d'inverser la tendance à la hausse du nombre de décès dus aux accidents

de la route et de sauver ainsi, selon les estimations, 5 millions de vies sur 10 ans.

1.2 Les différents acteurs

La sécurité routière fait intervenir de nombreux acteurs, dont les conducteurs, l'industrie automobile et les pouvoirs publics. L'acteur central est indéniablement le conducteur, utilisateur du véhicule en tant que pilote et usager de l'infrastructure. L'industrie automobile intervient en tant que concepteur et constructeur des véhicules. Elle doit répondre aux besoins de ses clients automobilistes. Les pouvoirs publics, quant à eux, sont en charge de l'infrastructure. Ils agissent en tant que régulateurs de l'activité des autres acteurs, à travers la réglementation de la conduite (code de la route, formation des conducteurs et politique de répression), de la réglementation automobile, et de la politique d'aménagement du territoire. A ce système socio-économique technique et politique déjà complexe, viennent se greffer d'autres acteurs comme les assureurs, les forces de l'ordre, les services d'urgence et hospitaliers, les réparateurs automobiles...

C'est principalement aux pouvoirs publics et à l'industrie automobile que revient la mission d'améliorer la sécurité routière, par des actions sur les trois composantes de la conduite automobile : le conducteur, le véhicule et l'infrastructure. Par leur rôle de régulateur, les pouvoirs publics ont les moyens réglementaires et sociaux d'agir sur les conducteurs et sur l'infrastructure et donc sur la sécurité par une politique de "prévention routière"'.

1.3 Evolution des mesures de sécurité routière

Depuis les années 70, de nombreuses mesures réglementant le comportement des usagers de la route ont été prises (figure 1.1). Les premières mesures qui suivirent l'année noire de 1972 portent sur le port de la ceinture de sécurité à l'avant du véhicule, le port du casque pour les deux-roues et sur la limitation de la vitesse. Dans les années 80 viennent les réglementations sur la vitesse maximale autorisée pour les cars, la vitesse par temps de pluie pour tous les véhicules, et la mise en place de limiteurs de vitesse dans les véhicules lourds (camions et autocars), la mise en place d'une législation pour lutter contre l'alcoolisme au volant, l'obligation du contrôle technique périodique. Les années 90 voient arriver le permis à points, la limitation de la vitesse à 50km/h dans les agglomérations, l'obligation du port de la ceinture aux places arrières, également l'obligation d'utiliser des moyens de retenue homologués pour le transport d'enfants de moins de dix ans à toutes les

places, le retrait de points pour non port de la ceinture, l'abaissement du taux d'alcool autorisé de 0.8g/l à 0.7g/l puis de 0.7g/l à 0.5g/l, la modification de l'accès à la conduite des motos, le brevet de sécurité routière obligatoire pour conduire un cyclomoteur entre 14 et 16 ans ainsi que l'institution d'un délit de mise en danger de la vie d'autrui et la création d'une police de la route. L'année 2001 voit naître un décret relatif à la recherche de stupéfiants pratiquée sur les conducteurs impliqués dans un accident mortel de la circulation. A partir de 2002, le volet repressif s'accentue avec la multiplication des radars automatiques sur les routes, l'aggravation des peines, la tolérance zéro, la mise en place d'un permis probatoire pour les plus jeunes et l'automatisation du traitement des infractions Et dernièrement en 2012, l'interdiction des avertisseurs de radars, le durcissement des sanctions contre l'usage d'un téléphone ou d'un appareil à écran en conduisant, la sécurisation renforcée des chantiers routiers et de la bande d'arrêt d'urgence, l'obligation d'installer des bandes d'alerte sonore sur les autoroutes pour lutter contre l'endormissement au volant et l'obligation pour tous les conducteurs de véhicules terrestres à moteur de posséder un éthylotest (sauf les cyclomoteurs de moins de 50 cm3).

Egalement depuis 1970, de nombreuses mesures ont été mises en place sur les voitures pour améliorer la protection des occupants et des usagers. Les premières actions ont visé à améliorer la partie sécurité passive. On peut citer parmi elles la ceinture, les airbags, les bosses anti-sous-marinage, les prétensionneurs, la rigidification des habitacles. Leurs contributions, au fil du temps, ont permis de réduire fortement ou d'éliminer certaines blessures graves ou mortelles (Labrousse et al., 2011). Par la suite, sont arrivés sur le marché les systèmes liés à la sécurité active tels que l'ABS, l'ESC, l'AFU,... Leur but, en amont de l'accident, est d'éviter tout simplement d'aller jusqu'à l'accident ou d'en réduire tout au moins les violences du choc, c'est ce qu'on appelle la sécurité active.

Le dernier domaine où l'on peut introduire de nouvelles mesures est la sécurité tertiaire. L'une d'elle consiste à la prise en charge précoce et efficace des polytraumatisés par les secours. Elle peut permettre de réduire le nombre de blessés graves et de tués de manière non négligeable. En effet, un retard lors de la prise en charge du polytraumatisé peut se traduire par une mortalité et une morbidité accrue. Un système d'appel automatique des secours est apparu sur certains véhicules en France depuis 2003 permettant la localisation géographique, l'identification du véhicule et du propriétaire et la possibilité de communiquer en direct avec les impliqués. Ce type de système permet une prise en charge immédiate et appropriée de l'accident, avec un déclenchement des secours les plus proches. Egalement un des besoins pour une prise en charge efficace est de pouvoir identifier rapidement

le niveau de gravité de l'accident afin d'aider les secours à agir de la meilleure manière.
Le problème d'identification de la gravité des blessures est également un problème économique. Le nombre de lits dans les hôpitaux est de plus en plus restreint. La question qui se pose alors se situe au niveau du triage des blessés qui arrivent soit aux urgences si les blessures ne sont pas jugées graves soit dans un service spécialisé dans le cas contraire. Cette estimation de la gravité est en général prise par le médecin responsable à partir d'un bilan médical fait sur place.

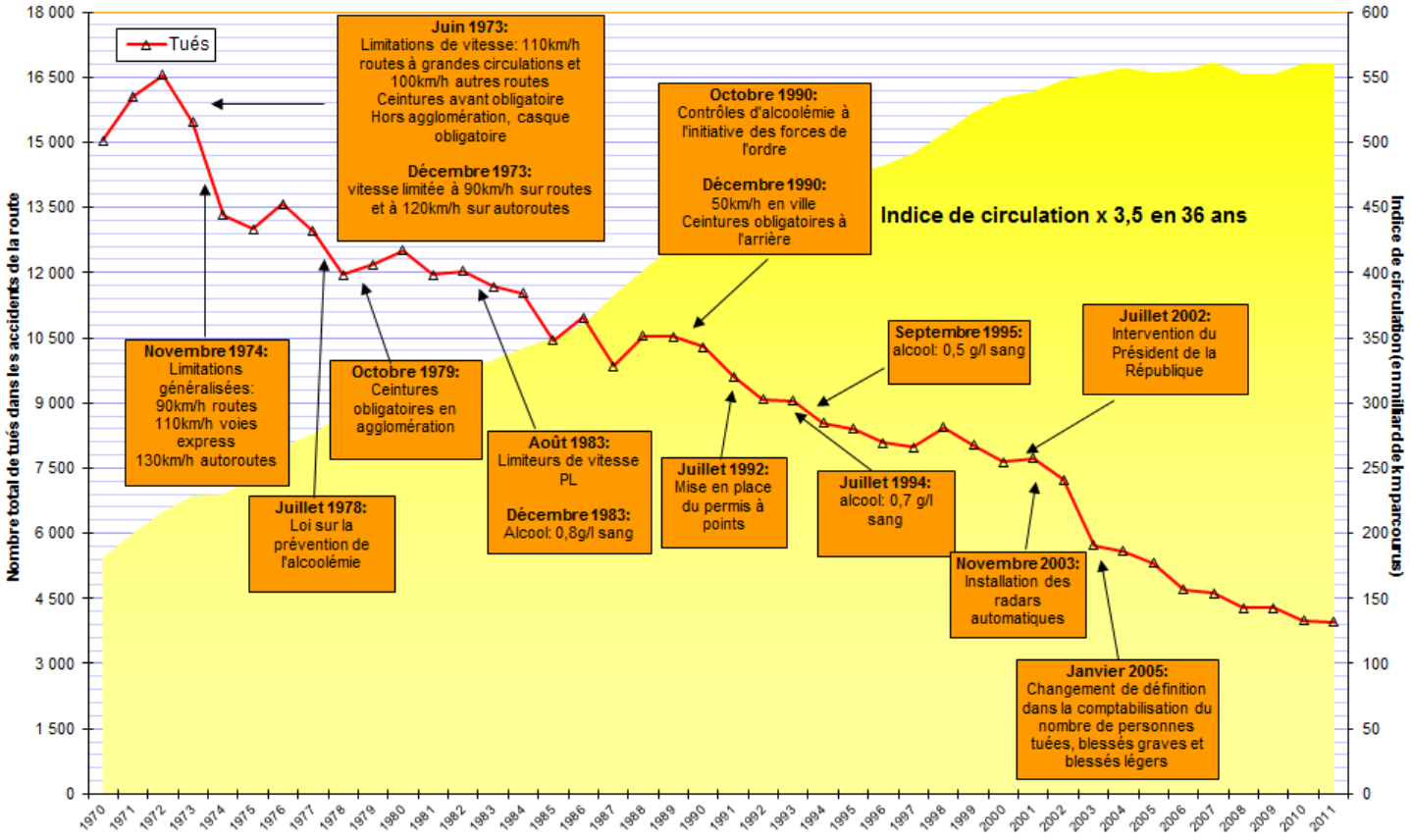

Figure 1.1: Mesures de sécurité routière

La sécurité routière reste toujours un enjeu capital aussi bien pour les pouvoirs publics que pour les constructeurs. En effet, la Commission Européenne a reconduit son objectif de diviser le nombre de tués par deux d'ici à 2020 et souhaite annoncer en 2014 un objectif chiffré de réduction du nombre de blessés graves sur la route pour 2015-2020 (Url, 2013c). Aussi, l'effort des constructeurs français d'automobiles en faveur de la recherche en sécurité routière passe essentiellement

par le Laboratoire d'accidentologie, de biomécanique et d'étude de comportement humain (LAB).

1.4 Le LAB

Le LAB est un laboratoire de recherche en accidentologie, en biomécanique des chocs et sur l'étude du comportement humain commun à PSA-Peugeot/Citroën et Renault et qui a été créé en 1969. Il a pour mission (1) d'acquérir et de transmettre de la connaissance sur les mécanismes accidentels et lésionnels des accidents de la route et (2) de spécifier et d'évaluer l'efficacité des contremesures en s'appuyant sur la réalité des accidents.

L'accidentologie est une science jeune (apparition fin des années 60) qui est définie comme l'étude scientifique des accidents de la route. Elle a pour objectif la compréhension des mécanismes accidentels et des conséquences corporelles sur les usagers. C'est un champs pluridisciplinaire qui nécessite entre autre des compétences en épidémiologie, en statistique, en dynamique du véhicule ou encore en médecine (pour la compréhension des bilans lésionnels).

La Biomécanique des chocs a pour objectifs la compréhension du comportement du corps humain en situation de choc et de spécifier le fonctionnement des stratégies en interaction avec le corps humain. Le travail de compréhension en situation de choc permet d'identifier les liens entre les lésions du corps humain et les efforts et/ou décélérations. Ce travail permet de caractériser des limites en dessous desquelles les blessures peuvent être évitées. Ces limites varient en fonction des caractéristiques humaines. Elles sont par exemple différentes en fonction de l'âge. La résistance des os est inférieure chez les personnes âgées que chez les adultes d'âge moyen. Cet exemple illustre le type de problème qui résulte de la grande diversité du corps humain. Les résultats varient selon l'âge, la taille, la morphologie, etc.

L'étude du comportement humain a pour but de comprendre et d'évaluer principalement le comportement du conducteur mais aussi des autres occupants lorsqu'ils utilisent les véhicules automobiles. En effet, plus de 80% des facteurs accentogènes sont liés au conducteur. Il est le principal acteur de l'accident, il est donc important de pouvoir décrire son comportement durant toutes les phases qui précèdent l'accident. Cela consiste à étudier ses perceptions (du véhicule adverse, du piéton qui traverse, ...), ses interprétations de la situation (évaluation de la situation, du danger, ...), ses décisions face à la situation (tenter un évitement, accélérer, freiner, ...) et ses actions effectives (vitesse et amplitude des coups de volant, blocage de

roues, ...)

A travers ses trois domaines d'expertise, le LAB intervient auprès de plusieurs intervenants de la sécurité routière. Il s'agit premièrement des deux constructeurs automobiles qui sont les membres du GIE pilotant le LAB : PSA et Renault. Le LAB traite les demandes provenant de ces constructeurs mais aussi celles qui sont de sa propre initiative. Il s'agit secondement des intervenants avec qui le LAB travaille en collaboration. Ce sont les instituts, les laboratoires ou les institutions liés au monde automobile, à la sécurité routière, à la recherche ou au monde universitaire. Il y a l'Institut National de Recherche sur les Transports et leur Sécurité (INRETS), La Fondation Sécurité Routière (FSR), les universités médicales et scientifiques, la Commission Européenne, le GHBM (consortium de modélisation de l'être humain) et les instances internationales pour la réglementation (CEVE7, ISO8 et ACEA9). Ces collaborations se font à travers des projets de recherche dont certains sont financés par les pouvoirs publics français (CACIAUP et WIPLASH) ou par les pouvoirs publics européens (COVER, THORAX, CASPER et DaCoTa).

1.5 L'accidentologie

L'étude des accidents est un phénomène assez récent et initié généralement par les constructeurs d'automobiles. L'accidentologie a pour but de comprendre les mécanismes des accidents, d'améliorer la sécurité des véhicules afin de protéger les occupants et réduire les traumatismes subis lors d'un accident de la route. C'est un enjeu majeur pour les constructeurs automobiles. Ces études permettent aussi, en terme de retour d'expérience, d'améliorer les infrastructures, de promouvoir des recommandations lorsqu'une mesure de sécurité s'est avérée efficace. Au LAB, l'accidentologie, étude des accidents de la route, traite des trois domaines de sécurité cités précédemment. Nous les appelons : accidentologie primaire, accidentologie secondaire et accidentologie tertiaire (voir figure 1.2).

- L'accidentologie primaire (avant accident) tente d'analyser les causes d'un accident et traite de l'étude des dispositifs de sécurité active comme l'ABS (contrôle longitudinal du véhicule), l'aide au freinage d'urgence, l'ESP (contrôle latéral du véhicule). Le but est de proposer des contre-mesures innovantes en matière de sécurité automobile pour éviter l'accident.

- L'accidentologie secondaire (une fois que l'accident a eu lieu) étudie les études des conséquences des accidents de la route. Le but est de proposer des contre-mesures pour protéger les occupants en cas de choc. Dans ce domaine, des

enquêtes détaillées sur des voitures accidentées sont menées avec l'aide des forces de police, des épavistes et des services d'urgence. Ces enquêtes permettent de connaître les circonstances de l'accident, d'effectuer des mesures sur les véhicules accidentés et d'obtenir des bilans précis des lésions des occupants.

- L'accidentologie tertiaire est l'ensemble des recherches ayant pour objectif d'améliorer les secours et la prise en charge des usagers ayant été impliqués dans l'accident. Cela consiste donc à l'amélioration de l'organisation des secours afin d'éviter un sur-accident mais également en l'optimisation de leur intervention (localisation, prise en charge des victimes, connaissance du véhicule impliqué, ...) afin d'éviter l'aggravation des blessures des impliqués. L'accidentologie tertiaire est un domaine d'études relativement récent.

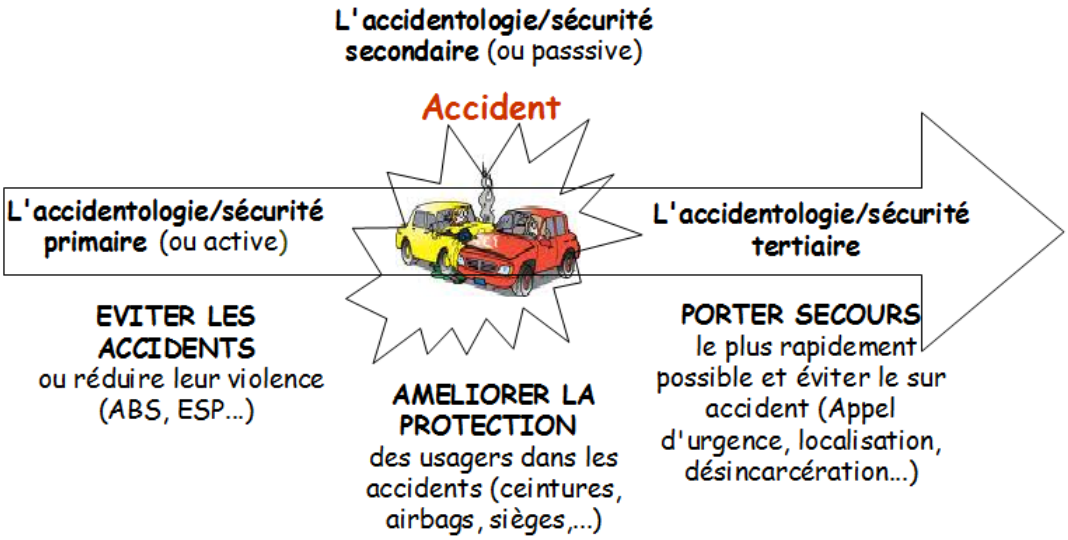

Figure 1.2: Les trois domaines de l'accidentologie

Les activités menées en accidentologie s'articulent autour des 3 principaux axes suivants :

- Le système d'information sur les accidents de la route (les bases de données).
- Le diagnostic et l'évaluation.
- Les recherches pour développer les méthodes et les outils.

1.5.1 Les bases de données

En France, il n'existe pas un seul et unique système d'information en sécurité routière regroupant l'ensemble des connaissances. Chaque acteur constitue son propre système d'information et l'oriente suivant ses propres aspirations. Depuis plus de 40 ans, le LAB a mis en place son système d'information qu'il entretient et fait évoluer année après année. Ce système d'information est aujourd'hui principalement basé autour du recueil de données d'accidents. Les bases de données du LAB reposent essentiellement sur les données d'accidents corporels. Ces bases se décomposent en trois grandes familles :

- **La macro-accidentologie** : Elle concerne l'ensemble des bases de données de type agrégée, c'est à dire que ces bases contiennent l'intégralité des accidents corporels d'une région, d'un pays ou d'une plaque géographique (Europe) mais avec un faible niveau de détails. Ce sont par exemple des bases de données telles que le BAAC (Bulletin d'Analyse des Accidents Corporels de la circulation) en France ou CARE (Community database on Accidents on the Roads in Europe) et IRTAD (International Road traffic and Accident Database) en Europe. L'avantage de ce type de bases c'est d'être représentatif de l'ensemble des accidents corporels du territoire qu'elle représente. On possède donc une grande quantité d'enregistrements (les accidents) avec cependant un nombre limité de caractéristiques. Ces bases de données servent à déterminer les enjeux en termes de sécurité routière et à réaliser des analyses descriptives du problème soulevé afin de donner de grands indicateurs.

- **La micro-accidentologie** : Ce sont les bases de données de type EDA (Études Détaillées d'Accidents). Les accidents sont collectés le plus rapidement possible après leur survenue, sur un territoire délimité et par une équipe spécialisée. En France, ce type d'enquête a été initié par l'INRETS (Ferrandez, 1995) et appliqué par le LAB/CEESAR. L'avantage de ces bases de données repose sur la grande quantité d'informations recueillies (à peu près un millier de variables) mais par contre sur un nombre limité d'accidents. Ce type d'enquête permet la réalisation de reconstructions mais aussi d'analyses fines sur les mécanismes lésionnels ou accidentels afin de répondre à des questions pointues. C'est à partir de ces données que nous travaillerons tout au long de la thèse.

- **Les données d'exposition** : Elles décrivent les données nationales qui permettent d'établir un support pour les analyses statistiques et d'estimer

les risques d'une population ciblée. Le nombre d'habitants en France, le nombre de ménages possédant un véhicule, la longueur du réseau routier national et/ou des différents types de route, le parc des différents véhicules, la consommation du carburant, le nombre de kilomètres parcourus par les différents types de véhicules ou en fonction du type de route etc. sont autant d'exemples de données d'exposition.

1.5.1.1 Le BAAC : Base de données nationale

Fichiers informatisés du Ministère des Transports et de la Gendarmerie :
Il s'agit du fichier national des accidents corporels enrichi par les forces de l'ordre et centralisé par l'ONISR. Tout accident corporel de la circulation routière fait l'objet d'un BAAC (Bulletin d'Analyse d'Accident Corporel de la circulation). Rempli par le service de police ou la gendarmerie, il regroupe quatre types d'information : caractéristiques, lieux de l'accident, véhicules et usagers impliqués. Ainsi, tous les accidents corporels de la circulation routière en France sont recensés puis informatisés. Ce fichier permet l'étude de la gravité et des implications selon des conditions générales (types d'usager, de collision, intempéries, éclairage, tracés en plan...) ainsi que l'évaluation des enjeux en termes de victimes. Il a pour avantage d'être la seule image complète et officielle des accidents en France. Néanmoins, il est orienté vers les statistiques générales et non l'accidentologie. Sa fiabilité est très variable selon la nature de l'information (ex : le port de la ceinture). Les types de chocs et de collisions doivent être interprétés. L'identification des modèles est délicate. Le délai de publication est de plusieurs mois. Beaucoup d'accidents corporels légers ne sont pas officiellement constatés. Il sert principalement à établir les statistiques nationales réalisées par l'Observatoire National Interministériel de la Sécurité Routière (ONISR).

Fichier mensuel des accidents corporels sur le réseau Gendarmerie :
Il contient les BAAC des accidents recensés par la Gendarmerie sur son réseau (rase campagne et petites agglomérations). Sa structure est identique à celle du fichier national dans lequel il est inclus et compte environ trois quarts des tués en France. Sa mise à jour est mensuelle et est transmise au Lab avec un décalage de deux mois. Il est utilisé pour la recherche des cas ciblés pour le fichier LAB secondaire. Il a pour avantage d'être disponible plus rapidement que le fichier national bien que le délai de mise à disposition ait doublé depuis 2004 (tués à 30jours).

1.5.1.2 Les Proces Verbaux d'accidents Mortels (PVM)

Les 2 objectifs principaux de la base de données PVM sont de :

- constituer une base de données spécifique et détaillée sur les accidents mortels survenus en France,
- comparer l'évolution des accidents mortels entre deux décennies : typologie d'accidents, violences d'impacts, lésions, ...

Pour avoir une idée la plus précise possible des différentes configurations des accidents mortels, chaque dix ans, tous les procès-verbaux d'accidents mortels de la France entière d'une ou de deux années complètes sont étudiés en détail. Cette banque de données, qui nécessite beaucoup de travail d'analyse à partir des photos, des plans et des auditions (estimation des vitesses, des angles de choc, des configurations...), permet d'avoir une représentation très fiable de la réalité de la mortalité routière. Elle permet également de mettre en évidence l'évolution des types et violences de choc avec un parc qui aura changé de façon importante durant plus de 10 ans mais aussi d'évaluer des contre-mesures pour les tués et d'effectuer des analyses par thèmes (ex : les enfants, les obstacles fixes, le choc latéral...)

Ces études d'accidents mortels donnent lieu à un codage spécifique : synthèse des données avec les circonstances de l'accident, fiche codage véhicule, fiche codage occupant, fiches médicales, fiches déformations véhicules, plans du site, planches photos, ... Ce codage constitue une base de données sur laquelle s'effectuent les traitements statistiques.

Cette base a pour avantage de donner une image complète des tués en France (y compris immersion, chutes dans ravin,...), de fournir l'évaluation des vitesses et la classification des types de chocs, des manœuvres avant choc etc. faites par des accidentologistes. Cependant, bien que complète cette image des tués dans le temps est fixe, la durée de constitution de la base est longue, le nombre de variables est limité et la base contient peu d'informations en sécurité primaire.

1.5.1.3 La base LAB

La base de données LAB a été conçue à la suite de la création du Laboratoire d'Accidentologie, de Biomécanique et d'étude du comportement humain en 1969. Son objectif est de constituer une base de données sur les accidents de la route, en France, pour pouvoir étudier les mécanismes lésionnels et comprendre comment mieux protéger les occupants des véhicules. Cette base est principalement orientée vers la sécurité secondaire. Les données d'accidents sont relevées en différé par des

experts en sécurité routière ou accidentologistes. Elle regroupe des informations provenant des forces de police, des hôpitaux (bilans médicaux) et des expertises propres à l'accidentologie concernant le véhicule accidenté, l'occupant et ses lésions. Seuls les accidents impliquant au moins un blessé dans un véhicule léger sont étudiés. Chaque année, la base est mise à jour par environ 400 nouveaux véhicules étudiés partout en France. Cette base contient environ 260 variables, recense 14 900 cas en tout depuis 1970. Ces études d'accidents s'effectuent soit sur une zone d'enquête représentative "statistiquement" des accidents en France (Yvelines), soit sur le territoire national après une sélection selon des critères spécifiques (véhicules neufs, chocs frontaux, chocs latéraux, accidents impliquant des enfants, etc.). Ainsi, deux méthodes d'obtention des cas d'accidents existent :

- **L'etude sur zone** : Il s'agit de l'étude systématique de tous les accidents de la zone nord-ouest des Yvelines, sans distinction de marques, ce qui représente environ 250 véhicules par an. L'objectif principal recherché pour cette base est la représentativité des accidents corporels de véhicules légers. Les enquêtes sont réalisées avec une identification des véhicules un à deux jours après l'accident. Dans ce cas, tous les véhicules impliqués en accident corporel sont analysés (toutes marques, toutes violences, tous types de chocs). Les études nécessitent un contact étroit et de confiance avec les forces de l'ordre, les hôpitaux et les épavistes. Les études sur zone donnent lieu à la constitution de dossiers d'accidents complets (synthèse des données BAAC, avec les circonstances de l'accident, fiche de codage véhicule, fiche de codage occupant, fiche médicale, fiche de déformation des véhicules, plans du site, planches photos, ...). Ils donnent également lieu à la constitution d'une base de données informatique multimédia. Le codage des cas est effectué, par chaque enquêteur, de manière périodique, afin d'avoir une base de données en sécurité passive la plus à jour possible.

- **La méthode ciblée** : Cette méthode a pour but d'analyser des accidents présentant un intérêt particulier et ce sur tout le territoire national. Près de 150 voitures accidentées sont analysées en France par an. Les critères de sélection reposent principalement sur :
  - le suivi de nouveaux véhicules commercialisés des deux groupes afin de donner un retour rapide en termes de protection vers les bureaux d'études,
  - certains véhicules de la concurrence,
  - sur des thématiques particulières comme les enfants ou certaines typologies de choc,

Les études de cas ciblés se font en différé et l'analyse repose sur la connaissance du choc. Elles n'ont pas pour vocation de connaître en détail les événements survenus lors de la pré-collision. Les accidents sont sélectionnés à partir des fichiers mensuels des accidents de la circulation fournis par la Gendarmerie Nationale : les fichiers BAAC. Les gendarmes sont démarchés par téléphone, ce qui permettra aux enquêteurs d'identifier précisément les véhicules impliqués et de trouver leurs localisations chez les épavistes et les garages.

Les enquêteurs, pour chaque cas, inspectent les véhicules en cause (mesures de déformations, photos, impacts occupants à l'intérieur du véhicule, utilisation et fonctionnement des moyens de retenue, ...) et obtiennent, auprès des hôpitaux, les bilans médicaux des victimes ainsi que toutes les informations nécessaires à la compréhension de l'accident auprès des forces de l'ordre (identité des victimes, plan de l'accident, localisation, météo, environnement, témoignages, photos prises lors de l'accident,...). Les études de cas ciblés d'accidents donnent lieu, ensuite, à la constitution de dossiers d'accidents complets (synthèses données BAAC avec les circonstances de l'accident, fiche codage véhicule, fiche codage occupant, fiches médicales, fiches déformations véhicules, plans du site, planches photos,...). Une fiche de synthèse comprenant les circonstances de l'accident, des tableaux de mesures (EES, VR, Dv, masse véhicule, enfoncement, intrusion, temps de déclenchement pyrotechnique,...), les bilans lésionnels des occupants, des commentaires sur l'accident et sur le comportement structurel des véhicules et des photos pertinentes (accompagnées de commentaires), est réalisée à chaque fois. Les études de cas ciblés d'accidents donnent également lieu à la constitution d'une base de données informatique multimédia. Le codage est effectué, par chaque enquêteur, de manière périodique, afin d'avoir une base de données la plus à jour possible.

1.5.1.4 Les EDA

L'objectif des EDA est de développer les connaissances sur les mécanismes accidentels (pourquoi et comment l'accident est survenu) et sur les mécanismes lésionnels (quelles sont les blessures des usagers impliqués dans de tels accidents ? Dans quelles conditions les usagers se sont blessés?) par l'analyse des accidents et de leurs conséquences. Pour cela, il s'agit, à partir d'un recueil de données de qualité :

- de reconstruire et de décrire le déroulement de l'accident,
- d'expliciter les enchaînements de causalité qui rendent compte de ce déroulement,

- d'identifier parmi les caractéristiques des usagers, des véhicules et des infrastructures, les facteurs dont le contrôle permettra des actions de prévention et/ou de protection.

La stratégie des EDA repose sur le recueil du maximum de données centrées sur le déroulement de l'accident, sur la scène même de l'accident, par une équipe pluridisciplinaire, intervenant le plus rapidement possible, en coordination avec les secours (OCDE, 1986). L'Etude Détaillée d'un Accident repose sur une méthode d'analyse de l'accident de la circulation routière adaptée à l'utilisation pratique dans les études de diagnostic de sécurité. Cette méthode relève du cadre méthodologique élaboré à l'INRETS (Ferrandez, 1995) dans le contexte de l'Etude détaillée d'accidents (EDA), qui s'appuie sur un modèle séquentiel de l'accident, associé à des approches événementielles, fonctionnelles et causales.

Après traitement, le cas est archivé sur support informatique : check-lists et codages relatifs aux impliqués, aux véhicules et à l'infrastructure, plan et reconstruction cinématique, photos, transcription des entretiens, synthèse sur les circonstances de l'accident et son déroulement.

Afin d'optimiser la collecte de données d'accidents, il existe deux méthodologies d'investigation menée conjointement : l'une sur la scène de l'accident en temps réel et l'autre en temps différé.

Les EDA en temps réel

Les études en temps réel ou investigation sur la scène (on the spot) de l'accident sont considérées comme la meilleure alternative pour la réalisation d'études accidentologiques approfondies car la présence d'un accidentologiste sur les lieux de l'accident, au moment où ce dernier survient, est quasiment impossible et les véhicules ne sont pas encore équipés de dispositifs enregistreurs embarqués exploitables (« boîte noire »).

Aussi le travail consiste à envoyer une équipe d'accidentologistes sur les lieux de l'accident en même temps que les pompiers et les forces de l'ordre puis de procéder rapidement à la collecte des données périssables (entretiens avec les impliqués, photographies des lieux et des véhicules, repérages et mesures des traces sur la chaussée, inspections des véhicules, ...). Une grande partie des informations importantes est recueillie sur le site. Les informations complémentaires sont obtenues par des entretiens avec les impliqués à l'hôpital ou à leur domicile, par l'inspection des véhicules chez les garagistes, et par des recherches au Conseil Général et à la DDE . Le procès-verbal est éventuellement consulté au commissariat, à la brigade

de gendarmerie, ou à la compagnie de CRS . Les bilans médicaux, établissant les lésions des impliqués lors de l'accident, sont obtenus auprès des services hospitaliers, avec l'accord des impliqués.

Cette méthode implique un système d'alerte, couvrant une zone géographique limitée, qui avise l'équipe accidentologiste en temps réel de la survenue d'un accident de la circulation dans cette zone. Cette équipe se rend sur les lieux aussi vite que possible avant que les véhicules ne soient déplacés.

Les EDA en temps différé
Une autre manière de réaliser une EDA est la méthode d'investigation en temps différé (investigation au maximum 24 heures après l'accident). Ce procédé consiste pour l'accidentologiste à se tenir informé, régulièrement, des accidents survenus sur le territoire d'enquête. En différé, l'accidentologiste analyse exclusivement des accidents ciblés définis dans le protocole de recherche.
Il doit recueillir auprès des forces de l'ordre les informations disponibles, notamment grâce aux procès-verbaux en cours de constitution, auprès des commissariats, compagnies de CRS ou des messages émis par les brigades de gendarmerie informant brièvement l'EDSR (Escadron Départemental de Sécurité Routière) des accidents traités (lieu et date de l'accident, configuration de l'accident, immatriculations et type des véhicules, nom et adresse des impliqués). L'investigation se poursuit en recherchant les données complémentaires. Pour cela, l'accidentologiste se déplace sur le site de l'accident, pour établir un plan de la scène de l'accident selon le marquage fait par les forces de l'ordre, pour prendre des photos de l'infrastructure (de l'approche au lieu de l'accident) puis se rend chez l'épaviste où ont été remisés les véhicules accidentés et procède à l'étude des déformations de ces véhicules. Il rencontre les impliqués et essaie de déterminer les conditions de l'accident et les manœuvres effectuées en phase de pré-collision.

Cette méthode donne, généralement, des résultats moins pertinents que la méthode en temps réel, en raison de la perte d'informations et plus particulièrement des données périssables qu'elle génère dans les trois composantes du système : infrastructure, véhicule et usager.

Les Etudes Détaillées d'Accidents (EDA) nécessitent de recueillir sur les lieux de l'accident de très nombreuses informations sur les impliqués, les routes, les véhicules, les conditions de circulation et l'environnement général afin d'obtenir toutes les données nécessaires à la compréhension du déroulement de l'accident et de proposer le scénario le plus probable de ce qui s'est passé.

C'est pourquoi compte tenu des contraintes de temps (informations périssables), ces investigations approfondies ne peuvent s'effectuer que sur un terrain d'observation restreint qu'il faut préalablement définir, et réclame une bonne coopération entre les services publics ou privés qui interviennent sur les accidents de la circulation ou soignent les victimes.
Ce type de recherche sur un territoire vaste, que l'on souhaiterait représentatif du pays pour pouvoir estimer par le calcul statistique des risques ou des risques relatifs d'accident ou de lésions associés à des populations spécifiques engendrerait des coûts trop importants.

A ce jour le territoire d'enquête EDA a été défini dans le nord-est du département autour d'un axe Longjumeau – Evry et couvre un total de 29 communes contiguës dans le département de l'Essonne (91), soit une zone de 9 kilomètres de rayon. L'antenne est implantée à Bondoufle ce qui permet d'être dans une position plus ou moins centrale par rapport au territoire d'enquête et d'être situés à proximité des grands axes routiers et ainsi de permettre un accès rapide à ces derniers à toute heure. La zone d'enquête EDA a été choisie du fait qu'elle permet de répondre aux critères suivants

- Un potentiel d'accidents suffisant,
- Obtention des autorisations du procureur pour consulter les procès-verbaux et pour se rendre sur les lieux de l'accident,
- Un réseau de connaissances suffisant pour s'assurer de la collaboration des services d'urgence et hospitaliers,
- Une bonne répartition des accidents en milieu urbain et rural,
- Une bonne répartition du réseau routier (nationales, départementales, autoroutes, voies express, etc.).

Cette base est plus récente que la base LAB et a vu le jour en 1995. Chaque année 60 nouveaux accidents corporels sont analysés. Cette base contient environ 1000 variables, recense 1100 cas en tout depuis 1995, est plus orientée sécurité primaire bien qu'elle contienne quelques éléments de sécurité secondaire et contient très peu de cas mortels.

1.5.1.5 Bases de données européennes

Egalement, en Europe il n'existe toujours pas de base de données commune, aussi en plus de ses propres bases le LAB a acheté les bases d'accidentologie CCIS (Cooperative Crash Injury Study) du Royaume-Uni et GIDAS (German In-Depth

Accident Study) pour l'Allemagne. Il est important pour le LAB de pouvoir se procurer d'autres bases de données afin d'avoir une vue différente de celle de la France en matière de sécurité routière. En effet, les infrastructures, les habitudes de conduite et les comportements sont différents à travers l'Union Européenne. De plus, chaque pays ayant des façons différentes d'étudier les accidents, ces achats ont aussi pour objectif de comprendre la logique de ces bases de données quant aux méthodes de travail de chaque équipe, mais aussi par rapport à la finalité recherchée de ces bases, aux informations que nous pourrions en tirer, aux analyses statistiques qui pourraient en être faites, à but accidentologique, stratégique et/ou politique pour contrer des avis ou des futurs règlements venant d'autres pays et qui ne seraient pas en faveur des constructeurs automobiles.

- **La base GIDAS** : Les enquêtes détaillées d'accidents en Allemagne ont été initiées en 1970 par les constructeurs d'automobiles allemands mais la base GIDAS (Url, b) n'est née qu'en 1999, résultat d'un projet commun entre l'Automotive Industry Research Association (Forschungsvereinigung Automobiltecnik : FAT) et le Federal Road Research Institute . Tous les types d'accidents avec au moins un impliqué de blessé sont étudiés. Les deux équipes de collecte, à Hanovre et Dresde, sont organisées de sorte à ce que toutes les périodes de la journée soient couvertes ainsi que toutes les périodes de l'année. De plus, la Police, les Pompiers et les services de secours, enregistrent tous les accidents et en informent les équipes. Les accidents sont sélectionnés selon un processus de sélection strict et les données collectées sont comparées aux statistiques officielles des accidents des deux zones d'enquêtes, et des facteurs de pondération sont calculés chaque année ce qui fait que la base de données GIDAS est représentative de la base nationale allemande. Ainsi, chaque année les équipes allemandes étudient environ 2000 cas. Les informations collectées concernent aussi bien l'accidentologie primaire que secondaire. Le recensement se fait en termes d'accidents et non en termes de véhicules impliqués (comme c'est le cas pour la base LAB). Ainsi, à chaque accident entre 500 et 3 000 informations sont collectées (Haviotte, 2007).

- **La base CCIS** : La base de données CCIS (Url, a) est découpée en plusieurs phases de collecte de trois ans. La base a été initiée en 1983, et en est maintenant à sa dixième phase. Cette base est sponsorisée et par le gouvernement et par les industries automobiles. Les phases de recueil 6 et 6a constituent les premières phases de la base de données CCIS. Elles correspondent aux accidents survenus entre 1998 et 2002 inclus. Sept équipes réparties sur sept zones géographiques différentes se chargent de collecter les données.

Les accidents sont collectés par les équipes selon les conditions suivantes : l'accident doit survenir dans la zone géographique de l'équipe d'intervention, dans cet accident au moins un véhicule doit avoir moins de sept ans d'âge et doit être dans un état qui nécessite un remorquage, ce véhicule est appelé : « véhicules cas », dans ce ou ces « véhicules cas » (selon le nombre de véhicules impliqués dans l'accident), il doit y avoir au moins un occupant blessé. De plus, tous les accidents mortels ou sérieux sont étudiés et une sélection est appliquée pour les accidents légers. Les cas ainsi sélectionnés peuvent être pondérés afin d'établir des liens avec les statistiques nationales et permettre une représentativité des bases de données détaillées d'accidents avec les statistiques nationales (Haviotte, 2007).

1.5.2 Le diagnostic et l'évaluation

Le diagnostic de sécurité routière consiste à étudier et établir les différents facteurs d'insécurité et ses conséquences, les principales situations accidentelles et leurs caractéristiques mais aussi les tendances concernant la mortalité et la gravité des accidents de la route. Ce diagnostic permet de mieux comprendre les différents mécanismes accidentels et lésionnels et permet également de dresser un panorama des problèmes accidentels et lésionnels qui restent à résoudre.

L'évaluation en sécurité routière consiste à mesurer les performances et l'efficacité de mesures ou de systèmes de sécurité présents ou à venir. Les résultats permettent aux constructeurs de déterminer l'orientation des choix en équipement de sécurité sur les futurs véhicules et ce d'autant plus dans un contexte économique particulièrement difficile.

Diagnostic et évaluation sont donc essentiels afin de pouvoir fournir aux constructeurs automobiles une expertise qui doit leur permettre d'anticiper les priorités à venir en termes de protection des occupants.

1.5.3 L'amélioration des méthodes et outils

Le diagnostic et l'évaluation requièrent sans cesse l'amélioration des méthodes et outils existants ainsi que le développement de nouveaux procédés. La majorité des travaux et études menés au LAB, quel qu'en soit le domaine d'expertise, s'appuient essentiellement sur l'exploitation et l'analyse des bases de données. Il est donc nécessaire de pouvoir mettre en place des méthodes qui soient capables d'extraire de ces bases des informations nouvelles et utiles pour le détenteur de ces données. Ma thèse entre donc dans cet objectif d'amélioration des méthodes et outils. Elle consiste à développer des méthodes statistiques applicables à l'accidentologie.

1.6 Problématique

1.6.1 Amélioration de la sécurité routière

La prévention des accidents nécessite leur compréhension. Elle permet d'identifier les points sur lesquels les stratégies de sécurité doivent être focalisées. Il s'agit pour cela d'expliquer la raison de la survenue des accidents (mécanismes accidentels) et également de comprendre les mécanismes lésionnels.

Il s'agit d'identifier les causes et facteurs qui ont contribué à provoquer un accident (alcool, vitesse excessive, mauvaises conditions climatiques, etc.). Ces facteurs, qui sont également nommés les facteurs de risque, augmentent la probabilité d'apparition d'un accident (Elvik and Vaa, 2004; Hollnagel, 2004). Cette définition rend alors possible leur identification par l'analyse des données sur les accidents. Les facteurs concernent l'environnement du véhicule et du conducteur (trafic, différentiel de vitesse, masque mobile, verglas, brouillard, etc.), les usagers de la route (âge, alcool, état de santé, défaillances fonctionnels, etc.), le véhicule (angle mort, défaillance, état des vitrages, usure des pneumatiques, etc.), l'infrastructure (signalisation, marquage au sol absent, revêtement usé, accotements non praticables, chaussée étroite, etc.) et la dynamique (distance inter-véhicules inadaptée, vitesse excessive, etc.). Les facteurs sont nombreux et leur détermination repose sur des méthodologies éprouvées.

L'amélioration de la sécurité routière en France, surtout depuis 2002, est en partie due aux changements survenus sur certains de ces facteurs. On observe tout d'abord une baisse des vitesses pratiquées, la vitesse moyenne diminue et les dépassements de vitesse de 10 km/h ont été réduits de 50% en 5 ans. Ensuite, il y a une amélioration du port de la ceinture ; pratiquement 100% aux places avant mais encore uniquement 85% à l'arrière. La réduction du trafic de 1.4% peut aussi être un facteur explicatif de la baisse. Enfin, l'amélioration constante des véhicules en termes de niveau de sécurité (structure des véhicules, conception et diffusion des systèmes de sécurité, etc.) a un effet sur la réduction du nombre de tués.

Aujourd'hui, on souhaiterait pouvoir améliorer l'efficacité d'intervention des secours afin de limiter les conséquences une fois l'accident produit. Une des possibilités est de permettre une meilleure prise en charge de l'accidenté par les secours mais également par les hôpitaux en leur fournissant rapidement des informations concernant la gravité des accidents. C'est dans cet axe que s'inscrit l'objectif de notre travail. En effet, nous essaierons de développer des modèles statistiques afin de prédire au mieux la gravité des accidents et évaluerons leurs performances.

1.6.2 La notion de gravité

Afin de modéliser la gravité des accidents, une première étape est de choisir le critère de gravité à modéliser. Différents critères et scores de gravité existent afin de représenter le degré de gravité d'une personne. En France par exemple le critère de gravité est basé sur la durée d'hospitalisation (voir ci-dessous), aux Etats-Unis l'échelle KABCO (KABCO: K = Killed; A = Incapacitating Injury; B = Non-Incapacitating Injury; C = Possible Injury; O=No Injury; and U=Injured, severity unknown) est utilisée (Url, c). Quant aux scores, certains s'appuient sur des critères anatomiques tel que l'AIS, l'ISS, le NISS, l'ICISS. D'autres sur des critères physiologiques tels que le GCS (Glascow Coma Score), OGS (Organ Grading Scales), RTS (Revised Trauma Score). D'autres scores comme le TRISS s'appuient sur les deux critères en même temps, anatomiques et physiologiques. Afin de décrire le degré de gravité le plus précisément possible il faudrait un critère qui comme le TRISS allie la composante anatomique et physiologique de la personne, mais qui est également capable de prendre en compte la capacité de la personne à répondre aux blessures (O'Keefe and Jurkovich, 2001). La difficulté réside bien évidemment dans l'obtention des informations permettant de calculer ces différents scores.

A ce jour dans nos bases de données nous disposons de trois critères de gravité différents, la gravité au sens de l'ONISR (Observatoire National Interministériel de Sécurité Routière), le MAIS (Maximum Abbreviated Injury Scale) et l'ISS (Injury Severity Score). Ces trois critères se veulent être une représentation de l'état de gravité de l'usager accidenté.

1.6.2.1 Gravité au sens de l'ONISR

La gravité au sens de l'ONISR est codifiée sur quatre niveaux (tué, blessé grave, blessé léger et indemne). Chacun de ces niveaux correspond à des définitions précises, néanmoins il faudra faire attention car ces définitions ont évolué et changé dans le temps. En effet, afin de faciliter les comparaisons internationales, en France, le comité interministériel de la sécurité routière du 7 juillet 2004 a adopté le principe d'une harmonisation des définitions de la gravité retenues dans le fichier national des accidents corporels avec celles adoptées par nos principaux voisins européens.

Aussi, depuis 2005 on adopte les définitions suivantes :

1. Les tués : toute personne qui décède sur le coup ou dans les trente jours qui suivent l'accident ;

2. Les blessés hospitalisés ou graves : victimes admises comme patients dans un hôpital plus de 24 heures ;
3. Les blessés légers : victimes ayant fait l'objet de soins médicaux mais n'ayant pas été admises comme patients à l'hôpital plus de 24 heures ;
4. Les indemnes : impliqués non décédés et dont l'état ne nécessite aucun soin médical.

Tandis qu'avant 2005, les définitions suivantes étaient utilisées :

1. Les tués : toute personne qui décède sur le coup ou dans les six jours qui suivent l'accident ;
2. Les blessés hospitalisés ou graves : victimes admises comme patients dans un hôpital plus de six jours ;
3. Les blessés légers : victimes ayant fait l'objet de soins médicaux mais n'ayant pas été admises comme patients à l'hôpital plus de six jours.
4. Les indemnes : impliqués non décédés, non hospitalisés et dont l'état ne nécessite aucun soin médical ;

Néanmoins, cette harmonisation se faisant seulement avec les principaux pays voisins européens, il n'y a pas de définition commune de la gravité à l'échelle européenne et encore moins à l'échelle internationale. A l'inverse, la définition des tués est relativement harmonisée en Europe, voire dans le monde. Il en résulte alors que la plupart des études internationales se basent alors essentiellement sur les tués.

Cette situation de non harmonisation de la gravité est cependant une préoccupation aussi bien dans l'Union Européenne (U.E) qu'aux Etats Unis. En effet, récemment la Commission Européenne et les Etats membres se sont accordés sur une politique de réduction des blessés graves des accidents de la route d'ici à 2020. Dans ce contexte, la définition commune est donc une première étape nécessaire à la réalisation de cet objectif. L'UE est parvenue, avec l'aide d'experts issus des États membres, à une proposition de définition commune de ce qu'est une blessure grave. Cette définition commune s'appuie sur le MAIS (Maximum Abbreviated Injury Scale) (Url, 2013b) défini ci-après. Egalement, aux Etats-Unis le NCHRP (National Cooperative Highway Research Program) se penche sur l'utilisation de l'AIS comme une mesure de blessure grave à la place de l'usuelle définition KABCO utilisée aux États-Unis.

1.6.2.2 AIS et MAIS

Avant de définir le MAIS nous définirons l'AIS puisque le MAIS est le maximum des AIS

L'AIS, Abbreviated Injury Scale (figure 1.3), développé par l'Association Américaine pour l'avancement de la médecine automobile (Url, 2013a) , a été conçu afin de fournir aux chercheurs une méthode numérique simple pour hiérarchiser et comparer les blessures par degré de sévérité ainsi que pour standardiser la terminologie décrivant les lésions (Ceesar et al., 2004).

L'échelle AIS se fonde sur la lésion anatomique et, en ce sens, elle diffère d'autres systèmes qui sont liés à des paramètres physiologiques. La conséquence de ce principe est qu'il y a une seule valeur AIS pour chaque lésion d'une victime, tandis que dans les échelles qui dépendent des paramètres physiologiques, de nombreux scores sont possibles chez la même personne au cours du temps.
Les valeurs d'AIS quantifient les blessures et non les conséquences de celles-ci. Grâce à ce principe l'AIS peut être utilisé comme une évaluation intrinsèque de la gravité de la lésion et non comme une mesure des incapacités qui peuvent en résulter.

L'échelle AIS est un système de notation qui résulte d'un consensus fondé sur un repère anatomique et basé sur le risque vital de la blessure (Lesko et al., 2010). Il a pour but de classer la sévérité d'une blessure au sein d'un territoire corporel (voir ci-dessous) selon une échelle de sévérité allant de 1 (blessure mineure) à 6 (au-delà de toute ressource thérapeutique).

L'AIS est l'une des échelles anatomiques les plus courantes pour les lésions traumatiques (Peitzman et al., 2007) et est internationalement utilisée dans les recherches en accidentologie (Ceesar et al., 2004).

Territoire corporel

1. Tête
2. Face
3. Cou
4. Thorax
5. Abdomen

6. Colonne vertébrale
7. Membres supérieurs
8. Membres inférieurs
9. Indéterminé

AIS

1. Mineure
2. Modérée
3. Sérieuse
4. Sévère
5. Critique
6. Maximale
7. Inconnu

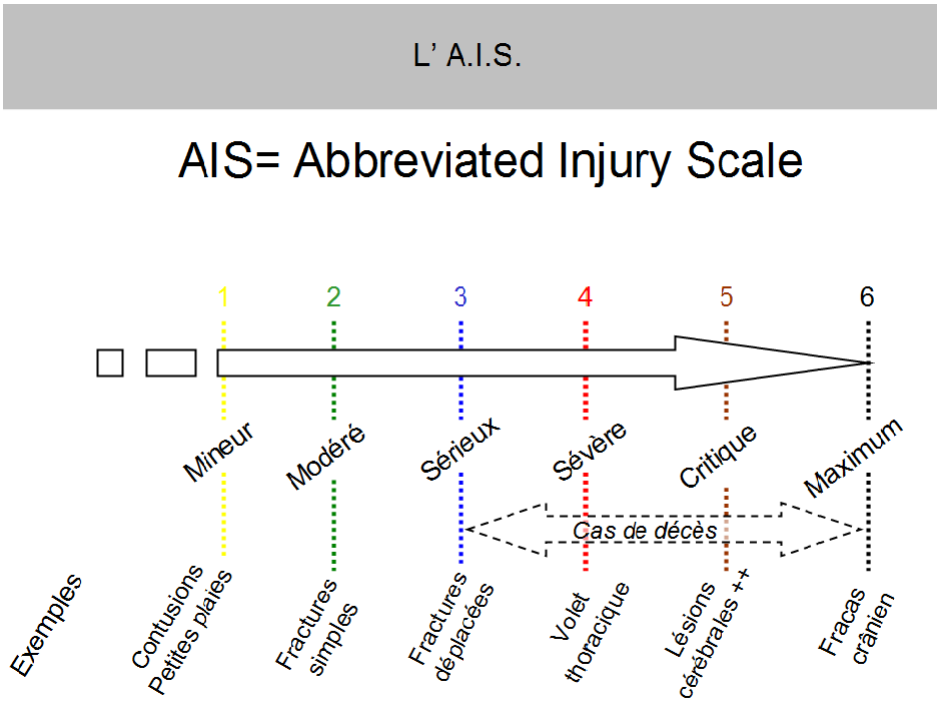

Figure 1.3: Abbreviated Injury Scale

Ainsi, chaque lésion identifiée chez un usager peut être codée. Cependant, l'AIS ne comprend pas d'évaluation des effets conjugués des associations lésionnelles chez les victimes. Il est donc nécessaire d'avoir une synthèse globale du niveau de blessure dont souffre l'occupant. Le MAIS est une solution permettant de répondre à ce besoin.

Le MAIS (pour Maximum AIS) est l'AIS le plus élevé recensé chez un blessé ayant subi des lésions multiples. Il est utilisé par les chercheurs pour définir le niveau global de sévérité des lésions. Son intérêt demeure important dans la recherche sur l'évolution des blessures accidentelles en fonction de la conception des véhicules. Cependant, dans les recherches en traumatologie, le MAIS a été jugé insuffisant en raison de sa relation non linéaire avec le risque de décès. De plus, les taux de mortalité varient significativement au sein de chaque valeur d'AIS pour chaque lésion principale en fonction de l'AIS de la seconde blessure la plus sévère.

L'obtention de l'AIS et donc du MAIS n'est possible que si un bilan médical existe, ce qui limite son utilisation. Pour cela, le LAB a développé d'autres valeurs MAIS pour les tués sans dossier médical (7=Tué non autopsié, 8=Tué incarcéré et 10= Tué, suspicion de malaise cardiaque ou décès indirectement lié à des lésions majeures). Egalement de nouvelles notions font leurs apparitions telles que le MAIS3+ qui englobe toutes les valeurs MAIS supérieurs ou égales à 3 ou le MAIS2+ qui englobe toutes les valeurs MAIS supérieurs ou égales à 2. Ces nouvelles notions permettent de diviser la gravité en deux grands groupes un peu à l'image de grave et non grave. Elles sont très utiles dans les études où nous ne sommes pas intéressés par chaque niveau de gravité mais seulement pas les niveaux de gravité les plus importants, ce qui est par ailleurs souvent le cas.

1.6.2.3 L'ISS

L'ISS (Injury Severity Score) publié en 1974 par Baker et al. (1974) fournit une bien meilleure corrélation entre la gravité globale des blessures et la probabilité de survie (Baker et al., 1974; Bull, 1975) que le MAIS. L'ISS est la somme des carrés des AIS les plus élevés des trois régions corporelles les plus atteintes. Les six régions corporelles utilisées dans l'ISS sont les suivantes :

Régions corporelles

1. Tête ou Cou
2. Face

3. Thorax
4. Abdomen et contenu pelvien
5. Membres ou ceinture pelvienne
6. Externes (toute la surface cutanée)

- Les lésions de la tête ou du cou comprennent les lésions cérébrales et de la colonne cervicale ainsi que les fractures du crâne et de la colonne cervicale.
- Les lésions de la face incluent celles intéressant la bouche, les oreilles, les yeux, le nez et les os de la face.
- Les lésions du thorax et du contenu de l'abdomen ou du bassin comprennent toutes les blessures des organes internes de l'une ou l'autre de ces cavités. Les blessures du thorax incluent aussi celles du diaphragme, de la cage thoracique, et de la colonne dorsale. Les lésions de la colonne lombaire sont incluses dans le contenu abdomino-pelvien.
- Les lésions des membres, de la ceinture scapulaire, ou de la ceinture pelvienne comprennent les entorses, fractures, luxations et amputations à l'exception de celles de la colonne vertébrale, du crâne et de la cage thoracique.
- Les lésions externes incluent les plaies, contusions, abrasions et brûlures indépendamment de leur localisation sur la surface du corps.

Il faut noter que les régions corporelles de l'ISS ne coïncident pas nécessairement avec les sections utilisées dans l'AIS. Par exemple la région COLONNE VERTÉBRALE de l'AIS est divisée en trois régions corporelles dans l'ISS ; la colonne cervicale est dans la région TÊTE ou COU de l'ISS, la colonne dorsale est dans la région THORAX et la colonne lombaire est dans la région ABDOMEN ET CONTENU PELVIEN.

Ces différences doivent être soigneusement prises en compte de façon à éviter des erreurs lors de l'attribution des lésions à la région corporelle adéquate pour le calcul de l'ISS.

Les valeurs d'ISS varient de 1 à 75. On obtient la valeur 75 de deux manières, soit par trois lésions d'AIS 5, soit par au moins une lésion d'AIS 6. Toute lésion

d'AIS 6 se voit automatiquement attribuer un ISS de 75. Cependant le codeur doit coder toutes les lésions même si elles n'augmentent pas la valeur de l'ISS. **Il n'est pas possible de calculer un ISS pour un blessé qui a au moins une lésion d'AIS 9 ;** d'où la nécessité d'obtenir un bilan médical complet (Ceesar et al., 2004).

1.7 Introduction aux modèles utilisés

Dans cette section, nous introduisons les méthodes permettant de modéliser et analyser la gravité des accidents. Tout d'abord les notions de bases nécessaires à la compréhension du travail sont définies. Ensuite, une revue des méthodes appliquées à l'étude des accidents est faite.

1.7.1 Définition des termes

1.7.1.1 Variable à expliquer et variable explicative

On appelle variable à expliquer (ou variable dépendante ou variable réponse ou variable d'intérêt) la variable que l'on cherche à analyser (à expliquer) et variables explicatives (ou variables indépendantes ou facteurs explicatifs) les variables qui ont une influence sur la variable à expliquer. Par exemple, expliquer l'hypertension artérielle des patients (variable à expliquer) à partir de leurs caractéristiques physiologiques, cliniques et comportementales tels que le sexe, le fait de fumer ou pas, la pratique régulière d'exercices physiques, etc. (variables explicatives).

1.7.1.2 Variable catégorielle

Une variable catégorielle (ou qualitative) est une variable possédant plusieurs catégories (ou modalités) c'est à dire qu'elle attribue à chaque sujet une étiquette. Par exemple les variables 'sexe' (masculin ou féminin), couleur (bleu ou vert ou jaune) ou satisfaction ('pas du tout', 'un peu', 'très') sont des variables catégorielles. Cependant on fait la distinction entre les variables dont l'ordre des modalités a un sens qu'on appelle variables ordinales et celles dont l'ordre n'a aucun sens qu'on appelle variable nominales. Par exemple, les modalités de la variable satisfaction ('pas du tout', 'un peu', 'très') ont un ordre, ce qui fait d'elle une variable ordinale. En revanche, les modalités des variables sexe (masculin ou féminin) et couleur (bleu ou vert ou jaune) n'ont pas d'ordre et ces variables sont donc nominales.

1.7.2 Revue des méthodes utilisées en analyse de gravité d'accidents

1.7.2.1 Modèles classiques

L'analyse de la gravité des accidents ou des blessures se fait généralement à l'aide de variables catégorielles. Les méthodes statistiques utilisées par les chercheurs dépendent en premier lieu de la nature de la variable catégorielle à expliquer. Celle-ci peut être de nature binaire (grave, non-grave par exemple) ou multinomial (indemne, léger, grave, fatal par exemple). Dans le cas multinomial, la variable à expliquer peut être alors soit nominale soit ordinale.
Lorsque la variable à expliquer catégorielle est binaire (0/1, oui/non, grave/non grave) on utilise les modèles dichotomiques. Il s'agit alors généralement d'expliquer et prédire la survenue ou non d'un évènement ou d'un choix à partir d'une collection de variables explicatives. On peut aussi voir cela comme un problème de classification ou de discrimination puisque l'on cherche à classer les observations dans un groupe ou dans l'autre. Les principaux modèles pour des variables binaires sont le modèle logit (plus communément appelé régression logistique binomiale) ou le modèle probit. Ces deux modèles possèdent des propriétés identiques. Cependant, dans de nombreux cas, on préfère utiliser le modèle logistique car d'un point de vue numérique, la transformation logistique est plus simple à manipuler (notamment pour l'écriture des estimateurs du maximum de vraisemblance), de plus on a une interprétation claire des coefficients en terme d'odds ratio pour la transformation logistique. Enfin le modèle logit possède la propriété que les estimateurs de pente (c'est à dire des paramètres relatifs aux variables explicatives) sont invariants à une sur-représentation des individus (observations) ayant la caractéristique définie par la variable réponse. Seule la constante du modèle est affectée par la sur-représentation (Leblanc et al., 2000). Cette propriété indique donc que si le modèle logistique est vrai dans la population, peu importe que l'on sur-représente dans l'échantillon l'une des modalités de la variable réponse (Droesbeke et al., 2005).
L'approche de la régression logistique peut être facilement généralisée à l'explication des valeurs prises par une variable explicative catégorielle nominale à K (K $>$ 2) modalités. Par exemple, si l'on souhaite expliquer les préférences que les enfants ont en matière de sucrerie, la variable à expliquer serait alors le type de sucrerie ('chocolat', 'caramel' et 'bonbon') en fonction d'une ou plusieurs variables explicatives telles que le sexe, l'âge etc. La prédiction et l'interprétation des coefficients de la régression se font de manière similaire au cas binaire en prenant une modalité de référence. On parle de régression logistique polytomique à variable dépendante nominale ou de régression logistique multinomiale en référence à la distribution utilisée pour modéliser la probabilité d'appartenance à un groupe.

Concernant ce genre de modèles, une multitude d'interprétations est possible. En effet, nous pouvons faire l'impasse sur le caractère ordinal afin de revenir simplement au modèle multinomial mais également assimiler la variable à expliquer à une variable quantitative ce qui reviendrait à faire une régression linéaire multiple. Entre ces deux cas extrêmes existent différentes approches comme les logits adjacents et les logits cumulatifs (Droesbeke et al., 2005). On parle alors de régression logistique polytomique à variable dépendante ordinale.

Ces différents modèles sont très utilisés dans le domaine de l'accidentologie. Par exemple, Chen et al. (2003) utilisent la régression logistique pour étudier les facteurs de risque associés à la gravité des blessures lors des accidents en intersection. Al-Ghamdi (2002) utilise la régression logistique pour estimer l'influence de certains facteurs sur la gravité de l'accident. Kong and Yang (2010) utilisent la régression logistique pour étudier le risque de blessures des piétons impliqués dans des accidents de véhicules en Chine. Haleema and Gana (2011) utilisent un modèle probit binaire et un modèle probit ordinal afin d'identifier des prédicteurs traditionnels et des prédicteurs moins traditionnels de la gravité des accidents sur les routes principales urbaines. Kockelman and Kweon (2002) utilisent un modèle probit ordinal pour étudier le degré de blessures de conducteurs impliqués dans différents types de collision. O'Donnell and Connor (1996) utilisent un modèle logit ordinal et un modèle probit ordinal pour prédire la sévérité des accidents de véhicules motorisés. Islam and Mannering (2006) utilisent un modèle multinomial pour analyser les différences de gravité de blessures entre les hommes et femmes et ce à travers différentes tranches d'âge. Binghama et al. (2008) utilisent un modèle logistique et modèle multinomial afin de prédire la conduite et ses conséquences après consommation d'alcool, de marijuana ou autres drogues. Pour une revue complète sue les modèles statistiques en accidentologie (voir Anastasopoulosa and Mannering, 2011).

1.7.2.2 Modèles mixtes ou à effets aléatoires

Bon nombre de chercheurs effectuent ainsi l'analyse des blessures par le biais de modèles à variables discrètes, le plus souvent par le modèle logistique binomial ou multinomial (Ulfarssoon and Mannering, 2004; Yau, 2004), voir également les références ci-dessus . Ces derniers considèrent les effets des variables explicatives identiques pour toutes les observations de blessures. Cependant, une hétérogénéité inobservée entre les différents usagers de la route peut exister et certains effets des variables explicatives varier. De tels effets, s'ils ne sont pas pris en compte peuvent mener à des potentiels biais ainsi qu'à des inférences statistiques erronées (Mc Fadden and Train, 2000). Les modèles logistiques mixtes récemment développés permettent de palier à ces problèmes par l'introduction d'effets aléatoires faisant ainsi

varier l'effet de certains facteurs. Par exemple, dans son analyse de la gravité des blessures des piétons, Kim et al. (2008) trouvent la présence d'hétérogénéité et trouvent que celle-ci est captée par l'âge du piéton. Ce résultat s'explique par le fait que l'état de santé des piétons les plus âgés présente une plus grande variance que chez les jeunes piétons. Finalement la présence d'hétérogénéité captée par l'âge est due à une information manquante qui est l'état de santé du piéton. Ainsi, les modèles mixtes sont utilisés pour palier à une information manquante qui impacte le niveau de gravité. On retrouve aussi cette idée chez Anastasopoulosa and Mannering (2011). En effet, ils montrent que la prédiction de la gravité des blessures chez les accidentés de la route est bien meilleure en utilisant un modèle à effets aléatoires associé à des données d'accidents générales que la prédiction obtenue à l'aide d'un modèle classique associé à des données d'accident détaillées.

Chapter 2

Modèles linéaires généralisés mixtes

2.1 Introduction

Dans ce chapitre nous exposons tout d'abord les éléments de la théorie de base nécessaire à une discussion sur les modèles linéaires généralisés mixtes, notés GL2M (Generalized Linear Mixed Models), utilisés dans ce travail. En effet, ces derniers résultent de la fusion des modèles linéaires généralisés, notés GLM (Generalized Linear Models) et des modèles linéaires mixtes, notés L2M (Linear Mixed Models). Ce chapitre expose donc les techniques d'inférence pour ces deux familles de modèles car elles sont à la base de certaines techniques d'inférence utilisées pour les GL2M. La famille des GL2M est aussi présentée ainsi que la problématique d'estimation qui lui est liée.

2.2 Modèle linéaire généralisé (GLM)

2.2.1 Définition

La classe des GLM est une extension des modèles linéaires classiques en termes de loi. Elle a une place importante dans la modélisation statistique, trouvant son intérêt dans différents domaines d'application. Dans leur ouvrage, McCullagh and Nelder (1989); Agresti (2002) en font une présentation complète.

On note $y = (y_1, \ldots, y_n)$ le vecteur de taille n des observations, réalisation du vecteur aléatoire Y, la variable à expliquer. Un GLM se caractérise par les trois hypothèses suivantes :

- (H1) Les composantes y_i sont supposées indépendantes et distribuées selon une loi appartenant à la famille exponentielle au sens de Nelder and Wed-

derburn (1972), autrement dit la fonction de densité de la variable aléatoire Y_i est donnée par :

$$f_{Y_i}(y_i, \gamma_i) = \exp\left(\frac{y_i\gamma_i - b(\gamma_i)}{a_i(\kappa)} + c(y_i, \kappa)\right) \quad \text{pour } i \in \{1, \ldots, n\} \tag{2.1}$$

où γ_i est un paramètre canonique et κ un paramètre de dispersion. Les fonctions b et c sont spécifiques à chaque distribution et la fonction a_i s'écrit : $a_i(\kappa) = \frac{\kappa}{\omega_i}$ avec ω_i un poids connu associé à l'observation i.

L'espérance et la variance de la variable associée s'expriment à l'aide des fonctions a_i et b et leurs dérivées :

$$E(Y_i) = b'(\gamma_i) \tag{2.2}$$

$$Var(Y_i) = a_i(\kappa)b''(\gamma_i) \tag{2.3}$$

Il existe donc une relation directe entre l'espérance de Y_i, notée μ_i, et sa variance :

$$\begin{aligned} Var(Y_i) &= a_i(\kappa)b''(b'^{-1}(\mu_i)) \\ &= \frac{\kappa}{\omega_i}v(\mu_i) \end{aligned} \tag{2.4}$$

Où $v = b'' \circ b'^{-1}$ désigne la fonction de variance.

- (H2) On définit le prédicteur linéaire

$$\eta = X\beta \tag{2.5}$$

où β est un vecteur de paramètres de taille p qu'on souhaite estimer, et X est la matrice connue de taille $n \times p$ contenant les variables explicatives.

- (H3) Le lien entre l'espérance de Y_i, μ_i, et la $i^{\text{ème}}$ composante du prédicteur linéaire η_i est réalisé par la fonction g (monotone et différentiable) appelée fonction de lien :

$$\eta_i = g(\mu_i). \tag{2.6}$$

Une fonction de lien pour laquelle $\eta_i = \gamma_i$ est appelée fonction de lien canonique.

Pour résumer, les modèles linéaires généralisés sont caractérisés par deux fonctions :

- la fonction de lien, spécifiant l'introduction de la linéarité,
- la fonction de variance, spécifiant la relation entre l'espérance et la variance.

2.2.2 Estimation

Pour n observations supposées indépendantes et en tenant compte du fait que γ le paramètre canonique dépend de β, la log-vraisemblance d'un GLM s'écrit :

$$L(\beta; y) = \sum_{i=1}^{n} \log(f_{Y_i}(y_i)) = \sum_{i=1}^{n} l_i$$

Calculons

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \gamma_i} \frac{\partial \gamma_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

Comme

$$\frac{\partial l_i}{\partial \gamma_i} = \frac{y_i - b'(\gamma_i)}{a_i(\kappa)} = \frac{y_i - \mu_i}{a_i(\kappa)}$$

En utilisant (2.3) nous avons

$$\frac{\partial \mu_i}{\partial \gamma_i} = \frac{\partial b'(\gamma_i)}{\partial \gamma_i} = b''(\gamma_i) = \frac{\mathrm{Var}(Y_i)}{a_i(\kappa)}$$

$$\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \sum_{k=1}^{p} X_{ik}\beta_k}{\partial \beta_j} = x_{ij}$$

$$\frac{\partial \eta_i}{\partial \mu_i} = \frac{\partial g(\mu_i)}{\partial \eta_i} = g'(\mu_i).$$

D'où

$$\begin{aligned} \frac{\partial L(\beta)}{\partial \beta} &= \sum_{i=1}^{n} \frac{(y_i - \mu_i) X_{ij}}{g'(\mu_i)} \\ &= \sum_{i=1}^{n} X_{ij} \frac{1}{g'(\mu_i)^2 \mathrm{Var}(Y_i)} g'(\mu_i)(y_i - \mu_i), \quad j = 1, \ldots, p. \end{aligned} \tag{2.7}$$

On considère les matrices diagonales définies par

$$W_\beta = \text{diag}\{\text{Var}(Y_i)(g'(\mu_i))^2\}_{i=1,\dots,n} = \text{diag}\{W_{i,\beta}\}_{i=1,\dots,n}$$

et

$$\frac{\partial \eta}{\partial \mu} = \text{diag}\{\frac{\partial \eta_i}{\partial \mu_i}\}_{i=1,\dots,n} = \text{diag}\{g'(\mu_i)\}_{i=1,\dots,n}$$

où W_β est la matrice de pondération.

Les équations du maximum de vraisemblance pour β s'écrivent alors

$$X' W_\beta^{-1} \frac{\partial \eta}{\partial \mu}(y - \mu) = 0. \tag{2.8}$$

Ce système d'équations n'étant pas linéaire en β, on utilise des méthodes itératives pour maximiser la log-vraisemblance telle que que la méthode de Polak Ribière, de Newton ou de Fisher scoring.

Comme $L(\beta; y) = \sum_{i=1}^{n} l_i$ il nous suffit de calculer pour une observation i la dérivée seconde et ensuite d'en faire la somme.

Pour une observation i, d'après (2.7) :

$$\frac{\partial l_i}{\partial \beta_j} = (y_i - \mu_i) W_{i,\beta}^{-1} g'(\mu_i) X_{ij}$$

d'où

$$\begin{aligned} \frac{\partial l_i^2}{\partial \beta_j \beta_k} &= \frac{\partial}{\partial \beta_k}\left((y_i - \mu_i) W_{i,\beta}^{-1} g'(\mu_i) X_{ij}\right) \\ &= \frac{\partial}{\partial \beta_k}\Big((y_i - \mu_i)\Big) W_{i,\beta}^{-1} g'(\mu_i) X_{ij} + (y_i - \mu_i) \frac{\partial}{\partial \beta_k}\left(W_{i,\beta}^{-1} g'(\mu_i) X_{ij}\right). \end{aligned}$$

Le dernier terme du membre de droite, que l'on notera K, possède une espérance nulle et il sera nul lorsqu'on évaluera non plus $\frac{\partial l_i^2}{\partial \beta_j \beta_k}$ mais son espérance (Droesbeke et al., 2005). Pour dériver le premier terme, on utilise encore la composition de fonction

$$\frac{\partial \mu_i}{\partial \beta_k} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_k}.$$

$$\frac{\partial l_i^2}{\partial \beta_j \beta_k} = -\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_k} W_{i,\beta}^{-1} \frac{\partial \eta_i}{\partial \mu_i} X_{ij} + K.$$

En passant à l'espérance nous avons

$$-\mathrm{E}(\frac{\partial l_i^2}{\partial \beta_j \beta_k}) = X_{ik} W_{i,\beta}^{-1} X_{ij} - \mathrm{E}(K)$$

avec $\mathrm{E}(K) = 0$.

d'où

$$-\mathrm{E}(\frac{\partial L^2}{\partial \beta \beta^T}) = X^T W_\beta^{-1} X \tag{2.9}$$

En combinant (2.8) et (2.9) et utilisant l'algorithme de Fisher, nous obtenons le système itératif suivant :

$$\begin{aligned} \beta^{(t+1)} &= \beta^{(t)} + (X^T W_{\beta^{(t)}}^{-1} X)^{-1} X^T \frac{\partial \eta}{\partial \mu} W_{\beta^{(t)}}^{-1} (Y - \mu) \\ &= \beta^{(t)} + (X^T W_{\beta^{(t)}}^{-1} X)^{-1} X^T W_{\beta^{(t)}}^{-1} \frac{\partial \eta}{\partial \mu} (Y - \mu) \end{aligned}$$

En prémultipliant (à gauche) par $(X^T W_{\beta^{(t)}}^{-1} X)$ nous obtenons une autre forme plus classique et plus facile à manipuler

$$\begin{aligned} (X^T W_{\beta^{(t)}}^{-1} X)\beta^{(t+1)} &= (X^T W_{\beta^{(t)}}^{-1} X)\beta^{(t)} + X^T W_{\beta^{(t)}}^{-1} \frac{\partial \eta^{(t)}}{\partial \mu} W_{\beta^{(t)}}^{-1} (Y - \mu^{(t)}) \\ &= X^T W_{\beta^{(t)}}^{-1} (X\beta^{(t)} + \frac{\partial \eta^{(t)}}{\partial \mu}(Y - \mu^{(t)})) \end{aligned}$$

Ce qui donne finalement

$$\begin{aligned} \beta^{(t+1)} &= (X^T W_{\beta^{(t)}}^{-1} X)^{-1} X^T W_{\beta^{(t)}}^{-1} \big(X\beta^{(t)} + \frac{\partial \eta^{(t)}}{\partial \mu}(Y - \mu^{(t)})\big) \\ &= (X^T W_{\beta^{(t)}}^{-1} X)^{-1} X^T W_{\beta^{(t)}}^{-1} Y^{*(t)} \end{aligned}$$

L'algorithme de Fisher scoring à l'étape t revient à faire une régression pondérée de matrice de poids $W_{\beta^{(t)}}$ de X sur $Y^{*(t)}$ où

$$Y^{*(t)} = X\beta^{(t)} + \frac{\partial \eta^{(t)}}{\partial \mu}(Y - \mu^{(t)}). \tag{2.10}$$

McCullagh and Nelder (1989) montrent que l'algorithme de Fisher scoring est équivalent à l'algorithme IRLS pour n'importe quelle fonction de lien et pour une distribution appartenant à la famille exponentielle. Ainsi, cette algorithme est également souvent appelé IRLS, signifiant "iterative reweighted least squares". Il est dit 'itérativement repondéré' dans le sens où les paramètres estimés sont déterminés pour une matrice de poids W fixée qui est ensuite mise à jour par les valeurs courantes des estimations et le processus est alors répété. Le point de départ est souvent η^0 plutôt que β^0 car il est plus facile de trouver l'expression d'un bon point de départ à partir de η que de β.

Les équations (2.8) s'écrivent alors

$$X^T W_\beta^{-1}\big(Y^* - X\beta\big). \tag{2.11}$$

Ainsi, le même algorithme est décrit en résolvant itérativement les équations (2.11) comme des équations normales. A chaque itération, la valeur courante de β est utilisée pour le calcul de la matrice des poids W_β et du vecteur dépendant Y^*. Cela permet ensuite par résolution de ce systéme linéarisé d'obtenir une nouvelle valeur de β.

Cette réécriture (2.11) permet une interprétation de type linéaire. A β fixé, en considérant Y^* comme un nouveau vecteur de données et W_β comme une matrice de poids fixés, on reconnait dans le système (2.11) les équations classiques des moindres carrés généralisés associées au modèle

$$Y^* = X\beta + e$$

où $E(e) = 0$ et $Var(e) = W_\beta$.

La $i^{\text{ème}}$ composante du vecteur aléatoire $Y^* = X\beta + g'(\mu)\big(Y - \mu\big)$ a pour variance : $\mathrm{Var}(Y_i^*) = g'(\mu)^2 \mathrm{Var}(Y_i)$. Ainsi W_β correspond bien à la matrice de variance de Y^*.

Pour conclure, l'estimation du maximum de vraisemblance de β dans le GLM écrit sous la forme :

$$Y = g^{-1}(\eta) + \epsilon$$

où

$$\begin{aligned} E(\epsilon) &= 0 \\ \mathrm{Var}(\epsilon) &= \mathrm{diag}\{\mathrm{Var}(Y_i)\}_{i=1,\dots,n} = \mathrm{diag}\{a_i(\kappa)v(\mu_i)\}_{i=1,\dots,n} \end{aligned}$$

est donc équivalente à l'estimation successive du maximum de vraisemblance dans le modèle linéaire défini à l'étape t par :

$$Y^{*(t)} = \eta^{(t)} + e^{(t)} \qquad E(e^{(t)}) = 0 \qquad \mathrm{Var}(e^{(t)}) = W_{\beta^{(t)}}$$

pour le vecteur de données $Y^{*(t)}$.

2.2.3 Simplification dans le cas d'un lien canonique

Le lien canonique se définit par :

$$\eta_i = \gamma_i = g(\mu_i) = x_i\beta$$

Dans le cas d'un lien canonique nous avons donc :

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \gamma_i} = \frac{\partial b'(\gamma_i)}{\partial \gamma_i} = b''(\gamma_i) = \frac{Var(Y_i)}{a_i(\kappa)},$$

et

$$\frac{\partial l_i}{\partial \beta_j} = X_{ij}\frac{1}{a_i(\kappa)}(y_i - \mu_i).$$

Par la suite les termes de la matrice hessienne s'écrivent :

$$\begin{aligned} -\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k} &= -\frac{\partial}{\partial \beta_t}\big(X_{ij}\frac{1}{a_i(\kappa)}(y_i - \mu_i)\big) \\ &= X_{ij}X_{ik}\frac{Var(Y_i)}{(a_i(\kappa))^2}, \end{aligned}$$

et ceux de la matrice d'information de Fisher s'écrivent :

$$\begin{aligned} -E\Big[\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k}\Big] &= -E\Big[(\frac{\partial l_i}{\partial \beta_j})(\frac{\partial l_i}{\partial \beta_k})\Big] \\ &= E\Big[X_{ij}X_{ik}\frac{1}{a_i(\kappa)^2}(y_i - \mu_i)^2\Big] \\ &= X_{ij}X_{ik}\frac{Var(Y_i)}{(a_i(\kappa))^2}. \end{aligned}$$

Ainsi, dans le cas d'un lien canonique, l'algorithme des scores de Fisher est identique à l'algorithme de Newton-Raphson et la matrice de poids s'écrit $W_\beta = \mathrm{diag}\{\frac{a_i(\kappa)^2}{\mathrm{Var}(Y_i)}\}$.

2.3 Modèles linéaires mixtes (L2M)

2.3.1 Définition

Les modèles linéaires mixtes, notés L2M (Linear Mixed Model), introduits pour l'analyse de données de génétique animale par Henderson et al. (1959), sont en fait une extension des modèles linéaires classiques : aux effets fixes de ces derniers, viennent s'ajouter des effets aléatoires (voir aussi Verbeke and Molenberghs, 2000; McCulloch et al., 2008).

L'introduction d'effets aléatoires dans les modèles linéaires permet d'être plus précis sur l'origine de la variabilité totale par rapport à la modélisation statistique sans effets aléatoires. En effet, cette variabilité se divise en deux parties : la variabilité due aux effets aléatoires et la variabilité due aux erreurs. On parle de composantes de la variance. Searle et al. (1992) consacrent leur ouvrage à la décomposition de la variabilité.

Les L2M se formalisent de la manière suivante :

$$Y = X\beta + ZU + \epsilon \tag{2.12}$$

Où

- $Y = (y_1, \ldots, y_n)$ est le vecteur des observations qui, conditionnellement à $(U = u)$, sont indépendantes et uniformément distribuées selon une loi normale.

- β est le vecteur des effets fixes à estimer de taille $(1 + p)$

- X est la matrice de dimension $[n \times (1 + p)]$ associée à β

- U est le vecteur des effets aléatoires de taille q. Plus généralement, ce vecteur se décompose en K parties $U = (u_1^T, \ldots, u_K^T)$ où K est le nombre d'effets aléatoires considérés dans le modèle. Chaque composante u_i^T $(i = 1, \ldots, K)$ est un vecteur aléatoire modélisant un effet aléatoire de dimension q_i tel que $\sum_{i=1}^{K} q_i = q$. On suppose que chaque effet aléatoire suit une distribution normale centrée, c'est-à-dire : $\forall i \in \{1, \ldots, K\}$, $u_i \sim \mathcal{N}_{q_i}(0, \sigma_i^2 A_i)$ avec A_i matrice de dimension $q_i \times q_i$ supposée connue et σ_i^2 à estimer. D'autre part, $\forall i, i' \in \{1, \ldots, K\}^2$, $i \neq i'$, u_i et $u_{i'}$ sont indépendants. Donc $U \sim \mathcal{N}_q(0, \Sigma_\theta)$ où Σ_θ avec $\theta = (\sigma_1, \ldots, \sigma_K)$ est une matrice diagonale par blocs telle que

$\Sigma_\theta = \text{diag}\{\sigma_i^2 A_i\}_{i=1,\ldots,K}$.

- Z est la matrice associée à U de dimension $[n \times q]$ connue et formée des matrices Z_i de dimension $n \times q_i$ associées à chaque effet aléatoire : $Z = [Z_i, \ldots, Z_K]$.

- ϵ est le vecteur aléatoire des erreurs de taille n tel que $\epsilon \sim \mathcal{N}_n(0, \sigma_0^2 V_0)$. On notera aussi $R = \sigma_0^2 V_0$. On suppose que $\forall j \in \{1, \ldots, K\}$, ϵ et u_j sont indépendants.

Sous ces différentes hypothèses, on notera aussi :

$$\begin{aligned} \text{Var}(Y) = \Gamma &= R + Z\Sigma_\theta Z^T \\ &= \sigma_0^2 V_0 + \sum_{j=1}^{K} \sigma_j^2 Z_j A_j Z_j^T \end{aligned}$$

Ce qui, avec $\forall j \in 1, \ldots, K$, $V_j = Z_j A_j Z_j^T$ donne :

$$\Gamma = \sum_{j=0}^{K} \sigma_j^2 V_j \tag{2.13}$$

La variance totale est donc scindée en plusieurs composantes que l'on appelle *composantes de la variance.*

Comme $U \sim \mathcal{N}_q(0, \Sigma_\theta)$, la densité du vecteur u des effets aléatoires s'écrit :

$$f_U(u) = \frac{\exp\Big(-u^T \Sigma_\theta^{-1} u/2\Big)}{(2\pi)^{q/2}\sqrt{|\Sigma_\theta|}} \tag{2.14}$$

Aussi, conditionnement à U de la loi normale, on a :

$$(Y|U=u) \sim \mathcal{N}_n(X\beta + Zu, R) \tag{2.15}$$

$$Y \sim \mathcal{N}_n(X\beta, \Gamma) \tag{2.16}$$

$$\begin{pmatrix} Y \\ U \end{pmatrix} \sim \mathcal{N}_{n+q}\left(\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \begin{pmatrix} R + Z\Sigma_\theta Z^T & Z\Sigma_\theta \\ \Sigma_\theta Z^T & \Sigma_\theta \end{pmatrix} \right)$$

$$(U|Y=y) \sim \mathcal{N}_q(\Sigma_\theta Z^T \Gamma^{-1}(y - X\beta), \Sigma_\theta - \Sigma_\theta Z^T \Gamma Z \Sigma_\theta)$$

Dans ces modèles, nous nous intéressons à la fois à l'estimation de l'effet fixe ainsi qu'à celle des composantes de la variance. De nombreux travaux sur les L2M ont été réalisés sous plusieurs formes et selon des approches différentes. Citons par exemple Pinheiro and Bates (2000) ou encore Searle et al. (1992) qui y consacrent leur ouvrage. Mais ce sont Hartley and Rao (1967) qui, les premiers, ont donné un formalisme à l'estimation des composantes de la variance. Par la suite, différentes méthodes d'estimation ont été proposées. L'approche ML (Maximum Likelihood) utilise le concept classique de la fonction de vraisemblance. Une autre approche est celle du maximum de vraisemblance restreint (REML) qui, comme son nom l'indique, reste apparentée à celle du maximum de vraisemblance. On se focalise davantage, dans ce cas, sur l'estimation des composantes de la variance : on fait disparaître momentanément les effets fixes pour ne maximiser que la partie de la vraisemblance concernant les composantes. Cette méthode REML (Restricted Maximum Likelihood) a été proposée par Anderson and Bancroft (1952) et Thompson (1962) pour l'analyse de dispositifs équilibrés, puis généralisée à un modèle mixte gaussien quelconque par Patterson and Thompson (1971) dans les années 70. Foulley et al. (2002) présentent une synthèse sur l'utilisation de ces méthodes du maximum de vraisemblance.

2.3.2 Estimation ML

2.3.2.1 Dérivation directe des équations

On note $\Theta = (\sigma_0, \sigma_1, \ldots, \sigma_K) = (\sigma_0, \theta)$. La fonction de vraisemblance connaissant le vecteur des observations y s'écrit

$$l(\beta, u, \Theta; y) = \frac{1}{(2\pi)^{n/2}|\Gamma|^{1/2}} \exp\Big(-\frac{1}{2}(y - X\beta)'\Gamma^{-1}(y - X\beta)\Big) \tag{2.17}$$

et la log-vraisemblance

$$l(\beta, u, \Theta; y) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|\Gamma|) - \frac{1}{2}(y - X\beta)'\Gamma^{-1}(y - X\beta) \tag{2.18}$$

En dérivant cette log-vraisemblance, on obtient [1]

[1] Rappelons pour cela les resultats (Searle et al. (1992) p456-457) :

$$\frac{\partial \log|A|}{\partial \sigma_j^2} = \mathrm{tr}(A^{-1}\frac{\partial A}{\partial \sigma_j^2}), \text{ et } \frac{\partial A^{-1}}{\partial \sigma_j^2} = -A^{-1}\frac{\partial A}{\partial \sigma_j^2}A^{-1}$$

$$\left\{ \frac{\partial l(\beta,u,\Theta;y)}{\partial \beta} = X^T\Gamma^{-1}(y-X\beta)\forall j \in 0,\ldots,K, \frac{\partial l(\beta,u,\Theta;y)}{\partial \sigma_j^2} = -\frac{1}{2}\mathrm{tr}\left(\Gamma^{-1}V_j\right) + \frac{1}{2}(y-X\beta)'\Gamma^{-1}V_j \right. \tag{2.19}$$

En annulant ces dérivées, on obtient alors le système suivant à résoudre

$$\left\{ X^T\Gamma^{-1}X\beta = X^T\Gamma^{-1}y\mathrm{tr}\left(\Gamma^{-1}V_j\right) = y'PV_jPy \qquad ,j=0,\ldots,K \right. \tag{2.20}$$

où $P = \Gamma^{-1}\left(I - X\left(X\Gamma^{-1}X\right)^{-1}X^T\Gamma^{-1}\right.$.

Les équations de ce système sont résolues simultanément. La matrice Γ intervient dans la première équation concernant β. Autrement dit, l'estimation de β est relativement transparente au fait que les composantes de la variance soient connues ou estimées. Si une composante est connue, on reportera dans Γ cette vraie valeur, sinon on la remplacera par son estimation. D'autre part, les $K+1$ autres estimations ne sont pas linéaires en Θ. Une solution itérative devra donc être envisagée pour le système (2.20).

Cependant, même itérativement, la résolution directe de ce système n'est pas aisée. Plusieurs transformations peuvent être envisagées (dont celle de Hartley-Rao (Searle et al., 1992)). L'une rapide et très lisible conduit au système suivant équivalent à (2.20)

$$\left\{ X^T\Gamma^{-1}X\beta = X^T\Gamma^{-1}y\left(\mathrm{tr}\left(\Gamma^{-1}V_i\Gamma^{-1}V_j\right)\right)_{i,j=0,\ldots,K}\left(\begin{array}{c}\sigma_0^2 \\ : \\ \sigma_K^2\end{array}\right) = \left(y'PV_jPy\right)_{j=0,\ldots,K} \right. \tag{2.21}$$

Pour cela on a utilisé le fait que

$$\begin{aligned} \mathrm{tr}\left(\Gamma^{-1}V_i\right) &= \mathrm{tr}\left(\Gamma^{-1}V_i\Gamma^{-1}\Gamma\right) \\ &= \sum_{j=0}^{K}\mathrm{tr}\left(\Gamma^{-1}V_i\Gamma^{-1}V_j\right) \end{aligned}$$

Malgré cette nouvelle présentation, les équations ne sont toujours pas linéaires en Θ, du fait notamment de la présence de Γ dans les membres des équations. Elles permettent cependant de mettre en place très rapidement un algorithme itératif (qui ne sera pas forcément optimum). A partir des valeurs initiales de Θ, on itère la résolution des équations concernant les composantes de la variance jusqu'à convergence, en résolvant le système linéaires à chaque étape. Puis à l'aide des valeurs alors obtenues, on résout la première équation de (2.21) pour trouver l'estimation

de β. Cependant rien n'assure la positivité des estimations des composantes. Pour cela, si au cours d'une itération, l'estimation obtenue est négative, on forcera cette composante à 0 (ou à une petite valeur positive) ; ce qui revient à annuler l'effet du facteur correspondant.

Nous retenons donc que les estimations ML β et σ^2 dans un L2M vérifient :

$$\left\{ X^T\hat{\Gamma}^{-1}X\hat{\beta} = X^T\hat{\Gamma}^{-1}y\Big(\mathrm{tr}\Big(\hat{\Gamma}^{-1}V_i\hat{\Gamma}^{-1}V_j\Big)\Big)_{i,j=0,\ldots,K} \begin{pmatrix} \hat{\sigma}_0^2 \\ \vdots \\ \hat{\sigma}_K^2 \end{pmatrix} = \big(y'\hat{P}V_j\hat{P}y\big)_{j=0,\ldots,K} \right. \tag{2.22}$$

dans lequel $\hat{\Gamma}$ designe Γ où l'on a remplacé Θ par son estimateur et $\hat{P} = \hat{\Gamma}^{-1}(y - X\hat{\beta})$ d'où $\hat{P}y = \hat{\Gamma}^{-1}(y - X\hat{\beta})$.

2.3.2.2 Méthode de Henderson

La méthode de Henderson (Henderson et al., 1959) propose des équations permettant d'obtenir simultanément l'estimation BLUE (Best Linear Unbiaised Estimator) de β (notée $\hat{\beta}$, équivalente au maximum de vraisemblance sous des hypothèses de normalité adéquates) et la prédiction BLUP de U. Pour former ce système d'équations, la distribution jointe de Y et U est maximisée en β et U. Ainsi, après avoir utilisé sa distribution pour construire la fonction de vraisemblance, U joue alors le rôle de paramètre.

Compte tenu de (2.15) et (2.14) la distribution jointe s'écrit :

$$\begin{aligned} f_{U,Y}(u,y) &= \frac{1}{(2\pi)^{\frac{n+q}{2}}|R|^{\frac{1}{2}}|\Sigma_\theta|^{\frac{1}{2}}} \\ &\quad \exp\Big(-\frac{1}{2}\big[(y - X\beta - Zu)'R^{-1}(y - X\beta - Zu) + u^T\Sigma_\theta^{-1}u\big]\Big) \end{aligned}$$

En dérivant le logarithme de cette distribution en u et β on en déduit alors le système d'équations :

$$\begin{pmatrix} Z^TR^{-1}Z + \Sigma_\theta^{-1} & Z^TR^{-1}X \\ X^TR^{-1}Z & X^TR^{-1}X \end{pmatrix} \begin{pmatrix} u \\ \beta \end{pmatrix} = \begin{pmatrix} Z^TR^{-1}y \\ X^TR^{-1}y \end{pmatrix}. \tag{2.23}$$

Ces équations sont souvent appelées équations du modèle mixte ou MME (Mixed Model Equations) ou encore équations de Henderson. Remarquons que sans la présence de Σ_θ^{-1} dans la partie inférieure droite de ce système, il correspondrait aux équations du maximum de vraisemblance lorsqu'on traite U comme un effet fixe. Donc par l'introduction de Σ_θ^{-1}, on prend en compte en partie la nature

aléatoire de U. Ce système est équivalent à :

$$\left\{ X^T\Gamma^{-1}X\beta = X^T\Gamma^{-1}yu = \Sigma_\theta Z^T\Gamma^{-1}(y - X\beta) = E(U|Y = y) \right. \tag{2.24}$$

Le système (2.24) permet d'obtenir l'estimation BLUE de β et la prédiction BLUP de U. Cependant, il nécessite l'inversion de Γ non diagonale et d'ordre n. Ainsi, les équations de Henderson qui ne nécessitent que l'inversion des matrices R et Σ_θ (souvent diagonales) et celle du système (d'ordre $p+q$ souvent plus petit que n) représentent une alternative intéressante à la résolution directe de ce système.

Ayant obtenu $\hat{\beta}$ et $\tilde{u}$, il reste à estimer les composantes de la variance.

Dans le système (2.23), les matrices R et Σ_θ dépendent respectivement des valeurs σ_0^2 et $\sigma_1^2, \ldots, \sigma_K^2$ toutes inconnues. L'estimation de ces composantes est donc nécessaire. Les valeurs de u et β, obtenues par résolution du système de Henderson, vont alors permettre de calculer les estimations ML et REML dans un schéma itératif (Harville, 1977). A partir des équations ML (Searle et al., 1992), on construit les procédures itératives suivantes :

$$\left\{ \sigma_j^{2(t+1)} = \frac{u_j^{T(t)} A_j^{-1} u_j^{(t)}}{q_j - \mathrm{tr}(P_{jj}^{(t)})}, \quad j = 1, \ldots, K \sigma_0^{2(t+1)} = \frac{y^T V_0^{-1}\big(y - X\beta^{(t)} - Zu^{(t)}\big)}{n} \right. \tag{2.25}$$

où $P = (I_q + Z^T R^{-1} Z \Sigma_\theta)^{-1}$ avec I_q la matrice identité de dimension q et P_{jj} la j$^{\text{ème}}$ sous matrice de P

De façon, équivalente, on peut utiliser les schémas itératifs suivants qui s'avèrent être plus utiles d'un point de vue pratique (Trottier, 1998) :

$$\left\{ \sigma_j^{2(t+1)} = \frac{u_j^{T(t)} A_j^{-1} u_j^{(t)}}{q_j - \dfrac{\mathrm{tr}(A_j^{-1} C_{jj}^{*(t)})}{\sigma_j^{2(t)}}}, \quad j = 1, \ldots, K \sigma_0^{2(t+1)} = \frac{\big(y - X\beta^{(t)} - Zu^{(t)}\big)' V_0^{-1}\big(y - X\beta^{(t)} - Zu^{(t)}}{n - \sum_{j=1}^{K}\left(q_j - \dfrac{\mathrm{tr}(A_j^{-1} C_j^{*(t)} j)}{\sigma_j^{2(t)}}\right)} \right. \tag{2.26}$$

où $C^* = (Z^T R^{-1} Z + \Sigma_\theta^{-1})^{-1}$ est l'inverse de la matrice formée par les q dernières lignes et colonnes de la matrice des coefficients du système de Henderson (2.23) et C_{jj}^* est la la j$^{\text{ème}}$ sous matrice de C^*, correspondant au j$^{\text{ème}}$ effet aléatoire.

La procédure d'estimation alterne alors entre :

1. la résolution de (2.23) pour des valeurs de σ_j^2 connues (fixant les valeurs de R et Σ_θ) ;

2. la résolution de (2.25) ou (2.26) pour des valeurs de β et u.

2.3.3 Estimation REML

Cette méthode d'estimation spécifique aux L2M a été développée par Patterson and Thompson (1971). On supprime provisoirement les effets fixes pour ne maximiser que la partie de la vraisemblance concernant les composantes de la variance. Pour éliminer la partie des effets fixes, on projette le modèle sur l'orthogonal du sous-espace vectoriel engendré par les colonnes de X. L'estimation par maximum de vraisemblance restreint n'est autre que l'estimation par maximum de vraisemblance dans le modèle projeté. Après une estimation itérative des composantes de la variance, on estime le vecteur des effets fixes β.
L'estimation REML a également une interprétation bayésienne (Harville, 1974). Elle repose sur le concept de vraisemblance marginale. Après élimination de β par intégration de la fonction de vraisemblance, on obtient la vraisemblance marginale des σ_j^2. Le système des $K+1$ équations obtenues par dérivation de cette log-vraisemblance marginale est identique au système d'équations obtenues par projection du modèle sur l'orthogonal de X.
Cette méthode d'estimation a l'avantage sur la méthode ML de tenir compte de la perte de degrés de liberté occasionnée par l'estimation des effets fixes et de donner ainsi des estimations non biaisées des paramètres de la variance.

2.3.3.1 Dérivation des équations

Afin d'éléminer la partie effet fixe, ce n'est pas directement sur le vecteur Y que l'on va travailler mais sur une transformation de ce vecteur. On s'intéresse à des combinaisons linéaires m^TY indépendantes et d'espérance nulle, aussi appelées *contrastes*. Si $\text{rang}(X) = p$, il existe $n-p$ combinaisons linéaires indépendantes de la sorte. Elles sont regroupées dans la matrice M^TY telle que $M = [m_1, \ldots, m_{n-p}]$ et donc

$$M^TY = \left[\begin{array}{c} m_1^TY \vdots m_{n-p}^TY \end{array} \right]$$

Puisque $M^TX = 0$, on a alors $E(M^TY) = M^TX\beta = 0$, $\text{Var}(M^TY) = M^T\Gamma M$, et $M^TY \sim \mathcal{N}_{n-p}(0, M^T\Gamma M)$.
Cela revient à projeter le modèle sur l'orthogonal du sous-espace vectoriel engendré par les colonnes de X noté $X^\perp$. Ainsi, la matrice M est la matrice dont les $n-p$ vecteurs colonnes constituent une base de $X^\perp$ alors $M^TX = 0$.

Le modèle projeté est donc :

$$\begin{aligned} M^TY &= M^TX\beta + M^TZU + M^T\epsilon \\ &= M^TZU + M^T\epsilon \end{aligned}$$

et

$$\mathrm{Var}(M^T Y) = M^T \Gamma M$$

Alors, l'estimation par maximum de vraisemblance restreint n'est autre que l'estimation par maximum de vraisemblance dans le modèle projeté. Pour établir ces équations du maximum de vraisemblance pour les composantes, il suffit de reprendre les équations du système (2.20), de supprimer la première ligne (celle des effets fixes) et d'effectuer le changement de notation :

$$\begin{array}{rcll} Y & \rightarrow & M^T Y & \\ Z_j & \rightarrow & M^T Z_j & j = 1, \ldots, K \\ V_j & \rightarrow & M^T V_j M & j = 1, \ldots, K \\ \Gamma & \rightarrow & M^T \Gamma M, & \end{array}$$

en sachant que $M(M^T \Gamma M)^{-1} M^T = P$. On obtient alors le système :

$$\mathrm{tr}\big(PV_j\big) = y^T P V_j P y \qquad j = 0, \ldots, K \tag{2.27}$$

De même que dans la section 2.3.2, on envisage plutôt une résolution itérative du système équivalent suivant :

$$\left\{ \left(\mathrm{tr}\big(PV_i PV_j\big)\right)_{i,j=0,\ldots,K} \left(\begin{array}{c} \sigma_0^2 \\ \vdots \\ \sigma_K^2 \end{array} \right) = \left(y^T P V_j P y\right)_{j=0,\ldots,K} \right. \tag{2.28}$$

Ce système permet donc d'obtenir les estimations REML dans un L2M. Cette méthode d'estimation a l'avantage sur la méthode ML de tenir compte de la perte de degrés de liberté occasionnée par l'estimation des effets fixes. Après avoir estimé les composantes de la variance, il suffit alors de former $\hat{\Gamma}$ pour obtenir directement l'estimation de β.
On note que pour passer du système ML au système REML, il a suffi de remplacer Γ^{-1} par P (voir (2.21)).

2.3.3.2 Méthode de Henderson

Tout comme pour la méthode du ML, les valeurs de u et β, obtenues par résolution du système de Henderson (2.23), vont permettre de calculer les estimations REML dans un schéma itératif (Harville, 1977). A partir des équations REML (Searle et al., 1992), on construit les procédures itératives suivantes :

$$\left\{ \sigma_j^{2(t+1)} = \frac{u^{T(t)}{}_j A_j^{-1} u_j^{(t)}}{q_j - \mathrm{tr}(Q_{jj}^{(t)})}, \quad j = 1, \ldots, K \sigma_0^{2(t+1)} = \frac{y' V_0^{-1}\big(y - X\beta^{(t)} - Zu^{(t)}\big)}{n - \mathrm{rg}(X)} \right. \tag{2.29}$$

où $Q = (I_q + Z^T S Z \Sigma_\theta)^{-1}$ avec I_q la matrice identité de dimension q, Q_{jj} la j$^{\text{ème}}$ sous matrice de Q et $S = R^{-1}(I_n - X(X^T R^{-1} X)^{-1} X^T R^{-1})$.

De façon équivalente, on peut utiliser les schémas itératifs suivants qui s'avèrent être plus utiles d'un point de vue pratique (Trottier, 1998) :

$$\left\{ \sigma_j^{2(t+1)} = \frac{u_j^{T(t)} A_j^{-1} u_j^{(t)}}{q_j - \frac{\text{tr}(A_j^{-1} C_{jj}^{*(t)})}{\sigma_j^{2(t)}}}, \quad j = 1, \ldots, K \sigma_0^{2(t+1)} = \frac{\left(y - X\beta^{(t)} - Zu^{(t)}\right)^T V_0^{-1} \left(y - X\beta^{(t)} - Zu^{(t)}\right.}{n - \text{rg}(X) - \sum_{j=1}^{K} \left(q_j - \frac{\text{tr}(A_j^{-1} C_{jj)}^{*(t)}}{\sigma_j^{2(t)}}\right.} \right. \tag{2.30}$$

où $C^* = (Z^T R^{-1} Z + \Sigma_\theta^{-1})^{-1}$ est est l'inverse de la matrice formée par les q dernières lignes et colonnes de la matrice des coefficients du système de Henderson (2.23) et C_{jj}^* est la j$^{\text{ème}}$ sous matrice de C^*, correspondant au j$^{\text{ème}}$ effet aléatoire.

La procédure d'estimation alterne alors entre :

1. la résolution de (2.23) pour des valeurs de σ_j^2 connues (fixant les valeurs de R et Σ_θ) ;
2. la résolution de (2.29) ou (2.30) pour des valeurs de β et u.

2.3.4 Approches orientées R

L'estimation des L2M dans le logiciel R (Bates, 2010; Pinheiro and Bates, 2000) se fait à l'aide de la fonction **lmer** du package `lme4` (R Development Core Team, 2012; Bates et al., 2011). Les deux méthodes, ML et REML, sont proposées.
La fonction **lmer** travaille sous l'hypothèse que $V_0 = I_n$ c'est à dire $\text{Var}(\epsilon) = \sigma_0^2 I_n$ d'où

$$(Y|U = u) \sim \mathcal{N}_n(X\beta + Zu, \sigma_0^2 I_n)$$
$$U \sim \mathcal{N}_q(0, \Sigma_\theta)$$

et sous l'hypothèse que $A_i = I_{q_i}$ où I_{q_i} est la matrice identité de taille $q_i \times q_i$ d'où Σ_θ est une matrice diagonale par blocs telle que $\Sigma_\theta = \text{diag}\{\sigma_i^2 I_{q_i}\}_{i=1,\ldots,K}$ et avec $\theta = (\sigma_1^2, \ldots, \sigma_K^2)$.

2.4 Modèles linéaires généralisés mixtes (GL2M)

2.4.1 Définition

De même que les effets aléatoires ont été introduits dans les modèles linéaires définissant ainsi les L2M, ils peuvent l'être au sein des modèles linéaires généralisés pour donner naissance aux modèles linéaires généralisés mixtes notés GL2M. C'est dans l'expression du prédicteur linéaire qu'une partie aléatoire vient s'ajouter à la partie fixe. En gardant les notations de la section précédente concernant le vecteur des paramètres d'effets fixes β et sa matrice associée X ainsi que le vecteur des effets aléatoires U et sa matrice associée Z, le prédicteur linéaire s'exprime de la façon suivante :

$$\eta_u = X\beta + Zu \tag{2.31}$$

où

$$\eta_u = g(\mu_u) \quad \text{avec} \quad \mu_u = E(Y|U=u)$$

tel que g est la fonction de lien et $Y = (y_1, \ldots, y_n)$ est le vecteur des observations qui, conditionnellement à U, sont supposées indépendantes et uniformément distribuées selon une loi appartenant à la famille exponentielle.

Pour bien insister sur l'introduction des effets aléatoires dans ce prédicteur, nous l'avons indicé par u. Nous maintenons, de plus, l'hypothèse de normalité sur U : $U \sim N_q(0, \Sigma_\theta)$ où $\Sigma_\theta = \text{diag}\{\sigma_i^2 A_i\}_{i=1,\ldots,K}$ et $\theta = (\sigma_1, \ldots, \sigma_K)$

De plus, la fonction de variance v intervient dans l'expression de la variance conditionnelle de la manière suivante :

$$\forall i \in \{1, \ldots, N\}, Var(Y_i|U=u) = a_i(\kappa)v(\mu_{u,i}) \quad \text{où} \quad a_i(\kappa) = \frac{\kappa}{\omega_i}.$$

Ainsi, conditionnellement à $(U = u)$, le GL2M conserve toutes les propriétés du GLM.

Dans le cas de la loi normale, nous retrouvons bien la définition précédente du L2M avec le lien identité. Ainsi, comme le modèle linéaire est un cas particulier des GLM, le L2M en est un des GL2M. Cependant, il est important de noter que pour les L2M, il y a conservation de loi lors du passage des lois de $(Y|U)$ et U à la loi marginale de Y. Cette propriété est spécifique à la loi normale. Elle ne se retrouve pas pour d'autres lois de façon générale avec une hypothèse gaussienne sur les effets aléatoires. Le L2M est donc un cas particulier de GL2M.

2.4.2 Vraisemblance

En prenant un lien canonique et en utilisant des notations matricielles la densité conditionnelle de Y_i sachant u, $f_{Y_i|U}(y|u)$, s'écrit donc:

$$f_{Y_i|U}(y_i|u) = \exp\left(\frac{y_i\eta_{i,u} - b(\eta_{i,u})}{a_i(\kappa)} + c(y_i,\kappa)\right) \quad \text{pour } i \in \{1,\dots,n\} \tag{2.32}$$

avec les mêmes notations que (2.1).

Le vecteur U des effets aléatoires se distribuant selon une loi normale multivariée sa densité, $f_U(u)$, s'écrit comme pour les L2M :

$$f_U(u) = \frac{\exp\left(-u^T\Sigma_\theta^{-1}u/2\right)^{1/2}}{(2\pi)^{q/2}|\Sigma_\theta|} \tag{2.33}$$

De manière similaire aux L2M, la fonction de vraisemblance du modèle s'écrit :

$$\begin{aligned} L(\beta,u;y_1,\dots,y_n) &= \int_{\mathbb{R}^q} f_{U,Y}(y_{\text{obs}},u)\,\mathrm{d}u \\ &= \int_{\mathbb{R}^q} f_{Y|U}(y_{\text{obs}}|u)f_U(u)\,\mathrm{d}u \qquad (2.34) \\ &= \int_{\mathbb{R}^q} \exp\left(\frac{y^T\eta_u - 1^Tb(\eta_u)}{a(\kappa)} + 1^Tc(y,\kappa)\right) \times \frac{\exp\left(-u^T\Sigma_\theta^{-1}u/2\right)}{(2\pi)^{q/2}|\Sigma_\theta|^{1/2}}\,\mathrm{d}b \end{aligned}$$

Cette intégrale n'est pas calculable formellement sauf pour quelques cas particuliers comme les L2M, les modèles béta-binomial ou les modèles de Poisson-gamma (Martinez, 2006). De ce fait, contrairement aux GLM ou L2M il n'existe pas de méthodes standards d'estimation des effets fixes et des composantes de la variance pour les GL2M. Plusieurs approches ont ainsi été développées afin de soit résoudre, soit contourner les difficultés liées au calcul.

Chapter 3

Estimation par approximation de la vraisemblance

3.1 Introduction et présentation générale

Stiratelli et al. (1984) ont estimé les paramètres d'un modèle logistique à effets aléatoires emboités en approximant la densité jointe à posteriori par une densité normale multivariée. Lee and Nelder (1996) se placent dans le cadre des HGLM (Hierachical Generalized Linear Mixed Models) et définissent la h-vraisemblance comme étant l'intégrande de (2.34) qu'ils maximisent en u et β. D'autres auteurs ont étendu l'approche d'estimation de manières différentes. Schall (1991) s'inscrit dans un raisonnement conditionnel basé sur le fait que, conditionnellement aux effets aléatoires, le modèle linéaire généralisé mixte s'apparente à un modèle linéaire généralisé. En effectuant une linéarisation du modèle celui-ci est replongé dans un cadre linéaire et le problème du calcul intégral est alors contourné. D'autres démarches s'inscrivant également dans un raisonnement conditionnel ont été développées (Engel and Keen, 1994; Wolfinger, 1993). La démarche de Breslow and Clayton (1993) nommée plus couramment par méthode PQL (Penalized Quasi Likelihood) s'inscrit aussi dans ce cadre bien qu'elle utilise l'approximation de Laplace (voir chapitre 4) pour approximer la fonction de quasi-vraisemblance. Une autre démarche consiste en l'approximation numérique de l'intégrale (2.34). Raudenbush et al. (2000) utilisent une approximation de Laplace d'ordre élevé. Anderson and Bancroft (1952) quant à eux utilisent la quadrature gaussienne dans le cas de données binaires. Hedeker et al. (1994) utilisent également la quadrature de Gauss-Hermite afin d'évaluer la vraisemblance d'un modèle probit ordinal et d'un modèle logistique avec des lois à priori gaussiennes multivariées. Pinheiro and Bates (1995) améliorent cette approche à l'aide de la quadrature gaussienne adaptative et approxime la vraisemblance d'un modèle à effets aléatoires emboîtés

pour des données normales sans lien linéaire. Les méthodes de Monte Carlo par chaînes de Markov sont également utilisées comme alternative. Zeger and Karim (1991) ont, ainsi, proposé un algorithme de Gibbs sampling pour l'estimation des paramètres. McCulloch (1997); Booth and Hobert (1999) proposent également une méthode s'appuyant sur une étape de Metropolis-Hastings conduisant à la construction d'un algorithme de type Monte Carlo EM. Globalement, les trois méthodes les plus connues et utilisées sont l'approche conditionnelle, l'approximation numérique de l'intégrale et les méthodes de Monte Carlo.

3.2 Approximation de Laplace

3.2.1 Approximation du second ordre

On rappelle que Σ_θ est la matrice de variance-covariance des effets aléatoires du modèle tel que $\theta = (\sigma_1, \ldots, \sigma_K)$. Bien que Σ_θ puisse être une matrice de grande taille, elle est déterminée par le paramètre θ dont la dimension est relativement petite. Les estimateurs du maximum de vraisemblance $\hat{\beta}$ et $\hat{\theta}$ maximisent la vraisemblance en les paramètres β et θ étant donné les données observées. La vraisemblance du modèle s'écrit selon l'équation (2.34) et comme dit plus haut cette intégrale n'est pas directement calculable. Cependant, elle peut être approchée en utilisant l'approximation de Laplace ou la quadrature gaussienne.

La méthode de Laplace (Tierny and Kadane, 1986) est utilisée pour approximer des intégrales de la forme

$$\int \exp\{l(t)\}\,\mathrm{d}t$$

où $l(t)$ est une fonction régulière. L'idée consiste à approximer $l(t)$ en effectuant un développement de Taylor de la fonction $l(t)$ autour d'un point de maximisation globale $\hat{t}$:

$$l(t) \simeq l(\hat{t}) + \frac{\partial l(t)^T}{\partial t}|_{t=\hat{t}}(t-\hat{t}) + \frac{1}{2}(t-\hat{t})^T l''(\hat{t})(t-\hat{t}) \tag{3.1}$$

où l' et l'' sont respectivement la dérivée première et la matrice hessienne de $l(.)$. Le second terme de l'équation (3.1) s'annule car au point maximum de $l(t)$, la dérivée première vaut zéro. En remplaçant $l(t)$ par son développement de Taylor à l'ordre 2 nous avons :

$$\begin{aligned}\int \exp\{l(t)\}\,\mathrm{d}t &\approx \int \exp\{l(\hat{t}) + \frac{1}{2}l''(\hat{t})(t-\hat{t})^2\}\,\mathrm{d}t \\ &= \exp\{l(\hat{t})\}\int \exp\{-\frac{1}{2}(-l''(\hat{t}))(t-\hat{t})^2\}\,\mathrm{d}t\end{aligned}$$

Dans la fonction à intégrer du terme de droite, on reconnait le noyau d'une densité normale multivariée de moyenne $\hat{t}$ et de matrice de covariance l'inverse de $-l''(\hat{t})$. L'intégrale est donc égale à $(2\pi)^{\frac{q}{2}}|-l''(\hat{t})|^{-\frac{1}{2}}$ et l'approximation de Laplace peut alors s'écrire :

$$\int \exp\{l(t)\}\,\mathrm{d}t \simeq (2\pi)^{\frac{q}{2}}|-l''(\hat{t})|^{-\frac{1}{2}}\exp\{l(\hat{t})\}$$

On réécrit alors la vraisemblance (2.34) de la manière suivante :

$$L(\beta,u,\theta;y)=\int_{\mathbb{R}^q}\exp(h(u))\,\mathrm{d}u$$

avec $h(u)=\log(f_{y|y}(y|u))+\log(f_u(u))$ et soit $\tilde{u}=\arg\max_u h(u)$ les modes conditionnels des effets aléatoires.

Alors en appliquant l'approximation de Laplace à $L(\beta,u,\theta;y)$ autour de $\tilde{u}$, nous obtenons

$$L(\beta,\tilde{u},\theta;y)=(2\pi)^{\frac{q}{2}}|-h''(\tilde{u})|^{-\frac{1}{2}}\exp(h(\tilde{u})).$$

Comme

$$\begin{aligned} h(u) &= \log(f_{Y|U}(y|u))+\log(f_u(u)) \\ &= \frac{y^T(X\beta+Zu)-1^Tb(X\beta+Zu)}{a(\kappa)}+1^Tc(y,\kappa)-\frac{q}{2}\log(2\pi)-\frac{1}{2}\log\left(|\Sigma_\theta|\right)-\frac{1}{2}u^T\Sigma_\theta^{-1}u \end{aligned}$$

en supposant un lien canonique pour $f_{Y|U}(y|u)$.

On a donc que

$$\frac{\partial h(u)}{\partial u}=\frac{y^TZ-1^T\mathrm{diag}\big(b^T(X\beta+Zu)\big)Z}{a(\kappa)}-u^T\Sigma_\theta^{-1} \tag{3.2}$$

et

$$h''(u)=\frac{\partial^2 h(u)}{\partial u\partial u^T}=-Z^TW^{-1}Z-\Sigma_\theta^{-1} \tag{3.3}$$

où $W=\mathrm{diag}\{\frac{a(\kappa)}{b''(X\beta+Zu)}\}=\mathrm{diag}\{\frac{a(\kappa)^2}{Var(Y|U=u)}\}$ est la matrice de poids (McCullagh and Nelder, 1989).

D'où

$$L(\beta,\tilde{u},\theta;y)=(2\pi)^{\frac{q}{2}}|Z^TW^{-1}Z+\Sigma_\theta^{-1}|^{-\frac{1}{2}}\log\{f_{Y|U}(y|\tilde{u})f_U(\tilde{u})\} \tag{3.4}$$

Des estimateurs du maximum de vraisemblance, notés $\hat{\beta}$ et $\hat{\theta}$, pour les paramètres β et θ sont alors obtenus à partir de cette approximation. L'estimateur du maximum de vraisemblance de u est alors obtenu en prenant $\tilde{u}$ en $\hat{\beta}$ et $\hat{\theta}$.

3.2.2 Approximation à des ordres supérieurs

Dans les applications standards, l'approximation de Laplace du second degré est très utilisée. En effet les termes de plus haut degré diminuent avec la taille de l'échantillon, rendant l'approximation précise pour les échantillons de grande taille (Raudenbush et al., 2000). Cependant, Shun and Mc Cullagh (1995) et Shun (1995) notent que dans certains cas, la dimension de l'intégrale augmente en fonction de la taille de l'échantillon. Dans ces cas, l'approximation de Laplace standard (second ordre) n'est pas valide car l'erreur d'approximation du second ordre ne diminue pas avec la taille de l'échantillon. Dans cette partie, nous exposons donc le principe d'approximation de Laplace basée sur le développement de Taylor pris à des ordres supérieurs à 2. Cette approximation est ensuite adaptée au cas des GL2M. Les différents résultats de cette partie sont issus de l'article de Raudenbush et al. (2000).

Soit $l(t)$ une fonction scalaire avec t un vecteur et telles que toutes les dérivées de l soient continues dans un voisinage $v(\hat{t})$ de $\hat{t}$, $\hat{t}$ étant le maximum de $l(t)$ en t. Le développement de Taylor pour t dans $v(\hat{t})$ s'écrit

$$\begin{aligned} l(t) &\approx l(\hat{t}) + l^{(1)}(\hat{t})(t-\hat{t}) + \frac{1}{2}(t-\hat{t})^T l^{(2)}(\hat{t})(t-\hat{t}) \\ &\quad + \frac{1}{3!}[(t-\hat{t})^T \otimes (t-\hat{t})^T] l^{(3)}(\hat{t})(t-\hat{t}) \\ &\quad + \ldots + \frac{1}{K!}\big[\otimes^{K-1} (t-\hat{t})^T \big] l^{(K)}(\hat{t})(t-\hat{t}) + \ldots \\ &= l(\hat{t}) + l^{(1)}(\hat{t})(t-\hat{t}) + \sum_{k=2}^{\infty} \frac{1}{k!}\Big[\otimes^{k-1} (t-\hat{t})^T \Big] l^{(k)}(\hat{t})(t-\hat{t}) \end{aligned} \tag{3.5}$$

où

$$l^{(k)}(\hat{t}) = \frac{\partial vec\, h^{(k)}(t)}{\partial t^T}|_{t=\hat{t}}$$

et $\otimes^k t = t \otimes t \otimes \ldots \otimes t$, k produits de kronecker sur t. Par exemple, $\otimes^3 t = t \otimes t \otimes t$.

Appliquons alors l'approximation de Laplace avec le développement de Taylor

ci-dessus à $\int \exp\{l(t)\}\,\mathrm{d}t$:

$$\int_{R^q} \exp\{l(t)\}\,\mathrm{d}t \approx (2\pi)^{\frac{q}{2}}|-l^{(2)}(\hat{t})|^{-\frac{1}{2}} \exp\{l(\hat{t})\}\times E\Big\{ \exp\Big[\sum_{k=3}^{\infty}\frac{1}{k!}\Big[\otimes^{k-1}(t-\hat{t})^T\Big]l^{(k)}(\hat{t})(t-\hat{(t)})\Big]\Big\}. \tag{3.6}$$

Ici $E(.)$ est l'espérance prise sur une distribution normale multivariée de moyenne le vecteur 0 et de matrice de variance-covariance $V = [-l^{(2)}(\hat{t})]^{-1}$.

Comme $\hat{t}$ maximise $l(t)$, $l^{(1)}(\hat{t})$ vaut zéro et $l^{(1)}(\hat{t})(t-\hat{t})$ disparait. Alors, l'expression (3.6) est le produit de la constante $\exp\{l(\hat{t})\}$, du noyau d'une loi normale multivariée

$$\exp\Big\{-\frac{1}{2}(t-\hat{t})^T(-l^{(2)}(\hat{t}))(t-\hat{t})\Big\}$$

et de l'exponentiel de la somme des termes restant du développement. L'intégrale a ainsi la forme d'une espérance prise sur une loi normale multivariée de densité $\mathcal{N}(0,V)$.

Maintenant, il s'agit d'évaluer l'espérance dans l'équation (3.6), c'est à dire $E(\exp(R))$ où

$$R = \sum_{k=3}^{\infty} T_k \quad \text{avec} \quad T_k = \frac{1}{k!}[\otimes^{k-1}(t-\hat{t})^T]l^{(k)}(\hat{t})(t-\hat{t}).$$

En posant $\exp(R) = 1+R+\frac{R^2}{2}+\ldots+\frac{R^k}{k!}+\ldots$, il faut donc évaluer $E(R), E(R^2), \ldots$.

En évaluant $E(R)$ et $E(R^2)$, (Raudenbush et al., 2000) trouvent les résultats suivants (voir annexes):

$$\begin{aligned} E(T_k) &= 0 \text{ pour k impair, } k>2; \\ E(T_k) &= \frac{(k-1)(k-3)\ldots 3}{k!} vec^T\Big(\otimes^{\frac{k}{2}} V\Big)\, vec[l^{(k)}(\hat{t})] \\ &\quad \text{pour k pair, } k>2; \end{aligned} \tag{3.7}$$

$$\begin{aligned} E(T_kT_m) &= 0 \text{ pour (k+m), k impairs et, } m>2; \\ E(T_kT_m) &= \frac{(k+m-1)(k+m-3)\ldots 3}{k!m!} vec^T\Big(\otimes^{\frac{(k+m)}{2}} V\Big)\, vec[l^{(k)}(\hat{t})\otimes l^{(m)}(\hat{t})] \\ &\quad \text{pour (k+m), k pairs et, } m>2; \end{aligned} \tag{3.8}$$

La vraisemblance (2.34) peut se réécrire sous la forme :

$$\begin{aligned} &= \int_{\mathbb{R}^{q_j}} \prod_{j=1}^{K} f_{Y_j|u_j}(y_j|u_j) f_{u_j}(u_j)\, \mathrm{d}u_j \qquad (3.9) \\ &= \prod_{j=1}^{K} (2\pi)^{-\frac{q_j}{2}} \sigma_j^{-q_j} |A_j|^{-\frac{1}{2}} \int_{\mathbb{R}^j} \exp\left(l_j - \frac{u_j^T A_j^{-1} u_j}{2\sigma_j^2}\right) \mathrm{d}u_j \end{aligned}$$

avec

$$l_j = \sum_{i=1}^{n_{ij}} \frac{y_{ij}\eta_{ij} - b(\eta_{ij})}{a_j(\kappa)} + c(y_{ij}, \kappa)$$

et où, pour rappel, $u_j \sim \mathcal{N}_{q_j}(0, \sigma_j^2 A_j)$ pour tout $j \in \{1, \ldots, K\}$, .

Afin d'appliquer l'approximation de Laplace, $h_j(u_j) = l_j - \frac{u_j^T A_j^{-1} u_j}{2\sigma_j^2}$ joue le rôle de $l(t)$, $\hat{u}_j$ maximise $h_j(u_j)$ en u_j et les dérivées de $h_j(u_j)$ par rapport à u_j jusque l'ordre six sont établies telles que:

1. $h_j(\hat{u}_j) = \hat{l}_j - \frac{\hat{u}_j^T A_j^{-1} \hat{u}_j}{2\sigma_j^2}$, où $\hat{l}_j$ est l_j évalué à $\hat{u}_j$.

2. $h_j^{(1)}(\hat{u}_j) = \hat{l}_j^{(1)} - \frac{\hat{u}_j^T}{\sigma_j^2} A_j^{-1}$, où

$$\hat{l}_j^{(1)} = \frac{\partial \hat{l}_j}{\partial u^T}|_{u=\hat{u}_j} = \frac{(y_j - \hat{\mu}_j)^T Z_j}{a_j(\kappa)} = \frac{(y_j^* - \hat{\eta}_{ij})^T \hat{W}_j Z_j}{a_j(\kappa)},$$

   avec $y_j^* = \hat{W}_j^{-1}(y_j - \hat{\mu}_j) + \hat{\eta}_j$ la variable dépendante linéarisée (McCullagh and Nelder, 1989), $\hat{W}_j = \mathrm{diag}\{\hat{w}_{ij}\}_{i=1,\ldots,n_{ij}}$ tel que $\hat{w}_{ij} = d\hat{\mu}_{ij}/d\hat{\eta}_{ij}$, la dérivée de μ_{ij} par rapport à η_{ij} évaluée en $\hat{u}_{ij}$.

3. $h_j^{(2)}(\hat{b}_j) = \hat{l}_j^{(2)} - \sigma_j^{-2} A_j^{-1}$, où

$$\hat{l}_j^{(2)} = -Z_j^T \hat{W}_j Z_j / a_j(\kappa),$$

   la dérivée seconde de l_j évaluée en $\hat{u}_j$

4. Pour $k \geq 3$, $h_j^{(k)}(\hat{u}_j) = \hat{l}_j^{(k)} = -\sum_{i=1}^{n_j} \hat{m}_{ij}^{(k)} (\otimes^k Z_{ij}^T)/a_j(\kappa)$,
   où $\hat{m}_{ij}^{(k)}$ est la $(k-1)^{\text{ième}}$ dérivée de μ_{ij} par rapport à η_{ij} évaluée en $\hat{u}_j$.

Dans le cas où y_{ij} est binaire avec le lien logit, $w_{ij} = \mu_{ij}(1-\mu_{ij})$, et la dérivée seconde jusque la cinquième s'écrit

$$\begin{aligned}
\hat{m}_{ij}^{(3)} &= \hat{w}_{ij}(1-2\hat{\mu}_{ij}) \\
\hat{m}_{ij}^{(4)} &= \hat{w}_{ij}(1-6\hat{\mu}_{ij}) \\
\hat{m}_{ij}^{(5)} &= \hat{m}_{ij}^{(3)}(1-12\hat{\mu}_{ij}) \\
\hat{m}_{ij}^{(6)} &= \hat{m}_{ij}^{(4)}(1-12\hat{\mu}_{ij}) - 12\hat{m}_{ij}^{(3)2}
\end{aligned} \tag{3.10}$$

Dans le cas des données de comptage telles que les observations ($y_{ij} \in \{0, 1, \ldots\}$) soient issues d'une distribution de Poisson conditionnelle avec lien log, $w_{ij} = m_{ij}^{(k)} = \mu_{ij}$ pour tout k.

Quand les y_{ij} sont conditionnellement distribuées selon une loi gamma avec le lien réciproque, $w_{ij} = \mu_{ij}^2$, $m_{ij}^{(k)} = (k-1)!\mu_{ij}^k$ pour tout k et dans le cas normal, $w_{ij} = 1$ et $m_{ij}^{(k)} = 0$ pour $k > 2$.

Les constantes $a_j(\kappa)$ valent un dans le cas binomial et de Poissson, $\text{var}(y_{ij}|b_j) = \sigma_j^2$ pour le cas normal, et $-1/v$ pour la loi gamma, où $\text{var}(y_{ij}|b_j) = \mu_{ij}^2/v$. Toutes les dérivées de l_j sont évaluées en $\hat{u}_j$. Afin d'éviter de lourdes notations, nous supprimerons le chapeau par la suite sauf pour $\hat{l}_j$ et $\hat{u}_j$.

En appliquant l'approximation de Laplace à l'équation (3.9),

$$\begin{aligned}
L = \prod_{j=1}^{K} = \Big\{ (2\pi\sigma^2)^{-\frac{q_j}{2}} |A_j|^{-\frac{1}{2}} \exp\Big(\hat{l}_j - \frac{1}{2\sigma_j^2} \hat{u}_j^T A_j^{-1} \hat{u}_j \Big) \\
\times \int_{R^q} \exp\Big\{ -\frac{1}{2}(u_j - \hat{u}_j)^T V_j^{-1} (u_j - \hat{u}_j) \Big\} \exp(R_j)\, \mathrm{d}u_j \Big\}
\end{aligned} \tag{3.11}$$

où le terme de correction $R_j = \sum_{k=3}^{\infty} T_{kj}$ avec

$$T_{kj} = \frac{1}{k!} [\otimes^{k-1} (u_j - \hat{u_j})^T] h_j^{(k)}(\hat{u}_j)(u_j - \hat{u}_j).$$

Notons que $h_j^{(1)}(\hat{u}_j)$ disparait de l'équation (3.11) car $\hat{u}_j$ maximise $h_j(u_j)$, c'est à dire que

$$\hat{u}_j = \sigma_j^2 A_j \hat{l}_j^{(1)} = (Z_j^T W_j Z_j + \sigma_j^{-2} A_j^{-1})^{-1} Z_j^T W_j (y_j^* - X_j \beta).$$

$\hat{u}_j$ est alors obtenu en résolvant itérativement cette équation et en substituant le nouveau $\hat{u}_j$ dans y_j^* et W_j.

Comme mentionné précédemment,

$$\exp(R_j) = 1 + R_j + \frac{1}{2}R_j^2 + \dots.$$

Dans leurs simulations Raudenbush et al. (2000) utilisent l'approximation avec

$$E\big(\exp(R_j)\big) \approx 1 + E(T_{4j}) + E(T_{6j}) + \frac{1}{2}E(T_{3j}^2)$$

et la trouvent très précise, bien que la méthode autorise à aller plus loin.

Notons également que Lindley (1980) a utilisé les mêmes termes pour approximer les dérivées jusque l'ordre six et que Liu and Pierce (1993) ont utilisé les dérivées jusque l'ordre quatre pour établir les approximations dans le cadre d'intégrales univariées. Le développement complet du terme de correction implique T_4, T_6, $T_3^2/2$,T_8, $T_3T_5/2$, $T_4^2/2$, ...

L'ordre de grandeur des plus hauts termes diminue pour deux principales raisons. Tout d'abord, dans le cas binomial et poissonnien, le dénominateur factoriel augmente rapidement. Ensuite, les termes diminuent en fonction du nombre d'observations par groupe j, noté n_j. Les termes T_4 et $T_3^2/2$ sont $O(n_j^{-1})$ tandis que T_6, $T_3T_5/2$, $T_4^2/2$ sont $O(n_j^{-2})$ et T_8 est $O(n_j^{-3})$. Les termes de plus hauts ordres sont $O(n_j^{-3})$ ou plus petits et ont de très grands dénominateurs factoriels. Cela implique que, pour la plupart des applications, ajouter T_4 sans ajouter $T_3^2/2$ n'améliorerait que très peu l'approximation. Pour certaines applications, il peut être utile d'ajouter $T_3T_5/2$, $T_4^2/2$, bien que dans l'expérience de Raudenbush et al. (2000) avec le cas logistique ces termes aient été négligés.

En utilisant l'approximation à l'ordre six et en appliquant les équations (3.5) et (3.6), l'approximation de (3.11) à une constante près est

$$L \approx \prod_{j=1}^{K} \sigma_j^{-q}|A_j|^{-\frac{1}{2}}|Z_j^T W_j Z_j + \sigma_j^{-2}A_j^{-1}|^{-\frac{1}{2}} \exp\left\{\hat{l}_j - \frac{\hat{u}_j^T A_j^{-1}\hat{u}_j}{2\sigma_j^2}\right\} \times \exp\left\{1 + E(T_{4j}) + E(T_{6j}) + \frac{1}{2}E(T_{3j}^2)\right\} \tag{3.12}$$

Une approximation alternative à (3.12) proposée par Shun and Mc Cullagh (1995) et Shun (1995) est

$$L \approx \prod_{j=1}^{K} \sigma_j^{-q}|A_j|^{-\frac{1}{2}}|Z_j^T W_j Z_j + \sigma_j^{-2}A_j^{-1}|^{-\frac{1}{2}} \exp\left\{\hat{l}_j - \frac{\hat{u}_j^T A_j^{-1}\hat{u}_j}{2\sigma_j^2}\right\} \times \exp\left\{E(T_{4j}) + E(T_{6j}) + \frac{1}{2}E(T_{3j}^2)\right\} \tag{3.13}$$

Dans les simulations de Raudenbush et al. (2000), les approximations (3.12) et (3.13) mènent à des résultats essentiellement identiques. En prenant le log de (3.12) et en appliquant des simplifications algébriques (Yang, 1998), la log-vraisemblance marginal s'écrit alors

$$\log(L) \approx \sum_{j=1}^{K} -q_j \log(\sigma_j) - \frac{1}{2}\log|A_j| + \frac{1}{2}\log|V_j| + \hat{l}_j - \frac{\hat{u}_j^T A_j^{-1}\hat{u}_j}{2\sigma_j^2} + \log(M_j) \quad (3.14)$$

où

$$\begin{aligned} M_j &= 1 + E(T_{4j}) + E(T_{6j}) + \frac{1}{2}E(T_{3j}^2) \\ &= 1 - \frac{1}{8}\sum_{i}^{n_j} m_{ij}^{(4)} B_{ij}^2 - \frac{1}{48}\sum_{i}^{n_j} m_{ij}^{(6)} B_{ij}^3 + \frac{15}{72} k_j^T V_j k_j \end{aligned}$$

avec

$$\begin{aligned} V_j^{-1} &= -h_j^{(2)}(\hat{u}_j) = Z_j^T W_j Z_j + \sigma^{-2} A_j^{-1}, \\ B_{ij} &= Z_{ij}^T V_j Z_{ij} \quad \text{et} \\ k_j &= \sum_{i}^{n_j} m_{ij}^{(3)} Z_{ij} B_{ij}. \end{aligned}$$

L'algorithme de Fisher Scoring (Green, 1984) est utilisé pour maximiser (3.14) et mène à une approximation très précise des estimateurs du maximum de vraisemblance pour β et σ_j pour tout j (voir annexes).

3.2.3 Determination des modes conditionnels

L'utilisation de l'approximation de Laplace et de Gauss-Hermite adaptative requiert donc le calcul de $\hat{u} = (\hat{u}_1, \ldots, \hat{u}_K)$ le vecteur des modes conditionnels des effets aléatoires. Pour des valeurs données de β et θ nous avons

$$\begin{aligned} \tilde{u}(\beta, \theta) &= \arg\max_u h(u) \\ &= \arg\max_u \log\Big(f_{Y|U}(y|u) f_U(u)\Big) \end{aligned}$$

Ces modes conditionnels sont les valeurs des effets aléatoires rendant maximum la densité conditionnelle des effets aléatoires sachant les données et les paramètres du modèle.

Comme $\tilde{u}$ maximise $h(u)$, il peut être trouvé en résolvant l'équation $\frac{\partial h(u)}{\partial u}|_{u=\tilde{u}} = 0$ par des algorithmes itératifs tel que celui de Newton-Raphson qui rappelons-le est équivalent à celui de Fisher dans le cas d'un lien canonique (voir chapitre 2).

Cet algorithme s'écrit :

$$\tilde{u}^{(t+1)} = \tilde{u}^{(t)} - (h''(\tilde{u}))^{-1} \frac{\partial h(u)}{\partial u^T} | u = \tilde{u}^{(t)}$$

où $h''(.)$ est la matrice hessienne de $h(.)$.

En combinant (3.2) et (3.3) et en supposant $a(\kappa) = 1$ on obtient que :

$$\begin{aligned} \tilde{u}^{(t+1)} &= \tilde{u}^{(t)} - \big(Z^T W^{-1}_{\tilde{u}^{(t)}} Z + \Sigma_\theta^{-1}\big)^{-1} \big(- Z^T y + Z^T diag\big(b'(X\beta + Z\tilde{u}^{(t)})\big) + \Sigma_\theta^{-1}\tilde{u}^{(t)}\big) \\ &= \tilde{u}^{(t)} - \big(Z^T W^{-1}_{\tilde{u}^{(t)}} Z + \Sigma_\theta^{-1}\big)^{-1} \big(- Z^T (y - \tilde{\mu}) + \Sigma_\theta^{-1}\tilde{u}^{(t)}\big) \end{aligned} \quad (3.15)$$

d'où

$$\big(Z^T W^{-1}_{\tilde{u}^{(t)}} Z + \Sigma_\theta^{-1}\big)\tilde{u}^{(t+1)} = Z^T W^{-1}_{\tilde{u}^{(t)}} Z\tilde{u}^{(t)} + \Sigma_\theta^{-1}\tilde{u}^{(t)} + Z^T y - Z^T\tilde{\mu}^{(t)} - \Sigma_\theta^{-1}\tilde{u}^{(t)}$$

ce qui donne finalement

$$\tilde{u}^{(t+1)} = \big(Z^T W^{-1}_{\tilde{u}^{(t)}} Z + \Sigma_\theta^{-1}\big)^{-1} Z^T W^{-1}_{\tilde{u}^{(t)}} \big(Z\tilde{u}^{(t)} + W_{\tilde{u}^{(t)}}(y - \tilde{\mu}^{(t)})\big)$$

où $\tilde{\mu}^{(t)}$ est μ évalué pour $u = \tilde{u}^{(t)}$. En notant $W_{\tilde{u}^{(t)}}(y - \tilde{\mu}^{(t)}) + \tilde{\eta}^{(t)} = y^*$ nous obtenons

$$\tilde{u}^{(t+1)} = \big(Z^T W^{-1}_{\tilde{u}^{(t)}} Z + \Sigma_\theta^{-1}\big)^{-1} Z^T W^{-1}_{\tilde{u}^{(t)}} (y^* - X\beta)$$

On est donc amenés à résoudre de manière itérative cette équation où Y^* est la version linéarisée de la variable réponse donnée par $Y^* = W_{\tilde{u}}(Y - \tilde{\mu}) + \tilde{\eta}$ et "(t)" est l'exposant indiquant que la matrice ou le vecteur est évalué à la t$^{\text{ième}}$ itération quand $u = \tilde{u}^{(t)}$.

A la t$^{\text{ième}}$ itération de l'algorithme la valeur courante du vecteur des effets aléatoires $\tilde{u}^{(t)}$ produit un prédicteur linéaire $\tilde{\eta}^{(t)} = X\beta + Z\tilde{u}^{(t)}$. La séquence d'itérations $\tilde{u}_0, \tilde{u}_1, \ldots$ est alors considérée converger vers les modes conditionnels $\tilde{u}(\beta, \theta, y)$ quand $\| \tilde{\eta}^{(t+1)} - \tilde{\eta}^{(t)} \| / \| \tilde{\eta}^{(t)} \|$ ne dépasse plus un certain seuil fixé. La matrice de variance covariance de u, conditionnellement à β et θ, est donc approximée par (Doran et al., 2007):

$$\mathrm{Var}(U|\beta, \theta, y) \approx \big(Z^T W_u^{-1} Z + \Sigma_\theta^{-1}\big)^{-1}$$

3.3 Approximation par quadrature gaussienne

3.3.1 Approximation de Gauss-Hermite

La quadrature de Gauss-Hermite (Liu and Pierce, 1993) est souvent utilisée pour les intégrations numériques en statistiques, en raison de sa relation avec les densités gaussiennes, mais demande souvent à être repensée afin de rendre son implémentation adéquate (Liu and Pierce, 1994). La quadrature gaussienne est définie pour des intégrales de la forme

$$\int_{\mathbb{R}} \exp(-t^2)\, f(t)\, \mathrm{d}t. \tag{3.16}$$

Dans de nombreuses applications statistiques, la densité gaussienne est un facteur explicite de l'intégrande. Lorsque ce n'est pas le cas, une transformation linéaire peut être faite afin de faire apparaitre le facteur $\exp(-t^2)$ dans l'intégrande.

La forme générale de la quadrature de Gauss-Hermite s'écrit

$$\int_{\mathbb{R}} \exp(-t^2)\, f(t)\, \mathrm{d}t \simeq \sum_{m=1}^{M} v_m f(t_m) \tag{3.17}$$

où $f(.)$ est une fonction lisse pouvant être approximée de manière précise par un polynôme, M est le nombre de points de quadrature, t_m et v_m sont respectivement les points de quadrature (noeuds ou abscisses) et les poids de quadrature associés à ces points. Les points de quadrature sont déterminés comme les M racines du $M^{\text{ème}}$ polynôme de Hermite. Pour $M = 1, \ldots, 20$ les valeurs des noeuds et abcisses de quadrature peuvent être trouvées dans Abramowitz and Stegun (1974) et pour $M > 20$ elles peuvent être calculées à l'aide de l'algorithme décrit par Golub and Welsch (1969).

La quadrature de Gauss-Hermite peut également être utilisée dans le cas multivarié en appliquant la quadrature à chaque intégrale univariée tour à tour (Pan and Thompson, 2003)

$$\begin{aligned}
\int_{\mathbb{R}^q} g(t)\, \mathrm{d}t &= \int_{\mathbb{R}} \ldots \int_{\mathbb{R}} g(t_1, \ldots, t_q)\, \mathrm{d}t_1 \ldots \mathrm{d}t_q \\
&= \int_{\mathbb{R}} \ldots \int_{\mathbb{R}} f(t_1, \ldots, t_q) e^{-t^T t}\, \mathrm{d}t_1 \ldots \mathrm{d}t_q \\
&\simeq \sum_{m_1=1}^{M_1} v_{m_1}^{(1)} \sum_{m_2=1}^{M_2} v_{m_2}^{(2)} \ldots \sum_{m_q=1}^{M_q} v_{m_q}^{(q)}\, f(t_{m_1}^{(1)}, t_{m_2}^{(2)}, \ldots, t_{m_q}^{(q)})
\end{aligned} \tag{3.18}$$

où $t_{m_j}^{(j)}$ et $v_{m_j}^{(j)}$ $(m_j = 1, \ldots, M_j)$ sont les noeuds et poids de quadrature pour M_j points de quadrature à la $j^{\text{ème}}$ coordonnée de t, $j = 1, \ldots, q$. Cependant, cette

expression requiert l'évaluation de $f(t_1, \ldots, t_q)$ en $M_1 M_2 \ldots M_q$ noeuds et peut alors demander énormément de temps de calcul.

3.3.2 Approximation de Gauss-Hermite et densité gaussienne

Si la fonction $\exp\{-t^2\}$ dans l'équation (3.17) est remplacée par une densité normale ϕ de moyenne μ et d'écart-type σ, alors les poids de quadrature sont linéairement transformés (Naylor and Smith, 1982). Ainsi, la quadrature de Gauss-Hermite(3.17) peut se réexprimer sous la forme :

$$\int_{\mathbb{R}} f(t)\phi(t|\mu, \sigma^2)\,\mathrm{d}t \simeq \sum_{m=1}^{M} f(t_m^*)v_m^* \tag{3.19}$$

où les noeuds et poids transformés $t_m^* = \mu + \sigma\sqrt{2}t_m$ et $v_m^* = \frac{v_m}{\sqrt{\pi}}$ constituent un nouvel ensemble de noeuds et poids qui sont dits basés sur le noyau normal.

En effet, en changeant la variable d'intégration par une variable normale standard, c'est à dire en posant $r = \frac{t-\mu}{\sqrt{2}\sigma}$, cela devient

$$\begin{aligned}
\int_{\mathbb{R}} f(t)\phi(t|\mu, \sigma^2)\,\mathrm{d}t &= \frac{1}{\sqrt{2\pi}\sigma}\int f(t)\exp\{-\frac{1}{2}\frac{t-\mu}{\sigma^2}\}\,\mathrm{d}t \\
&= \frac{\sqrt{2}\sigma}{\sqrt{2\pi}\sigma}\int f(\mu + \sigma\sqrt{2}r)\exp(-r^2)\,\mathrm{d}r \\
&= \frac{1}{\sqrt{\pi}}\int f(\mu + \sigma\sqrt{2}r)\exp(-r^2)\,\mathrm{d}r.
\end{aligned}$$

En appliquant ensuite la quadrature de Gauss-Hermite (3.17), cela donne

$$\frac{1}{\sqrt{\pi}}\int f(\mu + \sigma\sqrt{2}r)\exp(-r^2)\,\mathrm{d}r \simeq \sum_{m=1}^{M} f(\mu + \sigma\sqrt{2}t_m)\frac{v_m}{\sqrt{\pi}}$$

3.3.3 Approximation de Gauss-Hermite Adaptative et GL2M

Cependant l'utilisation de la méthode de Gauss-Hermite pose quelques problèmes dans le cas des GL2M. En effet, la qualité de l'approximation dépend grandement du nombre de points de quadrature, plus celui-ci est élevé plus l'approximation est améliorée. Or, pour un GL2M dont le vecteur des effets aléatoires est de dimension

q la méthode doit approximer q intégrales. En supposant que chaque intégrale est approximée par une somme de M termes, la méthode de Gauss-Hermite requiert alors l'évaluation de $M \times q$ termes, ce qui fait considérablement augmenter la charge de calcul. Un autre aspect critique qui a également une grande influence sur la qualité de l'approximation de Gauss-Hermite est l'emplacement des points de quadrature et ce même avec un nombre de quadrature élevé.

La méthode de Gauss-Hermite adaptative (Liu and Pierce, 1994; Pinheiro and Bates, 2000) permet de résoudre ces problèmes en adaptant les abscisses et les poids à la fonction à intégrer. Dans notre cas, les abscisses sont centrées autour des modes conditionnels des effets aléatoires notés $\tilde{u}$ et redimensionnés en multipliant leurs valeurs par l'inverse de la matrice hessienne de l'intégrande de la log-vraisemblance du GL2M considéré.

On suppose qu'on a un modèle avec un seul effet aléatoire tel que $A_1 = I_{q_1} = I_q$ (voir notations de la partie 3.1) où I_q désigne la matrice identité de taille q avec $\Sigma_\theta = \sigma^2 I_q$. Les composantes u_i $(i = 1, \ldots, q)$ de l'unique effet aléatoire U sont alors indépendantes et identiquement distribuées telles que $u_i \sim N(0, \sigma^2)$ pour tout i.

La vraisemblance d'un tel modèle s'écrit :

$$\prod_{i=1}^{q} \int_{\mathbb{R}} f_{u_i}(u_i) \prod_{j=1}^{n_i} f_{y_{ij}|u_i}(y_{ij}|u_i) \ \mathrm{d}u_i$$

où n_i est le nombre d'observations caractérisés par la $i^{\text{ème}}$ composante de l'effet aléatoire U.

Ainsi, la contribution apportée à la vraisemblance par la composante i s'écrit

$$\int_{\mathbb{R}} f_{u_i}(u_i) \prod_{j=1}^{n_i} f_{y_{ij}|u_i}(y_{ij}|u_i) \ \mathrm{d}u_i.$$

En changeant la variable d'intégration par une variable normale standard, c'est à dire en posant $u_i^* = u_i/\sigma$, cela devient

$$\int_{\mathbb{R}} \Phi(u_i^*) \prod_{j=1}^{n_i} f(y_{ij}|\sigma u_i^*) \ \mathrm{d}u_i^* \tag{3.20}$$

où $\Phi(.)$ est la fonction de densité d'une loi normale standard

$$\Phi(u_i^*) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u_i^{*2}}{2}).$$

En appliquant la quadrature de Gauss-hermite, cela donne

$$\int_{\mathbb{R}} \Phi(u_i^*) \prod_{j=1}^{n_i} f(y_{ij}|\sigma u_i^*) \ \mathrm{d}u_i^* \simeq \sum_{m=1}^{M} v_m^* \prod_{j=1}^{n_i} f(y_{ij}|\sigma t_m^*)$$

où

$$v_m^* \equiv v_m/\sqrt{\pi} \qquad t_m^* \equiv \sqrt{2} t_m.$$

L'intégrande dans (3.20)

$$h(u_i^*) = \Phi(u_i^*) \prod_{j=1}^{n_i} f(y_{ij}|\sigma u_i^*)$$

est proportionnelle à $f(u_i|y_{ij})$ qui est la densité à posteriori de u_i sachant les observations. Cette densité peut être approximée par une densité de loi normale $N(\mu_i, \tau_i^2)$ où $\mu_i = \tilde{u_i}$ et $\tau_i^2 = \left(\frac{\partial^2 h(u_i^*)}{\partial u_i^* \partial u_i^{*T}}\right)^{-1}$.

Au lieu de traiter la densité à priori $\Phi(u_i^*)$ comme étant la fonction de pondération pour la méthode de Gauss-Hermite, nous réécrivons (3.20) de la manière suivante

$$f(y_i) = \int_{\mathbb{R}} \varphi(u_i^*; \mu_i, \tau_i^2) \left(\frac{\Phi(u_i^*) \prod_{j=1}^{n_i} f(y_{ij}|\sigma u_i^*)}{\varphi(u_i^*; \mu_i, \tau_i^2)} \right) \mathrm{d}u_i^*$$

où $\varphi(u_i^*; \mu_i, \tau_i^2)$ est la densité d'une loi normale $N(\mu_i, \tau_i^2)$ et traitons alors la densité normale qui approxime la densité a posteriori comme la fonction de pondération pour la quadrature.

En effectuant le changement de variable $x_i = (u_i^* - \mu_i)/\tau_i$ et en appliquant l'approximation de Gauss-Hermite nous obtenons

$$\begin{aligned} f(y_i) &= \int_{\mathbb{R}} \frac{\Phi(x_i)}{\tau_i} \left\{ \frac{\Phi(\tau_i x_i + \mu_i) \prod_{j=1}^{n_i} f(y_{ij}|\sigma(\tau_i x_i^* + \mu_i))}{\exp(-x_i^2/2)/\sqrt{2\pi\tau_i^2}} \right\} \tau_i \ \mathrm{d}x_i \\ &= \sum_{m=1}^{M} v_m^* \left\{ \frac{\Phi(\tau_i t_m^* + \mu_i) \prod_{j=1}^{n_i} f(y_{ij}|\sigma(\tau_i t_m^* + \mu_i))}{\exp(-t_m^{*2}/2)/\sqrt{2\pi\tau_i^2}} \right\} \\ &= \sum_{m=1}^{M} \pi_{im} \prod_{j=1}^{n_i} f(y_{ij}|\sigma\alpha_{im}) \end{aligned}$$

où

$$\alpha_{im} \equiv \tau_i t_m^* + \mu_i$$

sont les abscisses déplacées et redimensionnées dont les poids correspondants sont

$$\pi_{im} \equiv \sqrt{2\pi}\tau_i \exp(t_m^{*2}/2) \Phi(\tau_i t_m^* + \mu_i) v_m^*$$

3.4 Utilisation du logiciel R

L'estimation des GL2M par les méthodes de Laplace et Adaptative Gauss-Hermite (AGH) sous R (R Development Core Team, 2012) se fait à l'aide de la fonction **glmer** du package `lme4` (Bates, 2010). L'approximation de Laplace est par défaut et pour obtenir l'approximation AGH il faut renseigner le nombre de points de quadrature à l'aide de la commande 'nAGQ'.

La fonction **glmer** travaille sous l'hypothèse que $A_i = I_{q_i}$ où I_{q_i} est la matrice identité de taille $q_i \times q_i$ d'où $U \sim \mathcal{N}_q(0, \Sigma_\theta)$ où Σ_θ est une matrice diagonale par blocs telle que $\Sigma_\theta = \text{diag}\{\sigma_i^2 I_{q_i}\}_{i=1,\ldots,K}$ et avec $\theta = (\sigma_1^2, \ldots, \sigma_K^2)$.

La fonction **glmer** est de la forme :

$$\text{glmer(formula, data, family = gaussian, start = NULL, verbose = FALSE, nAGQ = 1, subset, weights, na.action, ...)}$$

Les arguments 'formula', 'data' et 'family' sont obligatoires et représentent respectivement la forme du modèle, les données du modèle et la famille de GLM utilisée.

La forme du modèle est une formule linéaire ayant deux partie séparées par $\sim$. La partie gauche désigne la variable à expliquer et la partie de droite les variables explicatives (variables à effets fixes et à effets aléatoires).

Les autres arguments sont optionnels et les principaux 'start', 'verbose', 'nAGQ', 'subset', 'weights', 'na.action', 'offset', 'contrasts', 'model' et'control' représentent respectivement l'ensemble des valeurs de départ données aux paramètres, un argument logique indiquant s'il y a impression ou non des différentes itérations, le nombre de points de quadrature à utiliser, un sous-ensemble de données qui seront ajustées par le modèle, les poids ω_i associés à l'observation i définis dans la fonction de densité des GLM (2.1) et une fonction indiquant ce qu'il se passe quand les données contiennent des 'NA'..

3.5 Mise en oeuvre sous R et comparaison

3.5.1 Procédure de simulation

Nous souhaitons comparer les méthodes de Laplace et AGH. Pour cela, nous simulons 100 jeux de données de 800 observations et les analysons à l'aide des deux méthodes. Nous voulons savoir quelle est la proximité des estimations avec les vraies valeurs ainsi que le temps nécessaire à chacune de ses méthodes pour founir des estimations. La structure des données simulées est similaire à celle de la base de données d'accident qui a motivé ce travail.

Elle implique 5 variables à effets fixes, X_1, X_2, X_3, X_4, X_5 et une variable à effets aléatoires U ayant 6 niveaux. La variable X_0 est fixée à 1, les cinq autres variables à effets fixes sont générées à partir de distributions multinomiales $\mathcal{M}(n; \pi_1^{(0)}, \ldots, \pi_r^{(0)})$ telles que
$\mathbf{X}_1 \sim \mathcal{M}(n; 0.325,\ 0.237,\ 0.179,\ 0.136,\ 0.076,\ 0.047)$,
$\mathbf{X}_2 \sim \mathcal{M}(n; 0.2320, 0.2430, 0.2120, 0.1250, 0.0850, 0.0640, 0.0390)$,
$\mathbf{X}_3 \sim \mathcal{M}(n; 0.189,\ 0.561,\ 0.193,\ 0.058)$,
$\mathbf{X}_4 \sim \mathcal{M}(n; 0.3720,\ 0.6280)$,
$\mathbf{X}_5 \sim \mathcal{M}(n; 0.293,\ 0.707)$.
U est générée à partir d'une distribution multivariée normale $\mathcal{N}(\mathbf{0}_6; \sigma^2 \times I_6)$ où $\mathbf{0}_6$ est le vecteur de dimension 6 composé de zéro et I_6 la matrice identité de dimension 6×6. La variable dépendante y_{ij} est binaire (1 ou 0) et est générée à partir d'une loi de Bernoulli.

$$y_{ij} \sim \mathcal{B}(\text{logit}^{-1}(\beta_0 + \sum_{m=1}^{5} \mathbf{X}_{ijm}^T \beta_m + \mathbf{U}_i)), \quad (i = 1, \ldots, 6; j = 1, \ldots, n_i) \qquad (3.21)$$

où β_0 est une constante, les β_m sont des vecteurs de dimension $(p_m - 1)$ où p_m est le nombre de modalités de la $m^{\text{ième}}$ variable à effets fixes $\mathbf{X}_{ijm}$. Dans les simulations, les vraies valeurs des paramètres à effets fixes sont fixées à:
$\beta_0 = 6.440$, $\beta_1^T = (-8.448, -6.681, -5.625, -4.372, -3.338)$,
$\beta_2^T = (-1.220, -1.912, -1.659, -1.338, -2.233, -1.073)$, $\beta_3^T = (-0.757, -0.865, -0.988)$,
$\beta_4 = -0.701$, $\beta_5 = -0.561$
et l'écart-type de l'effet aléatoire est fixé à: $\sigma = 0.684$.

3.5.2 Analyse du jeu de données 1

La table 3.1 présente les estimations moyennnes, les biais moyens, les erreurs standards des paramètres à effets fixes et l'erreur quadratique moyenne pour chaque méthode. Les valeurs entre parenthèses représentent le pourcentage de réplications qui ont convergé sur les 100 effectuées. Les estimations des paramètres sont

effectuées en fonction de ce nombre de réplications convergentes. Nous pouvons voir que la méthode de Laplace a convergé 73 fois (73%) et AGH15 58 fois (58%).

Concernant les estimations moyennes, nous observons que le taux de sous-estimation (comparaison entre l'estimation moyenne et sa vraie valeur) des effets fixes et aléatoires est assez élevé pour les deux méthodes. Les méthodes d'estimation utilisées donnent des taux de sous-estimations élevés mais similaires et des estimations moyennes très proches des vraies valeurs des paramètres, et ce particulièrement en ce qui concerne les paramètres des effets fixes.

A chaque paramètre à effet fixe estimé, nous avons également calculé une estimation de son erreur standard par inversion numérique de la matrice hessienne au point de convergence. On en déduit alors une estimation de l'erreur standard en prenant la racine carrée de la diagonale principale de l'inverse. Nous observons que dans la plupart des cas les estimations moyennes des erreurs standards sont très proches pour chaque méthode et restent inférieures à un. Le calcul des étendues moyennnes (différences moyennes entre les estimations maximum et minimum) montre que la méthode de Laplace donne des estimations moins dispersées des erreurs standards que la méthode AGH15. Des calculs supplémentaires de la table 3.1 montrent que l'étendue des estimations des erreurs standards en moyenne est de 0.3777 pour la méthode de Laplace et de 0.4184 pour la méthode AGH15. Globalement, les deux méthodes fournissent des estimations moyennes des erreurs standards quasiment identiques.

Parallèlement aux valeurs estimées des paramètres et à leurs erreurs standards, nous avons étudié les méthodes d'approximation à travers le calcul des biais et de l'erreur quadratique moyenne. Le biais est obtenu en faisant la différence entre la valeur moyenne (colonne mean) et la vraie valeur (colonne true value) du paramètre et l'erreur quadratique moyenne spécifique aux paramètres à effets fixes est définie par l'expression suivante:

$$\mathrm{M}SE(\beta) = \frac{1}{J}\sum_{j=1}^{J}(\hat{\beta}_{n,j}^{(mean)} - \beta_j^{(0)})^2$$

où $\hat{\beta}_{n,j}^{(\mathrm{mean})}$ est la moyenne estimée pour chaque paramètre à effet fixe (colonne mean) et calculée en fonction du nombre de réplications convergentes, J est le nombre de paramètres à effet fixe réellement estimés et $\beta_j^{(0)}$ la vraie valeur associée (colonne True Value). Il s'agit d'une mesure empirique globale qui permet d'analyser la proximité entre les solutions issues des méthodes d'approximation et les vraies valeurs. Concernant les biais, nous pouvons voir que les deux méthodes fournissent des résultats quasi-identiques pour les paramètres à effets fixes.

En moyenne, le biais des paramètres à effets fixes vaut 0.009 pour la méthode de Laplace et 0.007 pour la méthode AGH15. La différence entre les deux méthodes se fait plus au niveau du biais de l'écart-type de la variable à effets aléatoires, en effet ce dernier est beaucoup plus petit avec la méthode AGH15 (en valeur absolue, 0.145 pour Laplace contre 0.049 pour AGH15). Pour finir, les résultats montrent que la méthode AGH15 tend à avoir un MSE relatif aux effets fixes plus petit que celui de la méthode de Laplace. Plus de détails concernant la comparaison de ces deux méthodes sont donnés au chapitre 5 notamment avec des jeux de données de plus grandes tailles.

		Laplace (73)					**AGH15 (58)**				
		Estim.	**Bias**	**Standard Errors**			**Estim.**	**Bias**	**Standards Errors**		
Parameter	**TV**	**mean**	**mean**	**min**	**mean**	**max**	**mean**	**mean**	**min**	**mean**	**max**
β_0	6,440	6,359	-0,081	0,863	1,126	1,425	6,269	-0,172	0,852	1,150	1,455
β_{11}	-8,448	-8,390	0,058	0,741	1,084	1,508	-8,575	-0,127	0,739	1,108	1,540
β_{12}	-6,681	-6,529	0,152	0,670	0,931	1,184	-6,701	-0,020	0,631	0,945	1,192
β_{13}	-5,625	-5,415	0,210	0,622	0,904	1,186	-5,547	0,078	0,582	0,907	1,159
β_{14}	-4,372	-4,113	0,260	0,613	0,888	1,143	-4,226	0,146	0,553	0,890	1,137
β_{15}	-3,338	-3,070	0,268	0,621	0,903	1,141	-3,167	0,171	0,576	0,902	1,160
β_{21}	-1,220	-1,390	-0,170	0,492	0,603	0,781	-1,108	0,112	0,497	0,618	0,888
β_{22}	-1,912	-2,044	-0,132	0,505	0,618	0,791	-1,856	0,056	0,509	0,635	0,911
β_{23}	-1,659	-1,802	-0,143	0,505	0,618	0,811	-1,740	-0,081	0,514	0,641	0,887
β_{24}	-1,338	-1,493	-0,155	0,549	0,650	0,844	-1,243	0,095	0,523	0,666	0,911
β_{25}	-2,233	-2,373	-0,140	0,598	0,754	1,018	-2,262	-0,029	0,620	0,780	1,116
β_{26}	-1,073	-1,225	-0,152	0,575	0,727	0,944	-0,970	0,103	0,592	0,741	0,988
β_{31}	-0,988	-1,000	-0,012	0,274	0,318	0,376	-1,065	-0,077	0,279	0,324	0,425
β_{32}	-0,865	-0,900	-0,035	0,333	0,397	0,482	-0,944	-0,079	0,340	0,402	0,477
β_{33}	-0,757	-0,800	-0,043	0,470	0,603	0,972	-0,847	-0,090	0,491	0,622	0,926
β_4	-0,701	-0,714	-0,013	0,228	0,276	0,332	-0,725	-0,025	0,227	0,280	0,347
β_5	-0,561	-0,586	-0,025	0,245	0,294	0,385	-0,500	0,061	0,249	0,296	0,368
σ	0,684	0,539	-0,145				0,636	-0,049			
		MSE: 2,08E-02					MSE: 9,57E-03				

TV: True Values; Estim.: Estimate; MSE: Mean Squared Error

Table 3.1: S1: Estimates, Biases, Standard Errors and Mean Squared Errors for Laplace and AGH15 methods

Pour conclure, nous pouvons dire que globalement les deux méthodes fournissent des estimations assez précises et très proches en ce qui concerne les paramètres à effets fixes. Néanmoins, en ce qui concerne l'estimation de l'écart-type de la

variable à effets aléatoires, la méthode AGH15 est plus précise que la méthode de Laplace . Cependant, au regard des temps d'estimation, 14 secondes pour la méthode de Laplace contre 43 secondes pour la méthode AGH15, une question serait de savoir si pour des tailles d'échantillon plus élévées la méthode de Laplace deviendrait plus précise quant à l'estimation de l'écart-type de la variable à effets aléatoires. Ce point sera abordé lors du chapitre 5

3.5.3 Analyse du jeu de données 2

Nous effectuons maintenons, une nouvelle analyse, semblable à la précédente mais avec un modèle mixte dont l'écart-type de la variable à effet alétoire est bien supérieur au cas précédent. Dans le cas précédent l'écart-type de l'effet aléatoire valait 0.684, dans ce nouveau cas d'étude il vaut 2.620.
De même, nous simulons 100 jeux de données de 800 observations et les analysons à l'aide des deux méthodes. Nous voulons savoir si l'analyse de ce nouveau jeux de données à travers les deux méthodes mène aux mêmes conclusions que précédemment. La structure des données simulées est également issue du jeu de données ayant motivé le travail. Cette structure est similaire à la précédente mais avec une différence quant à la variable $\mathbf{X}_1$ et avec une différence sur la valeur de l'écart-type de la variable à effets aléatoires. Maintenant,
$\mathbf{X}_1 \sim \mathcal{M}(n; 0.237,\ 0.039,\ 0.055,\ 0.248,\ 0.2470.325,\ 0.174)$
et $\sigma = 2.620$. On simule toujours selon (3.21).

Dans ces nouvelles simulations, les vraies valeurs des paramètres à effets fixes sont fixées à:
$\beta_0 = 2.748$, $\beta_1^T = (-1.165, -1.811, -1.599, -1.266, -2.169, -1.051)$,
$\beta_2^T = (-0.723, -0.885, -1.000)$,
$\beta_3^T = (-0.897, -2.119, -1.829, -0.605, -1.727)$,
$\beta_4^T = -0.750$, $\beta_5^T = -0.589$
et l'erreur standard de la variable à effet aléatoire est tel que $\sigma = 2.620$.

La table 3.2 présente les estimations moyennnes, les biais moyens, les erreurs standards des paramètres à effets fixes et l'erreur quadratique moyenne pour chaque méthode. Tout d'abord, nous pouvons voir que le taux de réplications convergentes est beaucoup plus élevé que pour le cas précédent, en effet la méthode de Laplace a convergé 99 fois (99%) et la méthode AGH15 96 fois (96%). Comme pour le cas précédent, la méthode de Laplace a un taux de convergence plus élévé.

Concernant les estimations moyennes, nous observons comme précédemment que le taux de sous-estimation est assez élevé pour les deux méthodes mais que les estimations moyennes restent cependant très proches des vraies valeures des paramètres,

et ce particulièrement en ce qui concerne les paramètres des effets fixes.

Concernant les erreurs standards des variables à effets fixes, là encore comme dans le cas précédent, nous observons que dans la plupart des cas les estimations moyennes sont très proches pour chaque méthode et restent inférieures à un. Cependant, contrairement au cas précédent, les estimations sont légèrement plus dispersées en moyenne pour la méthode de Laplace que pour la méthode AGH15. Des calculs supplémentaires de la table 3.2 montrent que l'étendue des estimations des erreurs standards en moyenne est de 0.5817 pour la méthode de Laplace et de 0.5616 pour la méthode AGH15. On note également que la sur-dispersion est plus grande que dans le cas précédent. Globalement, comme dans le cas précédent, les deux méthodes fournissent des estimations moyennes des erreurs standards quasiment identiques.

Concernant les biais, nous pouvons voir que la méthode de Laplace fournit en moyenne des biais plus petits que la méthode AGH15 pour les paramètres à effets fixes (0.039 pour Laplace contre 0.091 pour AGH15 en valeur absolue). Néanmoins comme poour le cas précédent la méthode de Laplace fournit un biais plus élevé que la méthode AGH15 en ce qui concerne l'écart-type de la variable à effets aléatoire (0.252 pour Laplace contre 0.161 pour AGH15 en valeur absolue). Nous notons également que ces biais sont en moyenne supérieurs aux cas précédents. Enfin concernant le MSE, les résultats montrent que contrairement au cas précédent la méthode AGH15 tend à avoir un MSE relatif aux effets fixes plus grand que celui de la méthode de Laplace. De même que pour le cas précédent, plus de détails concernant la comparaison de ces deux méthodes seront donnés au chapitre 5 notamment avec des jeux de données de plus grandes tailles.

Pour conclure, nous pouvons dire que la méthode de Laplace fournit des estimations plus précises que la méthode AGH15 en ce qui concerne les paramètres à effets fixes mais qu'elle fournit une estimation moins précise de l'écart-type de la variable à effets aléatoires que la méthode AGH15.

		Laplace (99)					AGH15 (96)				
		Estim.	Bias	Standard Errors			Estim.	Bias	Standards Errors		
Parameter	TV	mean	mean	min	mean	max	mean	mean	min	mean	max
β_0	2,748	2,758	0,010	0,649	1,200	2,212	2,843	0,096	0,565	1,231	2,341
β_{11}	-1,165	-1,233	-0,069	0,440	0,614	1,124	-1,313	-0,148	0,407	0,608	1,153
β_{12}	-1,811	-1,877	-0,065	0,457	0,621	1,124	-1,961	-0,150	0,433	0,616	1,147
β_{13}	-1,599	-1,669	-0,070	0,450	0,624	1,120	-1,763	-0,164	0,430	0,620	1,145
β_{14}	-1,266	-1,302	-0,036	0,477	0,652	1,146	-1,458	-0,192	0,456	0,647	1,167
β_{15}	-2,169	-2,243	-0,074	0,534	0,717	1,171	-2,382	-0,213	0,515	0,717	1,171
β_{21}	-1,051	-1,072	-0,020	0,534	0,728	1,217	-1,250	-0,199	0,485	0,716	1,290
β_{22}	-0,897	-0,904	-0,007	0,452	0,619	1,904	-0,900	-0,004	0,417	0,605	1,082
β_{23}	-2,119	-2,115	0,004	0,399	0,581	1,037	-2,202	-0,083	0,412	0,596	1,075
β_{24}	-1,829	-1,890	-0,061	0,252	0,340	0,522	-1,931	-0,102	0,241	0,342	0,487
β_{25}	-1,727	-1,822	-0,096	0,250	0,340	0,508	-1,833	-0,107	0,242	0,340	0,485
β_{26}	-0,605	-0,649	-0,043	0,257	0,344	0,514	-0,685	-0,080	0,255	0,342	0,525
β_{31}	-1,000	-1,041	-0,041	0,226	0,298	0,415	-1,032	-0,032	0,219	0,297	0,430
β_{32}	-0,885	-0,952	-0,067	0,271	0,365	0,511	-0,883	0,001	0,257	0,360	0,527
β_{33}	-0,723	-0,692	0,031	0,383	0,532	1,077	-0,798	-0,075	0,398	0,539	0,915
β_4	-0,750	-0,771	-0,021	0,175	0,241	0,326	-0,837	-0,087	0,175	0,241	0,347
β_5	-0,589	-0,622	-0,032	0,185	0,255	0,354	-0,605	-0,015	0,190	0,256	0,356
σ	2,620	2,368	-0,252				2,459	-0,161			
		MSE: 2,63E-03					MSE: 1,48E-02				

TV: True Values; Estim.: Estimate; MSE: Mean Squared Error

Table 3.2: S2: Estimates, Biases, Standard Errors and Mean Squared Errors for Laplace and AGH15 methods

Chapter 4

Quasi-vraisemblance pénalisée (PQL)

4.1 La méthode PQL

La méthode de la quasi-vraisemblance pénalisée, notée PQL pour Penalized Quasi-Likelihood est une methode d'estimation proposée par Breslow and Clayton (1993). Elle est basée sur une approximation de la vraisemblance marginale par approximation de Laplace d'ordre un.

Dans un premier temps, Breslow et Clayton définissent une fonction de quasi-vraisemblance marginale en intégrant la quasi-vraisemblance conditionnelle de Y sachant u par rapport à la loi de u :

$$Q(\beta,\theta;y) \propto |\Sigma_\theta|^{-\frac{1}{2}} \int_{\mathbb{R}^q} \exp\big(-\frac{1}{2}\sum_{i=1}^{n} d_i(y_i,\mu_{u,i}) - \frac{1}{2}u\Sigma_\theta^{-1}u\big)\,\mathrm{d}u \tag{4.1}$$

où

$$d_i(y_i,\mu_{u,i}) = -2\int_{y_i}^{\mu_{u,i}} \frac{y_i - t}{a_i(\kappa)v(t)}\,\mathrm{d}t \tag{4.2}$$

est le logarithme de la fonction de quasi-vraisemblance pour la loi conditionnelle de y_i sachant l'effet aléatoire à un facteur -2 près et où Σ_θ et θ sont définis comme précédemment c'est à dire que Σ_θ est la matrice de variance-covariance des effets aléatoires du modèle tel que $\theta = (\sigma_1,\ldots,\sigma_K)$.
Comme pour le cas de la vraisemblance, on ne peut calculer directement (4.1). On utilise donc l'approximation de Laplace pour l'équation :

$$PQL(\beta,\theta;y) = -\frac{1}{2}\sum_{i=1}^{n} d_i(y_i-\mu_{u,i}) - \frac{1}{2}u^T\Sigma_\theta^{-1}u$$

que l'on appelle log-quasi-vraisemblance pénalisée puisqu'un terme de pénalité $-\frac{1}{2}u\Sigma_\theta^{-1}u$ a été introduit à la quasi-vraisemblance.

L'équation (4.1) étant de la forme $c|\Sigma_\theta|^{-\frac{1}{2}}\int e^{-k(u)}\,\mathrm{d}u$, une log-quasi-vraisemblance pénalisée approchée est obtenue par approximation de Laplace d'ordre 1 en $\tilde{u}$ solution de $k'(u)=0$:

$$q(\beta,\theta) \simeq -\frac{1}{2}\log|\Sigma_\theta| - \frac{1}{2}\log|k''(\tilde{u})| - k(\tilde{u}) \tag{4.3}$$

avec

$$k(u) = \frac{1}{2}\sum_{i=1}^{n} d_i(y_i-\mu_{u,i}) + \frac{1}{2}u^T\Sigma_\theta^{-1}u \tag{4.4}$$

d'où

$$\begin{aligned} k'(u) &= -\sum_{i=1}^{n}\frac{(y_i-\mu_{u,i})z_i}{a_i(\kappa)v(\mu_{u,i})g'(\mu_{u,i})} + \Sigma_\theta^{-1}u \quad \text{où } Z_i^T \text{ est la } i^{\text{ème}} \text{ ligne de } Z \\ k''(u) &= \sum_{i=1}^{n}\frac{z_iz_i^T}{a_i(\kappa)v(\mu_{u,i})g'(\mu_{u,i})^2} + \Sigma_\theta^{-1} - \sum_{i=1}^{n}(y_i-\mu_{u,i})z_i\frac{\partial}{\partial u}\Big[\frac{1}{a_i(\kappa)v(\mu_{u,i})g'(\mu_{u,i})}\Big]^T \end{aligned}$$

Le dernier terme de $k''(u)$ est d'espérance nulle et est égal à zéro pour la fonction de lien canonique. En négligeant ce terme, on obtient l'approximation de $k''(u)$ suivante :

$$k''(u) = Z^TW_u^{-1}Z + \Sigma_\theta^{-1} \tag{4.5}$$

avec

$$W_u = \mathrm{diag}\big\{a_i(\kappa)v(\mu_{u,i})g'(\mu_{u,i})^2\big\}_{i=1,\ldots,n}$$

Finalement, en combinant (4.3)-(4.5), on obtient l'approximation de la log-quasi-vraisemblance pénalisée suivante :

$$\begin{aligned} q(\beta,\theta) &\simeq -\frac{1}{2}\log|\Sigma_\theta| - \frac{1}{2}\log|Z^TW_{\tilde{u}}^{-1}Z + \Sigma_\theta^{-1}| - \frac{1}{2}\sum_{i=1}^{n} d_i(y_i\mu_{\tilde{u},i}) - \frac{1}{2}u^T\Sigma_\theta^{-1}u \\ &\simeq -\frac{1}{2}\log|Z^TW_{\tilde{u}}^{-1}Z\Sigma_\theta + I_q| - \frac{1}{2}\sum_{i=1}^{n} d_i(y_i\mu_{\tilde{u},i}) - \frac{1}{2}u^T\Sigma_\theta^{-1}u \end{aligned} \tag{4.6}$$

En supposant que $W_{\tilde{u}}^{-1}$ varie de façon négligeable en fonction des paramètres, on néglige le premier terme et les paramètres qui maximisent (4.6) sont alors ceux qui maximisent la log-quasi-vraisemblance pénalisée de Green (1987) :

$$-\frac{1}{2}\sum_{i=1}^{n} d_i(y_i, \mu_{\tilde{u},i}) - \frac{1}{2}u^T\Sigma_\theta^{-1}u \tag{4.7}$$

Soit alors $(\hat{\beta}, \hat{u})$ avec $\hat{u} = \tilde{u}(\hat{\beta})$ le couple maximisant cette log-quasi-vraisemblance pénalisée.

En dérivant (4.7) par rapport à β et u nous obtenons les équations suivantes :

$$\frac{(y_i - \mu_{u,i})x_i}{a_i(\kappa)v(\mu_{u,i})g'(\mu_{u,i})} = 0 \tag{4.8}$$

et

$$\frac{(y_i - \mu_{u,i})z_i}{a_i(\kappa)v(\mu_{u,i})g'(\mu_{u,i})} = \Sigma_\theta^{-1}u \tag{4.9}$$

où z_i^T est la $i^{\text{ème}}$ ligne de Z.

Les équations (4.8) et (4.9) se résolvent de manière itérative avec des algorithmes tels que celui de Ficher ou Newton, lesquels sont équivalents dans le cas d'une fonction de lien canonique.

Nous obtenons donc le système itératif suivant :

$$\begin{pmatrix} Z^TW_u^{-1}Z + \Sigma_\theta^{-1} & Z^TW_u^{-1}X \\ X^TW_u^{-1}Z & X^TW_u^{-1}X \end{pmatrix} \quad \begin{pmatrix} u \\ \beta \end{pmatrix} = \begin{pmatrix} Z^TW_u^{-1}y^* \\ X^TW_u^{-1}y^* \end{pmatrix} \tag{4.10}$$

où y^* est la variable dépendante définie par $y^* = X\beta + ZU + (y - \mu_u)g'(\mu_u)$

Ce système correspond au système résolu itérativement de Henderson dans le modèle linéaire suivant

$$Y^* = X\beta + Zu + e \tag{4.11}$$

où

$$\mathrm{E}(Y^*|U = u) = X\beta + Zu$$

et

$$\begin{aligned} \mathrm{Var}(Y^*|U = u) &= var(e|U = u) \\ &= \mathrm{var}\big((y - \mu_u)g'(\mu_u)|U = u\big) \\ &= \mathrm{var}(\epsilon g'(\mu_u)|U = u) \\ &= \mathrm{diag}\{g'(\mu_{u,i})^2\mathrm{Var}(\epsilon_i|u)\}_{i=1,\ldots,n} \\ &= W_u. \end{aligned}$$

Dans le cas d'un lien canonique, W_u n'est autre que $\mathrm{diag}\{\frac{a_i(\kappa)^2}{\mathrm{Var}(Y_i|U=u)}\}_{i=1,\ldots,n}$.

4.2 Etape d'estimation

4.2.1 Présentation générale

Dans le modèle linéaire $Y^* = X\beta + Zu + e$, on adopte alors la structure d'un L2M où :

$$\begin{aligned} \mathrm{E}(Y^*) &= X\beta \\ \mathrm{Var}(Y^*) &= Z\Sigma_\theta Z^T + W_u = \Gamma_u \end{aligned}$$

La matrice de variance des erreurs de ce modèle linéaire mixte est donc W_u.

La résolution du système (4.10) est alors équivalente (Harville, 1977) à d'abord résoudre en β

$$(X^T V^{-1} X)\beta = X^T V^{-1} Y^* \tag{4.12}$$

où $V = W^{-1} + Z\Sigma_\theta Z^T$ et alors

$$\hat{u} = \Sigma_\theta Z^T V^{-1}(Y^* - X\hat{\beta}) \tag{4.13}$$

Cela suggère qu'on peut prendre comme estimation de la variance de $\hat{\beta}$ la matrice

$$(X^T V^{-1} X)^{-1}.$$

Afin de déterminer les composantes de la variance, Breslow and Clayton (1993) remplacent la quasidéviance $\sum_i d_i(y_i, \mu_{\tilde{u},i})$ de l'expression (4.1) par le χ^2 de Pearson généralisé
$\sum_i (y_i - \mu_{\tilde{u},i})2/a_i v(\mu_{\tilde{u},i})$ dans (4.6). Après simplification (Harville, 1977), cela mène à la log-quasi-vraisemblance pénalisée profilée de θ ce qui correspond à la log-vraisemblance profilée du modèle linéaire (4.11),

$$l_p(\hat{\beta}(\theta), \theta) = -\frac{1}{2}\log|V| - \frac{1}{2}(Y^* - X\hat{\beta})^T V^{-1}(Y^* - X\hat{\beta}) \tag{4.14}$$

Afin de prendre en compte la perte du nombre de degrés de libertés du fait de l'utilisation de $\hat{\beta}$ plutôt que β dans la forme quadratique (4.14), la version REML (Patterson and Thompson, 1971) ajoute un terme tel que :

$$l_{REML}(\hat{\beta}(\theta), \theta) = -\frac{1}{2}\log|V| - \frac{1}{2}\log|X^T V^{-1} X| - \frac{1}{2}(Y^* - X\hat{\beta})^T V^{-1}(Y^* - X\hat{\beta}) \tag{4.15}$$

La plupart des implémentations de la méthode PQL dans les logiciels utilise cette approximation.

En dérivant alors (4.15) par rapport aux composantes de $\theta = (\sigma_1, \ldots, \sigma_K)$ on obtient les équations pour les paramètres de la variance :

$$-\frac{1}{2}\Big[(Y^* - X\hat{\beta})^T V^{-1}\frac{\partial V}{\partial \sigma_j}V^{-1}(Y^* - X\hat{\beta}) - \mathrm{tr}\Big(P\frac{\partial V}{\partial \sigma_j}\Big)\Big] = 0 \tag{4.16}$$

où $P = V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}$.

La matrice de Fisher F correspondant a pour composantes (Breslow and Clayton, 1993; Harville, 1977)

$$F_{jk} = -\frac{1}{2}\mathrm{tr}\Big(P\frac{\partial V}{\partial \sigma_j}P\frac{\partial V}{\partial \sigma_k}\Big)$$

et donc une estimation de la variance de $\hat{\theta}$ est F^{-1}.

4.2.2 Utilisation du logiciel R

L'utilisation de la méthode PQL sous R se fait à l'aide de la fonction **glmmPQL** du package MASS (Venables, 2002). Cette fonction est de la forme

$$\mathrm{glmmPQL}(\mathrm{fixed}, \mathrm{random}, \mathrm{family}, \mathrm{data}, \mathrm{correlation}, \mathrm{weights}, \mathrm{control}, \mathrm{niter} = 10, \mathrm{verbose} = \mathrm{TRUE}, \ldots)$$

Les arguments 'fixed', 'random', 'family' et 'data' sont obligatoires et représentent respectivement les variables à effets fixes, les variables à effets aléatoires, la famille de GLM utilisée et les données du modèle. Les autres arguments sont optionnels et les principaux 'correlation', 'weights', 'control', 'niter' et 'verbose' représentent respectivement la structure de la matrice de corrélation des effets aléatoires, les poids ω_i associés à l'observation i définis dans la fonction de densité des GLM 2.1, des options de contrôle sur le LMM 4.11, le nombre d'itérations et un argument logique indiquant s'il y a impression ou non des différentes itérations. D'autres arguments liés à l'utilisation du LMM 4.11 peuvent être utilisés (voir pour cela les arguments de la fonction **lme** de R).

4.3 Différentes approches et discussions

Des approches différentes de celles de Breslow and Clayton (1993) justifient la méthode PQL. McGilchrist (1994) suggére la méthode PQL en arguant que la log-

quasivraisemblance conditionnelle doit avoir un maximum bien défini et peut alors être approximée par une log-vraisemblance de loi normale. Schall (1991) également approuve l'approch PQL, aussi cette approche est souvent appellée "'approche de Schall"' au lieu de PQL. Tous deux, Schall (1991) et McGilchrist (1994) affirment que la méthode PQL est robuste à la non-spécification de la loi des effets aléatoires, et peut alors être appliquée même quand l'hypothèse de normalité est invalide. Engel and Keen (1994) appelle cette approach "'IRREML"' (iterative reweighted REML), laquelle correspond à la macro GLMM du logiciel Genstat du même nom.Wolfinger (1993) prefera utiliser le nom de "'pseudo-vraisemblance"' (PL) ou "'pseudo-vraisemblance restreinte"' (REPL) quand l'ajustement du type "'REML"' est appliqué. Le terme "'Pseudo-vraisemblance"', dans le sens de Carroll et al. (1988), semble être plus approprié que "'quasi-vraisemblance pénalisée"'. En effet, il reflète l'hypothèse de normalité implicite de la variable dépendante y^*. Quant à l'article de Wolfinger (1993), il est associé à la macro SAS `glmmix` qui est une des implémentations de PQL les plus utilisées.

D'autres auteurs quant à eux sont moins affirmatifs concernant les bénéfices de la méthode PQL. McCulloch (1997) énonce le fait que pour la variable dépendante

$$y_i^* = x_i^T\beta + z_i^T u + g'(\mu_{u,i})(y_i - \mu_{u,i})$$

l'approche PQL suppose que

$$Var(y_i^*) = z_i^T\Sigma_\theta z_i + g'(\mu_{u,i})^2 v(\mu_{u,i})$$

mais que cette hypoyhèse ignore le fait que $\mu_{u,i}$ est une fonction des effets aléatoires u, donc non constante. Engel and Keen (1994) suggèrent une modification de la méthode PQL afin d'autoriser la dépendance des poids $w_i = a_i v(\mu_{u,i})g'(\mu_{u,i})^2$ sur l'estimation des effets aléatoires. Ils suggèrent l'utilisation de l'espérance des poids $E(w_i^{-1})$, au lieu de w_i^{-1}, où l'espérance est prise par rapport à la distribution de u en utilisant une technique de bootstrap. Cependant, les études de simulations dans Engel and Buist (1998) montrent que cette technique n'obtient pas de bonnes performances.

4.4 Estimation et correction du biais

Les estimations issues de la méthode PQL peuvent souffrir de biais dans le cas de certains GL2M, en particulier pour les données binaires avec des groupes de petits effectifs, ou plus généralement, quand le nombre d'observations par effet aléatoire est petit (Rodriguez and Goldman, 2001). Aussi, certaines techniques de correction de biais ont été proposées telles que le CPQL, le PQL2 et la correction par bootstrap itéré.

4.4.1 PQL corrigé ou CPQL

Breslow and Lin (1995); Lin and Breslow (1996) utilisent un développement de Taylor de la vraisemblance autour de $\theta = 0$ afin de déduire les expressions des biais de la méthode PQL. Ils fournissent alors les corrections pour les estimations de β et θ. Par simplicité, les corrections exposées ici ne concernent que le cas de GL2M avec une composante de variance, comme dans Breslow and Lin (1995). Soient les observations y_{ij}, $i = 1, \ldots, q$, $j = 1, \ldots, n_i$ où $\sum_i n_i = n$. Le modèle s'écrit alors $g(\mu^u_{ij}) = x^T_{ij}\beta + u_i$ avec $\mu^u_{ij} = E(y_{ij}|u_i)$ et où $u_i \sim N(0, \sigma^2_i)$ et $var(y_{ij}|u_i) = a_{ij}(\kappa)v(\mu^u_{ij})$.

Paramètres à effets fixes

Ils fournissent une correction de l'estimation de β en utilisant la densité jointe $f_{y,u} = f_{y|u}f_u$ plutôt que la densité marginale de y. La correction pour β a alors pour expression

$$\hat{\beta}_{cpql} = \hat{\beta}_{pql} - \frac{\sigma^2}{2}(X^T W^0 X)^{-1} X^T t^0$$

où $\hat{\beta}_{pql}$ est l'estimateur obtenu à l'aide de la méthode PQL, $\mu^0_{ij} = g^{-1}(x^T_{ij}\hat{\beta}_{pql})$, $W^0 = \text{diag}\{a_{ij}v(\mu^0_{ij})/\phi\}$, et t^0 est un $n\times 1$ vecteur avec pour éléments $a_{ij}(\kappa)v'(\mu^0_{ij})v(\mu^0_{ij})$, la puissance "'0"' indiquant que l'évaluation se fait en $\beta = \hat{\beta}_{pql}$ et $u_i = 0$, $i = 1, \ldots, q$. Cette correction est basée sur le développement de Taylor du premier ordre autour de $\sigma^2 = 0$ tandis que celle proposée par Lin and Breslow (1996) est basée sur un developpement de Taylor du second ordre.

Composantes de la variance

Breslow and Lin (1995) fournissent une correction de l'estimation de σ^2 afin de prendre en compte l'utilisation de la vraisemblance profilée (4.14) plutôt que la vraie vraisemblance. La correction pour σ^2 a alors pour expression

$$\hat{\sigma}^2_{cpql} = \frac{D}{B - C}\hat{\sigma}^2_{pql}$$

où $\hat{\sigma}^2_{pql}$ est l'estimation obtenue avec la méthode PQL et les quantités B, C et D données dans Breslow and Lin (1995) page 88 sont telles que

$$B = \sum_i \frac{l^{0(2)}_i}{2} - u^T X(X^T W^0 X)^{-1}\frac{u}{4}$$

$$C = \sum_i \frac{l^{0(4)}_i}{4} \quad \text{et} \quad D = \sum_i \frac{l^{0(2)2}_i}{2}$$

où $l_i^{0(k)}$ est la dérivée k$^{\text{ième}}$ de la log-vraisemblance évaluée en $\beta = \hat{\beta}_{pql}$ et $u_i = 0$, $i = 1, \ldots, q$

Lin et Breslow font des suggestions sur comment implémenter ces deux corrections dans la pratique. Il suggèrent d'appliquer tout d'abord la correction à σ^2, puis réestimer β, et enfin appliquer la correction à β. Pour les données où les composantes de la variance estimées par la méthode PQL sont plus grandes que un, ils recommandent d'ignorer la correction pour β.

Etant donné que les corrections établies utilisent une approximation autour de $\sigma^2 = 0$, l'application de ces corrections est limitée aux cas où la composante de la variance estimée est relativement petite. Pour des données binaires, par exemple, Lin and Breslow (1996), avancent que ces corrections sont satisfaisantes dans le cas où les composantes de la variance estimées sont inférieures à un. De plus, Lin and Breslow (1996) (section 5.2) notent que les corrections ne sont valables que lorsque Z, la matrice de design des effets aléatoires, est hautement creuse. La correction CPQL n'est en général pas disponible dans les logiciels statistiques, il revient donc à l'utilisateur de la programmer s'il veut pouvoir l'utiliser.

4.4.2 PQL2

Goldstein and Rasbash (1996) proposent une approximation du second ordre, nommée PQL2. Il s'agit d'une extension de la méthode de PQL mais en utilisant le développement de Taylor jusqu'au second ordre au lieu de s'arrêter au premier ordre. Comme précédemment pour la méthode PQL, soit $\eta_i^{(k-1)} = x_i^T \hat{\beta}^{(k-1)} + z_i^T \hat{u}^{(k-1)}$, où $\hat{\beta}^{(k-1)}$ et $\hat{u}^{(k-1)}$ sont les estimations courantes et $h = g^{-1}$ est la fonction de lien inverse. Un développement de Taylor du second ordre autour de $\eta_i = \eta_i^{(k-1)}$ donne

$$y_i \approx h\left(\eta_i^{(k-1)}\right) + h'\left(\eta_i^{(k-1)}\right)\left(\eta_i - \eta_i^{(k-1)}\right) + \frac{1}{2}h''\left(\eta_i^{(k-1)}\right)\left(\eta_i - \eta_i^{(k-1)}\right)^2 + e_i$$

qui, en négligeant les termes de second ordre impliquant β, devient

$$\begin{aligned} y_i \approx\ & h\left(\eta_i^{(k-1)}\right) + h'\left(\eta_i^{(k-1)}\right) x_i^T\left(\beta - \beta^{(k-1)}\right) + h'\left(\eta_i^{(k-1)}\right) z_i^T\left(u - u^{(k-1)}\right) \\ & + \frac{1}{2}h''\left(\eta_i^{(k-1)}\right) z_i^T\left(u - u^{(k-1)}\right)\left(u - u^{(k-1)}\right)^T z_i + e_i. \end{aligned}$$

Cette expression est de nouveau simplifiée en remplaçant le dernier terme par son espérance, et ainsi une variable dépendante modifiée peut être formée :

$$\begin{aligned} \psi_i^{(k)} =\ & \left(x_i^T \beta^{(k-1)} + z_i^T u^{(k-1)}\right) + g'\left(\mu_i^{u,(k-1)}\right)\left(y_i - \mu_i^{u,(k-1)}\right) \\ & - \frac{1}{2} g'(\mu_i^{u,(k-1)}) h''\left(\eta_i^{(k-1)}\right) z_i^T \mathrm{Var}\left(u - u^{(k-1)}\right) z_i. \end{aligned}$$

Comme pour la méthode PQL, la variable dépendante modifiée, $\psi^{(k)} = (\psi_1^{(k)}, \dots \psi_n^{(k)})^T$ peut alors être ajustée par un L2M (4.11) afin de générer les nouvelles estimations $\hat{\beta}^{(k)}$ et $\hat{u}^{(k)}$. La même procédure itérative que celle de la méthode PQL est utilisée, la seule différence réside dans la formation de la variable dépendante, comme décrit ci-dessus. La méthode PQL2 est implémentée dans le logiciel statistique MLwiN (Goldstein et al., 1998).

4.4.3 La méthode de bootstrap itéré

Kuk (1995) propose une approche par simulation afin de corriger le biais. Cette approche est une approche statistique générale pour corriger les biais d'estimation et pas seulement que pour la méthode PQL.

Soit $\varphi = (\beta^T, \theta^T)^T$. En utilisant l'estimateur PQL noté $\hat{\varphi}^{(0)}$, plusieurs jeux de données simulées sont alors générés à partir de la distribution $f_y(y; \hat{\varphi}^{(0)})$. L'estimation PQL est alors appliquée à chaque jeu de données simulées, avec pour estimateur PQL moyen $\tilde{\varphi}^{(0)}$ et biais estimé $b^{(0)} = \tilde{\varphi}^{(0)} - \hat{\varphi}^{(0)}$, ce qui donne alors un estimateur révisé $\hat{\varphi}^{(1)} = \hat{\varphi}^{(0)} - b^{(0)}$. De nouveau plusieurs jeux de données sont alors générés à partir de la distribution $f_y(y; \hat{\varphi}^{(1)})$, l'estimation PQL est alors appliquée à chaque jeu de données simulées, avec pour estimateur PQL moyen $\tilde{\varphi}^{(1)}$ et biais estimé $b^{(1)} = \tilde{\varphi}^{(1)} - \hat{\varphi}^{(1)}$, ce qui donne alors un nouvel estimateur révisé $\hat{\varphi}^{(2)} = \hat{\varphi}^{(0)} - b^{(1)}$. Ce processus est répété jusqu'à convergence du biais estimé, c'est à dire que $b^{(k)} = b^{(k-1)}$ pour une itération k donnée (avec une tolérance convenable). De manière équivalente, le processus est répété jusqu'à ce que l'estimateur PQL moyen pour une itération k donnée, $\tilde{\varphi}^{(k)}$, soit égal à l'estimateur PQL d'origine, $\hat{\varphi}^{(0)}$ (toujours avec une tolérance convenable).

Kuk (1995) montre qu'une fois la convergence atteinte, cette méthode donne des estimateurs consistants. Cependant, la charge informatique de cette approche peut être énorme, comme l'ont souligné Rodriguez and Goldman (2001). Goldstein and Rasbash (1996) discutent du calcul des erreurs standards et des tests d'hypothèse en utilisant cette technique, qui est d'ailleurs implémentée dans le logiciel MLwiN Goldstein et al. (1998).

4.5 Comparaison aux méthodes de Laplace et AGH

Nous utilisons ici les deux mêmes jeux de données que ceux analysés en partie 3.5. Nous les analyserons à l'aide de la méthode PQL et comparerons les résultats obtenus d'une part avec la méthode de Laplace et d'autre part avec la méthode AGH15.

4.5.1 Analyse jeu de données 1

La table 4.1 présente, pour le jeu de données 1, les estimations moyennnes, les biais moyens, les erreurs standards des paramètres à effets fixes et l'erreur quadratique moyenne pour chaque méthode. Tout d'abord, nous pouvons voir que la méthode PQL a convergé 64 fois, ce qui est mieux que la méthode AGH15 qui, pour rappel, a convergé 58 fois, mais moins bien que la méthode de Laplace qui elle a convergé 73 fois.

Concernant les estimations moyennes, nous observons que, tout comme pour les méthodes de Laplace et AGH15, le taux de sous-estimation est assez élevé mais que les estimations moyennes sont très proches des vraies valeurs des paramètres, et ce particulièrement en ce qui concerne les paramètres des effets fixes.

Concernant les erreurs standards des variables à effets fixes, nous observons que, comme pour les méthodes de Laplace et AGH15, les estimations moyennes dans la plupart des cas sont très proches pour chaque méthode et restent inférieures à un. On remarque également que les estimations sont plus dispersées en moyenne pour la méthode de PQL que pour les méthodes de Laplace et AGH15. Des calculs supplémentaires de la table 4.1 montrent que l'étendue des estimations des erreurs standards en moyenne est de 0.5774 pour la méthode PQL. Pour rappel cette étendue était de 0.3777 pour la méthode de Laplace et de 0.4184 pour la méthode AGH15. Ainsi, la méthode PQl est celle qui fournit les estimations des erreurs standards les plus dispersées, la méthode de Laplace les moins dipersées et la méthode AGH15 des estimations intermédiairement dispersées. Globalement, les trois méthodes fournissent des estimations moyennes des erreurs standards quasiment identiques.

Concernant les biais, la méthode PQL, fournit en moyenne des biais légèrement moins élevés pour les paramètres à effets fixes que la méthode AGH15 et Laplace, qui rappelons fournissaient en moyenne des estimations des biais quasiment identiques. En moyenne, le biais des paramètres à effets fixes vaut 0.001 pour la méthode PQL, 0.009 pour la méthode de Laplace et 0.007 pour la méthode AGH15 (en valeur absolue). Par contre, concernant le biais de l'écart-type de la variable à effets aléatoires, la méthode PQL fournit une estimation moyenne plus élevée que la méthode AGH15 mais lègèrement moins élevée que la méthode de Laplace. En moyenne, le biais de l'écart-type de la variable à effets aléatoires vaut (en valeur absolue) 0.145 pour Laplace, 0.112 pour PQL, et 0.049 pour AGH15. Pour finir, les résultats montrent que la méthode PQL a un MSE relatif aux effets fixes très proche de celui de la méthode AGH15, de l'ordre de 10^{-3} et plus petit que celui de la méthode de Laplace qui est de l'ordre de 10^{-2}.

Pour conclure, nous pouvons dire que globalement les trois méthodes fournissent des estimations assez précises et semblables concernant les paramètres à effets fixes. Cependant, concernant l'écart-type de la variable à effets aléatoires la méthode AGH15 semble en donner une estimations plus précise.

PQL (64)						
		Estim.	Bias	Standard Errors		
Parameter	TV	mean	mean	min	mean	max
β_0	6,440	6,342	-0,099	0,843	1,121	1,735
β_{11}	-8,448	-8,469	-0,021	0,768	1,071	1,658
β_{12}	-6,681	-6,623	0,058	0,579	0,927	1,444
β_{13}	-5,625	-5,499	0,126	0,573	0,900	1,393
β_{14}	-4,372	-4,195	0,177	0,551	0,887	1,381
β_{15}	-3,338	-3,152	0,186	0,567	0,902	1,393
β_{21}	-1,220	-1,242	-0,022	0,418	0,589	1,000
β_{22}	-1,912	-1,965	-0,053	0,444	0,608	1,087
β_{23}	-1,659	-1,705	-0,045	0,442	0,606	1,017
β_{24}	-1,338	-1,369	-0,031	0,477	0,638	1,065
β_{25}	-2,233	-2,342	-0,109	0,522	0,739	1,128
β_{26}	-1,073	-1,056	0,017	0,504	0,710	1,119
β_{31}	-0,988	-0,969	0,018	0,239	0,309	0,430
β_{32}	-0,865	-0,882	-0,017	0,311	0,383	0,501
β_{33}	-0,757	-0,886	-0,129	0,448	0,588	0,797
β_4	-0,701	-0,749	-0,048	0,221	0,269	0,379
β_5	-0,561	-0,548	0,013	0,223	0,284	0,423
σ	0,684	0,572	-0,112			
		MSE: 7,87E-03				

TV: True Values; Estim.: Estimate; MSE: Mean Squared Error

Table 4.1: S1: Estimates, Biases, Standard Errors and Mean Squared Errors for PQL

4.5.2 Analyse jeu de données 2

La table 4.2 présente, pour le jeu de données 2, les estimations moyennes, les biais moyens, les erreurs standards des paramètres à effets fixes ainsi que l'erreur quadratique moyenne associés à chaque méthode. Tout d'abord, nous pouvons voir que la méthode PQL a convergé 100 fois, ce qui va dans le même sens que les

méthodes de Laplace et AGH15 pour ce jeu de données, en effet ces dernières ont convergé presque 100% de fois (99 fois pour Laplace et 96 fois pour AGH15). Nous notons également que ce taux de convergence est bien supérieur à celui obtenu avec le jeu de données 1.

Concernant les estimations moyennes, nous observons que, tout comme pour les méthodes de Laplace et AGH15, le taux de sous-estimation est élevé mais que les estimations moyennes sont très proches des vraies valeurs des paramètres, et ce particulièrement en ce qui concerne les paramètres des effets fixes.

Concernant les erreurs standards des variables à effets fixes, nous observons que, comme pour les méthodes de Laplace et AGH15, les estimations moyennes dans la plupart des cas sont très proches pour chaque méthode et restent inférieures à un. On remarque également que contrairement au jeu de données 1, les estimations les plus dispersées en moyenne ne sont plus obtenues par la méthode PQL mais par la méthode de Laplace. Néanmoins, pour ce jeu de données les étendues des estimations des erreurs standards sont relativement proches pour chacune des méthodes. Des calculs supplémentaires de la table 4.2 montrent que l'étendue des estimations des erreurs standards en moyenne est de 0.5817 pour la méthode de Laplace, de 0.5692 pour la méthode PQL et de 0.5616 pour la méthode AGH15. Globalement, les trois méthodes fournissent des estimations moyennes des erreurs standards quasiment identiques.

Concernant les biais, contrairement au jeu de données 1, ce n'est plus la méthode PQL qui fournit en moyenne les biais les moins élevés pour les paramètres à effets fixes mais la méthode de Laplace. En moyenne, le biais des paramètres à effets fixes vaut (en valeur absolue) 0.039 pour la méthode de Laplace, 0.069 pour la méthode PQL et 0.091 pour la méthode AGH15. Ces biais sont plus élevés que pour le cas du jeu de données 1. Par contre, concernant le biais de l'écart-type de la variable à effets aléatoires, la méthode de Laplace fournit une estimation moyenne plus élevée que la méthode AGH15 et PQL. La méthode AGH15 est celle qui fournit les plus petits biais, suivie de la méthode PQL. En moyenne, le biais de l'écart-type de la variable à effets aléatoires vaut (en valeur absolue) 0.161 pour AGH15, 0.190 pour PQL et 0.252 pour Laplace. Pour finir, les résultats montrent que la méthode PQL a un MSE relatif aux effets fixes très proche de celui de la méthode AGH15, de l'ordre de 10^{-2} mais plus grand que celui de la méthode de Laplace qui est de l'orde de 10^{-3}.

Pour conclure, nous pouvons dire que concernant les paramètres à effets fixes, globalement les trois méthodes fournissent des estimations assez précises avec une

légère supériorité pour la méthode de Laplace. Cependant, concernant l'écart-type de la variable à effets aléatoires la méthode AGH15 semble donner l'estimation la plus précise.

Nous voyons ainsi que pour les deux jeux de données les trois méthodes fournissent des estimations similaires concernant les effets fixes et la méthode AGH15 donne des estimations légèrement plus précises concernant l'écart-type de l'effet aléatoire. La question est alors de savoir si, avec des tailles d'échantillons plus grandes mais également avec un plus grand nombre de réplications, ces conclusions sont conservées. Aussi, au chapitre 5 nous traiterons de cette question.

		PQL (100)				
		Estim.	**Bias**	**Standard Errors**		
Parameter	**TV**	**mean**	**mean**	**min**	**mean**	**max**
β_0	2,748	2,964	0,216	0,642	1,203	2,274
β_{11}	-0,897	-0,895	0,002	0,391	0,568	0,963
β_{12}	-2,119	-2,295	-0,176	0,377	0,566	1,524
β_{13}	-1,829	-1,906	-0,077	0,240	0,319	0,528
β_{14}	-1,727	-1,800	-0,073	0,244	0,312	0,496
β_{15}	-0,605	-0,635	-0,030	0,248	0,318	0,516
β_{21}	-1,165	-1,283	-0,118	0,414	0,573	1,139
β_{22}	-1,811	-1,942	-0,131	0,411	0,580	1,139
β_{23}	-1,599	-1,694	-0,095	0,414	0,583	1,141
β_{24}	-1,266	-1,418	-0,152	0,441	0,610	1,157
β_{25}	-2,169	-2,331	-0,162	0,480	0,670	1,164
β_{26}	-1,051	-1,189	-0,138	0,495	0,675	1,192
β_{31}	-1,000	-1,037	-0,037	0,222	0,277	0,418
β_{32}	-0,885	-0,944	-0,059	0,261	0,339	0,489
β_{33}	-0,723	-0,819	-0,096	0,379	0,500	0,795
β_4	-0,750	-0,759	-0,009	0,174	0,223	0,339
β_5	-0,589	-0,631	-0,042	0,184	0,236	0,419
σ	2,620	2,429	-0,190			
		MSE: 1,26E-02				

TV: True Values; Estim.: Estimate; MSE: Mean Squared Error

Table 4.2: S2: Estimates, Biases, Standard Errors and Mean Squared Errors for PQL

4.6 Avantages et inconvénients des méthodes

	Avantages	Inconvénients	Packages statistiques
GQ	• Approxime très précisément l'intégral quand un grand nombre de points de quadrature ($\geq$ 20) est utilisé (Lesaffre and Spiessens, 2001). • Les résultats obtenus avec les autres méthodes sont souvent comparés à ceux de la méthode GQ • La déviance peut être calculée ; l'adéquation du modèle peut être évaluée (Hedeker et al., 1994).	• L'utilisation d'un petit nombre de points de quadrature ($<$ 10) peut causer la convergence vers un minimum local (Lesaffre and Spiessens, 2001). • L'augmentation du nombre de points de quadrature fait augmenter rapidement le temps de calcul et l'intégration numérique devient très gourmande en calculs (Lesaffre and Spiessens, 2001; Hedeker et al., 1994). • La précision diminue à mesure que la corrélation entre les effets aléatoires augmente et que la valeur des paramètres à effets fixes augmente (González et al., 2006)	• SuperMix • SAS NLMIXED (method=gauss)
AGQ	• Les points de quadrature sont choisis optimalement afin de couvrir au mieux l'aire de la densité de la fonction à intégrer (Hedeker et al., 1994) • Nécessite moins de points de quadrature, est donc plus rapide que GQ (Lesaffre and Spiessens, 2001) • La déviance peut être calculée ; l'adéquation du modèle peut être évaluée (Hedeker et al., 1994).	• Toujours très gourmande en calculs surtout pour des modèles avec $>$ 3 effets aléatoires corrélés.	• SuperMix • SAS NLMIXED (method=gauss) • SAS GLIMMIX (method=quad) • R package lme4 • Stata xtmelogit
Laplace	• Nécessite moins de calculs, est donc numériquement moins lourde que GQ et AGQ (Raudenbush et al., 2000) • Produit des estimations plus précises que celles de PQL (Raudenbush et al., 2000; Diaz, 2007). • Les estimations obtenues par Laplace6 sont au moins aussi précises que celles obtenues par GQ avec 20 points de quadrature et AGQ avec 7 points de quadrature (Raudenbush et al., 2000) • La déviance peut être calculée ; l'adéquation du modèle peut être évaluée.	• Potentiellement moins précise que GQ et AGQ car elle est basée sur le développement de la log-vraisemblance autour d'un seul point (le mode) (Clarkson and Zhan, 2002). • Laplace6 fournit des erreurs standards moyennes plus élevées que celles de PQL. (Diaz, 2007) • Tend à produire des estimations biaisées pour de petites tailles d'échantillon (petit nombre de groupes ou petites tailles de groupe) (Raudenbush et al., 2000; Clarkson and Zhan, 2002; Diaz, 2007; Joe, 2008) • La MSE augmente pour les paramètres à effets fixes et la variance des effets aléatoires quand la proportion d'un évènement approche 0 (Diaz, 2007)	• SAS GLIMMIX (method=laplace) • R package lme4 (nAGQ=1) • HLM6 • Stata xtmelogit (Laplace)
PQL	• Converge facilement et numériquement faisable. • Estimation REML pour la matrix de covariance des effets aléatoires (Breslow and Clayton, 1993; Wolfinger, 1993)	• Produit des estimations très biaisées pour les paramètres à effets fixes et les variances des effets aléatoires quand la (les) variance(s) des effets aléatoires est (sont) larges et/ou quand la proportion d'un évènement est proche de 0 ou 1 (Breslow and Clayton, 1993; Zhou et al., 1999; Diaz, 2007) • La déviance est indisponible ; l'adéquation du modèle ne peut pas être évaluée (Hedeker et al., 1994)	• SAS GLIMMIX (method=rspl) • R package MASS • HLM6

GQ : Gauss-Hermite quadrature ; AGQ : Adaptive Gauss-Hermite quadrature ; PQL : Penalized quasi-likelihood

Table 4.3: Résumé des avantages et inconvénients de chaque méthode

Ce résumé des différentes méthodes a été établi par Yoonsang et al. (2013).

Chapter 5

Modélisation de la gravité corporelle maximum

5.1 Description des données utilisées

5.1.1 Base de données EDA

Parmi les différentes bases de données du LAB, le choix a été fait de travailler sur la base EDA. En effet, cette base étant la plus récente des bases du LAB, ce choix permettait notamment de voir quels étaient ses points forts et ses points faibles. Cette base contient 1145 accidents corporels impliquant 1722 véhicules de 1991 à 2009. Des informations concernant la configuration de l'accident, la localisation de l'accident, le conducteur, les caractéristiques de l'impact et les blessures des usagers sont collectées en temps réel par une équipe de spécialistes suivant une méthode bien précise établie par l'INRETS (Ferrandez, 1995) et ce sur plusieurs zones. De 1991 à 2004 dans la région d'Evreux (27) et Amiens (80), puis de 2005 à aujourd'hui dans la région de Bondoufle (91). Ces zones ont été sélectionnées en raison de leur représentativité en termes de caractéristiques d'accident (type de collision, localisation, gravité etc.)

Une grande partie des données d'accidents sont recueillies sur la scène de l'accident en temps réel (entretiens avec les impliqués, photographies des lieux et des véhicules, repérages et mesures des traces sur la chaussée, inspections des véhicules, ...). Par ailleurs, une enquête rétrospective est menée afin de confirmer les informations détaillées de l'accident mais également d'en recueillir d'autres. Les informations complémentaires sont obtenues par des entretiens avec les impliqués à l'hôpital ou à leur domicile, par l'inspection des véhicules chez les garagistes ou épavistes, et par des recherches au Conseil Général et à la DDE pour ce qui concerne l'environnement, le trafic et l'infrastructure. Le procès-verbal, dans la mesure du possible, est con-

sulté au commissariat de police, à la brigade de gendarmerie, ou à la compagnie de CRS . Les bilans médicaux, établissant les lésions des impliqués lors de l'accident, sont obtenus auprès des services hospitaliers, avec l'accord des impliqués.

Chaque accident est ainsi décrit par environ 1000 variables. Certaines variables donnent des informations sur les généralités de l'accident (localisation, l'heure, la météo, la luminosité, etc.), d'autres sur les véhicules (caractéristiques des conducteurs, caractéristiques des véhicules, etc.) et sur certains usagers de la route (l'âge, le sexe, la position dans le véhicule, blessures, etc.). Les variables de codage sont soit des variables descriptives (localisation de l'accident), soit des variables mesurées, calculées/estimées par le calcul (la vitesse d'impact), soit des variables d'analyse (choix de l'événement initiateur de l'accident), ou des variables de jugement/expertise (responsabilité du conducteur), la plupart d'entre elles étant des variables catégorielles.

Parmi ce grand nombre de variables, certaines présentent un fort taux de valeurs inconnues. Cela peut être dû à l'évolution des moyens utilisés pour recueillir ces données, à l'évolution des besoins à partir des études et analyses réalisées mais également à la difficulté d'obtenir certaines informations, en particulier celles classées confidentielles. Cela implique des informations parfois redondantes (même information mais sur des niveaux de détails différents) correspondant à des besoins différents.

5.1.2 Variable à expliquer

Dans ce travail, nous cherchons à modéliser la gravité des blessures lors des accidents automobiles. Comme décrit au chapitre un, les trois critères de gravité présents dans nos bases de données sont l'AIS (Abbreviated Injury Scale), le MAIS qui est le maximum des AIS chez un usager, l'ISS (Injury Severity Score) et la gravité au sens de l'Onisr (se reporter au chapitre 1 pour les différentes définitions).

L'échelle AIS a été développée afin de fournir un système de standardisation visant à classer les catégories de blessures et gravité (Ceesar et al., 2004). Comme vu au chapitre un cette échelle repose sur des règles précises visant à coder les lésions une par une. En ce sens elle s'avère être plus pertinente et rigoureuse que le codage de la gravité au sens de l'ONISR. De plus, l'AIS étant universellement reconnue, des comparaisons avec des études autres que françaises sont alors possibles.

Egalement l'Injury Severity Score (ISS) qui découle de l'AIS est un bien meilleur indicateur de la gravité du traumatisme dans son ensemble que le MAIS. En effet, le score a été démontré comme corrélé à la mortalité, la morbidité et la durée

d'hospitalisation (Baker et al., 1974; Bull, 1975).

Aussi, dans l'idéal, le choix du critère de gravité à modéliser devrait se porter sur l'ISS. Cependant, la base de données EDA utilisée ne contient pas le critère ISS. En effet, le souci rencontré est que pour bon nombre de cas, les différentes régions corporelles n'étaient pas spécifiées ou elles étaient erronées. Ainsi, pour pouvoir retrouver le détail des lésions pour chaque accidenté il aurait fallu retourner aux dossiers papiers. Cette alternative étant beaucoup trop longue et fastidieuse il a alors été décidé de modéliser la gravité par l'intermédiaire du MAIS.

Afin de s'assurer que le MAIS codé dans notre base de données est bien cohérent avec l'état de gravité de l'usager, nous allons les croiser avec la durée d'hospitalisation qui elle est une donnée fiable,

où MAIS6+ contient :

- MAIS 6 : Tué avec lésions connues ou maximales
- MAIS 7 : Tué non autopsié
- MAIS 8 : Tué incarcéré
- MAIS 10 : Tué (suspicion de malaise cardiaque ou décès indirectement lié à des lésions majeures)

D'après la figure 5.1 nous voyons que la répartition des différents groupes MAIS semble bien cohérente avec la durée d'hospitalisation. De plus, nous constatons que les décès font leurs apparitions à partir du niveau MAIS 3. Ce constat est corroboré par la littérature lésionnelle (Ceesar et al., 2004), à savoir que le pronostiv vital est engagé à partir du MAIS3. Nous décidons donc dans la suite de notre étude de modéliser la gravité des accidents à l'aide de cette variable MAIS et tout blessé dont le MAIS sera supérieur ou égale à 3 sera dit critique.

5.1.3 Variables et échantillon

Notre objectif est de modéliser la probabilité qu'un véhicule léger impliqué dans un accident corporel contienne au moins un blessé critique, c'est à dire avec un MAIS supérieur ou égale à 3 (MAIS3+). On définit donc la variable à modéliser comme suit. Pour un véhicule i :

$$y_i = \begin{cases} 1 & \text{si } \max\{\text{MAIS des passagers du véhicule i}\} \geq 3, \\ 0 & \text{si } \max\{\text{MAIS des passagers du véhicule i}\} < 3. \end{cases}$$

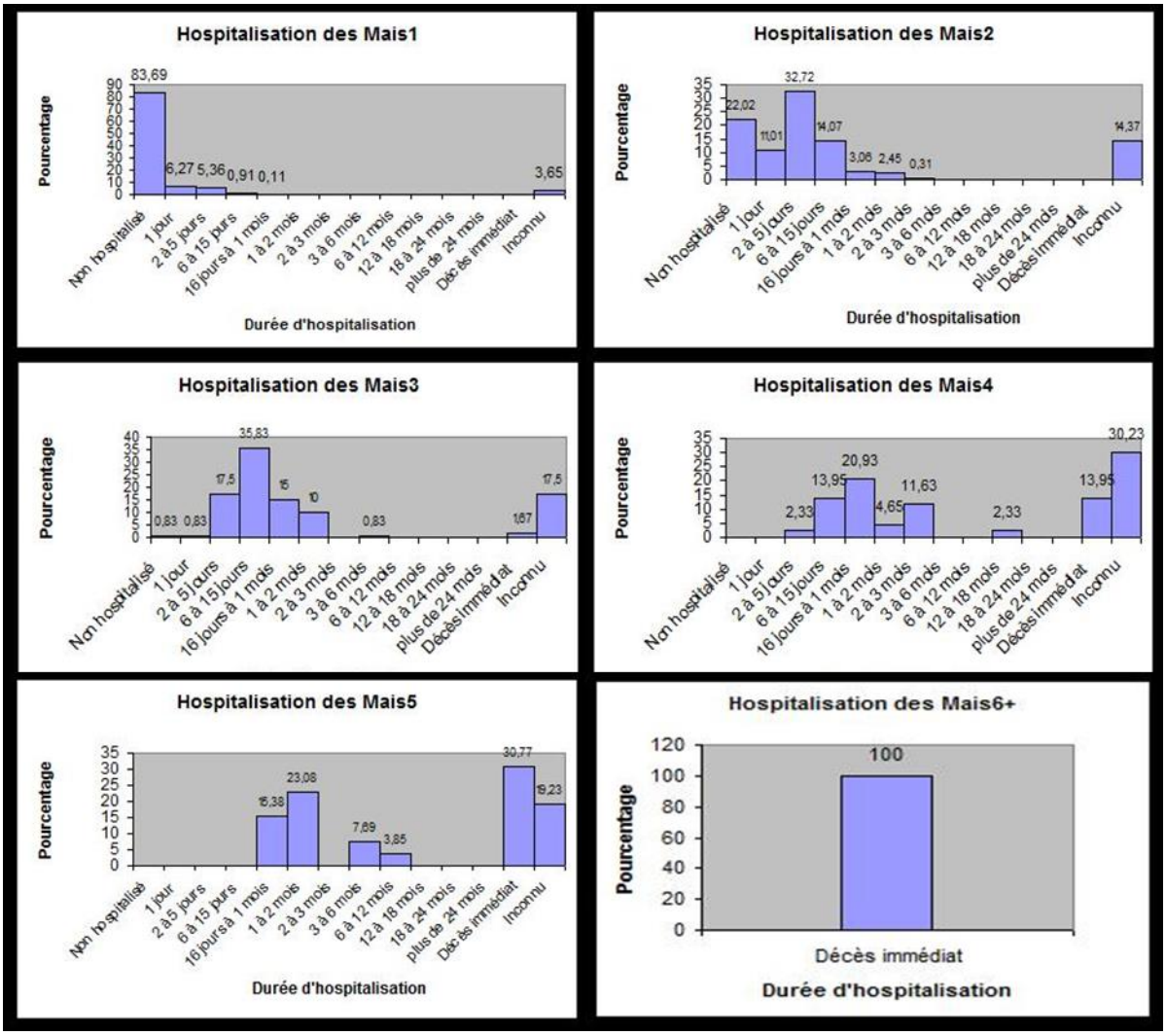

Figure 5.1: Durée d'hospitalisation en fonction des différents MAIS

Dans la suite nous nommerons cette variable 'MAIS3+'.

La base de données EDA contient 200 variables. Tout d'abord, toutes les variables avec un taux de non remplissage supérieur à 30% ont été omises. Cela concerne 98 variables, il en reste alors 102. Une seconde étape a consisté à omettre toutes variables présentant des incohérences et/ou problèmes de codage. Leur nombre étant de 62, il reste alors 40 variables étudiables. Une troisème étape fut d'étudier les associations univariées entre la variable à expliquer MAIS3+ et chacune de ces 40 variables. Pour cela, nous avons utilisé des test du chi-2 et des coefficients de Cramer. Les variables dont la p-value du test du chi-2 était inférieure à 0.05 ou le coefficient de Cramer supérieur à 0.1 (quand le test du chi-2 n'était pas fiable) ont été retenues. Il en résulta alors 12 variables. Parmi elles, le type de collision, la localisation de l'accident, l'Energie de déformation (Energy Equivalent Speed) la situation pré-conflictuelle, le type d'obstacle rencontré etc. Si certaines variables présentaient de fortes corrélations entre elles celles ayant le coefficient de Cramer le plus élévé ont été conservées, ce qui aboutit à 7 variables. A cet ensemble de 7 variables nous ajoutâmes 5 autres variables que nous suspections d'avoir un pouvoir explicatif sur la gravité et ce même si ces dernières ont été rejetées lors de l'étude de liaison univariée. Parmi ces 5 variables, il y a l'âge et le sexe du conducteur ainsi que l'année de l'accident. Il en résulte alors un total de 12 variables. L'étape finale fut alors d'appliquer un modèle logistique classique aux variables retenues. Cependant, les 12 variables ne sont pas injectées dans le modèle en même temps car cela menait systématiquement à un problème d'hyper séparation. Aussi, nous firent plusieurs essais de modélisation avec 5, 6 ou 7 variables, en fonction du sous-ensemble de variables testées, de sorte de ne pas avoir d'hyper séparation. A chaque essai les variables estimées comme significatives à un niveau 0.05 (test de Wald ou variante) ont été gardées et celles estimées comme non significatives ont été omises. Les variables non-testées ont alors été rajoutées au fur et à mesure à l'ensemble des variables gardées et le processus de sélection à l'aide du test de Wald réitéré.

La significativité des paramètres associés aux variables explicatives est testée à l'aide de la statistique $(z_k)_{(k=1,\dots,p)}$ qui sous l'hypothèse $[H_0 : \hat{\beta}_k = 0]$ suit asymptotiquement une loi normale centrée réduite. Cette statistique est une variante de la statistique de Wald W_k qui sous l'hypothèse $[H_0]$ suit asymptotiquement une loi du χ^2 à un degré de liberté. On a $z_k = \frac{\hat{\beta}_k}{\hat{\sigma}_{\beta_k}} = \text{signe}(\hat{\beta}_k) \times \sqrt{W_k}$, $k = 0, \dots, p$, où $\hat{\sigma}_{\beta_k}$ est une estimation de l'écart-type (erreur standard) du coefficient $\hat{\beta}_k$. Si les paramètres ne sont pas trouvés significativement différents de zéro au seuil de 0.05, ils sont fixés à zéro. Les résultats du modèle logit obtenus avec le logiciel R sont exposés dans le tableau en partie. Les erreurs standards des paramètres estimés

sont les racines carrées des termes diagonaux de la matrice de variance-covariance des estimateurs qui est obtenue en inversant la matrice hessienne. Cette dernière a pour expression $X'VX$ où X est la matrice des variables explicatives de dimension $n \times (1+p)$ et V est une matrice diagonale de taille $n \times n$, composée des valeurs de $\hat{p}_j \times (1-\hat{p}_j)$, les probabilités estimées $\hat{p}_j$ étant obtenues après estimation des paramètres.

Au final, les variables trouvées significativement différentes de zéro et ainsi incorporées à la modélisation sont au nombre de six : l'EES (Energy Equivalent Speed), le type de collision, l'année de l'accident, sa localisation, l'âge du conducteur, son sexe ainsi que sa réaction.

L'EES (Energy Equivalent Speed)

L'EES est estimée par expertise en comparant visuellement les dégâts subis par le véhicule accidenté et ceux d'un véhicule similaire testé dans une typologie de choc semblable lors de crash-tests. Cette variable mesure donc l'énergie de déformation absorbée par le véhicule lors d'une collision. Par conséquent, l'EES est une variable importante pour décrire et expliquer la gravité des blessures d'un accident (Zeidler et al., 1985). Miltner and Salwender (1971) ont montré la relation entre les variables de collision, y compris l'EES et la gravité des blessures. L'EES est une variable continue et s'agissant d'une vitesse elle s'exprime en km/h. Dans notre échantillon, les valeurs de l'EES varient de 2 à 83 km/h et 85% de ces valeurs sont inférieures à 50 km/h (tableau 1). Pour une meilleure visibilité des résultats de la modélisation, cette variable a été recodée en six tranches de vitesse: 0 - 20 km/h, 20 km/h - 30 km/h, 30 km/h - 40km/h, 40 km/h - 50 km/h, 50 km/h - 60km/h et 60 km/h - 90km/h.

Le type de collision : COL

La variable COL prend en compte le choc du véhicule concerné mais aussi celui du véhicule opposé s'il s'agit d'un accident à deux véhicules. Cette variable dispose de 6 niveaux : le véhicule concerné heurte avec son avant un véhicule par l'arrière (F-A), le véhicule concerné est heurté au niveau de l'arrière par l'avant d'un autre véhicule (A-F),les deux véhicules se heurtent face à face (F-F), le véhicule concerné heurte avec son avant le côté d'un autre véhicule (F-L), le véhicule concerné est heurté sur le côté par l'avant d'un autre véhicule (L-F), les véhicules seuls heurtent un obstacle fixe (V-O).

L'année de l'accident : YEAR

L'année de l'accident est codée sur 4 catégoories : '1991-1994', '1995-1999', '2000-2004' et '2005-2010'

La localisation de l'accident : LOC
La variable LOC est codée sur 2 niveaux: En agglomération (Inag) et Hors agglomération (Hoag).

L'âge du conducteur : AGE
La variable AGE est divisée en 7 tranches d'années: '16-24 ', '25-34', '35-44 ', '45-54', '55-64 ', '65-74' et '75+'.

Le sexe du conducteur : GENDER
Enfin la variable GENDER dispose de 2 modalités : Masculin (M) et Féminin (F).

Concernant les observations, nous sélectionnons uniquement les accidents n'impliquant que des véhicules légers. Les observations avec des données manquantes parmi les variables utilisées dans la modélisation sont supprimées. La plus grande partie des données manquantes concerne le MAIS en raison de difficultés à obtenir dans certains cas le dossier médical. Il en résulte donc un échantillon de 826 véhicules.

La table 5.1 donne un aperçu de la distribution de chacune des variables explicatives dans l'échantillon mais également de la distribution des blessés MAIS3+ pour chacune de ces variables.

	Echantillon		Gravité MAIS3+	
	N	%	N	%
Total	826	100,00	162	19,61
Variables	N	%	N	%
EES				
0-20	268	32,5	4	1,5
20-30	196	23,7	17	8,7
30-40	148	17,9	25	16,9
40-50	112	13,6	41	36,6
50-60	63	7,6	37	58,7
60-90	39	4,7	38	97,4
COL				
A-F	32	3,9	3	9,4
F-A	45	5,5	2	4,4
F-F	205	24,8	58	28,3
F-L	204	24,7	15	7,4
L-F	196	23,7	46	23,5
V-O	144	17,4	38	26,4
LOC				
Inag	307	37,2	23	7,5
Hoag	519	62,8	139	26,8
AGE				
16-24	192	23,2	48	25,0
25-34	201	24,3	40	19,9
35-44	175	21,2	31	17,7
45-54	103	12,5	17	16,5
55-64	70	8,5	5	7,1
65-74	53	6,4	12	22,6
75+	32	3,9	9	28,1
YEAR				
1991-1994	156	18,9	57	36,5
1995-1999	463	56,1	75	16,2
2000-2004	159	19,3	21	13,2
2005-2010	48	5,8	9	18,8
GENDER				
F	242	29,3	37	15,3
M	584	70,7	125	21,4

Table 5.1: Distribution des variables dans l'échantillon

5.2 Modélisation mixte et sélection des effets aléatoires

5.2.1 Introduction

Traditionnellement, une variable comme le MAIS est naturellement modélisée par un modèle logistique binaire classique en prenant la totalité des six facteurs décrits ci-dessus comme co-variables. Cependant, le phénomène accidentologique étant de nature complexe et les causes diverses et variées, certains facteurs explicatifs peuvent avoir des effets hétérogènes non observés sur les mécanismes accidentels et lésionnels.

Par l'introduction de variables à effets aléatoires nous ne cherchons pas à évaluer l'effet de chacun des niveaux observés de ces facteurs (comme cela est le cas pour les facteurs à effets fixes) mais plutôt à estimer la variabilité due à ces facteurs. En effet, l'introduction de facteurs à effets aléatoires permet de prendre en compte des effets inobservés liés à ce facteur et qui influent de manières différentes sur le niveau de blessure (Milton et al., 2008; Kim et al., 2008; Malyshkina and Maneering, 2010; Kim et al., 2010). Par exemple, dans son analyse de la gravité des blessures des piétons, Kim et al. (2008) détectent la présence d'hétérogénéité et trouvent que celle-ci est captée par l'âge du piéton. Ce résultat s'explique par le fait que l'état de santé des piétons les plus âgés présente une plus grande variance que chez les jeunes piétons.

Aussi il convient, au cours de la construction du modèle, de pouvoir déterminer les variables explicatives qu'il faut inclure dans les composantes à effets fixes et celles qu'il faut affecter aux composantes à effets aléatoires. Une question difficile est donc de savoir comment décider lesquels des prédicteurs possèdent des coefficients qui varient significativement à travers les sujets.

5.2.2 Test de rapport de vraisemblance et sélection

Un grand nombre de travaux existe sur la sélection des paramètres à effets aléatoires (voir par exemple Kinney and Dunson, 2007; Ibrahim et al., 2011; Bondell et al., 2010, et les références incluses). Des auteurs comme Lee and Nelder (1996) ont proposé une méthode pour tester si certaines variables à effets aléatoires pouvaient être supprimées. D'autres par contre (Commenges and Jacqmin-Gadda, 1997; Lin and Breslow, 1996) ont proposé l'utilisation d'un test de score pour évaluer si une ou plusieurs variables à effets aléatoires peuvent être incluses dans un modèle. Dans un cadre bayésien, Chen et al. (2003) ont développé une ap-

proche stochastique plus générale pour la sélection de variables à effets aléatoires. Kinney and Dunson (2007) ont étendu cette approche bayésienne au cadre du modèle logistique mixte pour données binaires. Yang (2012) a proposé une approche bayésienne non paramétrique pour sélectionner simultanément les effets fixes et aléatoires. Cependant, dans la pratique quotidienne de l'appréciation de la gravité corporelle, on a besoin d'un critère de sélection simple, pragmatique et accessible à tout utilisateur. De ce fait, nous avons opté, dans ce travail, pour un test paramétrique du rapport de vraisemblance (voir par exemple Zhang and Lin, 2008; Self and Liang, 1987; Stram and Lee, 1994; Goldman and Whelan, 2000; Gao, 2004) pour évaluer si la variabilité de chacune des variables explicatives à travers les véhicules est significativement différente de zéro ou non.

Ce test de rapport de vraisemblance est alors utilisé afin de comparer le modèle avec effet aléatoire et celui sans effet aléatoire. Habituellement la statistique du rapport de vraisemblance suit asymptotiquement une loi du χ^2 à l degré de liberté (χ^2_l), l étant la différence de degrés de liberté entre les deux modèles testés. Cependant, dans notre cas le test suppose l'hypothèse nulle $\sigma^2 = 0$, où σ^2 est la variance de la variable testée, et place donc la variance sur la borne inférieure de son domaine de définition. Il en résulte que la statistique du test de vraisemblance habituellement utilisée ne suit pas asymptotiquement la traditionnelle distribution du χ^2_l qui est trop conservatrice (Zhang and Lin, 2008) mais un mélange de lois du χ^2_l et χ^2_{l-1} dans les cas les plus simples (Self and Liang, 1987; Stram and Lee, 1994; Goldman and Whelan, 2000). Pour un modèle n'impliquant qu'un seul paramètre de variance une approche suppose $0.5 \times \chi^2_1 + 0.5 \times \chi^2_0$ (Self and Liang, 1987).

Nous testons différents modèles en leur attribuant à chaque fois une variable à effets aléatoires différente. Sur les six variables testées, en prenant un risque de se tromper de 5%, trois sont trouvées comme ayant une variabilité significativement différente de zéro. Il s'agit de l'année de l'accident, le type de collision et l'EES. Les p-values des tests statistiques effectués sont listées dans la table 5.2. Nous obtenons alors trois modèles différents à un effet aléatoire. Le premier ayant pour effet aléatoire le type de collision, le second l'EES et le troisième l'année de l'accident. Nous nommons respectivement ces trois modèles 'modèle 1', 'modèle 2' et 'modèle 3' (voir table 5.3).

Variables à effets fixes	Variables à effets aléatoires	p-value du test
Age Col Year Loc Gender	EES	<0,001
EES Col Year Loc Gender	Age	0,176
EES Age Year Loc Gender	Col	<0,001
EES Age Col Loc Gender	Year	0,015
EES Age Col Year Gender	Loc	0,065
EES Age Col Year Loc	Gender	0,158

Table 5.2: Tests de sélection des effets aléatoires

De plus, nous considérons les modèles faisant intervenir toutes les combinaisons possibles à partir des trois variables à effets aléatoires sélectionnées précédemment. Nous obtenons alors, en plus des trois modèles à un effet aléatoire, trois modèles à deux effets aléatoires et un modèle à trois effets aléatoires. Nous nommons ces modèles 'modèle 4', 'modèle 5', 'modèle 6' et 'modèle 7' (voir table 5.3). En effet, le test de rapport de vraisemblance ne s'effectuant que sur des modèles emboîtés, c'est-à-dire avec même partie fixe ou même partie aléatoire, nous ne pouvons en toute rigueur utiliser ce test pour comparer les modèles 1, 2 et 3 avec les modèles 4, 5 et 6 et justifier du fait que les variances des deux variables sont significativement différentes de zéro simultanément. Il en est de même pour le modèle 7 à trois effets aléatoires. Aussi, lors de l'estimation des paramètres nous verrons si cette demarche est bien justifiée. Nous nommons également le modèle logit classique 'modèle 0'.

	Variables à effets fixes	Variable à effets aléatoires
Modèle 0	EES Age Col Year Loc Gender	Aucune
Modèle 1	EES Age Year Loc Gender	Col
Modèle 2	Age Col Year Lo cGender	EES
Modèle 3	EES Age Col Loc Gender	Year
Modèle 4	EES Age Loc Gender	Col Year
Modèle 5	Age Col Loc Gender	EES Year
Modèle 6	Age Year Loc Gender	Col EES
Modèle 7	Age Loc Gender	Col EES Year

Table 5.3: Les différents modèles

5.3 Modèles logistiques mixtes à un effet aléatoire

5.3.1 Présentation générale et notation

Nous considérons n unités (véhicules) et nous supposons qu'ils sont partitionnés en q groupes (ou niveaux). Les observations sur chaque véhicule consistent en une variable réponse binaire y_{ij} (0 ou 1), $i = 1, \ldots, q;\ j = 1, \ldots, n_i$ tel que $\sum_{i=1}^{q} n_i = n$, en un vecteur de variables $X_{ij}^T = (1, X_{ij1}^T, X_{ij3}^T, X_{ij4}^T, \ldots, X_{ijp-1}^T)$ associé aux effets fixes et en un niveau u_i de l'effet aléatoire $U = (u_i, \ldots, u_q)$. On note alors $Y = (Y_1^T, \ldots, Y_q^T)^T$ le vecteur de dimension n où $Y_i = (y_{i1}, \ldots, y_{in_i})^T$ est un vecteur de dimension n_i tel que $\sum_{i=1}^{q} n_i = n$.On suppose alors $U \sim N(0_q, \sigma^2 I_q)$.

Le modèle logistique-normal considéré est le suivant :

$$\begin{aligned} \text{logit}(P(y_{ij} = 1|X_{ij}, \beta, u_i)) &= \beta_0 + \sum_{m=1}^{p-1} X_{ijm}\beta_m + u_i \\ & \quad i = 1, \ldots, q;\ j = 1, \ldots, n_i \end{aligned} \tag{5.1}$$

où $\beta = (\beta_0, \beta_1^T, \ldots, \beta_{p-1}^T)$, β_m vecteur de dimension $p_m - 1$ $(p_m > 1)$ tel que $p = 1 + \sum_{m=1}^{p-1}(p_m - 1)$.

5.3.2 Modèles à un effet aléatoire

Modèle 1 : modèle avec effet aléatoire le type de collision 'COL'

Le Modèle 1 a pour effet aléatoire le type de collision et pour effets fixes l'EES, l'année de l'accident, sa localisation, l'âge du conducteur ainsi que son sexe.

La probabilité $P(y_{ij} = 1|u_i)$ qu'un véhicule accidenté j impliqué dans une collision de type i contienne au moins un blessé grave est reliée aux variables explicatives et aux effets aléatoires par le modèle suivant :

$$\text{logit}(P(y_{ij} = 1|u_i)) = (x_u)_{ij}\beta_u + u_i, \quad i = 1, \ldots, 6 \quad j = 1, \ldots, n_i \tag{5.2}$$

Où, $(x_u)_{ij}^T$ est une matrice $[1 \times (1+p)]_{p=16}$ des variables à effets fixes, β_u le vecteur de taille $(1+p)_{p=16}$ des coefficients associés aux effets fixes à estimer , u_i un scalaire désignant la $i^{ème}$ composante de l'effet aléatoire U avec $\sum_{i=1}^{6} n_i = n = 826$ (nombre total des véhicules analysés).

Modèle 2 : modèle avec effet aléatoire l'EES

Le Modèle 2 a pour effet aléatoire l'EES et pour effets fixes l'année de l'accident, sa localisation, le type de collision, l'âge du conducteur ainsi que son sexe.

La probabilité$P(y_{ij} = 1|v_i)$ qu'un véhicule accidenté j ayant un EES dans la tranche i contienne au moins un blessé grave est reliée aux variables explicatives et aux effets aléatoires par le modèle suivant :

$$\text{logit}(P(y_{ij} = 1|v_i)) = (x_v)_{ij}\beta_v + v_i, \quad i = 1, \ldots, 6 \quad j = 1, \ldots, n_i \tag{5.3}$$

Où, $(x_v)_{ij}^T$ est une matrice $[1 \times (1+p)]_{p=16}$ des variables à effets fixes, β_v le vecteur de taille $(1+p)_{p=16}$ des coefficients associés aux effets fixes à estimer , v_i un scalaire désignant la i^{ème} composante de l'effet aléatoire V, avec $\sum_{i=1}^{6} n_i = n = 826$.

Modèle 3 : modèle avec effet aléatoire l'année de l'accident 'YEAR'

Le Modèle 3 a pour effet aléatoire l'année d'accident et pour effets fixes l'EES, le type de collision, la localisation de l'accident, l'âge du conducteur ainsi que son sexe.

La probabilité$P(y_{ij} = 1|v_i)$ qu'un véhicule accidenté j durant l'année i contienne au moins un blessé grave est reliée aux variables explicatives et aux effets aléatoires par le modèle suivant :

$$\text{logit}(P(y_{ij} = 1|w_i)) = (x_w)_{ij}\beta_w + w_i, \quad i = 1, \ldots, 4 \quad j = 1, \ldots, n_i \tag{5.4}$$

Où, $(x_w)_{ij}^T$ est une matrice $[1 \times (1+p)]_{p=18}$ des variables à effets fixes, β_w le vecteur de taille $(1+p)_{p=18}$ des coefficients associés aux effets fixes à estimer , w_i un scalaire désignant la i^{ème} composante de l'effet aléatoire W, avec $\sum_{i=1}^{4} n_i = n = 826$.

5.3.3 Approximations de la vraisemblance et revue

Conditionnellement aux effets aléatoires, les observations y_{ij} sont supposées être indépendantes. Ainsi, la log-vraisemblance aux paramètres (β, σ) pour le modèle (5.1) est :

$$\text{logLik}(\text{Y}|\beta, \sigma) = \sum_{i}^{q} \log \int_{\mathbb{R}} \prod_{j=1}^{n_i} [\frac{e^{y_{ij}(\text{X}_{ij}^T\beta + \sigma \text{u}_i)}}{1 + e^{(\text{X}_{ij}^T\beta + \sigma \text{u}_i)}}] \, \Phi(u_i) \, \text{d}u_i \tag{5.5}$$

où $\Phi(.)$ est la function de densité d'une loi normal standard.

Dans certains cas importants, la vraisemblance peut être évaluée analytiquement par des procédés de maximisation tels que l'algorithme EM (Dempster et al., 1981),

des scores de Fisher(Goldstein, 1986; Longford, 1987), ou de Newton-Raphson (Lindstrom and Bates, 1988). Ici, il s'agit de la vraisemblance d'une loi non-normale et de fonction de lien qui n'est pas l'identité. Dans ce cas, l'estimation et l'inférence sont souvent difficiles du fait de la présence d'une intégrale impropre dans l'expression de la log-vraisemblance. Différentes méthodes ont été développées ces deux dernières décennies. Afin de déterminer les estimateurs du maximim de vraisemblance, l'expression 5.5 doit être évaluée et maximisée en ces paramètres ce qui implique l'évaluation de l'intégral. Une façon classique de calculer l'intégral est via les polynômes de Gauss-Hermite et la méthode de Laplace (voir Breslow and Clayton, 1993; Breslow and Lin, 1995; Raudenbush et al., 2000, et les références incluses)

D'autres auteurs ont étendu l'approche de manières différentes : Stiratelli et al. (1984) ont estimé les paramètres d'un modèle de régression logistique avec des effets aléatoires emboîtés et de lois normales en approximant la densité jointe à posteriori par une densité normal multivariée. Anderson and Aitkin (1985) ont appliqué la quadrature de Gauss-Hermite afin d'évaluer et maximiser la vraisemblance dans le cas d'un modèle de regression logistique avec un effet aléatoire par groupe d'observations. Hedeker et al. (1994) ont également appliqué la quadrature de gauss-Hermite afin d'évaluer l'intégral intervenant dans la vraisemblance dans le cas de modèles logistique et probit ordinal à l'aide d'une loi normale multivariée. Pinheiro and Bates (1995) quant à lui a utilisé la quadrature de Gauss-Hermite adaptative pour approximer l'estimation du maximum de vraisemblance dans le cas d'un modèle à effets aléatoires emboîtés avec des données normales et un lien non-linéaire.

Dans tous les cas et quelle que soit la manière utilisée pour obtenir les estimations du maximum de vraisemblance de 5.5, les trois principales méthodes sont : (a) l'inférence par quasi-vraisemblance pénalisée ; (b) l'approximation de Gauss-Hermite ; et (c) les méthodes de Monte Carlo. Dans ce travail, nous utilisons les deux premières méthodes afin d'estimer les paramètres et comparons les résultats obtenus à la méthode de Laplace. Les fonctions glmer et glmmPQL des packages lme4 et MASS du logiciel R (voir R Development Core Team, 2012) sont utilisées.

5.3.4 Méthodes d'estimation

5.3.4.1 Approximation de Laplace

L'approximation du maximum de vraisemblance via la méthode de Laplace a été très utilisée afin de trouver les distributions à posteriori (voir par exemple Kass et al., 1990) et d'approximer les fonctions de vraisemblance (Solomon and Cox,

1992; Breslow and Lin, 1995; Lin and Breslow, 1996). Dans son application standard, le logarithme naturel de l'intégrand est développé à l'aide d'une série de Taylor du second-ordre. Le fait que les termes de plus hauts degrés diminuent avec la taille de l'échantillon rend l'approximation précise pour des échantillons de grandes tailles (Raudenbush et al., 2000, et citations incluses). Nous utilisons ces méthodes ici pour la vraisemblance marginale de y_i dans l'expression (5.5). L'intégrale que nous voulons approximer par la méthode de Laplace s'écrit :

$$\int \Phi(u_i) \prod_{j=1}^{n_i} \frac{e^{y_{ij}(X_{ij}^T\beta+\sigma u_i)}}{1+e^{X_{ij}^T\beta+\sigma u_i}} \, \mathrm{d}u_i = \int \exp\{h(u_i,\theta;y_i)\} \, \mathrm{d}u_i \tag{5.6}$$

avec $\theta = (\beta^T, \sigma)^T$ et

$$h(u_i,\theta;y_i) = \text{constant} - \frac{u_i^2}{2} + \sum_{j=1}^{n_i} \left[y_{ij}\left(X_{ij}^T\beta + \sigma u_i\right) - \log\left(1+e^{X_{ij}^T\beta+\sigma u_i}\right)\right] \tag{5.7}$$

L'idée de base est d'approximer $h(u_i,\theta;y_i)$ par un développement de Taylor du second degré de la fonction $h(u_i,\theta;y_i)$ sachant θ au voisinage de $\hat{u}_i$. Nous supposons alors que $h(u_i,\theta;y_i)$ est une fonction lisse, bornée et unimodale par rapport à u_i. Ainsi, en remplaçant $h(u_i,\theta;y_i)$ par son développement de Taylor au second ordre (voir par exemple Tierny and Kadane, 1986; Raudenbush et al., 2000) la log-vraisemblance (5.5) peut être écrite telle que (voir l'appendix pour les détails)

$$\mathrm{l}(\beta,\sigma,u;y) \approx \sum_{i=1}^{q} \log\left(\sqrt{2\pi}\hat{\tau}_i(\theta)\exp\{h(\hat{u}_i,\theta;y_i)\}\right) \tag{5.8}$$

où $\hat{\tau}_i(\theta) = \left(1+\sigma^2\sum_{j=1}^{n_i}\frac{e^{X_{ij}^T\beta+\sigma\hat{u}_i}}{\left(1+e^{X_{ij}^T\beta+\sigma\hat{u}_i}\right)^2}\right)^{-\frac{1}{2}}$.

Le maximum de vraisemblance $\hat{\theta}$ utilise $\mathrm{l}(\beta,\sigma,u;y)$ et ses deux premières dérivées partielles. L'approximation de Laplace est implémentée dans la macro glmer du logicel R.

5.3.4.2 Approximation de Gauss-Hermite adaptative

La quadrature de Gauss-Hermite adaptative est souvent utilisée pour l'approximation numérique de l'intégrale (5.5) en raison de sa relation avec les densités gaussiennes et aussi du fait que les points quadrature sont concentrés autour des modes. Quand une densité gaussienne (ou noyau) n'est pas un facteur de l'intégrand, l'intégrale est parfois mise sous la forme $\int f(t)e^{-t^2}\,\mathrm{d}t$ où $f(t)$ est une fonction régulière appropriée ou en divisant et en multipliant l'intégrale d'origine par une autre densité gaussienne arbitraire $\phi(t;\mu,\tau)$ de moyenne de μ et d'écart-type τ (voir par exemple

Liu and Pierce, 1994).
Ainsi, en utilisant la quadrature de Gauss-Hermite adaptative suggérée par Pinheiro and Bates (1995), l'intégrale de la contribution du i$^{\text{ème}}$ groupe dans (5.5) est approximée de la manière suivante (voir l'appendix)

$$\int \Phi(u_i) \prod_{j=1}^{n_i} \frac{e^{y_{ij}(X_{ij}^T\beta+\sigma u_i)}}{1+e^{X_{ij}^T\beta+\sigma u_i}} \, \mathrm{d}u_i \simeq \hat{\tau}_i(\theta) \sum_{m=1}^{NQ} w_{im} \exp(t_{im}^2) g(\hat{\mu}_i + \hat{\tau}_i(\theta) t_{im}) \tag{5.9}$$

où NQ est le nombre de points de quadrature, pour i fixé, t_{im} et w_{im} $(m = 1, \ldots, NQ)$ sont respectivement les racines du m$^{\text{th}}$ polynôme d'Hermite et les poids correspondants (voir par exemple Abramowitz and Stegun, 1974; Golub and Welsch, 1969) et $g(t) = f(t)e^{-t^2}$ avec

$$f(t) = \frac{1}{\sqrt{\pi}} \prod_{j=1}^{n_i} \frac{e^{y_{ij}(X_{ij}^T\beta+\sqrt{2}\sigma t)}}{1+e^{X_{ij}^T\beta+\sqrt{2}\sigma t}} \tag{5.10}$$

Ainsi, la log-vraisemblance (5.5) est approximée telle que

$$\mathrm{l}(\beta, \sigma, u; y) \approx \sum_{i=1}^{q} \log \left[\hat{\tau}_i(\theta) \sum_{m=1}^{NQ} w_{im} \exp(t_{im}^2) g(\hat{\mu}_i + \hat{\tau}_i(\theta) t_{im}) \right] \tag{5.11}$$

Pour la quadrature de Gauss-Hermite, les points où l'intégrande est évaluée sont fixés et ne dépendent pas des caractéristiques de la fonction à intégrer.
Le centrage et la mise à l'échelle usuels de l'approximation Gauss-Hermite peuvent donner une très mauvaise approximation car les noeuds (points de quadrature) peuvent ne pas être situés dans la région la plus intéressante de l'intégrale [voir par exemple] [] Pinheiro1995. La quadrature de Gauss-Hermite adaptative résout ce problème en adaptant les noeuds et les poids à la fonction à intégrer. Elle utilise les mêmes poids et les nœuds que la quadrature Gauss-Hermite, mais afin d'accroître la qualité de l'approximation, elle centre les noeuds par rapport au mode de la fonction à intégrer et les redimensionne en fonction de la courbure de la fonction à intégrer prise en son mode.

5.3.4.3 Méthode de quasi-vraisemblance pénalisée

Breslow and Clayton (1993) traitent l'estimation des modèles linéaires généralisés mixtes en utilisant l'approche de la quasi-vraisemblance pénalisée (PQL). Pour notre modèle logistique mixte, la fonction de quasi-vraisemblance est proportionnelle à :

$$\sigma^{-q} \int_{\mathbb{R}^q} \exp\Big(-\frac{1}{2} \sum_{i=1}^{q} \sum_{j=1}^{n_i} d_{ij}(y_{ij}, p_{ij}) - \frac{1}{2\sigma^2} u^T u \Big) \, \mathrm{d}u \tag{5.12}$$

où $p_{ij} = Pr(y_{ij} = 1|u_i)$, $u = (u_1, \ldots, u_q)^T$ et

$$d_{ij}(y_{ij}, p_{ij}) = -2 \int_{y_{ij}}^{p_{ij}} \frac{y_{ij} - t}{t(1-t)} \, \mathrm{d}t. \tag{5.13}$$

Ils appliquent l'approximation de Laplace à la fonction de quasi-vraisemblance et choisissent β et u maximisant conjointement

$$PQL(\beta, u) = -\frac{1}{2} \sum_{i=1}^{q} \sum_{j=1}^{n_i} d_{ij}(y_{ij}, p_{ij}) - \frac{1}{2\sigma^2} u^T u. \tag{5.14}$$

Ils déduisent l'équation pour l'estimation REML de σ^2 en prenant le vector dépendant $Y_i^* = (y_{i1}^*, \ldots, y_{in_i}^*)^T$ déterminé par

$$y_{ij}^* = X_{ij}^T \beta + u_i + \frac{(y_{ij} - p_{ij})}{p_{ij}(1 - p_{ij})}.$$

Ainsi, les matrices de covariance asymptotique pour $\hat{\beta}$ et $\hat{\sigma}^2$ peuvent être calculées (voir Gao, 2004) par $\mathrm{Cov}(\hat{\beta}) = \left(X^T V^{-1} X\right)^{-1}$ et $\mathrm{Var}(\hat{\sigma}^2) = \mathrm{trace}(RR)$ où

$$R = V^{-1} - V^{-1} X^T \left(X^T V^{-1} X\right)^{-1} X^T V^{-1},$$

$V = (W^{-1} + \sigma^2 I \otimes M_{n_i})$ et W^{-1} est une matrice diagonale avec $p_{ij}(1 - p_{ij})$ sur la diagonale, et $I \otimes M_{n_i}$ une matrice diagonale par blocs avec M_{n_i} sur la i$^{\text{ème}}$ diagonale où M_{n_i} est une matrice de dimension $n_i \times n_i$ composée que de uns. Les estimateurs de la méthode PQL aussi bien pour les effets fixes qu'aléatoires sont connus pour être biaisés dans le cas de données binaires. Des procédures de correction ont été proposées (voir par exemple Lin and Breslow, 1996) afin d'améliorer la performance des estimateurs. L'approche PQL est implémentée dans la macro glmmPQL du logiciel R

5.4 Etude des modèles 1 et 2

Dans les parties suivantes, nous décidons de nous focaliser sur l'étude des modèles 1 et 2 à un effet aléatoire chacun car tous deux ont obtenus des p-value$<$ 0.001 au test de détection de variable à effets aléatoires. Le modèle 1 est par ailleurs le support de deux articles qui se complètent. Le premier article (Meguellati et al., 2014) traite de l'aspect accidentologique ainsi que de l'apport de ces modèles à ce domaine. Quant au second (en cours), il est plus axé sur les méthodes d'estimation.

5.4.1 Etude de la précision des méthodes d'estimation à l'aide de simulations

5.4.1.1 Procédure de simulation

Pour évaluer la performance des algorithmes proposés, nous avons conduit plusieurs simulations. Nous exposons un extrait relatif à trois bases de données de taille $n \in \{800, 2000, 5000\}$ puis effectuons sur chaque base n_R réplications avec $n_R \in \{100, 500\}$. Les résultats de l'approximation de Laplace, de Gaus-Hermite Adaptative (AGH) avec différents points de quadrature sont alors comparés à ceux de la méthode (PQL). La structure des données simulées est similaire à celle de la base de données d'accidents qui a motivé ce travail. Notons que la comparaison entre les différentes méthodes a été amorcée aux chapitres 3 et 4. Nous proposons donc ici d'étendre les comparaisons effectuées dans ces chapitres à des bases simulées de plus grandes tailles afin de voir si les conclusions sont inchangées.

Pour le modèle avec effet aléatoire le type de collision 'COL', elle implique 5 variables à effet fixe: l'EES ($\mathbf{X}_{\text{EES}}$) avec 6 catégories, l'age du conducteur ($\mathbf{X}_{\text{AGE}}$) avec 7 catégories, l'année de l'accident ($\mathbf{X}_{\text{YEAR}}$) avec 4 catégories, la localisation ($\mathbf{X}_{\text{LOC}}$) avec 2 catégories, le sexe ($\mathbf{X}_{\text{GENDER}}$) du conducteur avec 2 catégories et une variable à effet aléatoire collision ($\mathbf{X}_{\text{COL}}$) avec 6 types de collision. La variable ($\mathbf{X}_0$) est fixée à 1, les cinq autres variables à effet fixe sont générées à partir de la distribution multinomiale $\mathcal{M}(n; \pi_1^{(0)}, \ldots, \pi_r^{(0)})$ de dimension égale aux nombres de catégories et où les probabilités de classe $\pi_l^{(0)}$ $(l = 1, \ldots, r)$ sont fixées à des proportions proches de celles observées dans la base de données analysée. Ainsi, on génère

- $\mathbf{X}_{\text{EES}} \sim \mathcal{M}(n; 0.325,\ 0.237,\ 0.179,\ 0.136,\ 0.076,\ 0.047)$,
- $\mathbf{X}_{\text{AGE}} \sim \mathcal{M}(n; 0.232, 0.243, 0.212, 0.125, 0.085, 0.064, 0.039)$,
- $\mathbf{X}_{\text{YEAR}} \sim \mathcal{M}(n; 0.189,\ 0.561,\ 0.193,\ 0.058)$,
- $\mathbf{X}_{\text{LOC}} \sim \mathcal{M}(n; 0.372,\ 0.628)$,
- $\mathbf{X}_{\text{GENDER}} \sim \mathcal{M}(n; 0.293,\ 0.707)$.
- $\mathbf{X}_{\text{COL}}$ est généré à partir d'une distribution multivariée normale $\mathbf{U} \sim \mathcal{N}(\mathbf{0}_6; \sigma^2 \times I_6)$ où $\mathbf{0}_6$ est le vecteur de dimension 6 composé de zéro et I_6 la matrice identité de dimension 6×6. La variable dépendante y_{ij} est binaire (1 si on a détecté une gravité corporelle maximale dans un des véhicules impliqués dans

l'accident et 0 sinon) et est générée à partir d'une Bernoulli

$$y_{ij} \sim \mathcal{B}(\text{logit}^{-1}(\beta_0 + \sum_{m=1}^{5} \mathbf{X}_{ijm}^T \beta_m + \mathbf{U}_i)), \quad (i = 1, \ldots, 6; j = 1, \ldots, n_i)$$

où les variables à effet fixe $\mathbf{X}_{ij1}, \mathbf{X}_{ij2}, \mathbf{X}_{ij3}, \mathbf{X}_{ij4}$ et $\mathbf{X}_{ij5}$ représentent respectivement $\mathbf{X}_{\text{EES}}, \mathbf{X}_{\text{AGE}}, \mathbf{X}_{\text{YEAR}}, \mathbf{X}_{\text{LOC}}$ et $\mathbf{X}_{\text{GENDER}}$, β_0 est une constante, les β_m sont des vecteurs de dimension $(p_m - 1)$ où p_m est le nombre de modalités de la m^{eme} variable à effet fixe $\mathbf{X}_{ijm}$. Par exemple β_1 est un vecteur de dimension 5 car une modalité (la dernière) de la variable $\mathbf{X}_{\text{EES}}$ est prise pour référence. Dans les simulations, les vraies valeurs des paramètres à effet fixe sont fixées à:
$\beta_0 = 6.440$, $\beta_1^T = (-8.448, -6.681, -5.625, -4.372, -3.338)$,
$\beta_2^T = (-1.220, -1.912, -1.659, -1.338, -2.233, -1.073)$, $\beta_3^T = (-0.757, -0.865, -0.988)$,
$\beta_4 = -0.701$, $\beta_5 = -0.561$ et la variance de l'effet aléatoire COL est fixée à:
$\sigma = 0.684$.

Pour le modèle avec effet aléatoire l'EES, elle implique également 5 variables à effet fixe: l'age du conducteur ($\mathbf{X}_{\text{AGE}}$) avec 7 catégories, l'année de l'accident ($\mathbf{X}_{\text{YEAR}}$) avec 4 catégories, le type de collision ($\mathbf{X}_{\text{COL}}$) avec 6 categories, la localisation de l'accident ($\mathbf{X}_{\text{LOC}}$) avec 2 catégories, le sexe ($\mathbf{X}_{\text{GENDER}}$) du conducteur avec 2 catégories et une variable à effet aléatoire EES ($\mathbf{X}_{\text{EES}}$) avec 6 tranches de valeurs. La variable ($\mathbf{X}_0$) est fixée à 1, les cinq autres variables à effet fixe sont générées à partir de la distribution multinomiale $\mathcal{M}(n; \pi_1^{(0)}, \ldots, \pi_r^{(0)})$ de dimension égale aux nombres de catégories et où les probabilités de classe $\pi_l^{(0)}$ $(l = 1, \ldots, r)$ sont fixées à des proportions proches de celles observées dans la base de données analysée. Ainsi, on génère

- $\mathbf{X}_{\text{AGE}} \sim \mathcal{M}(n, 0.232, 0.243, 0.212, 0.125, 0.085, 0.064, 0.039)$,
- $\mathbf{X}_{\text{YEAR}} \sim \mathcal{M}(n, 0.189, 0.561, 0.193, 0.058)$,
- $\mathbf{X}_{\text{COL}} \sim \mathcal{M}(n, 0.237, 0.039, 0.055, 0.248, 0.247, 0.174))$,
- $\mathbf{X}_{\text{LOC}} \sim \mathcal{M}(n, 0.372, 0.628)$,
- $\mathbf{X}_{\text{GENDER}} \sim \mathcal{M}(n, 0.293, 0.707)$,
- $\mathbf{X}_{\text{EES}}$ est généré à partir d'une distribution multivariée normale $U \sim \mathcal{N}(0_6; \sigma^2 \times I_6)$ où 0_6 est le vecteur de dimension 6 composé de zéro et I_6 la matrice identité de dimension 6×6. La variable dépendante y_{ij} est binaire et est générée à partir d'une Bernoulli

$$y_{ij} \sim \mathcal{B}(\text{logit}^{-1}(\beta_0 + \sum_{m=1}^{5} X_{ijm}^T \beta_m + u_i)), \qquad (i = 1, \ldots, 6; j = 1, \ldots, n_i)$$

les variables à effet fixe X_{ij1}, X_{ij2}, X_{ij3}, X_{ij4} et X_{ij5} représentent respectivement X_{AGE}, X_{YEAR}, X_{COL}, X_{LOC} et X_{GENDER}. β_0 est une constante et les β_m sont des vecteurs de dimension $(J_m - 1)$ où J_m est le nombre de modalités de la m^{eme} variable à effet fixe $\mathbf{X}_{ijm}$. Par exemple, β_1 est un vecteur de dimension 6 car une modalité (la dernière) de la variable X_{AGE} est prise comme référence. Dans les simulations, les vraies valeurs des paramètres à effet fixe sont fixées à :
$\beta_0 = 2.748$, $\beta_1^T = (-1.165, -1.811, -1.599, -1.266, -2.169, -1.051)$,
$\beta_2^T = (-0.723, -0.885, -1.000)$,
$\beta_3^T = (-0.897, -2.119, -1.829, -0.605, -1.727)$,
$\beta_4^T = -0.750$, $\beta_5^T = -0.589$ et l'écart-rype de la variable à effets aléatoires est fixé à $\sigma = 2.620$.

5.4.1.2 Comparaison entre les différentes méthodes d'estimation

Modèle avec effet aléatoire le type de collision 'COL'

Les tables C.1, C.2 et C.3 en annexe concernent les valeurs estimées des paramètres pour trois (800, 2000 et 5000) tailles d'échantillons de véhicules et 100 réplications. Les résultats relatifs aux méthodes de Laplace, Adaptative Gauss-Hermite avec 15 (AGH15), 30 (AGH30) points de quadrature et Quasi-Vraisemblance Pénalisée (PQL) y sont présentés. Les valeurs entre parenthèses autour des méthodes indiquent le nombre de réplications au cours desquels les algorithmes ont convergé. Les estimations ponctuelles des effets fixes (minimum, moyenne, maximum) et de l'erreur-type de l'effet aléatoire sont évaluées par rapport au nombre de réplications convergentes. D'après la figure 5.2, nous voyons que sur 100 réplications, AGH15 a convergé 58 fois (58%) alors que PQL, Laplace et AGH30 ont convergé respectivement 64%, 73% et 76% pour $n = 800$. Cette proportion de réplications convergentes varie de 72% à 83% pour $n = 2000$ et de 94% à 97% pour $n = 5000$.

En gardant les mêmes tailles d'échantillons de véhicules, nous avons augmenté le nombre de réplications. Les résultats pour 500 réplications sont consignés dans les tables C.4 à C.6 en annexe. On note alors (voir figure 5.2) que le taux de réplications convergentes augmente. On passe d'une proportion de réplications convergentes située entre 64% et 70% pour un échantillon de 800 véhicules (table C.4) à des proportions comprises entre 95% et 97% pour un échantillon de 5000 véhicules (table C.6). Ainsi, les procédures d'estimation utilisées dans ce travail ont, globalement, un comportement similaire quant aux nombres de réplications convergentes. Plus la taille de l'échantillon (nombre de de véhicules impliqués) augmente, plus la proportion de réplications convergentes par algorithme tend vers des valeurs voisines de 100% quelque soit la procédure d'estimation. Les quatre méthodes d'approximation ont un taux de réplications convergentes similaire

en dehors du cas relatif à 100 réplications de 800 véhicules impliqués (voir également la table C.1).

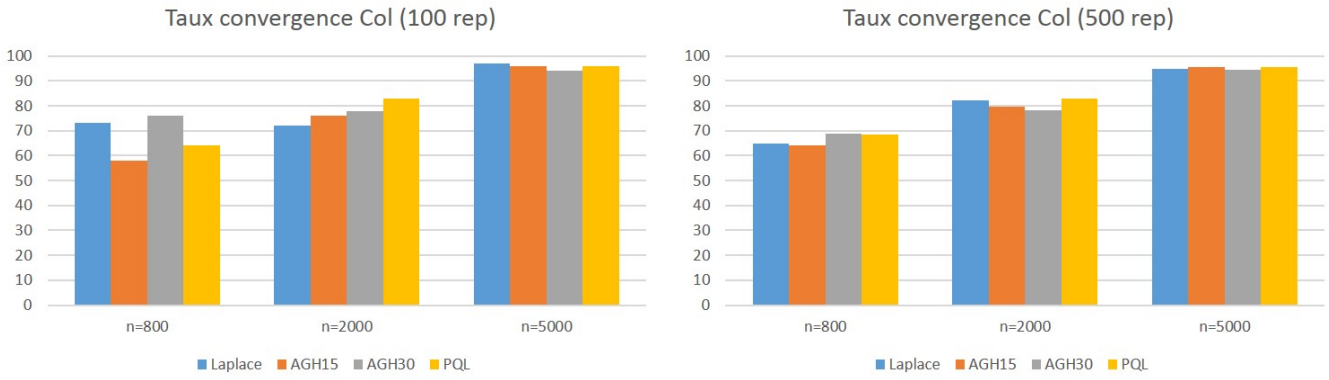

Figure 5.2: Taux de convergence M1

Par rapport aux solutions moyennes, on observe que le taux de sous-estimation (comparaison des colonnes mean et true value) des effets fixes et aléatoires reste élevé pour l'ensemble des méthodes d'approximation (voir figure 5.3). Par exemple pour $n = 800$ et 100 réplications cette sous-estimation varie entre 56% et 89%. Pour des nombres de véhicules impliqués élevés, la sous-estimation se situe entre 33% et 89% pendant que la sur-estimation. Quoi qu'il en soit, les méthodes d'estimation utilisées produisent des taux de sous-estimation élevés mais similaires et des solutions moyennes très proches des vraies valeurs des paramètres, en particulier en ce qui concerne les effets fixes.

Figure 5.3: Taux de sous-estimation M1

Les tables C.7 à C.10 en annexe présentent les valeurs minimales, maximales et moyennes des erreurs standards associées à chaque paramètre effet fixe. En dehors de la méthode PQL (table C.9) et AGH30 (table C.11) où on observe des valeurs maximales légèrement élevées pour des tailles d'échantillons de 800 et 5000 véhicules impliqués, l'ordre de grandeur des erreurs standards reste similaire et dans la majorité des cas en dessous de l'unité. Nous voyons également que plus la taille de l'échantillon augmente plus les estimations des erreurs standards deviennent petites. Lors du chapitre 3 pour le cas de 100 réplications avec 800 véhicules, nous avions vu que la méthode PQL était celle qui fournissait les estimations moyennes des erreurs standards relatives aux effets fixes les plus dispersées, l'approximation de Laplace les estimations les moins dispersées et les méthodes AGH des dispersions intermédiaires. Néanmoins, en faisant augmenter le nombre de véhicules on s'aperçoit que ce constat n'est plus vrai. En effet, pour 5000 véhicles et toujours 100 réplications, les quatre méthodes donnent des étendues moyennes des estimations des erreurs standards très proches, 0.0031 pour Laplace, 0.037pour AGH15, 0.039 pour AGH30 et 0.043 pour PQL. Le constat est similaire pour 500 réplications. Dans le cas de 800 véhicules la méthode PQL donne les estimations les plus dispersées (0.564 pour Laplace, 0.592 pour AGH15, 0,598 pour AGH30 et 0,847 pour PQL) tandis que pour 500 véhicules, les étendues moyennes des estimations des erreurs standards sont très proches, 0.205 pour Laplace, 0.210 pour AGH15, 0.217 pour AGH30 et PQL. Dans l'ensemble, plus la taille de l'échantillon (nombre de véhicules impliqués) augmente plus la précision des solutions issues des différentes méthodes d'approximation est meilleure et semblable.

Les tables C.13 et C.14 en annexe recensent les biais des estimations moyennes pour 100 et 500 réplications. Ces tables sont synthétisées par la figure 5.4. Si au niveau de l'effet aléatoire, l'ordre de grandeur du biais oscille entre 10^{-1} et 10^{-2}, celui des effets fixes varie de 10^{-1} à 10^{-4} voire plus. Concernant les paramètres à effets fixes nous pouvons faire le même constat que celui établi aux chapitres 3 et 4. Les quatres méthodes donnent des estimations des paramètres à effets fixes relativement précises et similaires. Nous voyons également que plus la taille de l'échantillon est grande plus les estimations entre les différentes méthodes sont proches. Concernant le paramètre à effets aléatoires, la méthode AGH15 est toujours celle qui donne le plus petit biais pour les cas de 100 réplications. Toutefois, on note que plus la taille d'échantillon augmente plus les biais entre les différentes méthodes tendent à se rapprocher. En effet, la différence entre le biais le plus élevé et celui le moins élevé est de 0.097 pour $n = 800$, 0.082 pour $n = 2000$ et 0.043 pour $n = 5000$. Pour les cas de 500 réplications, aucune des méthodes ne se dégage par rapport aux autres et ce d'autant plus que les biais entre chacune

des méthodes sont de plus en plus rapprochés à mesure que le nombre de véhicules augmente. En effet, la différence entre entre le biais le plus élevé et celui le moins élevé est de 0.0020 pour $n = 800$, 0.0018 pour $n = 2000$ et 0.006 pour $n = 5000$. Dans l'ensemble des simulations menées, les quatre méthodes d'approximations produisent des solutions numériques asymptotiquement non biaisées.

Figure 5.4: Biais M1

Le graphique 5.5 expose les résultats relatifs à l'erreur quadratique moyenne pour les paramètres à effets fixes et pour le paramètre à effet aléatoire (l'écart-type). Quelque soit la méthode d'approximation, l'ordre de grandeur de la MSE varie de 10^{-2} à 10^{-3} voir parfois 10^{-4} (voir table C.15 en annexe pour plus de détails). Pour 100 réplications, la méthode PQL fournit les plus petits MSE relatifs aux effets fixes, néanmoins concernant l'écart-type de l'effet aléatoire c'est la méthode AGH15 qui fournit les plus petits MSE. Nous notons cependant que les différences entre les différentes méthodes sont relativement petites et ce d'autant plus pour une grande taille d'échantillon. Pour le cas de 500 réplications, tout comme pour le biais, aucune méthode ne se dégage plus qu'une autre. Les MSE

sont pour la plus grande majorité de l'ordre de 10^{-3} et très proches les uns des autres et ce d'autant plus que la taille de l'échantillon augmente.

Pour conclure, nous pouvons dire que globalement les quatre méthodes fournissent des estimations assez précises et proches en ce qui concerne les paramètres à effets fixes mais également en ce qui concerne l'écart-type de l'effet aléatoire, et ce d'autant plus pour des grandes tailles d'échantillons et un nombre élevé de réplications.

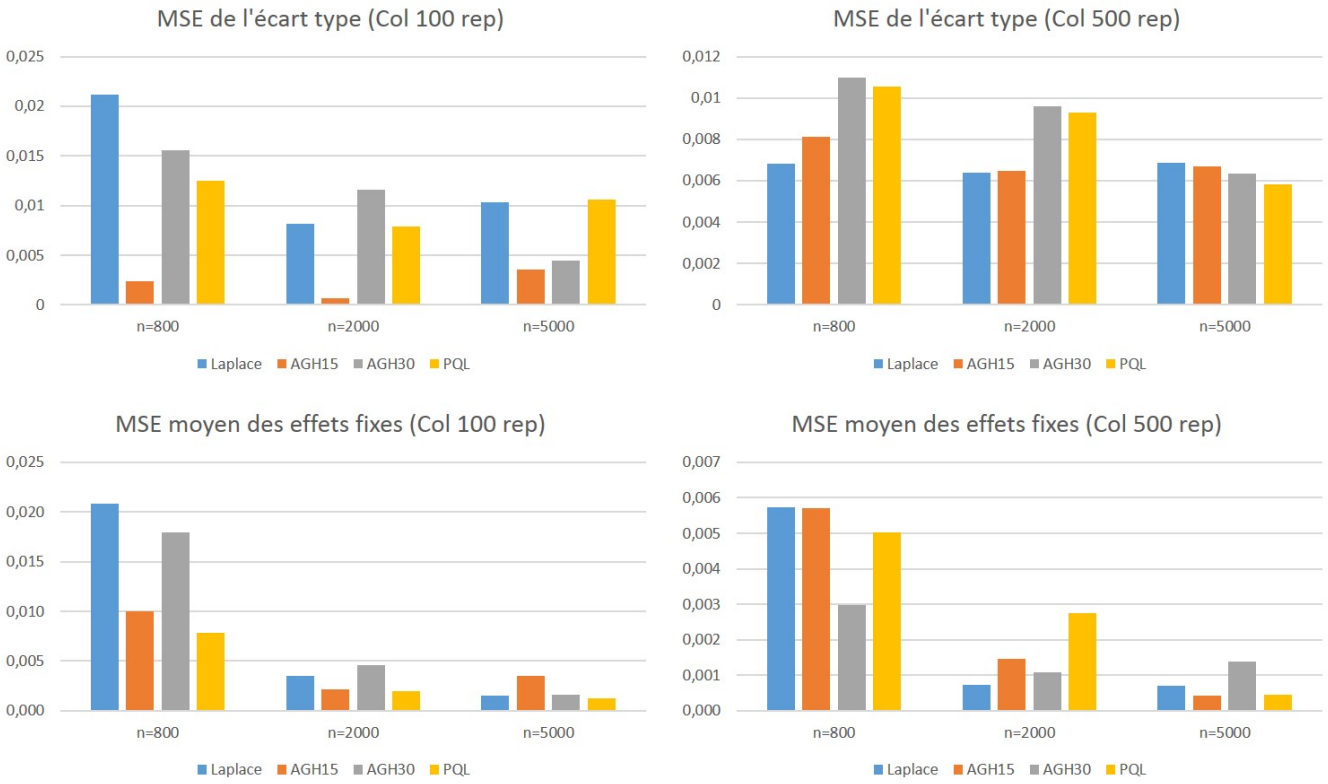

Figure 5.5: MSE M1

Modèle avec effet aléatoire l'EES

Les tables C.16, C.17 et C.18 en annexe concernent les valeurs estimées des paramètres pour trois tailles d'échantillons de véhicules (800, 2000 et 3000) et 100 réplications. Les résultats relatifs aux méthodes de Laplace, Adaptative Gauss-Hermite avec 15 et 30 points de quadrature (AGH15 et AGH30) et à la méthode de Quasi-Vraisemblance Pénalisée (PQL) y sont présentés. Les valeurs entre parenthèses sous le nom des méthodes indiquent le nombre de réplications au cours desquels

les algorithmes ont convergé. Les estimations moyennes des effets fixes et de l'écart type (SD pour Standard Deviation) de l'effet aléatoire sont évaluées par rapport à ce nombre de réplications convergentes. On note par exemple (voir figure 5.6), que pour 100 réplications et avec une taille d'échantillon de 800 véhicules, Laplace a convergé 99 fois (99%) alors que AGH15, AGH30 et PQL ont convergé respectivement 96%, 99% et 100%. Cette proportion de réplications convergentes est de 100% quelque soit la méthode utilisée pour les autres tailles d'échantillon (2000 ou 3000).

En gardant les mêmes tailles d'échantillons de véhicules, nous avons augmenté le nombre de réplications. Les résultats pour 500 réplications sont consignés dans les tables C.19, C.20 et C.21 en annexe. Nous obtenons (voir figure 5.6) des proportions de réplications convergentes situées aux alentours de 98% pour un échantillon de 800 véhicules. En passant à des échantillons de 2000 et 3000 véhicules ces proportions augmentent et sont quasiment égales à 100%. On voit donc que les procédures d'estimation utilisées dans ce travail ont, globalement, un comportement similaire quant au nombre de réplications convergentes. Plus la taille de l'échantillon (nombre de véhicules impliqués) augmente, plus la proportion de réplications convergentes par algorithme tend vers des valeurs voisines de 100% et ce quelque soit le nombre de réplications (100 ou 500) mais aussi quelle que soit la procédure d'estimation. De plus, les quatre méthodes d'approximation ont des taux de réplications convergentes similaires.

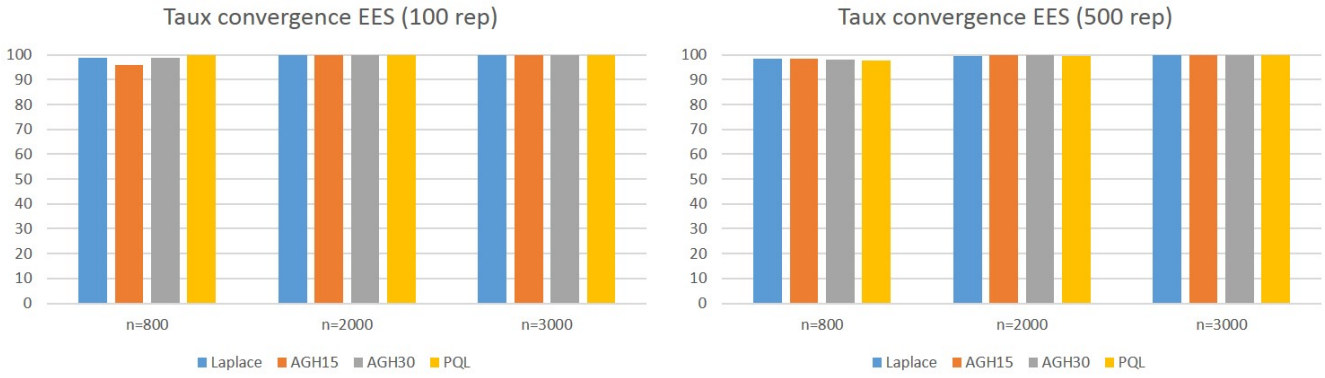

Figure 5.6: Taux de convergence M2

Par rapport aux solutions moyennes, on observe (voir figure 5.7) que le taux de sous-estimation des effets fixes et aléatoires reste élevé pour l'ensemble des méthodes d'approximation. Par exemple pour une taille d'échantillon de 800 véhicules

et 100 réplications, cette sous-estimation varie entre 83% et 94%. Pour des nombres de véhicules impliqués élevés, la sous-estimation se situe entre 44% et 94%. Quoi qu'il en soit, les méthodes d'estimation utilisées produisent des taux de sous-estimation élevés mais des solutions moyennes très proches des vraies valeurs des paramètres, en particulier en ce qui concerne les effets fixes.

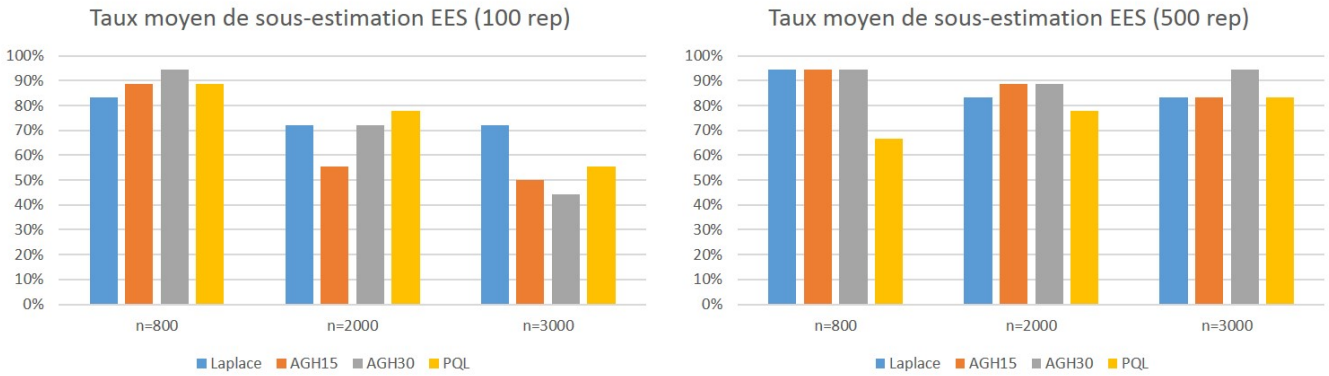

Figure 5.7: Taux de de sous-estimation M2

A chaque paramètre à effet fixe estimé, nous avons ajouté une estimation de son erreur standard. Les tables C.22, C.23, C.24, C.25, C.26 et C.27 en annexe présentent les valeurs moyennes estimées des erreurs standards associées à chaque paramètre à effet fixe. Globalement l'ordre de grandeur des erreurs standards entre les différentes méthodes reste similaire et dans la majorité des cas en dessous de l'unité, exception faite pour l'intercept. L'estimation de l'erreur standard de l'effet aléatoire n'est pas présentée ici car le logiciel R ne la fournit pas. En effet, son auteur Bates (2010) indique que résumer la précision des composantes de la variance n'est pas adéquat car les estimations des composantes de la variance ont des distributions assymétriques. Cette affirmation est renforcée par Anderson and Aitkin (1985).

Parallèlement aux valeurs estimées des paramètres et à leurs erreurs standards, nous avons évalué les méthodes d'approximation à travers le calcul des biais. Les résultats consignés dans les tables C.28 et C.29 en annexe sont obtenus en faisant la différence entre la valeur moyenne (colonne mean) et la vraie valeur (colonne true value) du paramètre. Si au niveau de l'écart-type de l'effet aléatoire, l'ordre de grandeur du biais est de 10^{-1}, celui des effets fixes varie de 10^{-1} à 10^{-4} voire plus. Comme on peut le voir sur la figure 5.8 toutes les méthodes fournissent des estimations des coefficients des effets fixes relativement proches des vraies valeurs

et aucune méthode n'a l'air de se dégager plus qu'une autre sur ce point. Quant aux estimations de l'écart type de l'effet aléatoire, la méthode PQL s'avère être celle qui fournit les estimateurs les plus biaisés tandis que les méthodes AGH30 puis AGH15, (exception faite pour le cas ou n=2000 et $n_R = 100$) sont celles qui se rapprochent le plus de la vraie valeur de l'écart-type de l'effet aléatoire.

Figure 5.8: Biais M2

La figure 5.9 expose les MSE relatifs aux paramètres à effets fixes et à l'écart-type de l'effet aléatoire, et ce à travers trois ensembles de données simulées et deux réplications (voir C.30 en annexe). Concernant les effets fixes on peut voir que l'ordre de grandeur des MSE pour la majeure partie des cas (sauf AGH15 pour n=800 et $n_R = 100$) est de 10^{-3}. Toutes les méthodes fournissent donc des estimateurs relativement proches des valeurs concernant les paramètres des effets fixes. Egalement, on remarque pour ces derniers, que plus la taille de l'échantillon augmente, plus la MSE diminue. Néanmoins, pour l'estimation de l'écart-type de l'effet aléatoire, cette dernière affirmation n'est pas vraie. Par exemple pour le cas $n_R = 100$ avec la méthode de Laplace, la MSE de l'estimation de l'écart-type de

l'effet aléatoire passe de 0.063 pour n=800 à 0.124 pour n=2000 puis à 0.068 pour n=3000. Il en est de même pour les autres méthodes que ce soit avec un nombre de réplications égal à 100 ou 500, la MSE de l'écart type de l'effet aléatoire ne diminue pas avec la taille de l'échantillon. D'autre part en comparant les différentes méthodes entre elles, toujours en ce qui concerne l'estimation de l'écart-type de l'effet aléatoire, on peut voir que la méthode PQL est celle qui fournit généralement les MSE les plus élevées et que les méthodes AGH30, puis AGH15 sont celles qui fournissent les plus petites MSE. On obtient ainsi les mêmes conclusions que celles portant sur les biais.

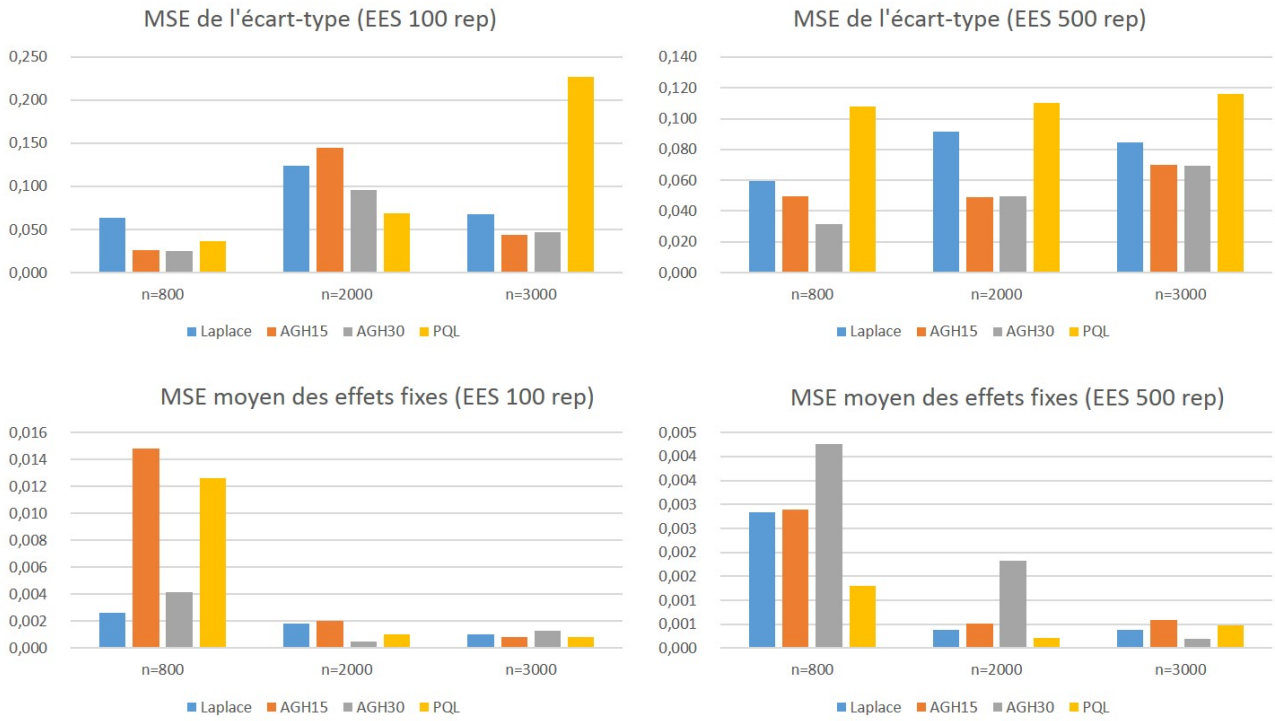

Figure 5.9: MSE M2

Les figures 5.10-5.15 exposent les densités lissées des estimations de β_3, β_6 et σ (notées $\hat{\beta}_3$, $\hat{\beta}_6$ et $\hat{\sigma}$) pour $n = 800$ et $n = 3000$ respectivement dans le cas de 500 réplications. Nous pouvons voir que les distributions empiriques de $\hat{\beta}_3$, $\hat{\beta}_6$ sont raisonnablement normales. Cela est aussi vrai pour les estimations des autres paramètres. Cependant, les distributions de $\hat{\sigma}$ sont toutes positivement assymétriques suggérant que la normalité assymptotique pour $\hat{\sigma}^2$ n'est pas vraie.

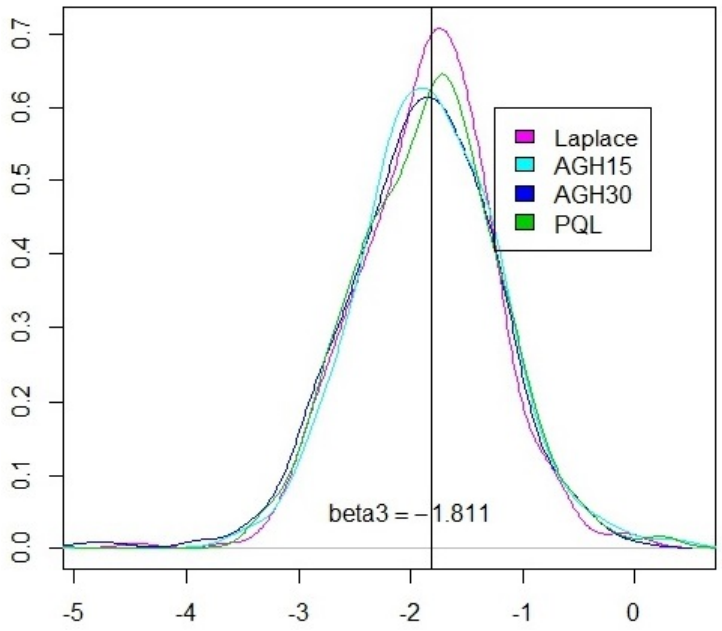

Figure 5.10: Densité de $\hat{\beta}_3$ pour n=800 et 500 réplications

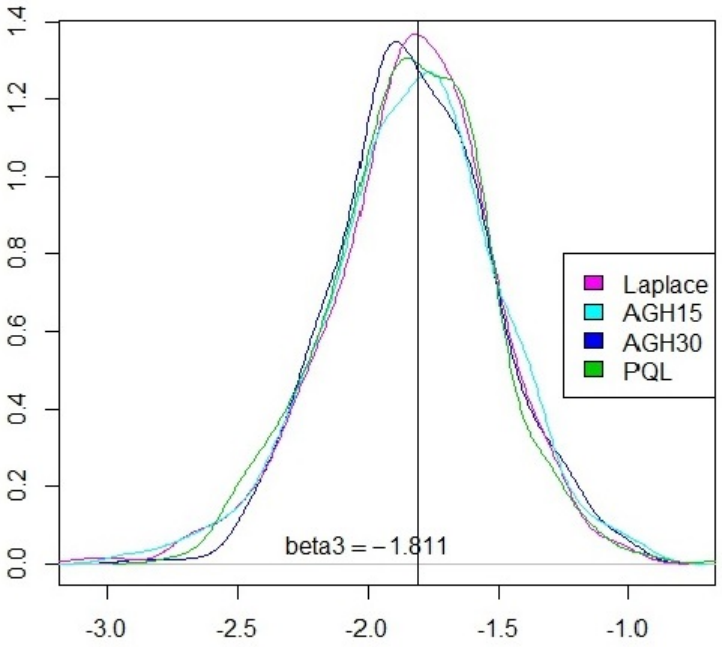

Figure 5.11: Densité de $\hat{\beta}_3$ pour n=3000 et 500 réplications

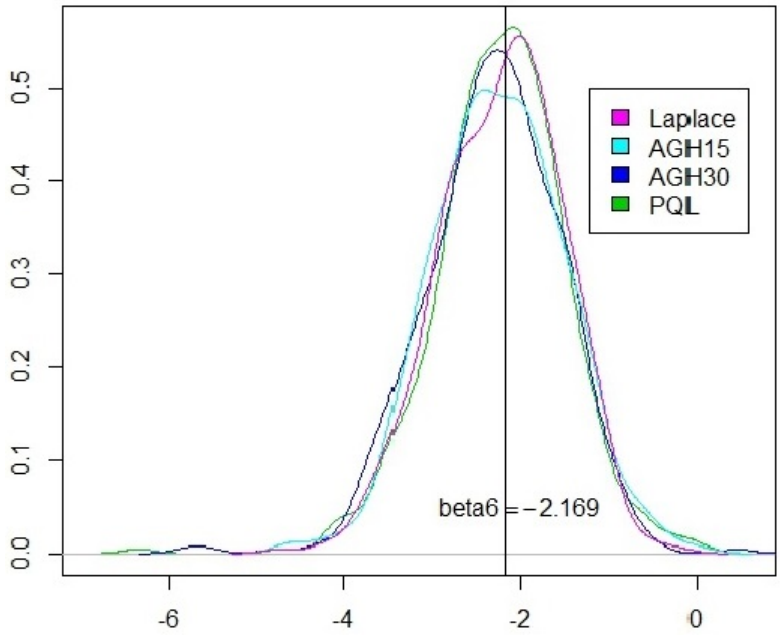

Figure 5.12: Densité de $\hat{\beta}_6$ pour n=800 et 500 réplications

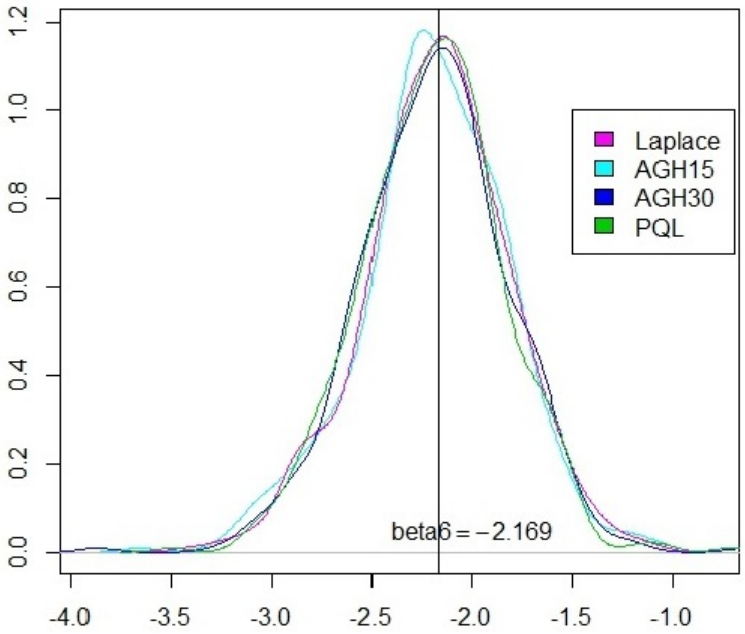

Figure 5.13: Densité de $\hat{\beta}_6$ pour n=3000 et 500 réplications

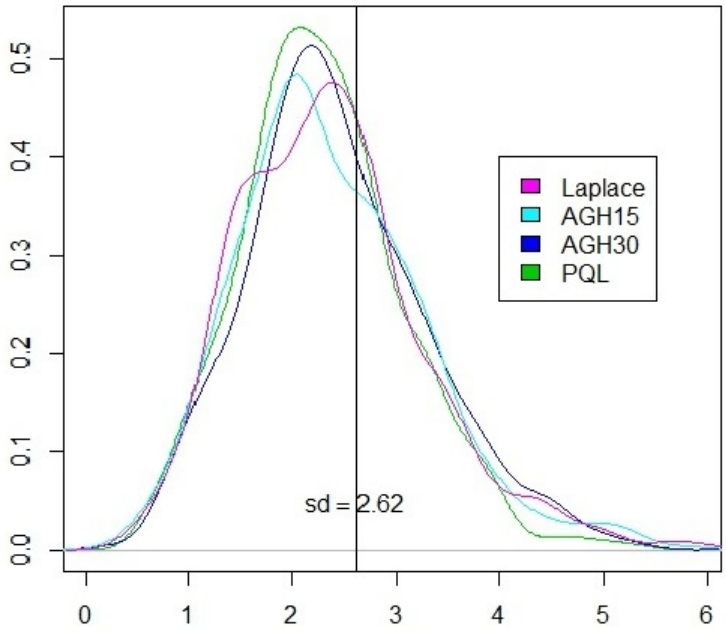

Figure 5.14: Densité de $\hat{\sigma}$ pour n=800 et 500 réplications

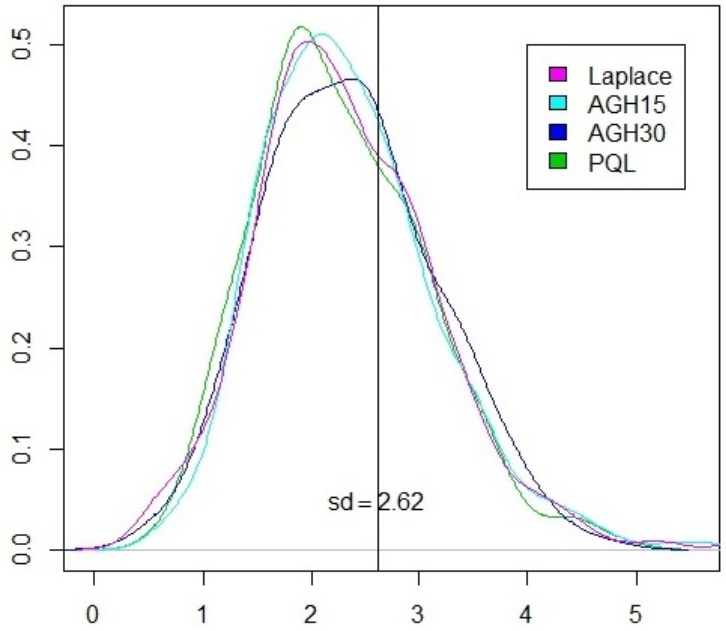

Figure 5.15: Densité de $\hat{\sigma}$ pour n=3000 et 500 réplications

Les tables 5.4 et 5.5 recensent les probabilités de couverture des intervalles de confiance des estimations moyennes des paramètres à effets fixes. Les intervalles de confiance sont construits en utilisant l'hypothèse de normalité assymptotique des estimations des paramètres à effets fixes. Tout d'abord, nous pouvons observer que dans la plupart des cas la plus mauvaise couverture est établie par la méthode PQL, excepté pour le cas avec $n = 3000$ et 500 réplications. Néanmoins, nous notons que pour cette dernière configuration, la différence entre les résultats obtenus par la méthodes PQL et les autres méthodes est très petite. En effet, pour ce cas la méthode PQL a une probabilité de couverture de 94.32% (la meilleure) tandis que la plus mauvaise est de 93.99% pour la méthode AGH15, aussi la différence n'est que de 0.33. Egalement, nous observons que plus la taille de l'échantillon est grande et le nombre de réplications élevé, plus les probabilités de couverture moyennes des différentes méthodes tendent à se rapprocher. Au final, la meilleure probabilité de couverture est obtenue pour la méthode AGH15.

Pour conclure, nous pouvons donc dire que concernant les estimations des paramètres à effets fixes, toutes les méthodes donnent des résultats similaires. Néanmoins, concernant l'estimation de l'écart-type de l'effet aléatoire les méthodes AGH15 et AGH30 sont celles qui semblent être les plus précises.

	Laplace			AGH15			AGH30			PQL		
Parameter	**800**	**2000**	**3000**	**800**	**2000**	**3000**	**800**	**2000**	**3000**	**800**	**2000**	**3000**
Intercept	94,95	92,00	91,00	94,79	90,00	90,00	91,92	94,00	94,00	91,00	93,00	87,00
Age												
16-24	92,93	94,00	93,00	97,92	92,00	96,00	93,94	94,00	97,00	91,00	89,00	93,00
25-34	94,95	95,00	93,00	98,96	93,00	96,00	98,99	94,00	98,00	89,00	89,00	97,00
35-44	95,96	95,00	93,00	98,96	94,00	95,00	95,96	96,00	96,00	87,00	91,00	92,00
45-54	92,93	92,00	94,00	96,88	92,00	95,00	94,95	96,00	97,00	93,00	93,00	95,00
55-64	95,96	94,00	92,00	94,79	95,00	93,00	95,96	94,00	97,00	93,00	97,00	97,00
65-74	96,97	96,00	96,00	97,92	93,00	95,00	97,98	98,00	99,00	88,00	87,00	90,00
Col												
A-F	87,88	91,00	95,00	95,83	97,00	93,00	94,95	94,00	93,00	92,00	92,00	94,00
F-A	96,97	94,00	96,00	96,88	96,00	96,00	97,98	94,00	98,00	95,00	92,00	93,00
F-F	89,90	92,00	94,00	96,88	95,00	94,00	92,93	95,00	93,00	92,00	88,00	90,00
F-L	96,97	94,00	93,00	94,79	95,00	92,00	93,94	91,00	97,00	93,00	95,00	93,00
V-O	90,91	93,00	93,00	96,88	96,00	94,00	92,93	92,00	94,00	96,00	94,00	96,00
Year												
1995-1999	91,92	93,00	97,00	97,92	98,00	92,00	97,98	98,00	93,00	95,00	95,00	94,00
2000-2004	94,95	99,00	95,00	92,71	96,00	92,00	95,96	95,00	94,00	91,00	92,00	95,00
2005-2010	93,94	94,00	98,00	94,79	92,00	94,00	92,93	95,00	93,00	95,00	94,00	93,00
Loc												
Inag	92,93	95,00	91,00	96,88	93,00	98,00	95,96	99,00	92,00	91,00	89,00	93,00
Gender												
F	94,95	93,00	93,00	96,88	92,00	94,00	94,95	97,00	93,00	92,00	96,00	93,00

Table 5.4: M2: Probabilités de couvertures pour 100 réplications

	Laplace			AGH15			AGH30			PQL		
Parameter	**800**	**2000**	**3000**	**800**	**2000**	**3000**	**800**	**2000**	**3000**	**800**	**2000**	**3000**
Intercept	92,68	89,56	92,00	94,72	91,60	90,60	94,50	92,00	91,60	92,64	93,37	92,99
Age												
16-24	95,53	93,98	93,80	93,90	95,80	93,40	93,89	93,80	95,60	93,66	93,17	95,19
25-34	96,34	93,98	93,80	94,92	94,20	93,20	94,50	93,60	95,00	93,46	93,57	93,99
35-44	94,72	94,38	93,20	93,29	95,60	94,80	94,70	93,80	94,20	94,89	92,57	94,19
45-54	95,33	93,78	94,00	94,92	96,00	92,00	93,89	93,60	94,80	94,27	93,17	93,39
55-64	96,34	95,38	93,60	94,92	95,20	93,00	94,91	94,80	94,80	93,46	93,57	94,99
65-74	95,73	93,98	94,40	94,11	94,00	94,00	94,30	93,00	94,60	92,43	93,57	94,59
Col												
A-F	94,92	95,18	94,80	94,72	94,40	94,40	92,87	94,60	94,20	92,84	94,38	93,79
F-A	93,90	94,18	94,40	93,50	92,60	95,60	94,30	94,00	92,40	93,05	91,97	93,79
F-F	93,50	95,18	92,40	94,11	95,00	94,60	93,89	94,40	94,20	94,27	94,98	95,19
F-L	93,90	93,98	94,20	93,29	96,20	95,00	93,69	95,20	94,40	93,46	93,78	94,59
V-O	93,09	95,18	96,20	93,70	95,60	94,00	95,32	95,80	97,00	92,43	94,18	94,79
Year												
1995-1999	93,70	90,76	95,80	95,73	94,20	94,20	93,69	93,00	93,20	92,84	93,57	94,19
2000-2004	94,72	94,18	94,20	95,12	92,80	93,80	94,09	95,80	93,80	93,66	91,77	94,79
2005-2010	92,89	94,38	94,80	93,70	92,20	94,80	91,85	95,00	94,60	92,23	93,57	94,19
Loc												
Inag	95,12	93,98	95,40	95,12	94,40	95,80	95,72	95,20	94,60	91,41	92,97	93,79
Gender												
F	93,29	94,58	94,40	93,09	94,00	94,60	93,48	93,40	93,60	93,25	94,58	94,99

Table 5.5: M2: Probabilités de couvertures pour 500 réplications

5.4.2 Analyse des données d'accident détaillées

5.4.2.1 Estimation des paramètres avec différentes méthodes

Dans cette section nous cherchons à comparer les différentes méthodes d'estimation. Nous réécrivons les modèles 1 et 2 (voir (5.2) et (5.3)) de la manière suivante afin de visualiser explicitement la structure des modèles ainsi que les différentes variables explicatives y intervenant.

Le modèle d'application 1 se présente sous la forme suivante :

$$\begin{aligned}\text{logit}(P(y_{ij}=1|(\text{COL})_i)) &= \beta_0 + (\text{EES})_{ij}^T\beta_1 + (\text{AGE})_{ij}^T\beta_2 + (\text{YEAR})_{ij}^T\beta_3 + (\text{LOC})_{ij}\beta_4 \\ &\quad +(\text{GENDER})_{ij}\beta_5 + (\text{COL})_i,\ i=1,\ldots,6; j=1,\ldots,n_i\end{aligned}$$

où $(\text{COL})_i \sim \mathcal{N}(0;\sigma^2)$ pour tout i dans $\{1,\ldots,q\}$, $\beta_1 = (\beta_{11},\beta_{12},\beta_{13},\beta_{14},\beta_{15})^T$, $\beta_2 = (\beta_{21},\beta_{22},\beta_{23},\beta_{24},\beta_{25},\beta_{26})^T$, $\beta_3 = (\beta_{31},\beta_{32},\beta_{33})^T$ sont des vecteurs de dimensions respectives 5, 6, 3 et β_4, β_5 des scalaires. Ces dimensions proviennent du fait qu'une des modalités des covariables est prise comme référence. Pour les covariables EES, AGE, nous prenons la dernière modalité comme référence alors que pour la covariable YEAR, la première année est prise comme référence. Pour la localisation (LOC) de l'accident nous prenons la modalité "'hors agglomération"'

comme référence et pour le conducteur (GENDER) du véhicule dans lequel la gravité maximale est observée, nous prenons la modalité "'homme"' comme référence.

Quant au modèle d'application 2 il se présente sous la forme :

$$\begin{aligned}\text{logit}(P(y_{ij}=1|(\text{EES})_i)) &= \beta_0 + (\text{AGE})_{ij}^T\beta_1 + (\text{YEAR})_{ij}^T\beta_2 + (\text{COL})_{ij}^T\beta_3 + (\text{LOC})_{ij}\beta_4 \\ &\quad +(\text{GENDER})_{ij}\beta_5 + (\text{EES})_i, i=1,\ldots,6; j=1,\ldots,n_i\end{aligned}$$

où $(EES)_i \sim \mathcal{N}(0,\sigma^2)$ pour tout i dans $\{1,\ldots,q\}$, $\beta_1 = (\beta_{11},\beta_{12},\beta_{13},\beta_{14},\beta_{15},\beta_{16})^T$, $\beta_2 = (\beta_{21},\beta_{22},\beta_{23})^T$, $\beta_3 = (\beta_{31},\beta_{32},\beta_{33},\beta_{34},\beta_{35})^T$ sont des vecteurs de dimensions respectives 6, 3, 5 et β_4, β_5 des scalaires. Ces dimensions proviennent du fait qu'une des modalités des covariables est prise comme référence. Pour la covariable AGE nous prenons la dernière modalité comme référence tandis que pour la variable YEAR, la première est prise comme référence. Pour le type de collision, COL, nous prenons la modalité 'L-F' comme référence, pour le lieu de l'accident, LOC, nous prenons la modalité 'Inag' (hors agglomération) comme référence et pour le sexe du conducteur (GENDER) nous prenons la modalité 'm' (masculin)comme référence.

L'estimation du maximum de vraisemblance des paramètres du modèle se fait en utilisant les méthodes d'approximation proposées dans la chapitre 3 à savoir (1) la méthode de Laplace (2) la méthode de Gauss-Hermite Adaptative (AGH) et (3) la méthode de quasi-vraisemblance pénalisée (PQL). Nous cherchons alors à savoir quellle est la proximité entre les estimations des différentes méthodes, quels sont les niveaux de précision des solutions proposées et quel est le niveau de signification de chaque variable (à travers ses modalités) par rapport à la probabilité de détecter une gravité corporelle importante. Pour cela nous comparons les résultats issus de l'approximation de Laplace avec ceux issus de la méthode PQL et de la méthode AGH. Pour la méthode AGH nous avons également fait varier le nombre de points de quadrature. L'estimation des paramètres se fait alors à l'aide du logiciel R (R Development Core Team, 2012). Nous utilisons le package lme4 (Bates et al., 2011) pour les méthodes de Laplace et AGH et le package MASS (Venables, 2002) pour la méthode PQL.

La table 5.6 compare les estimations du modèle 1 obtenues avec la méthode de Laplace avec celles obtenues avec la méthode AGH avec 15 et 30 points de quadrature (notées AGH15, AGH30) d'une part et avec la méthode PQL d'autre part. Les estimations obtenues avec les méthodes de Laplace et AGH sont quasiment similaires. On note quelques différences seulement à partir de la quatrième décimale pour l'estimation des effets fixes et à partir de la troisième décimale pour l'estimation de l'écart-type de la variable à effets aléatoires. Nous notons également que le fait d'augmenter le nombre de points de quadrature n'a pas changé les

résultats dans notre cas. Les estimations obtenues à l'aide de la méthode PQL sont globalement plus petites (en valeur absolue) comparativement aux solutions issues des autres approximations. De façon générale, les résultats d'estimation, à travers les différentes méthodes d'approximation, sont d'un aperçu général similaire.
La colonne Pv (P value) permet de mesurer la significativité des effets fixes. Etant donné les modalités de référence de chaque covariable, on note que les estimations ont le même signe et sont dans l'ensemble très significatives avec des risques faibles quelle que soit la méthode d'approximation. L'EES est de loin la variable la plus significative avec des risques associés à chaque modalité plus petite que 10^{-2} voire 10^{-3} si l'on considère la dernière modalité comme référence. Si l'on se fixe un risque de 5% et qu'on prend les conducteurs de plus de 75 ans comme référence, alors les conducteurs âgés de 25 à 34 ans contribuent très significativement à la modélisation, suivis des 35 à 44 ans puis des 55 à 65 ans. Les jeunes conducteurs ferment la marche. Les véhicules accidentés entre 1995 et 1999 sont les plus significatifs en référence à ceux des années 1991 à 1994. Par contre le coefficient à effet fixe estimé des véhicules accidentés entre 2005 et 2010 n'est pas significatif comparativement à ceux observés entre 1991 et 1994. Comparés aux véhicules accidentés hors agglomération, les véhicules accidentés en agglomération sont significatifs avec un risque plus petit de 5%. Il en est de même pour les véhicules conduits par les femmes si l'on prend les conducteurs hommes comme référence.

La table 5.7 compare les estimations obtenues par l'approximation de Laplace avec d'une part celles obtenues par l'approximation AGH avec 15 et 30 points de quadrature (noté AGH15 et AGH30) et d'autre part avec les estimations obtenues par la méthode PQL. Pour cela, nous utilisons la fonction glmer du package `lme4` (Bates et al., 2011) pour les approximations de Laplace et AGH et pour la méthode PQL nous utilisons la fonction glmmPQL du package MASS (Venables, 2002) du logiciel R (R Development Core Team, 2012). Les estimations fournies par les approximations de Laplace et AGH (AGH15 ou AGH30) sont quasiment similaires. Les différence se trouvent seulement au niveau de la quatrième décimale pour les estimations des effets fixes et au niveau de la troisième décimale pour l'écart-type de la variable à effets aléatoires. Nous notons également que le fait d'augmenter le nombre de points de quadrature pour l'approximation AGH ne change pas les résultats dans cette étude de cas. Les estimations fournies par la méthode PQL sont globalement plus petites (en valeur absolue) que les solutions fournies par les autres approximations (Laplace et AGH). De plus, la différence avec les autres méthodes se situe maintenant au niveau de la seconde décimale pour l'estimation des effets fixes et au niveau de la première décimale pour l'estimation de l'écart-type de la variable à effet aléatoires. Dans l'ensemble, les estimations de la méthode PQL restent légèrement décalées comparativement aux méthodes de Laplace et AGH.

Toutes les méthodes fournissent une forte association entre la gravité de l'accident et les variables explicatives en prenant un risque de 5%, sauf pour les conducteurs agés de 65-74 ans, l'année d'accident 2005-2010 et les collisions A-F et V-O.

	Laplace			**AGH15**			**PQL**			**Based on Laplace/AGH15**		
											95% Confidence Intervals	
	Estimate	SE	PV	Estimate	SE	PV	Estimate	SE	PV	OR	Lower	Upper
Variables												
Intercept	6,4404	1,2454	0,0000	6,4405	1,2454	0,0000	6,4000	1,1067	0,0000			
EES												
0-20	-8,4481	1,1903	0,0000	-8,4481	1,1903	0,0000	-8,4012	1,0464	0,0000	0,000	0,000	0,002
20-30	-6,6809	1,1039	0,0000	-6,6810	1,1039	0,0000	-6,6498	0,9711	0,0000	0,001	0,000	0,011
30-40	-5,6250	1,0825	0,0000	-5,6250	1,0825	0,0000	-5,5959	0,9518	0,0000	0,004	0,000	0,030
40-50	-4,3721	1,0706	0,0000	-4,3721	1,0707	0,0000	-4,3450	0,9409	0,0000	0,013	0,002	0,103
50-60	-3,3377	1,0769	0,0019	-3,3378	1,0770	0,0019	-3,3176	0,9464	0,0005	0,036	0,004	0,293
Age												
16-24	-1,2202	0,5397	0,0238	-1,2202	0,5397	0,0238	-1,2033	0,4798	0,0123	0,295	0,102	0,850
25-34	-1,9116	0,5625	0,0007	-1,9116	0,5625	0,0007	-1,8857	0,4997	0,0002	0,148	0,049	0,445
35-44	-1,6594	0,5567	0,0029	-1,6594	0,5567	0,0029	-1,6362	0,4947	0,0010	0,190	0,064	0,567
45-54	-1,3380	0,5919	0,0238	-1,3380	0,5919	0,0238	-1,3165	0,5260	0,0125	0,262	0,082	0,837
55-64	-2,2331	0,7311	0,0023	-2,2331	0,7311	0,0023	-2,2032	0,6486	0,0007	0,107	0,026	0,449
65-74	-1,0729	0,6420	0,0947	-1,0730	0,6420	0,0947	-1,0624	0,5709	0,0631	0,342	0,097	1,204
Year												
2005-2010	-0,7571	0,5343	0,1565	-0,7571	0,5343	0,1565	-0,7407	0,4742	0,1187	0,469	0,165	1,337
2000-2004	-0,8651	0,4102	0,0349	-0,8651	0,4102	0,0349	-0,8582	0,3644	0,0187	0,421	0,188	0,941
1995-1999	-0,9879	0,2915	0,0007	-0,9879	0,2915	0,0007	-0,9812	0,2591	0,0002	0,372	0,210	0,659
Loc												
Inag	-0,7009	0,3146	0,0259	-0,7009	0,3146	0,0259	-0,7027	0,2792	0,0120	0,496	0,268	0,919
Gender												
F	-0,5611	0,2827	0,0472	-0,5611	0,2827	0,0472	-0,5634	0,2511	0,0251	0,571	0,328	0,993
SD	0,684			0,684			0,687					

Pv: P-value, OR: Odd ratio

Table 5.6: M1: Estimations et Odds ratio basés sur les données réelles

	Laplace			AGH15			PQL			Based on Laplace/AGH15		
											95% Confidence Intervals	
	Estimate	SE	PV	Estimate	SE	PV	Estimate	SE	PV	OR	Lower	Upper
Variables												
Intercept	2,7476	1,2255	0,0250	2,7472	1,2252	0,0249	2,7188	1,1774	0,0212			
Age												
16-24	-1,1645	0,5393	0,0308	-1,1645	0,5393	0,0308	-1,1585	0,4778	0,0155	0,3121	0,1084	0,8981
25-34	-1,8111	0,5597	0,0012	-1,8111	0,5597	0,0012	-1,8029	0,4958	0,0003	0,1635	0,0546	0,4896
35-44	-1,5991	0,5556	0,0040	-1,5991	0,5556	0,0040	-1,5885	0,4920	0,0013	0,2021	0,0680	0,6004
45-54	-1,2659	0,5911	0,0322	-1,2659	0,5911	0,0322	-1,2551	0,5234	0,0167	0,2820	0,0885	0,8981
55-64	-2,1691	0,7312	0,0030	-2,1691	0,7312	0,0030	-2,1504	0,6460	0,0009	0,1143	0,0273	0,4791
65-74	-1,0512	0,6439	0,1026	-1,0512	0,6439	0,1026	-1,0462	0,5701	0,0669	0,3495	0,0989	1,2348
Year												
2005-2010	-0,7233	0,5350	0,1764	-0,7232	0,5350	0,1764	-0,7165	0,4732	0,1304	0,4852	0,1700	1,3845
2000-2004	-0,8849	0,4087	0,0304	-0,8849	0,4087	0,0304	-0,8749	0,3619	0,0159	0,4128	0,1853	0,9196
1995-1999	-1,0001	0,2918	0,0006	-1,0001	0,2918	0,0006	-0,9912	0,2582	0,0001	0,3678	0,2076	0,6517
Col												
A-F	-0,8965	0,7693	0,2439	-0,8966	0,7693	0,2439	-0,8883	0,6805	0,1921	0,4080	0,0903	1,8428
F-A	-2,1190	0,8747	0,0154	-2,1189	0,8747	0,0154	-2,1039	0,7721	0,0066	0,1202	0,0216	0,6673
F-F	-1,8290	0,3775	0,0000	-1,8290	0,3775	0,0000	-1,8268	0,3344	0,0000	0,1606	0,0766	0,3366
V-O	-0,6054	0,3897	0,1203	-0,6054	0,3897	0,1203	-0,6080	0,3453	0,0786	0,5459	0,2543	1,1716
F-L	-1,7266	0,4138	0,0000	-1,7267	0,4138	0,0000	-1,7170	0,3658	0,0000	0,1779	0,0791	0,4003
Loc												
Inag	-0,7502	0,3146	0,0171	-0,7503	0,3146	0,0171	-0,7394	0,2782	0,0080	0,4723	0,2549	0,8749
Gender												
F	-0,5894	0,2822	0,0368	-0,5894	0,2822	0,0368	-0,5845	0,2497	0,0195	0,5547	0,3190	0,9644
SD	2,620			2,619			2,548					

Pv: P-value, OR: Odd ratio

Table 5.7: M2: Estimations et Odds ratio basés sur les données réelles

5.4.2.2 Résultats et Interprétation

Les tables 5.6 et 5.7 exposent les odd-ratios associés aux estimations de Laplace/AGH15. Tout d'abord, nous pouvons voir que les estimations des effets fixes en commun sont relativement les mêmes pour chaque modèle. Les valeurs de la statistique z sont également très similaires pour chacun des modèles présentés. Il en résulte que ces modèles sont très semblables en ce qui concerne les effets fixes en commun. Nous choisissons alors de faire les interprétations concernant les effets fixes en commun à partir du modèle 1 (table 5.6).

EES

En prenant pour modalité de référence '60-90' on voit (table 5.6) que tous les signes des coefficients estimés sont négatifs. De plus, plus la valeur de l'EES diminue plus ce coefficient est petit. On peut alors en déduire que plus l'EES est bas plus le risque d'être blessé grave devient bas. En regardant les odds ratios on voit qu'un véhicule d'EES compris entre 60 et 90 km/h a 4800 fois (1/OR) plus de chances

de contenir un blessé grave qu'un véhicule d'EES compris entre 0 et 20 km/h, 850 fois plus de chances qu'un véhicule d'EES compris entre 20 et 30 km/h, 290 fois plus de chances qu'un véhicule d'EES compris entre 30 et 40 km/h, 78 fois plus de chances qu'un véhicule d'EES compris entre 40 et 50 km/h et 27 fois plus de chances qu'un véhicule dont l'EES est compris entre 50 et 60 km/h. Ces OR sont donnés en considérant tous les autres paramètres égaux par ailleurs. L'EES étant l'énergie de déformation absorbée par le véhicule lors d'une collision, il semble logique que plus l'énergie de déformation est élevée plus le choc est violent et donc la gravité sévère.
Lorsque l'EES est modélisé par un effet aléatoire, l'écart-type de sa distribution est estimé à 2.620 (table 5.7). En simulant une telle distribution on peut en déduire que d'après ce modèle l'EES contribue à faire augmenter la probabilité d'être MAIS3+ dans 49.8% des cas tandis que dans 50.2% des cas il contribue à la faire diminuer.

Collision

En prenant pour modalité de référence 'L-F' on voit (table 5.7) que tous les signes des coefficients estimés sont négatifs. Il en résulte que tous les autres types de collision sont moins sévères que les collisions latéro-frontales. Rappelons que tous les odds ratios donnés sont à interpréter en supposant tous les autres coefficients fixés par ailleurs. D'après les estimations des odds-ratios on peut voir que les véhicules en collision fronto-arrière ont la plus faible probabilité de contenir des blessés graves. En effet, ils ont environ 8.3 fois plus de chance que les véhicules impliqués en collision latéro-frontale de ne pas avoir de blessés de niveau MAIS3+. Ce rapport est de 6.6 pour les véhicules en collision fronto-frontale et de 5.6 pour les véhicules en collision fronto-latérale. Le coefficient estimé des modalités 'arrière-frontale' et 'véhicule seul contre obstacle' n'étant pas significativement différent de 0 nous ne pouvons les interpréter. En fixant alors tous les paramètres, les collisions latéro-frontales sont donc les plus graves. Ce résultat n'est cependant pas à confondre avec le fait que les collisions fronto-frontales sont généralement les plus sévères (Kockelman and Kweon, 2002) et qui s'explique par des vitesses de collision élevées dues aux vitesses des deux véhicules et donc des EES élevés. Or, en considérant tous les paramètres égaux et donc notamment l'EES, les collisions latéro-frontales sont les plus vulnérables du fait d'une mauvaise protection latérale tandis que les protections frontales elles sont bien développées.
Lorsque le type de collision est modélisé par un effet aléatoire, l'écart-type de sa distribution est estimé à 0.684 (table 5.6). En simulant une telle distribution on peut en déduire que d'après ce modèle le type de collision contribue à faire augmenter la probabilité d'être MAIS3+ dans 49.4% des cas tandis que dans 50.6% des cas il contribue à la faire diminuer.

Age du conducteur

En prenant pour modalité de référence '75+' on voit (table 5.6) là encore que tous les signes des coefficients estimés sont négatifs, il en résulte que les véhicules conduits par les plus de 75 ans sont estimés être les plus dangereux et ce toujours en considérant les autres paramètres fixés. De plus, les véhicules conduits par les 55 à 64 ans sont les moins dangereux, ils ont environ 8.7 (1/OR) fois plus de chances de ne pas contenir de blessés graves comparés aux véhicules conduits par les plus de 75 ans. Ce rapport est ensuite de 6.4 pour les 25-34 ans, 5 pour les 35-44 ans, 3.6 pour les 45-54 ans et 3.3 pour les 16-24 ans. La catégorie des 65-74 ans n'étant pas significative, nous n'en tirons aucune interprétation. D'après ces résultats nous voyons donc que les véhicules conduits par des jeunes (16-24) ou des personnes âgées (plus de 75 ans) ont plus de chances d'être impliqués dans des accidents graves que les autres catégories d'âge. Ces résultats sont en accord avec le fait que d'une part les jeunes conducteurs prennent plus de risques (Gregersen and Berg, 1994) notamment en violant le code de la route (Blockley and Hartley, 1995) et sont moins conscients du risque encouru (Dejoy, 1992) et que d'autre part les conducteurs les plus âgés subissent le déclin de leurs capacités perceptivo-motrices et cognitives (Brouwer and Ponds, 1994) ce qui implique qu'ils ont moins de réflexe et mettent donc plus de temps à réagir (Stelmach and Nahom, 1992; Islam and Mannering, 2006).

Sexe du conducteur

En prenant pour modalité de référence 'M' on voit (table 5.6) que le signe du coefficient estimé de la modalité 'F' est négatif. Cela implique que le modèle estime que les véhicules conduits par les femmes sont moins dangereux que ceux conduits par les hommes. En regardant les estimations des OR les véhicules conduits par les femmes ont environ 1.8 (1/OR) fois plus de chances de ne pas contenir de blessés graves comparés aux véhicules conduits par les hommes. Un certain nombre d'études montre que les conducteurs hommes ont une conduite plus agressive (Paleti et al., 2010) que les femmes et qu'ils ont tendance à prendre plus de risques que ces dernières et ce en particulier chez les jeunes conducteurs (Sivak et al., 1989; Dejoy, 1992; Yagil, 1998; Iversen and Rundmo, 2004).

Localisation de l'accident

Concernant la localisation de l'accident, l'agglomération semble être plus sûre. En effet, les véhicules dont l'accident a eu lieu en agglomération ont environ 2 (1/OR) fois plus de chances de ne pas contenir de blessés graves par rapport aux véhicules dont l'accident a eu lieu hors agglomération. Ce résultat confirme le fait qu'il y a généralement moins d'accidents fatals dans les zones urbanisées comparées aux

zones rurales (Jones et al., 2007) bien que les zones urbaines comportent plus d'accidents mais de faible sévérité (Eiksund, 2009). Ces différences peuvent être attribuées aux différences d'infrastructure entre les zones urbanisées et rurales mais également aux différences de vitesse pratiquée étant données les limites de vitesse moins élevées en zone urbaine qu'en zone rurale (Nordfjaerna et al., 2010)

L'année de l'accident
Enfin, concernant l'année de l'accident, toujours en supposant les autres paramètres fixés, les véhicules qui ont eu un accident entre 1995 et 1999 ont 2.7 fois plus de chances de ne pas contenir de blessés graves comparés à ceux dont l'accident a eu lieu entre 1991 et 1994. Quant à ceux dont l'accident a eu lieu entre 2000 et 2004 ils ont 2.4 fois plus de chances de ne pas contenir de blessés graves par rapport aux véhicules accidentés entre 1991 et 1994. Enfin, la modalité '2005-2010' n'étant pas significativement différente de 0 à un seuil de 5% ne peut être exploitée. Nous voyons donc qu'au fil des années la gravité des accidents diminue et ce en supposant être toujours dans les mêmes conditions (paramètres fixés). Cela peut s'expliquer qu'au fil des années les véhicules protègent mieux. Également les différentes dispositions de sécurité routière prises au fil des années peuvent expliquer ce phénomène.

5.5 Les autres modèles et la performance de prédiction

5.5.1 Présentation du modèle classique

Le modèle 0 est le modèle logistique habituellement utilisé c'est à dire sans effets aléatoires.

La probabilité $P(y_i = 1|\beta)$ qu'un véhicule accidenté i contienne au moins un blessé grave est reliée aux variables explicatives par le modèle suivant :

$$\text{logit}(P(y_i = 1|x)) = x_i\beta,\ i = 1, \ldots, 826 \tag{5.15}$$

Où x_i est la matrice de dimension $[1 * (1 + p)]_{p=21}$ des variables explicatives et β le vecteur de taille $(1 + p)_{p=23}$ des coefficients à estimer associés aux variables explicatives.

5.5.2 Présentation des modèles à deux effets aléatoires

Modèle 4 : modèle avec effets aléatoires 'COL' et 'YEAR'

Le Modèle 4 a pour effets aléatoires le type de collision et l'année de l'accident et pour effets fixes l'EES, la localisation de l'accident, l'âge du conducteur ainsi que son sexe.

La probabilité $P(y_{ijk} = 1|u_i, w_j)$ qu'un véhicule accidenté k impliqué dans une collision de type i durant l'année j contienne au moins un blessé grave est reliée aux variables explicatives et aux effets aléatoires par le modèle suivant :

$$\text{logit}(P(y_{ijk} = 1|u_i, w_j)) = (x_{uw})_{ijk}\beta_{uw}+u_i+w_j, i = 1, \ldots, 6 \quad j = 1, \ldots, 4 \quad k = 1, \ldots, n_{ij} \tag{5.16}$$

Où, $(x_{uw})_{ijk}$ est une matrice $[1 \times (1 + p)]_{p=13}$ des variables à effets fixes, β_{uw} le vecteur de taille $(1+p)_{p=13}$ des coefficients associés aux effets fixes à estimer , u_i et w_j deux scalaires désignant respectivement la i^{ème} composante de l'effet aléatoire U et la j^{ème} composante de l'effet aléatoire W tel que $\sum_{i=1}^{6}\sum_{j=1}^{4} n_{ij} = n = 826$ où n_{ij} désigne le nombre de véhicules accidentés impliqués dans une collision de type i durant l'année j.

Modèle 5 : modèle avec effets aléatoires 'YEAR' et 'EES'

Le Modèle 5 a pour effets aléatoires l'EES et l'année de l'accident et pour effets fixes le type de collision, la localisation de l'accident, l'âge du conducteur ainsi que son sexe.

La probabilité $P(y_{ijk} = 1|v_i, w_j)$ qu'un véhicule accidenté k d'EES i durant l'année j contienne au moins un blessé grave est reliée aux variables explicatives et aux effets aléatoires par le modèle suivant :

$$\text{logit}(P(y_{ijk} = 1|v_i, w_j)) = (x_{vw})_{ijk}\beta_{vw}+v_i+w_j, i = 1, \ldots, 6 \quad j = 1, \ldots, 4 \quad k = 1, \ldots, n_{ij} \tag{5.17}$$

Où, $(x_{vw})_{ijk}$ est une matrice $[1 \times (1 + p)]_{p=13}$ des variables à effets fixes, β_{vw} le vecteur de taille $(1+p)_{p=13}$ des coefficients associés aux effets fixes à estimer , v_i et w_j deux scalaires désignant respectivement la i^{ème} composante de l'effet aléatoire V et la j^{ème} composante de l'effet aléatoire W tel que $\sum_{i=1}^{6}\sum_{j=1}^{4} n_{ij} = n = 826$ où n_{ij} désigne le nombre de véhicules accidentés d'EES i durant l'année j.

Modèle 6 : modèle avec effets aléatoires 'COL' et 'EES'

Le Modèle 6 a pour effets aléatoires le type de collision et l'EES et pour effets fixes l'année de l'accident, sa localisation, l'âge du conducteur ainsi que son sexe.

La probabilité $P(y_{ijk} = 1|u_i, v_j)$ qu'un véhicule accidenté k impliqué dans une collision de type i et d'EES j contienne au moins un blessé grave est reliée aux variables explicatives et aux effets aléatoires par le modèle suivant :

$$\text{logit}(P(y_{ijk} = 1|u_i, v_j)) = (x_{uv})_{ijk}\beta_{uv}+u_i+v_j, i = 1, \ldots, 6 \quad j = 1, \ldots, 6 \quad k = 1, \ldots, n_{ij} \tag{5.18}$$

Où, $(x_{uv})_{ijk}$ est une matrice $[1 \times (1 + p)]_{p=11}$ des variables à effets fixes, β_{uv} le vecteur de taille $(1 + p)_{p=11}$ des coefficients associés aux effets fixes à estimer , u_i et v_j deux scalaires désignant respectivement la i^{ème} composante de l'effet aléatoire U et la j^{ème} composante de l'effet aléatoire V tel que $\sum_{i=1}^{6}\sum_{j=1}^{6} n_{ij} = n = 826$ où n_{ij} désigne le nombre de véhicules accidentés impliqués dans une collision de type i et d'EES j.

5.5.3 Présentation du modèle à trois effets aléatoires

Le Modèle 7 a pour effets aléatoires le type de collision, l'EES et l'année de l'accident et pour effets fixes la localisation de l'accident, l'âge du conducteur ainsi que son sexe.

La probabilité$P(y_{ijkl} = 1|u_i, v_j, w_k)$ qu'un véhicule accidenté l impliqué dans une collision de type i, d'EES j et durant l'année k contienne au moins un blessé grave est reliée aux variables explicatives et aux effets aléatoires par le modèle suivant :

$$\text{logit}(P(y_{ijkl} = 1|u_i, v_j, w_k)) = (x_{uvw})_{ijkl}\beta_{uvw}+u_i+v_j+w_k, i = 1, \ldots, 6 \quad j = 1, \ldots, 6 \quad k = 1, \ldots, 4$$

Où, $(x_{uvw})_{ijkl}$ est une matrice $[1 \times (1 + p)]_{p=11}$ des variables à effets fixes, β_{uvw} le vecteur de taille $(1 + p)_{p=11}$ des coefficients associés aux effets fixes à estimer, u_i, v_j et w_k trois scalaires désignant respectivement la i^{ème} composante de l'effet aléatoire U, la j^{ème} composante de l'effet aléatoire V et la k^{ème} composante de l'effet aléatoire W tel que $\sum_{i=1}^{6}\sum_{j=1}^{6}\sum_{k=1}^{4} n_{ijk} = n = 826$ où n_{ijk} désigne le nombre de véhicules accidentés impliqués dans une collision de type i, d'EES j et durant l'année k.

5.5.4 Estimation

Nous estimons les modèles 3, 4, 5, 6 et 7 à l'aide de la méthode de Laplace. Le modèle 0 est également estimé à l'aide de l'algorithme de Newton. Les résultats des estimations de l'ensemble des modèles sont exposés dans les tables 5.8 et 5.9.

	Modèle 0			Modèle 1			Modèle 2			Modèle 3		
Variables	Paramètre estimé	Erreur standard		Paramètre estimé	Erreur standard		Paramètre estimé	Erreur standard		Paramètre estimé	Erreur standard	
Intercept	7,599	1,232	***	6,887	1,222	***	6,440	1,245	***	2,748	1,226	*
EES												
0-20	-8,491	1,184	***	-8,488	1,195	***	-8,448	1,190	***			
20-30	-6,745	1,099	***	-6,745	1,118	***	-6,681	1,104	***			
30-40	-5,665	1,076	***	-5,679	1,098	***	-5,625	1,083	***			
40-50	-4,355	1,061	***	-4,313	1,081	***	-4,372	1,071	***			
50-60	-3,311	1,066	**	-3,281	1,088	**	-3,338	1,077	**			
60-90	0,000						0,000					
Age												
16-24	-1,189	0,545	*	-1,134	0,536	*	-1,220	0,540	*	-1,165	0,539	*
25-34	-1,851	0,567	**	-1,756	0,551	**	-1,912	0,563	***	-1,811	0,560	**
35-44	-1,619	0,561	**	-1,525	0,549	**	-1,659	0,557	**	-1,599	0,556	**
45-54	-1,274	0,597	*	-1,225	0,585	*	-1,338	0,592	*	-1,266	0,591	*
55-64	-2,167	0,735	**	-2,119	0,706	**	-2,233	0,731	**	-2,169	0,731	**
65-74	-1,073	0,650	.	-1,010	0,637		-1,073	0,642	,	-1,051	0,644	
75+	0,000			0,000			0,000			0,000		
Col												
A-F	-0,891	0,776		-0,931	0,757					-0,897	0,769	
F-A	-2,118	0,879	*	-2,105	0,849	*				-2,119	0,875	*
F-F	-1,888	0,385	***	-1,902	0,372	***				-1,829	0,378	***
V-O	-0,637	0,396		-0,684	0,379	.				-0,605	0,390	
F-L	-1,739	0,418	***	-1,743	0,404	***				-1,727	0,414	***
L-F	0,000			0,000						0,000		
Year												
2005-2010	-0,726	0,539					-0,757	0,534		-0,723	0,535	
2000-2004	-0,877	0,414	*				-0,865	0,410	*	-0,885	0,409	*
1995-1999	-1,000	0,295	***				-0,988	0,292	***	-1,000	0,292	***
1991-1994	0,000						0,000			0,000		
Loc												
Inag	-0,727	0,318	*	-0,739	0,306	*	-0,701	0,315	*	-0,750	0,315	*
Hoag	0,000			0,000			0,000			0,000		
Gender												
F	-0,590	0,285	*	-0,613	0,278	*	-0,561	0,283	*	-0,589	0,282	*
M	0,000			0,000			0,000			0,000		
effet aléatoire	Aucun			Year			Col			EES		
écart type de la distribution de l'effet aléatoire				0.342			0.684			2.62		
Log-vraisemblance	-227			-230.8			-234.7			-242.3		

Table 5.8: Estimations des modèles 0, 1, 2 et 3

	Modèle 4			Modèle 5			Modèle 6			Modèle 7		
Variables	Paramètre estimé	Erreur standard		Paramètre estimé	Erreur standard		Paramètre estimé	Erreur standard		Paramètre estimé	Erreur standard	
Intercept	5,724	1,231	***	2,059	1,223	.	1,614	1,254		0,915	1,248	
EES												
0-20	-8,449	1,195	***									
20-30	-6,683	1,109	***									
30-40	-5,640	1,088	***									
40-50	-4,336	1,075	***									
50-60	-3,313	1,082	**									
60-90	0,000											
Age												
16-24	-1,170	0,538	*	-1,122	0,537	*	-1,192	0,534	*	-1,148	0,531	*
25-34	-1,828	0,557	**	-1,739	0,554	**	-1,863	0,555	***	-1,791	0,549	**
35-44	-1,579	0,552	**	-1,527	0,551	**	-1,634	0,551	**	-1,564	0,546	**
45-54	-1,292	0,590	*	-1,229	0,589	*	-1,319	0,586	*	-1,281	0,583	*
55-64	-2,186	0,727	**	-2,132	0,727	**	-2,220	0,726	**	-2,181	0,721	**
65-74	-1,020	0,639		-1,003	0,641		-1,054	0,636	.	-1,006	0,633	
75+	0,000			0,000			0,000			0,000		
Col				-0,928	0,764							
A-F				-2,110	0,866	*						
F-A				-1,841	0,376	***						
F-F				-0,642	0,387	.						
V-O				-1,728	0,412	***						
F-L				0,000								
L-F												
Year												
2005-2010							-0,750	0,529				
2000-2004							-0,874	0,404	*			
1995-1999							-0,989	0,288	***			
1991-1994							0,000					
Loc	-0,715	0,313	*	-0,760	0,313	*	-0,728	0,311	*	-0,739	0,309	*
Inag	0,000			0,000			0,000			0,000		
Hoag												
Gender												
F	-0,582	0,281	*	-0,606	0,281	*	-0,564	0,280	*	-0,582	0,278	*
M	0,000			0,000			0,000			0,000		
effets aléatoires	Col Year			EES Year			EES Col			EES Col Year		
écarts types des distributions des effets aléatoires	0.703 0.382			2.634 0.413			2.625 0.754			2.637 0.752 0.410		
Log-vraisemblance	-237.9			-245.4			-248.7			-251.8		

Table 5.9: Estimations des modèles 4, 5, 6 et 7

Tout d'abord, d'après ces tableaux, tout comme pour les modèles 1 et 2, nous pouvons voir que les estimations des effets fixes en commun sont relativement les mêmes pour chaque modèle. Les valeurs de la statistique z sont également très similaires pour chacun des modèles présentés. Il en résulte que ces modèles sont très semblables en ce qui concerne les effets fixes en commun. Nous voyons également que les estimations des variances des variables à effets aléatoires des différents modèles sont assez proches pour une même variable. Nous en concluons alors que tout comme pour les modèles 1, 2 et 3 les variances des variables à effets aléatoires des modèles 4, 5, 6 et 7 sont significativement différentes de zéro et ces derniers sont donc bien justifiés.

5.5.5 Performance de prédiction

Dans cette section, nous considérons tous les modèles. L'objectif est ici de savoir si un des modèles présentés parmi les sept (5.3) est plus performant que les autres dans sa capacité à distinguer les véhicules accidentés graves des non graves.
Un premier critère d'évaluation simple est le taux d'erreur. Il consiste à comparer les vraies valeurs observées de la variable à expliquer avec celles prédites par le modèle et comptabiliser le nombre de mauvais classements. Cependant, ce critère ne permet pas de rendre compte de la structure de l'erreur c'est-à-dire de la manière de se tromper. Dans bon nombre de problèmes et notamment le notre l'importance de se tromper est différente en fonction de la situation. En effet, prédire un accident comme étant non grave alors qu'il est en réalité grave est plus problématique que l'inverse. Pour cela, il est plus judicieux de construire des indicateurs tels que la sensibilité (taux de positifs bien reclassés par le modèle), la précision (proportion de vrais positifs parmi les classés positifs), la spécificité (taux de négatifs bien reclassés par le modèle) et le taux de faux positifs (taux de négatifs mal reclassés par le modèle). Étant donné que le modèle construit fournit une probabilité en sortie, il est nécessaire pour pouvoir construire ces indicateurs de définir un seuil de probabilité 's' (probabilité de césure) au dessus duquel le véhicule accidenté sera classé comme contenant des passagers gravement blessés. En l'absence de raisons particulières on prend généralement ce seuil égal à 0.5. Cependant, se baser uniquement sur les probabilités de prédictions est un peu gênant car que dire d'un véhicule ayant une probabilité prédite de 0.495 et un autre ayant une probabilité prédite de 0.505. Le premier sera classé comme ayant un accident non grave et le second comme ayant un accident grave alors que pourtant d'un point de vue probabilité il sont très proches. L'aire sous la courbe ROC (AUC) généralise alors l'idée de la matrice de confusion en faisant varier le seuil s sur l'ensemble des valeurs possibles, c'est-à-dire entre 0 et 1. Un des avantages de son utilisation est de permettre la comparaison de plusieurs types de modèles entre eux, tels des modèles de nature différente, imbriqués ou non, etc.

Un "bon" modèle doit présenter des valeurs faibles de taux d'erreur et de taux de faux positifs (proche de 0); des valeurs élevées de sensibilité, précision et spécificité. En règle générale, lorsqu'on améliore la sensibilité on dégrade souvent la précision et la spécificité, il faut donc surveiller ces indicateurs simultanément. De plus, ces deux indicateurs jouent un rôle important dans l'évaluation de tels modèles car tous deux ne dépendent pas du schéma d'échantillonnage. Même si l'échantillon n'est pas représentatif ces deux indicateurs n'en sont pas affectés. Enfin, un "bon" modèle doit l'être indépendamment de la probabilité de césure. L'AUC (aire sous la courbe ROC) permet de répondre à toutes ces exigences simultanément. En effet, l'AUC estime la probabilité que le modèle place un positif devant un négatif (dans le meilleur des cas AUC=1) pour toute probabilité de césure.

On estime le taux d'erreur (donc également le taux de bon reclassement), la sensibilité, la précision, la spécificité, le taux de faux positifs (table 5.10), ainsi que l'AUC (table 5.11) par validation croisée pour chacun des modèles proposés. Pour cela on coupe l'échantillon en huit sous échantillons, les estimations du modèle se font sur sept sous échantillons et le calcul d'efficacité du modèle se fait sur le huitième sous échantillon. Puis on alterne de sorte que tous les huit sous échantillons aient servis tour à tour à construire le modèle puis calculer son efficacité et éviter ainsi un phénomène de sur-apprentissage.

La table 3 présente le taux d'erreur, la sensibilité, la précision, la spécificité, le taux de faux positif ainsi que l'AUC estimés par validation croisée pour le modèle logistique mixte et le modèle logistique classique.
En regardant tout d'abord les taux d'erreur de chaque modèle on s'aperçoit que le modèle mixte 4 (ayant pour effets aléatoires la collision et l'année) et le modèle mixte 7 (ayant pour effets aléatoires la collision, l'année et l'EES) ont les taux d'erreur les plus faibles (13.47%). Ces modèles sont donc ceux qui se trompent le moins dans le reclassement, viennent ensuite les modèles mixtes 1 et 2 (ayant pour effets aléatoires respectifs la collision et l'EES) pas très loin derrière avec des taux d'erreur de 13.59% puis le modèle 3avec un taux d'erreur de 13.84%. Enfin, le modèle classique et les modèles mixtes 6 et 5 sont ceux qui se trompent le plus dans le reclassement des accidents avec des taux d'erreurs respectifs de 13,71%, 13.71% et 13.96%. Finalement, ces modèles ont globalement des taux d'erreurs assez proches puisque la différence entre celui qui se trompe le plus et celui qui se trompe le moins n'est que de 0.5%. En regardant ensuite les autres indicateurs de performance on peut voir néanmoins que chacun des modèles est différent dans son efficacité de prédire et/ou de se tromper.

Par rapport à la sensibilité on se rend compte que le modèle modèle mixte 2, ayant pour effet aléatoire l'EES, est celui qui détecte le plus d'accidents graves, il en détécte 51.85%. Le second modèle à détecter le mieux les accidents graves est le modèle classique, il en détecte 51.23% soit une différence de 0.62%. On remarque également que les modèles détectant le moins d'accidents graves font intervenir l'année comme variable à effet aléatoire et ce qu'il s'agisse de modèles à un, deux ou trois effets aléatoires.
Concernant la détection des accidents non graves (Spécificité), les modèle mixtes 4 et 7 sont ceux qui les détectent le mieux, ils en détectent respectivement 95.47% et 95.32%. En revanche, le modèle mixte 2 et le modèle classique sont ceux qui en détectent le moins, ils en détectent 94.86%, soit une différence de 0.61% avec le modèle 4. Pour rappel ces deux modèles sont ceux qui détectent le mieux les accidents graves.
Pour finir, en ne faisant aucune priorisation et en faisant varier la probabilité de césure entre 0 et 1 on peut dire que tous ces deux modèles possèdent une excellente qualité discriminante. En effet, les courbes ROC de ces modèles se croisent et se superposent quasiment d'une part et d'autre part les valeurs d'AUC (aire sous la courbe ROC) pour tous les modèles sont estimées proches de 1 et proches les unes des autres.
Après étude de ces différents indicateurs, on peut donc conclure qu'à un seuil de 0.5 le modèle mixte 2 (ayant l'EES pour effets aléatoires) est plus performant que le modèle classique dans la reconnaissance des accidents graves. Ces deux modèles ayant la même capacité à détecter les accidents non graves et dans un objectif de détecter au mieux les accidents graves, le modèle mixte 2 sera préféré au modèle classique et ce d'autant plus que la précision du modèle 2 est légèrement supérieure à celle du modèle classique (71.19% contre 70.94% soit 0.25% de plus).

Modèle	Effets aléatoires	Taux d'erreur	TVP ou sensibilté	Précision	Spécificité	TFP= 1-Spécifité
Modèle 0	aucun	13,71%	51,23%	70,94%	94,86%	5,14%
Modèle 1	Col	13,59%	50,62%	71,93%	95,17%	4,83%
Modèle 2	EES	13,59%	51,85%	71,19%	94,86%	5,14%
Modèle 3	Year	13,84%	50,00%	71,05%	95,02%	4,98%
Modèle 4	Col/Year	13,47%	50,62%	72,57%	95,32%	4,68%
Modèle 5	EES/Year	13,96%	49,38%	70,80%	95,02%	4,98%
Modèle 6	EES/Col	13,71%	50,62%	71,30%	95,02%	4,98%
Modèle 7	EES/Col/Year	13,47%	50,00%	72,97%	95,47%	4,53%

Table 5.10: Indicateurs de performance pour chaque modèle

Modèle	Effets aléatoires	AUC	Intervalle de confiance 95%	
			Borne inf.	Borne sup
Modèle 0	aucun	0,899	0,810	0,990
Modèle 1	Col	0,900	0,809	0,989
Modèle 2	EES	0,899	0,807	0,988
Modèle 3	Year	0,898	0,806	0,988
Modèle 4	Col/Year	0,901	0,810	0,989
Modèle 5	EES/Year	0,899	0,807	0,988
Modèle 6	EES/Col	0,900	0,809	0,989
Modèle 7	EES/Col/Year	0,900	0,809	0,988

Table 5.11: AUC pour chaque modèle

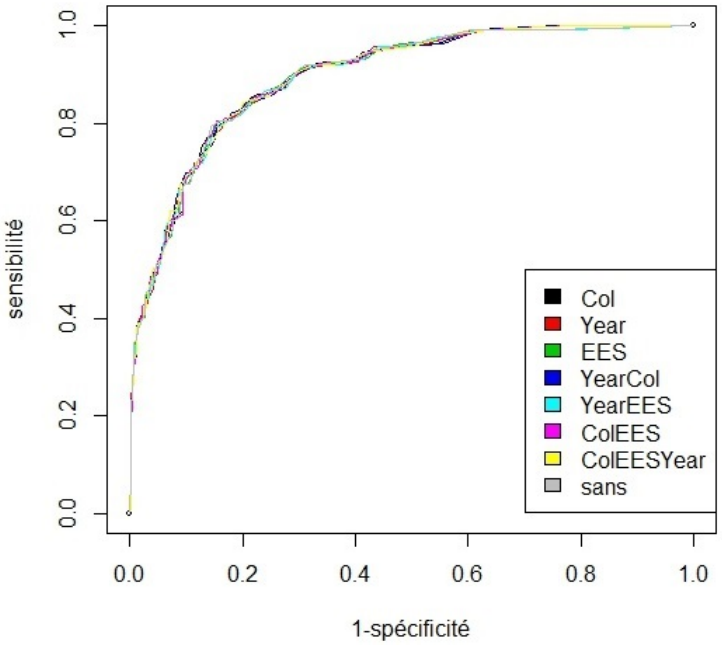

Figure 5.16: Courbes ROC

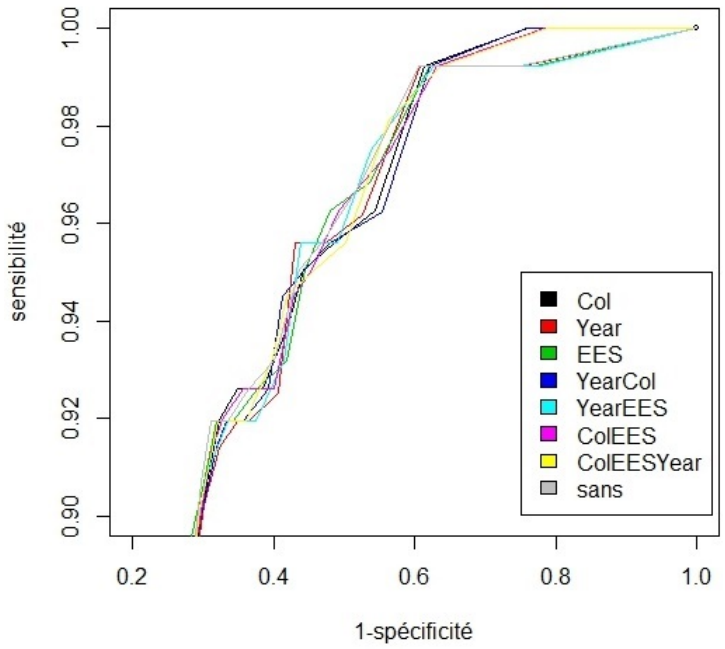

Figure 5.17: Zoom sur les courbes ROC

5.5.6 Synthèse des résultats et contributions

Ce dernier chapitre étant dense et contenant les principales contributions de la thèse, nous proposons dans cette section d'en faire une synthèse. Tout d'abord, dans une première contribution, nous avons construit un modèle logistique-Normal avec « Type de collision » comme variable à effet aléatoire pour analyser la gravité corporelle maximale observée dans un échantillon de véhicules accidentés. Des méthodes d'estimation fondées sur l'approximation de Laplace de la log-vraisemblance sont proposées pour estimer et analyser le modèle. Nous avons alors ensuite comparé, par simulation, cette approximation Laplacienne à celle basée sur l'adaptation des polynômes de Gauss-Hermite (AGH) ainsi qu'à la méthode de quasi-vraisemblance pénalisée. Nous montrons que les trois méthodes sont équivalentes par rapport à la précision de l'estimation bien qu'AGH soit légèrement supérieure. Une deuxième contribution a consisté à la construction d'un deuxième modèle logistique-Normal avec « EES» comme variable à effet aléatoire. Là également des simulations ont été menées afin de comparer les trois méthodes d'estimation. Contrairement au premier modèle les trois méthodes sont équivalentes seulement en ce qui concerne l'estimation des effets fixes. Concernant l'estimation de l'écart-type de la variable à effets aléatoires la méthode PQL est

celle qui donne les résultats les plus biaisés tandis que la méthode AGH est celle qui fournit les estimations les plus précises. Ces deux modèles se distinguant principalement par la valeur de l'écart-type de la variable à effets aléatoires. En effet, dans le premier modèle celle-ci est inférieure à un tandis que dans le second elle est supérieure à un. Dans une troisième et dense contribution, nous avons identifié plusieurs modèles logistique-normaux mixtes avec plus d'un effet aléatoire et avons analysé la performance de ces modèles à détecter des véhicules contenant des usagers ayant des blessures graves. Cette analyse de performance s'est également faite comparativement au modèle classique sans effets aléatoires. Il en a alors résulté que tous les modèles logistique-normaux mixtes étaient plus performant que le modèle classique dans la reconnaissance des accidents graves mais étaient moins performant dans la reconnaissance des accidents non graves et qu'ils avaient une légère tendance à surestimer la gravité des accidents par rapport au modèle classique.

Conclusions et perspectives

Dans ce travail, nous avons proposé de modéliser la gravité des accidents suivant un modèle particulier : le modèle logistique mixte-normal. L'utilisation de ce type de modèles vient du fait que la présence d'une hétérogénéité inobservée entre les observations est susceptible d'entraîner une variation des paramètres selon les observations mais également du fait que la variable à modéliser est de type binaire.

Tout d'abord les variables "'type de collision"', "'EES"' et "'année de l'accident"' ont été trouvées comme pouvant être modélisées par des variables à effets aléatoires tandis que les variables "'sexe du conducteur"', "'age du conducteur"' et "'localisation de l'accident"' ont été trouvées comme pouvant être modélisées par des variables à effets fixes. Ces six variables sont donc les variables explicatives de la gravité corporelle maximale des accidents de la route. Sept modèles ont alors fait suite à cette sélection de variables : trois modèles à une variable à effets aléatoires, trois modèles à deux variables à effets aléatoires et un modèle à trois variables à effets aléatoires.

Ensuite, concernant l'estimation des modèles logistiques mixtes-normaux, et ce aprés plusieurs simulations, nous avons pu constater pour des modèles impliquant une variable à effets aléatoires que, lorsque l'écart-type de la variable à effets aléatoires était inférieur à un les trois méthodes se valaient avec une légère supériorité pour AGH puis Laplace ce qui mène à recommander l'utilisation de la méthode de Laplace. Cependant, lorsque l'écart-type de la variable à effets aléatoires est supérieure à deux nous recommandons l'utilisation de la méthode AGH car plus précise.

L'étude des différents modèles construits a également permis de montrer et ainsi confirmer certains résultats de la littérature accidentologique. Notamment que plus l'EES est élevé plus les accidents sont sévères, que les collisions latéro-frontales sont les plus dangereuses en considérant tous les autres paramètres égaux par ailleurs, que les véhicules conduits par les jeunes (16-24) et les personnes agées (plus de 75 ans) ont plus de chances d'être impliqués dans des accidents graves que les autres

catégories d'âge, que les véhicules conduits par les femmes sont moins dangereux que ceux conduits par les hommes, qu'il y a généralement moins d'accidents graves dans les zones urbanisées comparées aux zones rurales et enfin qu'au fil des années la gravité des accidents diminue et ce, pour toutes les assertions, en supposant tous les autres paramètres égaux par ailleurs.

Enfin, à travers l'étude de différents indicateurs de performance, nous avons également montré dans ce travail que le modèle logistique mixte-normal avec pour variable à effets aléatoires l'EES est plus apte à détecter les accidents graves à un seuil de 0.5 comparativement aux autres modèles mixtes et notamment par rapport au modèle logistique classique.

Aussi, après l'étude des différents indicateurs de performance vient la question de la sélection du 'meilleur' modèle. Les critères que nous avons utilisé évaluent le pouvoir prédictif des modèles, or ce pouvoir prédictif ne doit pas être le seul critère à être pris en compte pour la sélection du 'meilleur' modèle. En effet, un compromis entre biais (qui diminue avec le nombre de paramètres) et parcimonie (nécessité de décrire les données avec le plus petit nombre de paramètres possible) doit être trouvé. Cette recherche d'équilibre entre biais et parcimonie se fait habituellement à l'aide de critères probabilistes tels que l'AIC ou le BIC. Cependant l'utilisation de ces critères est confrontée ici et de façon générale dans les GL2M au problème de la non accessibilité à la vraisemblance marginale. Nous ne pouvons donc pas en toute rigueur appliquer directement ces critères dans le cadre de tels modèles. Aussi, des critères tels que le CAIC ont été récemment développés pour la sélection de GL2M (Donohue et al., 2011; Yu and Yau, 2012).

De plus, dans ce travail la gravité a été modélisée à l'aide d'une variable dichotomique valant 1 s'il y a au moins une personne blessée avec un MAIS supérieur ou égale à 3 dans le véhicule étudié et 0 sinon. Comme l'indice AIS, et donc l'indice MAIS, est une échelle de gravité allant de 1 à 6 voire plus, une extension de ce travail serait d'étendre l'analyse des données par le biais d'une approche multivariée afin de mieux prendre en compte la structure multi-classes de la variable à modéliser 'MAIS'.

Egalement, concernant la comparaison des différentes méthodes d'estimation, celle-ci n'a été faite que pour le cas du modèle à une variable à effets aléatoires. Il serait intéressant d'établir cette comparaison pour des modèles ayant plus d'une variable à effets aléatoires et voir si les conclusions trouvées restent inchangées.

Pour finir, la modélisation des variables à effets aléatoires s'est faite à l'aide d'une

distribution normale comme c'est le plus souvent le cas. Dans de futurs travaux nous pourrions chercher à savoir si d'autres distributions conviennent et voir quels en sont les résultats.

Appendix A

Approximation de Laplace à des ordres supérieurs

A.1 Notation vec

La notation **vec(A)** réfère à la vectorisation de la matrice A. La vectorisation de la matrice A de taille $n \times m$ est la transformation linéaire qui convertit la matrice A en un vecteur colonne composé de $n \times m$ tels que:

$$vec(A) = [a_{1,1}, \ldots, a_{m,1}, a_{1,2}, \ldots, a_{m,2}, \ldots, a_{1,n}, \ldots, a_{m,n}^T]$$

où a_{ij} représente le $(i,j)^{\text{ième}}$ élément de la matrice A c'est à dire l'élement se trouvant à la $i^{\text{ème}}$ ligne et à la $j^{\text{ième}}$. colonne.

A.2 Calcul du terme de correction

Comme $T_k = \frac{1}{k!}[\otimes^{k-1}(t-\hat{t})^T]l^{(k)}(\hat{t})(t-\hat{t})$,

$$\begin{aligned} E(T_k) &= E\Big[\frac{1}{k!}\Big[\otimes^{k-1}(t-\hat{t})^T\Big]l^{(k)}(\hat{t})(t-\hat{t})\Big] \\ &= \frac{1}{k!}vec^T E\Big\{\Big[\otimes^{k-1}(t-\hat{t})^T\Big\}(t-\hat{t})\Big]vec\{l^{(k)}(\hat{t})\} \end{aligned}$$

(voir Magnus and Neudecker (1988), p. 30, eq. (3)).

Ensuite, $E\{[\otimes^{k-1}(t-\hat{t})^T](t-\hat{t})\} = \mu_{(k)}$ est le kième moment centré de la distribution normale multivariée de dimension q avec pour matrice de variance-covariance $V = [-l^{(2)}(\hat{t})]^{-1}$. Pour k impaire $\mu_{(k)} = 0$. En utilisant la fonction génératrice des moments pour une distribution normale multivariée, Yang (1998) a montré

que pour k paire, $\mu_{(k)}$ est composé de commutations de produits de Kronecker de $(k-1)(k-3)\ldots 3$ matrices ayant la forme V ou $vecV$. Cependant, l'intérêt ne porte pas sur $\mu_{(k)}$ lui-même mais sur le scalaire $vec^T(\mu_{(k)})vec(S)$, où S est toute matrice conforme de constantes. Yang (1998) a montré que les commutations pouvaient être ignorées car l'intérêt porte sur la trace qui est un scalaire. Par exemple,

$$E(T_4) = \frac{1}{4!}vec^T(\mu_{(4)}vec(l^{(4)}(\hat{t})) = \frac{3}{4!}vec^T\{V \otimes V\}vec\big[l^{(4)}(\hat{t})\big]$$

en dépit du fait que

$$\mu_{(4)} = (V \otimes vecV) + (vecV \otimes V) + (K_{qq} \otimes I_q)(V \otimes vecV)$$

où K_{qq} est la matrice de commutation de taille $q^2 \times q^2$ (Magnus and Neudecker, 1988) et I_q est la matrice identité de dimension $q \times q$. En général, alors, $vec^T(\mu_{(k)})vec(S) = (k-1)(k-3)\ldots 3\ vec^T\Big(\otimes^{\frac{k}{2}} V\Big)vec(S)$. La substitution $S = vec(l^{(t)}(\hat{t})$ produit l'équation (3.7). La démonstration pour $E(T_kT_m)$ est similaire.

A.3 Vecteurs scores

L'approximation de Fisher Scoring requiert seulement les dérivées premières, évitant ainsi le calcul fastidieux des dérivées secondes. L'algorithme de Fisher Scoring permet de trouver itérativement $\Theta = (\beta^T, \theta^T)^T$ (où $\theta = (\sigma_1, \ldots, \sigma_K)$) en utilisant l'équation $\Theta^{\text{new}} - \Theta^{\text{old}} = \sum_{j=1}^{K}(S_jE_j^T)^{-1}S_j$, où $S_j = (S_{\beta_j}^T, S_{\sigma_j}^T)^T$ est le vecteur score du $j^{\text{ième}}$ effet aléatoire, S_β et $S_{\sigma_1} \ldots, S_{\sigma_K}$ étant les dérivées de la log-vraisemblance marginal (3.14) par rapport à β, $\sigma_1, \ldots, \sigma_K$ respectivement.

Obtenant ces dérivées, il est néanmoins nécessaire de prendre en considération le fait que $h_j(\hat{u}_j$ est évaluée à $\hat{u} = \hat{u}(\beta, \sigma_1, \ldots, \sigma_K) = V_jZ_j^TW_j(y_j^* - X_j\beta)$. Cette interdépendance est résolue avec la différenciation implicite de $\hat{u}_j$ par rapport à β et $\sigma_1, \ldots, \sigma_K$, respectivement :

$$\frac{\partial \hat{u}_j}{\beta^T} = -V_jZ_j^TW_jX_j \tag{A.1}$$

$$\frac{\partial \hat{u}_j}{\sigma_j} = [(y_j^* - \eta_j)^TW_jZ_j \otimes V_j\sigma_j^{-2}A_j^{-1}] \tag{A.2}$$

Après l'utilisation répétée de matrice algébrique dans Magnus and Neudecker (1988), les vecteurs scores (Yang, 1998) peuvent alors s'écrire

$$S_{\beta_j} = X_j^TW_j(y_j^* - X_j\beta - Z_j\hat{u}_j) + \sum_i^{n_j} f_{ij}v_{ij} + \frac{1}{M_j}\sum_i^{n_j} c_{ij}v_{ij} \tag{A.3}$$

et

$$S_{\sigma_j} = \frac{1}{2}E^T[\sigma_j^{-2}A_j^{-1}(\hat{D}_j\sigma_j^{-2} - \sigma_j^2 A_j)\sigma_j^{-2}A_j^{-1}] - \frac{1}{2}\sum_i^{n_j} m_{ij}^{(3)} B_{ij} vec[Q_{ij}] + \frac{1}{M_j}\left\{ \sum_i^{n_j} c_{ij}E^T vec[Q_{ij}] + E^T \left[\right.\right.$$

où $f_{ij} = -1/2 m_{ij}^{(3)} B_{ij}$, $G_j = X_j^T W_j Z_j$, $v_{ij} = X_{ij} - G_j Z_{ij}$, et $c_{ij} = -1/8 m_{ij}^{(5)} B_{ij}^2 + 1/4 m_{ij}^{(3)} Z_{ij}^T H_j Z_{ij} - 1/48 m_{ij}^{(7)} \beta_{ij}^3 + 1/16 m_{ij}^{(3)} Z_{ij}^T p_j Z_{ij} + 15/72 m_{ij}^{(3)} a_{ij}^2 - 15/36 m_{ij}^{(3)} Z_{ij}^T h_j Z_{ij}$, avec

$$\begin{aligned}
H_j &= \sum_i^{n_j} B_{ij} m_{ij}^{(4)} V_j Z_{ij} Z_{ij}^T V_j \\
m_{ij}^{(7)} &= m_{ij}^{(5)}(1 - 12 w_{ij}) - 36 m_{ij}^{(3)} m_{ij}^{(4)} \\
p_j &= \sum_i^{n_j} m_{ij}^{(6)} B_{ij}^2 V_j Z_{ij} Z_{ij}^T V_j \\
a_{ij} &= Z_{ij}^T k_j \\
h_j &= \sum_i^{n_j} m_{ij}^{(3)} a_{ij} V_j Z_{ij} Z_{ij}^T V_j \\
\hat{D}_j &= \hat{u}_j \hat{u}_j^T + V_j \\
E &= \frac{\partial \sigma_j^2 A_j}{\partial \sigma_j} \\
Q_{ij} &= \sigma_j^{-2} A_j^{-1} V_j Z_{ij} (y_j^* - \eta_j)^T W_j Z_j \\
F_{ij} &= \sigma_j^{-2} A_j^{-1} V_j Z_{ij} Z_{ij}^T V_j \sigma_j^{-2} A_j^{-1}
\end{aligned}$$

Les calculs pour les modèles de Poisson-normal et gamma-normal sont plus simples que ceux du cas binomial-normal car les expressions des dérivées $m_{ij}^{(k)}$ sont plus simples (voir (3.10)).

Appendix B

Approximations de la vraisemblance

B.1 Approximation de Laplace

Considérons l'approximation de la vraisemblance pour la contribution du $i^{\text{ème}}$ groupe

$$\int \Phi(u_i) \prod_{j=1}^{n_i} \frac{e^{y_{ij}(X_{ij}^T\beta+\sigma u_i)}}{1+e^{X_{ij}^T\beta+\sigma u_i}}\, \mathrm{d}u_i \tag{B.1}$$

où $\Phi(.)$ la fonction de densité de la loi normale standard. L'intégrale que nous voulons évaluer par l'approximation de Laplace s'écrit

$$\int \exp\{h(u_i,\theta;y_i)\}\, \mathrm{d}u_i$$

où $\theta = (\beta^T,\sigma)^T$ et $h(u_i,\theta;y_i) = \log\Big(\Phi(u_i)\prod_{j=1}^{n_i}\frac{e^{y_{ij}(X_{ij}^T\beta+\sigma u_i)}}{1+e^{X_{ij}^T\beta+\sigma u_i}}\Big)$. On suppose que $h(u_i,\theta;y_i)$ est une fonction lisse, bornée et unimodale par rapport à u_i. En remplaçant $h(u_i,\theta;y_i)$ par son développement de Taylor au second degré au voisinage de son maximum $\hat{u}_i$ of $h(u_i;\theta;y_i)$ nous obtenons

$$\int \exp\{h(u_i,\theta;y_i)\} \simeq \exp\{h(\hat{u}_i,\theta;y_i)\} \int \exp\{-\frac{1}{2}\Big[\Big(\frac{u_i-\hat{u}_i}{\hat{\tau}_i(\theta)}\Big)^2\Big]\}\, \mathrm{d}u_i$$

où $\hat{u}_i = \arg\max_{u_i} h(u_i,\theta;y_i)$ et $\hat{\tau}_i(\theta) = \Big[-\frac{\partial^2 h(u_i,\theta;y_i)}{\partial^2 u_i}|_{u=\hat{u}_i}\Big]^{-\frac{1}{2}}$. Après calculs, nous avons que $\hat{\tau}_i(\theta) = \Big(1+\sigma^2\sum_{j=1}^{n_i}\frac{e^{X_{ij}^T\beta+\sigma\hat{u}_i}}{\big(1+e^{X_{ij}^T\beta+\sigma\hat{u}_i}\big)^2}\Big)^{-\frac{1}{2}}$. Ainsi, l'integrale (B.1) s'écrit

$$\int \exp\{h(u_i,\theta;y_i)\} \simeq \sqrt{2\pi}\hat{\tau}_i(\theta)\exp\{h(\hat{u}_i,\theta;y_i)\}$$

De cette dernière expression nous pouvons en déduire l'approximation de Laplace de (5.5).

B.2 Approximation de Gauss-Hermite adaptative

Nous montrons que

$$\int \Phi(u_i) \prod_{j=1}^{n_i} \frac{e^{y_{ij}(X_{ij}^T\beta+\sigma u_i)}}{1+e^{X_{ij}^T\beta+\sigma u_i}} \, \mathrm{d}u_i = \int f(v_i) e^{-v_i^2} \, \mathrm{d}v_i$$

where $f(v_i) = \frac{1}{\sqrt{\pi}} \prod_{j=1}^{n_i} \frac{e^{y_{ij}(X_{ij}^T\beta+\sqrt{2}\sigma v_i)}}{1+e^{X_{ij}^T\beta+\sqrt{2}\sigma v_i}}$. Ainsi, en appliquant la méthode de Gauss-Hermite adaptative suggérée par (Pinheiro and Bates, 1995), nous obtenons

$$\begin{aligned} \int f(v_i) e^{-v_i^2} \, \mathrm{d}v_i &= \int \frac{f(v_i) e^{-v_i^2}}{\phi(v_i; \hat{v}_i, \hat{\tau}_i(\theta))} \times \phi(v_i; \hat{v}_i, \hat{\tau}_i(\theta)) \, \mathrm{d}v_i \\ &\simeq \hat{\tau}_i(\theta) \sum_{m=1}^{NQ} w_{im} \exp(t_{im}^2) g(\hat{\mu}_i + \hat{\tau}_i(\theta) t_{im}), \end{aligned}$$

où $\phi(v_i; \hat{v}_i, \hat{\tau}_i(\theta))$ est la densité de probabilté pour une loi normale de moyenne $\hat{\mu}_i$ et écart-type $\hat{\tau}_i(\theta)$, et $g(t) = f(t)e^{-t^2}$.

Appendix C

Tables

C.1 Modèle 1 ayant la collision pour variable à effets aléatoires

		Laplace (73)			AGH15(58)			AGH30(76)			PQL(64)		
Parameter	**TV**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	6,440	4,983	6,359	8,797	4,313	6,269	8,388	4,426	6,565	9,160	3,790	6,342	9,119
EES													
0-20	-8,448	-10,800	-8,390	-6,663	-11,173	-8,575	-6,529	-11,044	-8,672	-7,087	-10,491	-8,469	-6,393
20-30	-6,681	-8,261	-6,529	-5,296	-8,320	-6,701	-5,033	-8,211	-6,730	-5,327	-7,902	-6,623	-4,802
30-40	-5,625	-7,723	-5,415	-4,289	-7,209	-5,547	-3,496	-7,177	-5,691	-4,343	-6,680	-5,499	-3,931
40-50	-4,372	-5,703	-4,113	-2,723	-5,921	-4,226	-2,628	-6,210	-4,366	-2,923	-5,394	-4,195	-2,356
50-60	-3,338	-4,748	-3,070	-1,675	-4,899	-3,167	-1,281	-5,434	-3,349	-1,781	-4,412	-3,152	-1,439
Age													
16-24	-1,220	-3,073	-1,390	-0,201	-2,443	-1,108	0,511	-2,666	-1,361	0,360	-2,793	-1,242	0,625
25-34	-1,912	-3,524	-2,044	-0,751	-3,365	-1,856	-0,411	-3,612	-2,094	-0,502	-3,781	-1,965	-0,810
35-44	-1,659	-3,397	-1,802	-0,202	-3,166	-1,740	0,099	-3,224	-1,831	-0,416	-3,657	-1,705	-0,379
45-54	-1,338	-4,159	-1,493	0,122	-2,814	-1,243	0,206	-3,232	-1,504	0,216	-3,227	-1,369	0,673
55-64	-2,233	-4,283	-2,373	-1,037	-4,948	-2,262	-0,241	-4,703	-2,472	-0,480	-4,737	-2,342	-0,648
65-74	-1,073	-2,771	-1,225	0,056	-2,625	-0,970	0,675	-3,107	-1,150	0,593	-4,137	-1,056	0,374
Year													
1995-1999	-0,988	-2,090	-1,000	-0,306	-2,050	-1,065	-0,205	-2,050	-1,086	-0,257	-1,646	-0,969	-0,316
2000-2004	-0,865	-2,323	-0,900	0,011	-1,717	-0,944	-0,078	-1,820	-1,014	-0,259	-1,592	-0,882	0,001
2005-2010	-0,757	-2,997	-0,800	0,322	-2,385	-0,847	0,472	-3,142	-0,920	0,838	-2,327	-0,886	0,272
Loc													
Inag	-0,701	-1,633	-0,714	-0,059	-1,765	-0,725	-0,146	-1,845	-0,720	-0,039	-1,470	-0,749	0,060
Gender													
F	-0,561	-1,186	-0,586	0,152	-1,236	-0,500	0,139	-1,313	-0,573	0,206	-1,761	-0,548	0,068
SD	0,684	0,000	0,539	1,130	0,000	0,636	1,540	0,000	0,559	1,332	0,000	0,572	1,014

TV: True Values; SD: Standard Deviation

Table C.1: M1: Mean estimates for n=800 and 100 replications

		Laplace(72)			AGH15(76)			AGH30(78)			PQL(83)		
Parameter	**TV**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	6,440	4,769	6,444	8,493	5,174	6,433	7,995	5,030	6,271	7,652	4,232	6,317	7,784
EES													
0-20	-8,448	-11,267	-8,643	-6,824	-10,048	-8,533	-7,221	-11,245	-8,549	-7,164	-9,762	-8,421	-7,211
20-30	-6,681	-9,262	-6,751	-5,445	-8,263	-6,735	-5,772	-7,585	-6,680	-5,242	-8,305	-6,691	-5,687
30-40	-5,625	-7,702	-5,651	-4,375	-7,267	-5,657	-4,633	-6,547	-5,627	-4,230	-7,023	-5,594	-4,477
40-50	-4,372	-6,214	-4,402	-3,019	-6,015	-4,394	-3,273	-5,462	-4,351	-3,074	-5,846	-4,317	-3,268
50-60	-3,338	-5,339	-3,315	-1,905	-4,754	-3,350	-2,357	-4,301	-3,305	-1,978	-4,822	-3,301	-2,225
Age													
16-24	-1,220	-2,651	-1,266	-0,433	-1,913	-1,160	-0,370	-2,248	-1,129	-0,435	-2,262	-1,186	-0,288
25-34	-1,912	-2,929	-1,973	-0,916	-2,775	-1,876	-0,891	-3,419	-1,870	-0,893	-2,683	-1,881	-0,866
35-44	-1,659	-2,550	-1,671	-0,698	-2,647	-1,609	-0,770	-2,554	-1,595	-0,810	-2,592	-1,655	-0,808
45-54	-1,338	-2,135	-1,393	-0,550	-2,146	-1,294	-0,169	-2,321	-1,283	-0,227	-2,647	-1,286	-0,131
55-64	-2,233	-3,460	-2,258	-1,181	-3,577	-2,193	-1,183	-3,406	-2,119	-0,960	-3,135	-2,175	-0,670
65-74	-1,073	-2,552	-1,092	-0,334	-2,356	-1,042	-0,006	-2,253	-1,044	-0,048	-2,399	-1,034	0,159
Year													
1995-1999	-0,988	-1,365	-1,019	-0,660	-1,523	-1,026	-0,556	-1,335	-0,961	-0,391	-1,348	-0,961	-0,437
2000-2004	-0,865	-1,424	-0,880	-0,321	-1,581	-0,938	-0,461	-1,560	-0,839	0,089	-1,410	-0,859	-0,442
2005-2010	-0,757	-1,611	-0,815	0,035	-1,952	-0,828	-0,203	-1,714	-0,705	0,044	-1,653	-0,761	-0,023
Loc													
Inag	-0,701	-1,037	-0,712	-0,231	-1,058	-0,690	-0,246	-1,172	-0,742	-0,374	-1,180	-0,726	-0,414
Gender													
F	-0,561	-1,068	-0,553	-0,075	-1,034	-0,530	-0,050	-1,284	-0,594	-0,255	-1,082	-0,581	-0,009
SD	0,684	0,000	0,594	1,159	0,130	0,658	1,126	0,082	0,577	1,064	0,005	0,595	1,580

TV: True Values; SD: Standard Deviation

Table C.2: M1: Mean estimates for n=2000 and 100 replications

		Laplace(97)			AGH15(96)			AGH30(94)			PQL(96)		
Parameter	**TV**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	6,440	5,128	6,456	7,806	5,539	6,560	8,207	5,266	6,487	7,923	5,041	6,481	7,947
EES													
0-20	-8,448	-9,867	-8,557	-7,422	-10,231	-8,534	-7,624	-10,024	-8,536	-7,742	-10,041	-8,525	-7,538
20-30	-6,681	-8,002	-6,756	-5,738	-8,277	-6,799	-5,948	-8,245	-6,704	-5,870	-7,831	-6,737	-5,713
30-40	-5,625	-6,792	-5,685	-4,834	-7,314	-5,700	-4,953	-7,245	-5,690	-4,852	-7,023	-5,678	-4,709
40-50	-4,372	-5,483	-4,408	-3,435	-5,959	-4,458	-3,598	-5,884	-4,405	-3,694	-5,628	-4,432	-3,421
50-60	-3,338	-4,386	-3,367	-2,650	-4,966	-3,389	-2,587	-4,927	-3,363	-2,527	-4,665	-3,377	-2,259
Age													
16-24	-1,220	-1,900	-1,210	-0,615	-1,870	-1,251	-0,730	-1,836	-1,260	-0,807	-1,687	-1,222	-0,475
25-34	-1,912	-2,515	-1,903	-1,304	-2,480	-1,928	-1,405	-2,625	-1,953	-1,396	-2,458	-1,910	-1,202
35-44	-1,659	-2,370	-1,663	-1,087	-2,308	-1,678	-1,157	-2,284	-1,695	-1,226	-2,260	-1,660	-0,896
45-54	-1,338	-1,889	-1,350	-0,583	-1,935	-1,362	-0,721	-1,921	-1,377	-0,844	-1,921	-1,346	-0,690
55-64	-2,233	-3,157	-2,250	-1,514	-2,942	-2,287	-1,734	-3,143	-2,273	-1,616	-2,852	-2,244	-1,214
65-74	-1,073	-1,669	-1,091	-0,439	-1,694	-1,083	-0,560	-1,886	-1,119	-0,241	-1,766	-1,062	-0,414
Year													
1995-1999	-0,988	-1,268	-0,969	-0,707	-1,444	-1,007	-0,670	-1,277	-0,981	-0,611	-1,393	-1,017	-0,731
2000-2004	-0,865	-1,286	-0,854	-0,413	-1,276	-0,895	-0,448	-1,221	-0,878	-0,469	-1,191	-0,865	-0,495
2005-2010	-0,757	-1,458	-0,764	-0,308	-1,772	-0,777	-0,265	-1,369	-0,759	-0,288	-1,310	-0,787	-0,320
Loc													
Inag	-0,701	-0,994	-0,703	-0,339	-0,981	-0,684	-0,414	-0,954	-0,708	-0,409	-1,131	-0,715	-0,481
Gender													
F	-0,561	-0,779	-0,567	-0,235	-0,861	-0,585	-0,226	-0,763	-0,569	-0,290	-0,913	-0,547	-0,343
SD	0,684	0,194	0,583	1,224	0,104	0,624	1,022	0,243	0,617	1,112	0,139	0,581	1,114

TV: True Values; SD: Standard Deviation

Table C.3: M1: Mean estimates for n=5000 and 100 replications

		Laplace(325)			AGH15(320)			AGH30(344)			PQL(342)		
Parameter	**TV**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	6,440	3,457	6,485	9,440	3,571	6,342	9,004	3,582	6,349	8,741	4,097	6,314	9,696
EES													
0-20	-8,448	-10,993	-8,625	-6,730	-10,898	-8,527	-6,549	-11,124	-8,448	-6,009	-11,601	-8,519	-6,150
20-30	-6,681	-8,703	-6,749	-4,743	-10,163	-6,718	-4,764	-9,146	-6,697	-4,330	-10,261	-6,647	-4,643
30-40	-5,625	-7,553	-5,673	-3,816	-7,747	-5,579	-3,654	-7,573	-5,577	-3,595	-7,188	-5,543	-3,517
40-50	-4,372	-6,285	-4,366	-2,404	-6,565	-4,260	-2,593	-6,053	-4,307	-2,500	-5,960	-4,251	-2,316
50-60	-3,338	-4,942	-3,291	-1,149	-5,057	-3,159	-1,496	-5,375	-3,196	-1,245	-4,805	-3,168	-1,392
Age													
16-24	-1,220	-2,854	-1,212	0,505	-3,189	-1,219	0,612	-2,796	-1,224	0,634	-3,295	-1,200	0,873
25-34	-1,912	-3,823	-1,947	-0,178	-4,061	-1,931	-0,023	-3,870	-1,940	-0,166	-4,459	-1,930	0,478
35-44	-1,659	-3,489	-1,716	0,013	-3,912	-1,700	0,332	-3,431	-1,662	0,371	-3,950	-1,693	0,465
45-54	-1,338	-3,395	-1,386	0,832	-3,513	-1,362	0,484	-3,213	-1,312	0,974	-3,082	-1,334	1,130
55-64	-2,233	-4,699	-2,340	-0,098	-4,891	-2,265	0,105	-4,577	-2,302	-0,546	-4,554	-2,316	0,634
65-74	-1,073	-2,957	-1,074	0,837	-3,414	-1,034	1,322	-2,912	-1,052	0,737	-3,652	-1,102	1,569
Year													
1995-1999	-0,988	-1,951	-1,071	0,274	-2,261	-1,034	0,283	-1,823	-1,028	-0,155	-2,064	-0,972	-0,103
2000-2004	-0,865	-2,489	-0,939	0,117	-2,057	-0,921	0,326	-1,971	-0,911	0,312	-2,477	-0,839	0,194
2005-2010	-0,757	-3,461	-0,902	0,520	-4,762	-0,890	1,168	-3,732	-0,814	0,638	-3,908	-0,753	0,918
Loc													
Inag	-0,701	-1,675	-0,742	-0,002	-2,048	-0,740	0,058	-1,626	-0,740	-0,013	-1,528	-0,732	0,016
Gender													
F	-0,561	-1,642	-0,601	0,128	-1,622	-0,617	0,379	-1,508	-0,561	0,377	-1,618	-0,600	0,285
SD	0,684	0,000	0,602	1,417	0,000	0,594	1,477	0,000	0,579	1,769	0,000	0,582	1,419

Table C.4: M1: Mean estimates for n=800 and 500 replications

		Laplace(411)			AGH15(398)			AGH30(392)			PQL(414)		
Parameter	**TV**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	6,440	4,741	6,431	8,441	4,724	6,438	8,139	4,670	6,379	8,577	4,149	6,311	8,198
EES													
0-20	-8,448	-10,975	-8,522	-7,020	-10,396	-8,516	-7,190	-11,090	-8,478	-7,059	-10,867	-8,409	-6,666
20-30	-6,681	-8,479	-6,710	-5,258	-8,317	-6,709	-5,490	-8,438	-6,676	-5,384	-8,547	-6,623	-5,106
30-40	-5,625	-7,486	-5,613	-4,220	-7,262	-5,641	-4,429	-7,126	-5,593	-4,400	-7,164	-5,545	-4,158
40-50	-4,372	-6,360	-4,359	-2,852	-6,040	-4,351	-3,165	-6,018	-4,329	-3,158	-5,692	-4,284	-3,015
50-60	-3,338	-5,319	-3,299	-2,061	-4,867	-3,307	-2,202	-4,870	-3,265	-2,101	-4,729	-3,246	-1,870
Age													
16-24	-1,220	-2,466	-1,238	-0,288	-2,734	-1,254	0,009	-2,248	-1,202	-0,135	-2,715	-1,233	0,136
25-34	-1,912	-3,096	-1,950	-0,830	-3,326	-1,965	-0,705	-3,277	-1,918	-0,753	-3,137	-1,928	-0,845
35-44	-1,659	-3,057	-1,689	-0,697	-3,070	-1,712	-0,451	-2,663	-1,659	-0,605	-2,652	-1,663	-0,584
45-54	-1,338	-2,674	-1,345	-0,379	-2,855	-1,363	0,010	-2,486	-1,324	-0,271	-2,708	-1,365	-0,015
55-64	-2,233	-3,929	-2,251	-0,911	-4,070	-2,315	-0,913	-3,583	-2,245	-0,777	-3,604	-2,260	-0,932
65-74	-1,073	-2,535	-1,077	0,323	-2,708	-1,117	0,231	-2,299	-1,061	0,007	-2,570	-1,098	0,159
Year													
1995-1999	-0,988	-1,491	-1,000	-0,367	-1,814	-0,996	-0,353	-1,611	-1,005	-0,401	-1,738	-0,978	-0,386
2000-2004	-0,865	-1,587	-0,867	0,013	-1,715	-0,887	0,072	-1,719	-0,909	-0,258	-1,573	-0,865	-0,208
2005-2010	-0,757	-1,901	-0,784	0,501	-2,040	-0,781	0,429	-1,912	-0,796	0,370	-2,107	-0,751	0,370
Loc													
Inag	-0,701	-1,218	-0,702	-0,180	-1,219	-0,717	-0,321	-1,440	-0,721	-0,226	-1,307	-0,709	-0,236
Gender													
F	-0,561	-1,184	-0,579	-0,086	-1,166	-0,574	-0,055	-1,087	-0,584	-0,097	-1,082	-0,561	-0,131
SD	0,684	0,000	0,605	1,375	0,000	0,604	1,339	0,000	0,586	1,383	0,000	0,588	1,304

Table C.5: M1: Mean estimates for n=2000 and 500 replications

		Laplace(474)			AGH15(478)			AGH30(472)			PQL(478)		
Parameter	**TV**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	6,440	5,005	6,442	8,199	5,027	6,457	7,867	4,882	6,493	8,093	5,088	6,448	8,051
EES													
0-20	-8,448	-10,289	-8,517	-7,451	-10,274	-8,505	-7,288	-10,341	-8,525	-7,330	-10,071	-8,499	-7,406
20-30	-6,681	-8,011	-6,731	-5,733	-8,140	-6,706	-5,797	-8,571	-6,742	-5,812	-7,939	-6,725	-5,694
30-40	-5,625	-7,198	-5,661	-4,689	-6,902	-5,648	-4,652	-7,182	-5,686	-4,756	-6,872	-5,649	-4,684
40-50	-4,372	-5,839	-4,403	-3,520	-5,640	-4,388	-3,504	-6,035	-4,425	-3,479	-5,672	-4,407	-3,499
50-60	-3,338	-4,985	-3,359	-2,495	-4,552	-3,360	-2,237	-4,903	-3,382	-2,351	-4,514	-3,356	-2,499
Age													
16-24	-1,220	-1,839	-1,197	-0,568	-2,055	-1,234	-0,545	-1,900	-1,218	-0,486	-1,998	-1,226	-0,570
25-34	-1,912	-2,613	-1,884	-1,227	-2,754	-1,922	-1,318	-2,668	-1,914	-1,181	-2,568	-1,909	-1,129
35-44	-1,659	-2,461	-1,649	-0,950	-2,441	-1,669	-0,850	-2,402	-1,659	-1,091	-2,319	-1,663	-1,013
45-54	-1,338	-2,244	-1,326	-0,690	-2,264	-1,350	-0,525	-2,211	-1,333	-0,565	-2,167	-1,342	-0,637
55-64	-2,233	-3,176	-2,230	-1,425	-3,154	-2,258	-1,436	-3,147	-2,245	-1,365	-2,958	-2,244	-1,418
65-74	-1,073	-2,090	-1,059	0,006	-1,984	-1,091	-0,210	-1,897	-1,057	-0,271	-1,916	-1,092	-0,145
Year													
1995-1999	-0,988	-1,403	-0,979	-0,652	-1,311	-0,988	-0,655	-1,387	-1,008	-0,610	-1,318	-0,988	-0,572
2000-2004	-0,865	-1,456	-0,865	-0,339	-1,390	-0,869	-0,365	-1,259	-0,880	-0,369	-1,449	-0,868	-0,359
2005-2010	-0,757	-1,412	-0,760	-0,258	-1,646	-0,764	-0,129	-1,628	-0,796	-0,098	-1,430	-0,779	-0,154
Loc													
Inag	-0,701	-1,068	-0,698	-0,250	-1,074	-0,707	-0,412	-1,048	-0,702	-0,351	-1,069	-0,697	-0,380
Gender													
F	-0,561	-0,906	-0,561	-0,238	-0,854	-0,557	-0,205	-0,907	-0,567	-0,242	-0,949	-0,564	-0,264
SD	0,684	0,119	0,602	1,144	0,000	0,603	1,259	0,000	0,605	1,173	0,164	0,608	1,211

Table C.6: M1: Mean estimates for n=5000 and 500 replications

	Laplace			AGH15			AGH30			PQL		
Parameter	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	0,863	1,126	1,425	0,852	1,150	1,455	0,835	1,165	1,511	0,843	1,121	1,735
EES												
0-20	0,741	1,084	1,508	0,739	1,108	1,540	0,765	1,133	1,532	0,768	1,071	1,658
20-30	0,670	0,931	1,184	0,631	0,945	1,192	0,629	0,963	1,331	0,579	0,927	1,444
30-40	0,622	0,904	1,186	0,582	0,907	1,159	0,600	0,936	1,321	0,573	0,900	1,393
40-50	0,613	0,888	1,143	0,553	0,890	1,137	0,573	0,919	1,303	0,551	0,887	1,381
50-60	0,621	0,903	1,141	0,576	0,902	1,160	0,597	0,931	1,296	0,567	0,902	1,393
Age												
16-24	0,492	0,603	0,781	0,497	0,618	0,888	0,509	0,618	0,828	0,418	0,589	1,000
25-34	0,505	0,618	0,791	0,509	0,635	0,911	0,509	0,637	0,843	0,444	0,608	1,087
35-44	0,505	0,618	0,811	0,514	0,641	0,887	0,514	0,637	0,862	0,442	0,606	1,017
45-54	0,549	0,650	0,844	0,523	0,666	0,911	0,569	0,669	0,853	0,477	0,638	1,065
55-64	0,598	0,754	1,018	0,620	0,780	1,116	0,631	0,783	1,193	0,522	0,739	1,128
65-74	0,575	0,727	0,944	0,592	0,741	0,988	0,624	0,738	0,899	0,504	0,710	1,119
Year												
1995-1999	0,274	0,318	0,376	0,279	0,324	0,425	0,285	0,327	0,395	0,239	0,309	0,430
2000-2004	0,333	0,397	0,482	0,340	0,402	0,477	0,352	0,411	0,512	0,311	0,383	0,501
2005-2010	0,470	0,603	0,972	0,491	0,622	0,926	0,485	0,619	1,422	0,448	0,588	0,797
Loc												
Inag	0,228	0,276	0,332	0,227	0,280	0,347	0,243	0,285	0,347	0,221	0,269	0,379
Gender												
F	0,245	0,294	0,385	0,249	0,296	0,368	0,256	0,303	0,376	0,223	0,284	0,423

Table C.7: M1: Standard errors of fixed effects for n=800 and 100 replications

	Laplace			AGH15			AGH30			PQL		
Parameter	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	0,604	0,773	1,245	0,575	0,791	1,206	0,575	0,755	1,188	0,587	0,744	1,227
EES												
0-20	0,531	0,734	1,246	0,537	0,717	1,119	0,502	0,713	1,173	0,487	0,683	1,245
20-30	0,442	0,631	1,109	0,443	0,640	1,088	0,395	0,611	1,055	0,393	0,603	1,105
30-40	0,415	0,615	1,090	0,424	0,627	1,081	0,372	0,595	1,051	0,361	0,587	1,098
40-50	0,403	0,606	1,080	0,408	0,617	1,075	0,356	0,585	1,048	0,346	0,578	1,092
50-60	0,412	0,614	1,079	0,411	0,626	1,076	0,379	0,593	1,055	0,359	0,586	1,097
Age												
16-24	0,318	0,373	0,434	0,309	0,372	0,471	0,326	0,376	0,442	0,298	0,356	0,509
25-34	0,328	0,384	0,453	0,320	0,382	0,474	0,339	0,388	0,452	0,306	0,366	0,518
35-44	0,326	0,385	0,461	0,316	0,383	0,489	0,336	0,387	0,455	0,306	0,367	0,524
45-54	0,347	0,404	0,468	0,346	0,401	0,491	0,355	0,406	0,468	0,337	0,385	0,545
55-64	0,410	0,469	0,576	0,390	0,466	0,650	0,411	0,471	0,546	0,373	0,449	0,610
65-74	0,393	0,448	0,538	0,383	0,444	0,534	0,399	0,455	0,521	0,366	0,431	0,626
Year												
1995-1999	0,176	0,200	0,228	0,171	0,198	0,224	0,177	0,201	0,235	0,171	0,195	0,267
2000-2004	0,219	0,248	0,288	0,211	0,247	0,293	0,223	0,250	0,290	0,209	0,241	0,336
2005-2010	0,303	0,376	0,476	0,288	0,370	0,499	0,306	0,373	0,471	0,295	0,361	0,508
Loc												
Inag	0,149	0,174	0,204	0,141	0,172	0,206	0,153	0,175	0,207	0,146	0,168	0,227
Gender												
F	0,158	0,185	0,227	0,147	0,182	0,217	0,159	0,186	0,234	0,153	0,180	0,235

Table C.8: M1: Standard errors of fixed effects for n=2000 and 100 replications

	Laplace			AGH15			AGH30			PQL		
Parameter	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	0,397	0,532	0,736	0,402	0,544	0,836	0,411	0,534	1,141	0,374	0,529	0,744
EES												
0-20	0,325	0,459	0,637	0,346	0,462	0,754	0,351	0,454	1,038	0,320	0,455	0,696
20-30	0,280	0,411	0,607	0,313	0,419	0,734	0,309	0,405	1,029	0,269	0,408	0,670
30-40	0,272	0,401	0,600	0,296	0,410	0,730	0,298	0,397	1,028	0,254	0,399	0,661
40-50	0,263	0,396	0,598	0,283	0,405	0,727	0,287	0,391	1,027	0,244	0,393	0,660
50-60	0,268	0,400	0,601	0,286	0,410	0,729	0,290	0,396	1,030	0,249	0,398	0,662
Age												
16-24	0,209	0,231	0,263	0,207	0,227	0,280	0,199	0,230	0,263	0,189	0,229	0,280
25-34	0,213	0,237	0,275	0,212	0,234	0,287	0,210	0,237	0,274	0,191	0,235	0,287
35-44	0,215	0,238	0,271	0,213	0,234	0,287	0,209	0,237	0,270	0,192	0,235	0,287
45-54	0,228	0,249	0,281	0,222	0,246	0,296	0,219	0,249	0,289	0,204	0,248	0,298
55-64	0,252	0,289	0,355	0,254	0,286	0,339	0,251	0,289	0,339	0,231	0,288	0,345
65-74	0,253	0,277	0,330	0,246	0,272	0,325	0,243	0,276	0,303	0,234	0,275	0,347
Year												
1995-1999	0,112	0,124	0,143	0,109	0,124	0,139	0,111	0,124	0,142	0,109	0,124	0,144
2000-2004	0,139	0,155	0,177	0,137	0,154	0,173	0,139	0,155	0,174	0,133	0,154	0,187
2005-2010	0,208	0,230	0,297	0,197	0,228	0,288	0,205	0,230	0,286	0,193	0,231	0,287
Loc												
Inag	0,094	0,108	0,125	0,093	0,107	0,122	0,094	0,108	0,122	0,092	0,108	0,132
Gender												
F	0,101	0,115	0,132	0,099	0,114	0,136	0,100	0,115	0,140	0,098	0,114	0,141

Table C.9: M1: Standard errors of fixed effects for n=5000 and 100 replications

	Laplace			AGH15			AGH30			PQL		
Parameter	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	0,808	1,173	1,532	0,811	1,153	1,552	0,759	1,152	1,569	0,744	1,124	1,781
EES												
0-20	0,719	1,120	1,533	0,680	1,105	1,540	0,668	1,091	1,584	0,630	1,073	2,071
20-30	0,596	0,971	1,274	0,549	0,952	1,374	0,547	0,955	1,337	0,538	0,928	1,562
30-40	0,560	0,944	1,246	0,550	0,921	1,273	0,533	0,927	1,281	0,530	0,900	1,539
40-50	0,548	0,927	1,226	0,516	0,903	1,258	0,513	0,910	1,261	0,515	0,884	1,501
50-60	0,561	0,941	1,242	0,536	0,918	1,271	0,523	0,925	1,283	0,529	0,899	1,515
Age												
16-24	0,480	0,610	1,087	0,472	0,613	1,080	0,484	0,604	1,058	0,421	0,591	1,145
25-34	0,480	0,629	1,094	0,498	0,631	1,088	0,493	0,622	1,075	0,432	0,609	1,159
35-44	0,486	0,630	1,109	0,482	0,631	1,098	0,500	0,622	1,081	0,442	0,609	1,168
45-54	0,523	0,661	1,105	0,508	0,662	1,120	0,521	0,651	1,130	0,458	0,640	1,223
55-64	0,620	0,773	1,267	0,588	0,770	1,191	0,583	0,764	1,429	0,527	0,751	1,555
65-74	0,569	0,736	1,157	0,558	0,732	1,147	0,561	0,726	1,170	0,511	0,715	2,215
Year												
1995-1999	0,268	0,325	0,424	0,271	0,325	0,422	0,272	0,322	0,447	0,255	0,313	0,549
2000-2004	0,331	0,403	0,535	0,339	0,404	0,554	0,338	0,402	0,487	0,305	0,390	0,752
2005-2010	0,458	0,615	1,413	0,445	0,625	1,561	0,448	0,617	1,394	0,407	0,591	1,343
Loc												
Inag	0,230	0,282	0,408	0,230	0,282	0,408	0,230	0,280	0,374	0,208	0,271	0,496
Gender												
F	0,238	0,300	0,407	0,247	0,300	0,415	0,238	0,297	0,413	0,226	0,289	0,510

Table C.10: M1: Standard errors of fixed effects for n=800 and 500 replications

	Laplace			AGH15			AGH30			PQL		
Parameter	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	0,548	0,771	1,230	0,562	0,764	1,204	0,557	0,763	1,221	0,543	0,740	1,180
EES												
0-20	0,479	0,713	1,251	0,474	0,704	1,120	0,475	0,706	1,245	0,476	0,684	1,227
20-30	0,427	0,628	1,105	0,418	0,620	1,065	0,403	0,621	1,058	0,398	0,599	1,041
30-40	0,395	0,612	1,098	0,380	0,604	1,059	0,393	0,605	1,050	0,363	0,582	1,032
40-50	0,381	0,603	1,091	0,356	0,593	1,054	0,377	0,596	1,045	0,340	0,573	1,025
50-60	0,389	0,611	1,093	0,367	0,601	1,062	0,382	0,604	1,050	0,350	0,581	1,031
Age												
16-24	0,312	0,369	0,460	0,312	0,370	0,463	0,307	0,368	0,485	0,283	0,362	0,526
25-34	0,321	0,380	0,484	0,315	0,381	0,476	0,315	0,379	0,500	0,291	0,373	0,538
35-44	0,321	0,380	0,470	0,317	0,381	0,479	0,319	0,380	0,495	0,293	0,373	0,540
45-54	0,340	0,399	0,482	0,326	0,400	0,503	0,336	0,399	0,537	0,315	0,393	0,558
55-64	0,385	0,462	0,638	0,383	0,467	0,601	0,370	0,465	0,598	0,361	0,458	0,655
65-74	0,373	0,443	0,548	0,378	0,445	0,591	0,373	0,443	0,615	0,360	0,437	0,655
Year												
1995-1999	0,169	0,199	0,235	0,175	0,200	0,240	0,173	0,199	0,238	0,167	0,196	0,268
2000-2004	0,209	0,246	0,297	0,216	0,249	0,309	0,211	0,248	0,297	0,208	0,244	0,345
2005-2010	0,305	0,370	0,495	0,289	0,372	0,534	0,307	0,373	0,501	0,283	0,364	0,581
Loc												
Inag	0,146	0,172	0,215	0,147	0,174	0,221	0,144	0,173	0,232	0,143	0,170	0,243
Gender												
F	0,151	0,183	0,226	0,156	0,184	0,244	0,151	0,184	0,238	0,146	0,181	0,259

Table C.11: M1: Standard errors of fixed effects for n=2000 and 500 replications

	Laplace			AGH15			AGH30			PQL		
Parameter	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	0,388	0,533	0,860	0,382	0,530	0,825	0,385	0,535	0,882	0,388	0,526	0,812
EES												
0-20	0,336	0,454	0,754	0,325	0,452	0,748	0,333	0,455	0,782	0,313	0,446	0,760
20-30	0,280	0,408	0,731	0,286	0,404	0,738	0,280	0,409	0,763	0,275	0,400	0,731
30-40	0,269	0,399	0,729	0,273	0,395	0,736	0,260	0,400	0,756	0,260	0,391	0,726
40-50	0,257	0,393	0,726	0,262	0,390	0,733	0,253	0,395	0,754	0,250	0,385	0,724
50-60	0,262	0,398	0,728	0,269	0,395	0,736	0,258	0,400	0,755	0,254	0,390	0,728
Age												
16-24	0,202	0,229	0,274	0,202	0,229	0,285	0,197	0,229	0,274	0,192	0,227	0,288
25-34	0,207	0,236	0,279	0,208	0,236	0,296	0,207	0,236	0,283	0,195	0,234	0,299
35-44	0,208	0,236	0,287	0,206	0,236	0,294	0,204	0,236	0,284	0,197	0,234	0,295
45-54	0,220	0,248	0,295	0,218	0,248	0,305	0,220	0,248	0,291	0,208	0,246	0,313
55-64	0,248	0,288	0,354	0,253	0,288	0,370	0,249	0,288	0,348	0,244	0,286	0,378
65-74	0,245	0,275	0,331	0,238	0,275	0,328	0,241	0,275	0,323	0,233	0,274	0,335
Year												
1995-1999	0,109	0,124	0,146	0,110	0,124	0,150	0,108	0,124	0,146	0,106	0,123	0,153
2000-2004	0,135	0,154	0,184	0,135	0,154	0,187	0,134	0,154	0,183	0,130	0,153	0,193
2005-2010	0,192	0,230	0,294	0,192	0,230	0,309	0,195	0,230	0,299	0,190	0,229	0,302
Loc												
Inag	0,094	0,107	0,126	0,092	0,107	0,133	0,092	0,108	0,132	0,091	0,107	0,136
Gender												
F	0,099	0,114	0,137	0,097	0,114	0,143	0,097	0,115	0,139	0,096	0,114	0,145

Table C.12: M1: Standard errors of fixed effects for n=5000 and 500 replications

	Laplace			AGH15			AGH30			PQL		
Parameter	**800**	**2000**	**5000**	**800**	**2000**	**5000**	**800**	**2000**	**5000**	**800**	**2000**	**5000**
Intercept	-0,081	0,003	0,016	-0,172	-0,007	0,119	0,124	-0,170	0,047	-0,099	-0,123	0,040
EES												
0-20	0,058	-0,195	-0,109	-0,127	-0,085	-0,086	-0,224	-0,101	-0,088	-0,021	0,027	-0,077
20-30	0,152	-0,070	-0,075	-0,020	-0,054	-0,118	-0,049	0,001	-0,023	0,058	-0,010	-0,056
30-40	0,210	-0,026	-0,060	0,078	-0,032	-0,075	-0,066	-0,002	-0,065	0,126	0,031	-0,053
40-50	0,260	-0,030	-0,036	0,146	-0,022	-0,086	0,006	0,021	-0,033	0,177	0,055	-0,060
50-60	0,268	0,023	-0,030	0,171	-0,013	-0,052	-0,011	0,033	-0,025	0,186	0,037	-0,039
Age												
16-24	-0,170	-0,046	0,010	0,112	0,060	-0,031	-0,141	0,091	-0,040	-0,022	0,034	-0,002
25-34	-0,132	-0,061	0,009	0,056	0,036	-0,016	-0,182	0,042	-0,041	-0,053	0,031	0,001
35-44	-0,143	-0,012	-0,003	-0,081	0,050	-0,018	-0,172	0,064	-0,035	-0,045	0,004	-0,001
45-54	-0,155	-0,055	-0,012	0,095	0,044	-0,024	-0,166	0,055	-0,039	-0,031	0,052	-0,008
55-64	-0,140	-0,025	-0,017	-0,029	0,040	-0,053	-0,239	0,114	-0,040	-0,109	0,058	-0,011
65-74	-0,152	-0,019	-0,018	0,103	0,030	-0,010	-0,077	0,029	-0,046	0,017	0,039	0,011
Year												
1995-1999	-0,012	-0,032	0,018	-0,077	-0,038	-0,019	-0,099	0,027	0,007	0,018	0,027	-0,029
2000-2004	-0,035	-0,015	0,011	-0,079	-0,073	-0,029	-0,149	0,026	-0,013	-0,017	0,006	0,000
2005-2010	-0,043	-0,058	-0,007	-0,090	-0,071	-0,020	-0,163	0,052	-0,002	-0,129	-0,004	-0,030
Loc												
Inag	-0,013	-0,011	-0,002	-0,025	0,010	0,017	-0,019	-0,042	-0,007	-0,048	-0,025	-0,014
Gender												
F	-0,025	0,008	-0,006	0,061	0,031	-0,024	-0,012	-0,033	-0,008	0,013	-0,020	0,014
SD	-0,145	-0,091	-0,102	-0,049	-0,026	-0,060	-0,125	-0,108	-0,067	-0,112	-0,089	-0,103

SD: Standard Deviation

Table C.13: M1: Bias of mean estimates for 100 replications

	Laplace			AGH15			AGH30			PQL		
Parameter	**800**	**2000**	**5000**	**800**	**2000**	**5000**	**800**	**2000**	**5000**	**800**	**2000**	**5000**
Intercept	0,044	-0,010	0,002	-0,099	-0,002	0,017	-0,092	-0,061	0,052	-0,126	-0,129	0,008
EES												
0-20	-0,177	-0,074	-0,069	-0,079	-0,068	-0,057	0,000	-0,030	-0,077	-0,071	0,039	-0,051
20-30	-0,068	-0,029	-0,050	-0,037	-0,028	-0,025	-0,016	0,005	-0,061	0,034	0,058	-0,044
30-40	-0,048	0,012	-0,036	0,046	-0,016	-0,023	0,048	0,032	-0,061	0,082	0,080	-0,024
40-50	0,006	0,013	-0,030	0,112	0,021	-0,016	0,065	0,043	-0,053	0,121	0,088	-0,035
50-60	0,047	0,038	-0,021	0,179	0,031	-0,022	0,142	0,072	-0,044	0,170	0,092	-0,019
Age												
16-24	0,008	-0,018	0,024	0,002	-0,034	-0,014	-0,003	0,018	0,003	0,020	-0,012	-0,006
25-34	-0,036	-0,038	0,027	-0,020	-0,053	-0,010	-0,029	-0,007	-0,002	-0,018	-0,017	0,003
35-44	-0,057	-0,030	0,010	-0,041	-0,052	-0,010	-0,003	0,001	0,000	-0,034	-0,003	-0,003
45-54	-0,048	-0,007	0,012	-0,024	-0,025	-0,012	0,026	0,014	0,005	0,004	-0,027	-0,004
55-64	-0,107	-0,018	0,003	-0,031	-0,082	-0,025	-0,069	-0,012	-0,012	-0,083	-0,027	-0,011
65-74	-0,001	-0,004	0,014	0,039	-0,044	-0,018	0,021	0,012	0,016	-0,029	-0,025	-0,019
Year												
1995-1999	-0,083	-0,012	0,008	-0,047	-0,009	0,000	-0,040	-0,017	-0,020	0,016	0,010	0,000
2000-2004	-0,074	-0,002	0,000	-0,056	-0,021	-0,004	-0,046	-0,043	-0,015	0,026	0,000	-0,003
2005-2010	-0,145	-0,027	-0,003	-0,133	-0,024	-0,007	-0,057	-0,039	-0,039	0,005	0,006	-0,022
Loc												
Inag	-0,041	-0,001	0,002	-0,040	-0,017	-0,006	-0,040	-0,021	-0,002	-0,031	-0,008	0,004
Gender												
F	-0,040	-0,018	0,001	-0,055	-0,013	0,004	0,000	-0,023	-0,005	-0,039	0,001	-0,003
SD	-0,083	-0,080	-0,083	-0,090	-0,081	-0,082	-0,105	-0,098	-0,080	-0,103	-0,096	-0,076

SD: Standard Deviation

Table C.14: M1: Bias of mean estimates for 500 replications

		Laplace			AGH15			AGH30			PQL		
nR	**param.**	**800**	**2000**	**5000**	**800**	**2000**	**5000**	**800**	**2000**	**5000**	**800**	**2000**	**5000**
	all	0,021	0,004	0,002	0,010	0,002	0,003	0,018	0,005	0,002	0,008	0,002	0,002
100	fixed	0,021	0,004	0,001	0,010	0,002	0,003	0,018	0,005	0,002	0,008	0,002	0,001
	SD	0,021	0,008	0,010	0,002	0,001	0,004	0,016	0,012	0,004	0,013	0,008	0,011
	all	0,006	0,001	0,001	0,006	0,002	0,001	0,003	0,002	0,002	0,005	0,003	0,001
500	fixed	0,006	0,001	0,001	0,006	0,001	0,000	0,003	0,001	0,001	0,005	0,003	0,000
	SD	0,007	0,006	0,007	0,008	0,006	0,007	0,011	0,010	0,006	0,011	0,009	0,006

Table C.15: M1: MSE of mean estimates

C.2 Modèle 2 ayant l'EES pour variable à effets aléatoires

		Laplace (99)			AGH15(96)			AGH30(99)			PQL(100)		
Parameter	**TV**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	2,748	0,300	2,758	6,681	0,119	2,843	6,735	0,106	2,879	5,846	0,355	2,964	6,920
Age													
16-24	-1,165	-3,001	-1,233	0,356	-2,561	-1,313	-0,040	-2,815	-1,206	-0,004	-3,452	-1,283	0,751
25-34	-1,811	-3,715	-1,877	-0,400	-3,096	-1,961	-0,694	-3,264	-1,881	-0,771	-3,687	-1,942	-0,201
35-44	-1,599	-3,163	-1,669	-0,398	-2,922	-1,763	-0,318	-2,981	-1,613	-0,147	-3,727	-1,694	0,082
45-54	-1,266	-3,297	-1,302	0,398	-2,787	-1,458	0,019	-2,928	-1,273	0,421	-3,332	-1,418	0,873
55-64	-2,169	-3,903	-2,243	-0,660	-4,774	-2,382	-0,556	-4,714	-2,259	0,406	-4,798	-2,331	-0,304
65-74	-1,051	-3,033	-1,072	0,603	-2,785	-1,250	-0,021	-2,697	-1,068	0,284	-4,082	-1,189	1,530
Col													
A-F	-0,897	-3,691	-0,904	1,331	-2,297	-0,900	1,007	-2,422	-0,965	0,601	-2,311	-0,895	0,785
F-A	-2,119	-3,483	-2,115	-0,999	-3,976	-2,202	-1,092	-3,598	-2,232	-1,037	-4,578	-2,295	-1,298
F-F	-1,829	-3,026	-1,890	-1,053	-2,986	-1,931	-1,120	-2,975	-1,870	-0,850	-2,934	-1,906	-1,267
F-L	-1,727	-2,838	-1,822	-1,255	-2,773	-1,833	-0,786	-2,746	-1,806	-1,097	-2,582	-1,800	-1,127
Vehic /obst	-0,605	-1,699	-0,649	0,441	-1,442	-0,685	0,106	-1,653	-0,662	0,151	-1,742	-0,635	0,128
Year													
1995-1999	-1,000	-1,879	-1,041	-0,183	-1,644	-1,032	-0,429	-1,839	-1,046	-0,475	-2,063	-1,037	-0,311
2000-2004	-0,885	-1,847	-0,952	0,067	-1,865	-0,883	-0,052	-1,854	-0,923	-0,013	-2,442	-0,944	-0,193
2005-2010	-0,723	-2,349	-0,692	0,774	-2,611	-0,798	0,270	-2,184	-0,769	0,819	-2,357	-0,819	0,527
Local													
In agglo	-0,750	-1,273	-0,771	-0,168	-1,419	-0,837	-0,386	-2,004	-0,809	-0,394	-1,329	-0,759	-0,091
Gender													
F	-0,589	-1,548	-0,622	-0,016	-1,752	-0,605	-0,137	-1,384	-0,600	-0,132	-1,198	-0,631	0,311
SD	2,620	0,578	2,368	4,938	0,704	2,459	5,301	0,703	2,462	6,990	0,254	2,429	5,186

Table C.16: M2: Mean estimates for n=800 and 100 replications

		Laplace(100)			AGH15(100)			AGH30(100)			PQL(100)		
Parameter	**TV**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	2,748	-0,286	2,705	5,332	-0,035	2,734	7,252	0,264	2,717	4,934	0,461	2,642	5,236
Age													
16-24	-1,165	-2,263	-1,149	-0,180	-2,359	-1,119	-0,055	-2,225	-1,220	-0,168	-2,335	-1,169	-0,323
25-34	-1,811	-2,674	-1,835	-0,774	-2,846	-1,748	-0,877	-2,976	-1,812	-0,862	-3,427	-1,786	-0,647
35-44	-1,599	-2,392	-1,592	-0,674	-2,787	-1,564	-0,329	-2,501	-1,610	-0,800	-2,919	-1,614	-0,724
45-54	-1,266	-2,232	-1,250	0,114	-2,382	-1,201	0,553	-2,219	-1,266	-0,345	-2,135	-1,270	0,012
55-64	-2,169	-3,053	-2,170	-0,845	-3,614	-2,085	-0,251	-3,168	-2,183	-0,934	-3,324	-2,178	-1,348
65-74	-1,051	-2,206	-1,069	-0,136	-2,114	-0,995	0,251	-2,198	-1,067	-0,351	-2,304	-1,072	0,305
Col													
A-F	-0,897	-2,434	-1,038	-0,050	-1,971	-0,888	-0,237	-1,985	-0,909	-0,042	-2,096	-0,924	0,207
F-A	-2,119	-4,160	-2,175	-1,466	-2,905	-2,137	-1,387	-4,121	-2,152	-1,101	-3,845	-2,132	-1,312
F-F	-1,829	-2,454	-1,875	-1,066	-2,449	-1,859	-1,141	-2,381	-1,842	-1,447	-2,477	-1,836	-1,354
F-L	-1,727	-2,190	-1,732	-1,234	-2,255	-1,766	-1,294	-2,488	-1,707	-1,194	-2,210	-1,745	-1,289
Vehic /obst	-0,605	-1,386	-0,637	0,038	-1,155	-0,638	-0,161	-1,275	-0,598	-0,111	-1,251	-0,637	-0,045
Year													
1995-1999	-1,000	-1,535	-0,972	-0,511	-1,649	-1,048	-0,550	-1,400	-0,995	-0,589	-1,414	-0,983	-0,396
2000-2004	-0,885	-1,271	-0,859	-0,308	-1,485	-0,954	-0,478	-1,383	-0,910	-0,380	-1,381	-0,852	-0,252
2005-2010	-0,723	-1,454	-0,726	-0,057	-1,802	-0,742	0,253	-1,477	-0,686	-0,004	-1,698	-0,725	0,119
Local													
In agglo	-0,750	-1,193	-0,762	-0,294	-1,311	-0,745	-0,422	-1,018	-0,749	-0,401	-1,103	-0,725	-0,225
Gender													
F	-0,589	-1,020	-0,605	-0,262	-1,059	-0,617	-0,265	-0,985	-0,593	-0,263	-1,215	-0,618	-0,340
SD	2,620	0,660	2,268	4,434	0,922	2,240	4,628	0,322	2,311	5,553	0,682	2,358	4,178

Table C.17: M2: Mean estimates for n=2000 and 100 replications

		Laplace(100)			AGH15(100)			AGH30(100)			PQL(100)		
Parameter	**TV**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	2,748	-0,843	2,741	5,823	0,074	2,830	6,674	-0,047	2,803	5,249	-0,614	2,778	6,659
Age													
16-24	-1,165	-2,398	-1,168	-0,249	-2,626	-1,152	-0,203	-2,091	-1,122	-0,491	-2,235	-1,120	-0,229
25-34	-1,811	-2,836	-1,830	-1,005	-2,910	-1,780	-1,081	-2,532	-1,784	-1,114	-2,798	-1,769	-0,919
35-44	-1,599	-2,477	-1,614	-0,812	-2,518	-1,594	-0,911	-2,519	-1,553	-0,855	-2,543	-1,559	-0,923
45-54	-1,266	-2,285	-1,326	-0,568	-2,230	-1,274	-0,564	-1,918	-1,187	-0,344	-2,089	-1,225	-0,328
55-64	-2,169	-3,102	-2,236	-1,123	-3,281	-2,162	-1,316	-2,907	-2,098	-1,331	-3,338	-2,134	-1,369
65-74	-1,051	-1,772	-1,068	-0,184	-2,654	-1,068	-0,397	-1,659	-1,047	0,467	-2,232	-0,992	-0,207
Col													
A-F	-0,897	-1,838	-0,896	-0,357	-1,630	-0,893	0,869	-1,648	-0,867	-0,117	-2,352	-0,899	-0,022
F-A	-2,119	-3,055	-2,194	-1,536	-2,843	-2,121	-1,252	-2,751	-2,133	-1,502	-4,445	-2,105	-1,000
F-F	-1,829	-2,586	-1,860	-1,487	-2,413	-1,853	-1,195	-2,268	-1,838	-1,200	-2,313	-1,838	-1,375
F-L	-1,727	-2,469	-1,751	-1,288	-2,230	-1,776	-1,361	-2,165	-1,731	-1,224	-2,330	-1,731	-1,265
Vehic /obst	-0,605	-1,049	-0,609	-0,196	-1,129	-0,634	-0,175	-1,036	-0,609	0,064	-1,084	-0,620	-0,285
Year													
1995-1999	-1,000	-1,390	-0,988	-0,569	-1,382	-0,993	-0,436	-1,363	-1,013	-0,438	-1,336	-1,007	-0,665
2000-2004	-0,885	-1,385	-0,871	-0,472	-1,403	-0,883	-0,284	-1,392	-0,897	-0,382	-1,425	-0,896	-0,284
2005-2010	-0,723	-1,388	-0,702	-0,239	-1,706	-0,695	0,057	-1,313	-0,718	-0,049	-1,438	-0,757	-0,160
Local													
In agglo	-0,750	-1,096	-0,747	-0,383	-1,488	-0,757	-0,515	-1,169	-0,725	-0,259	-1,082	-0,754	-0,409
Gender													
F	-0,589	-1,002	-0,590	-0,173	-1,000	-0,613	-0,073	-0,981	-0,612	-0,289	-1,200	-0,598	-0,323
SD	2,620	0,482	2,359	4,796	0,675	2,410	6,442	0,496	2,402	4,879	0,629	2,143	3,894

Table C.18: M2: Mean estimates for n=3000 and 100 replications

		Laplace(492)			AGH15(492)			AGH30(491)			PQL(489)		
Parameter	**TV**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	2,748	-1,381	2,829	7,629	-1,094	2,793	7,694	-1,686	2,844	6,441	-1,056	2,826	6,062
Age													
16-24	-1,165	-4,238	-1,193	0,591	-3,447	-1,170	0,841	-4,583	-1,227	0,815	-4,918	-1,167	0,850
25-34	-1,811	-4,432	-1,855	0,060	-5,396	-1,834	0,371	-4,901	-1,896	0,016	-5,367	-1,845	0,248
35-44	-1,599	-4,290	-1,622	0,185	-5,434	-1,647	0,651	-4,985	-1,681	0,037	-5,623	-1,627	0,882
45-54	-1,266	-3,725	-1,267	0,390	-3,484	-1,291	0,920	-4,276	-1,349	0,536	-4,987	-1,290	0,991
55-64	-2,169	-4,734	-2,213	-0,188	-4,667	-2,240	0,044	-5,779	-2,296	0,462	-6,354	-2,216	0,073
65-74	-1,051	-3,372	-1,101	1,515	-2,984	-1,068	2,384	-4,036	-1,093	1,811	-4,988	-1,014	1,485
Col													
A-F	-0,897	-4,750	-0,938	1,890	-3,313	-0,985	1,257	-3,262	-0,914	1,005	-3,693	-0,885	1,323
F-A	-2,119	-4,629	-2,247	-0,571	-4,309	-2,233	-0,677	-5,781	-2,209	-0,520	-5,081	-2,189	-0,482
F-F	-1,829	-3,334	-1,911	-0,771	-3,490	-1,899	-0,825	-3,131	-1,913	-0,943	-3,363	-1,869	-0,623
F-L	-1,727	-3,341	-1,770	-0,784	-3,792	-1,800	-0,804	-2,870	-1,774	-0,490	-3,164	-1,770	-0,538
Vehic /obst	-0,605	-1,807	-0,624	0,231	-2,346	-0,645	0,271	-1,576	-0,615	0,457	-1,984	-0,610	0,223
Year													
1995-1999	-1,000	-2,198	-1,035	0,151	-2,143	-1,020	-0,205	-2,258	-1,004	0,094	-2,050	-0,983	0,046
2000-2004	-0,885	-2,351	-0,943	0,240	-2,289	-0,919	0,504	-1,987	-0,904	0,206	-2,425	-0,883	0,323
2005-2010	-0,723	-3,844	-0,749	0,800	-2,952	-0,778	1,203	-2,728	-0,739	1,262	-3,018	-0,716	0,928
Local													
In agglo	-0,750	-1,666	-0,771	-0,018	-1,671	-0,756	-0,123	-1,848	-0,772	-0,141	-1,586	-0,762	0,107
Gender													
F	-0,589	-1,806	-0,622	0,215	-1,553	-0,617	0,233	-1,711	-0,626	0,319	-1,662	-0,617	0,172
SD	2,620	0,477	2,375	6,998	0,291	2,397	7,152	0,206	2,442	6,441	0,602	2,291	5,320

Table C.19: M2: Mean estimates for n=800 and 500 replications

		Laplace(498)			AGH15(500)			AGH30(500)			PQL(498)		
Parameter	**TV**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	2,748	-0,996	2,787	5,860	-0,445	2,729	6,187	-2,187	2,791	7,227	-0,037	2,744	5,720
Age													
16-24	-1,165	-2,666	-1,165	-0,129	-3,400	-1,186	-0,202	-2,905	-1,229	-0,054	-2,845	-1,160	-0,013
25-34	-1,811	-3,233	-1,822	-0,775	-4,521	-1,836	-0,616	-3,530	-1,871	-0,782	-3,293	-1,830	-0,860
35-44	-1,599	-3,133	-1,621	-0,544	-3,946	-1,617	-0,479	-3,425	-1,670	-0,330	-3,291	-1,605	-0,509
45-54	-1,266	-2,719	-1,278	0,226	-3,996	-1,305	-0,253	-3,151	-1,334	0,047	-3,025	-1,281	-0,004
55-64	-2,169	-4,226	-2,176	-0,888	-4,664	-2,209	-0,593	-3,982	-2,233	-0,922	-3,523	-2,196	-0,888
65-74	-1,051	-2,460	-1,043	0,274	-3,056	-1,067	0,296	-3,005	-1,112	0,228	-2,673	-1,074	0,163
Col													
A-F	-0,897	-3,028	-0,941	0,318	-2,434	-0,917	0,256	-1,921	-0,883	0,362	-1,991	-0,885	0,552
F-A	-2,119	-3,593	-2,131	-0,867	-3,598	-2,155	-0,793	-4,116	-2,156	-1,084	-3,658	-2,143	-1,164
F-F	-1,829	-2,735	-1,845	-1,091	-2,605	-1,852	-1,222	-2,479	-1,848	-1,197	-2,624	-1,845	-1,055
F-L	-1,727	-2,613	-1,763	-1,002	-2,574	-1,755	-1,036	-2,592	-1,757	-0,964	-2,458	-1,747	-0,930
Vehic /obst	-0,605	-1,391	-0,611	0,191	-1,189	-0,616	0,144	-1,518	-0,606	0,247	-1,715	-0,607	0,348
Year													
1995-1999	-1,000	-1,659	-1,009	-0,495	-1,591	-1,002	-0,286	-1,738	-1,013	-0,489	-1,959	-0,994	-0,406
2000-2004	-0,885	-1,917	-0,896	-0,056	-1,777	-0,881	-0,104	-1,764	-0,907	-0,272	-1,593	-0,887	-0,219
2005-2010	-0,723	-2,460	-0,712	0,744	-1,731	-0,716	0,215	-2,035	-0,733	0,350	-1,883	-0,715	0,360
Local													
In agglo	-0,750	-1,284	-0,761	-0,330	-1,147	-0,757	-0,326	-1,531	-0,756	-0,126	-1,442	-0,760	-0,271
Gender													
F	-0,589	-1,435	-0,594	-0,009	-1,355	-0,603	-0,015	-1,109	-0,601	-0,146	-1,205	-0,599	-0,080
SD	2,620	0,421	2,317	5,192	0,447	2,398	6,049	0,736	2,397	6,160	0,501	2,287	4,642

Table C.20: M2: Mean estimates for n=2000 and 500 replications

		Laplace(500)			AGH15(500)			AGH30(500)			PQL(499)		
Parameter	**TV**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	2,748	-0,707	2,682	5,776	-0,429	2,819	5,630	-1,723	2,736	6,368	-0,340	2,786	6,717
Age													
16-24	-1,165	-2,182	-1,161	-0,043	-2,490	-1,176	-0,022	-2,045	-1,177	-0,284	-2,000	-1,190	-0,161
25-34	-1,811	-3,083	-1,834	-0,992	-2,936	-1,828	-0,935	-2,864	-1,819	-0,985	-2,749	-1,840	-0,614
35-44	-1,599	-2,665	-1,610	-0,628	-2,809	-1,614	-0,587	-2,668	-1,610	-0,674	-2,595	-1,617	-0,377
45-54	-1,266	-2,350	-1,273	-0,284	-2,404	-1,265	-0,220	-2,622	-1,299	-0,137	-2,175	-1,282	-0,114
55-64	-2,169	-3,410	-2,175	-1,124	-3,659	-2,187	-1,060	-3,870	-2,191	-0,639	-3,499	-2,198	-0,708
65-74	-1,051	-2,235	-1,062	0,274	-2,163	-1,056	0,400	-2,047	-1,066	0,233	-2,173	-1,080	0,168
Col													
A-F	-0,897	-1,814	-0,891	-0,048	-2,191	-0,919	0,384	-2,150	-0,899	0,355	-2,133	-0,917	0,002
F-A	-2,119	-3,140	-2,131	-1,118	-3,146	-2,150	-1,307	-3,277	-2,135	-0,771	-3,399	-2,147	-1,156
F-F	-1,829	-2,555	-1,844	-1,221	-2,566	-1,862	-1,338	-2,428	-1,834	-1,265	-2,706	-1,843	-1,292
F-L	-1,727	-2,634	-1,744	-1,133	-2,410	-1,750	-1,228	-2,276	-1,735	-1,083	-2,737	-1,744	-0,919
Véhic /obst	-0,605	-1,061	-0,606	-0,111	-1,266	-0,619	-0,017	-1,241	-0,602	-0,080	-1,171	-0,609	0,023
Year													
1995-1999	-1,000	-1,423	-1,009	-0,541	-1,655	-1,013	-0,435	-1,562	-1,014	-0,410	-1,562	-1,015	-0,527
2000-2004	-0,885	-1,410	-0,895	-0,356	-1,833	-0,891	-0,346	-1,416	-0,894	-0,261	-1,374	-0,901	-0,331
2005-2010	-0,723	-1,984	-0,712	0,154	-1,649	-0,717	-0,001	-1,778	-0,744	0,252	-1,798	-0,751	0,111
Local													
In agglo	-0,750	-1,356	-0,761	-0,199	-1,121	-0,763	-0,436	-1,195	-0,754	-0,299	-1,202	-0,749	-0,290
Gender													
F	-0,589	-0,994	-0,594	-0,190	-1,006	-0,601	-0,009	-1,080	-0,592	-0,201	-0,995	-0,587	-0,087
SD	2,620	0,469	2,329	6,039	0,622	2,356	5,853	0,311	2,356	4,851	0,557	2,279	4,935

Table C.21: M2: Mean estimates for n=3000 and 500 replications

	Laplace			AGH15			AGH30			PQL		
Parameter	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	0,649	1,200	2,212	0,565	1,231	2,341	0,595	1,228	3,059	0,642	1,203	2,274
Age												
16-24	0,440	0,614	1,124	0,407	0,608	1,153	0,405	0,595	1,086	0,414	0,573	1,139
25-34	0,457	0,621	1,124	0,433	0,616	1,147	0,405	0,601	1,078	0,411	0,580	1,139
35-44	0,450	0,624	1,120	0,430	0,620	1,145	0,408	0,603	1,100	0,414	0,583	1,141
45-54	0,477	0,652	1,146	0,456	0,647	1,167	0,453	0,634	1,140	0,441	0,610	1,157
55-64	0,534	0,717	1,171	0,515	0,717	1,171	0,461	0,697	1,378	0,480	0,670	1,164
65-74	0,534	0,728	1,217	0,485	0,716	1,290	0,454	0,695	1,263	0,495	0,675	1,192
Col												
A-F	0,452	0,619	1,904	0,417	0,605	1,082	0,425	0,602	1,268	0,391	0,568	0,963
F-A	0,399	0,581	1,037	0,412	0,596	1,075	0,410	0,593	1,838	0,377	0,566	1,524
F-F	0,252	0,340	0,522	0,241	0,342	0,487	0,242	0,333	0,576	0,240	0,319	0,528
F-L	0,250	0,340	0,508	0,242	0,340	0,485	0,241	0,332	0,701	0,244	0,312	0,496
Vehic /obst	0,257	0,344	0,514	0,255	0,342	0,525	0,243	0,337	0,603	0,248	0,318	0,516
Year												
1995-1999	0,226	0,298	0,415	0,219	0,297	0,430	0,214	0,291	0,573	0,222	0,277	0,418
2000-2004	0,271	0,365	0,511	0,257	0,360	0,527	0,269	0,354	0,751	0,261	0,339	0,489
2005-2010	0,383	0,532	1,077	0,398	0,539	0,915	0,355	0,522	0,809	0,379	0,500	0,795
Local												
In agglo	0,175	0,241	0,326	0,175	0,241	0,347	0,175	0,236	0,619	0,174	0,223	0,339
Gender												
F	0,185	0,255	0,354	0,190	0,256	0,356	0,184	0,248	0,501	0,184	0,236	0,419

Table C.22: M2: Standard errors of fixed effects for n=800 and 100 replications

	Laplace			AGH15			AGH30			PQL		
Parameter	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	0,432	1,021	1,886	0,487	1,012	1,965	0,339	1,040	2,388	0,452	1,057	1,801
Age												
16-24	0,283	0,359	0,526	0,282	0,374	0,633	0,269	0,369	0,565	0,255	0,369	0,578
25-34	0,286	0,364	0,525	0,287	0,378	0,634	0,275	0,372	0,577	0,260	0,373	0,576
35-44	0,288	0,365	0,531	0,286	0,380	0,640	0,281	0,375	0,572	0,259	0,375	0,580
45-54	0,305	0,383	0,551	0,301	0,397	0,655	0,290	0,392	0,595	0,281	0,391	0,597
55-64	0,331	0,420	0,579	0,333	0,433	0,675	0,326	0,431	0,709	0,322	0,426	0,606
65-74	0,338	0,424	0,607	0,335	0,438	0,711	0,328	0,432	0,676	0,345	0,431	0,652
Col												
A-F	0,284	0,372	0,754	0,273	0,369	0,627	0,259	0,376	0,537	0,280	0,364	0,544
F-A	0,258	0,358	1,022	0,262	0,357	0,700	0,277	0,371	0,792	0,249	0,349	0,671
F-F	0,160	0,204	0,301	0,160	0,207	0,335	0,155	0,210	0,365	0,155	0,205	0,300
F-L	0,159	0,201	0,309	0,159	0,206	0,329	0,151	0,209	0,342	0,155	0,203	0,307
Vehic /obst	0,159	0,205	0,323	0,163	0,210	0,325	0,161	0,211	0,358	0,158	0,206	0,321
Year												
1995-1999	0,140	0,178	0,253	0,144	0,182	0,279	0,140	0,183	0,301	0,140	0,178	0,259
2000-2004	0,171	0,216	0,325	0,174	0,221	0,341	0,168	0,224	0,380	0,169	0,217	0,311
2005-2010	0,249	0,316	0,534	0,249	0,327	0,538	0,252	0,324	0,535	0,241	0,318	0,486
Local												
In agglo	0,113	0,143	0,226	0,113	0,146	0,230	0,111	0,147	0,245	0,107	0,143	0,204
Gender												
F	0,121	0,152	0,256	0,118	0,154	0,239	0,118	0,156	0,268	0,113	0,151	0,223

Table C.23: M2: Standard errors of fixed effects for n=2000 and 100 replications

	Laplace			AGH15			AGH30			PQL		
Parameter	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	0,350	1,031	2,018	0,394	1,052	2,724	0,330	1,047	2,064	0,402	0,945	1,660
Age												
16-24	0,232	0,311	0,554	0,232	0,312	0,637	0,223	0,306	0,631	0,200	0,302	0,660
25-34	0,237	0,315	0,549	0,238	0,315	0,636	0,226	0,309	0,617	0,204	0,306	0,656
35-44	0,238	0,316	0,553	0,240	0,317	0,642	0,231	0,311	0,622	0,210	0,307	0,660
45-54	0,250	0,331	0,568	0,252	0,331	0,653	0,240	0,325	0,639	0,217	0,321	0,673
55-64	0,286	0,364	0,592	0,283	0,362	0,662	0,264	0,355	0,669	0,267	0,349	0,682
65-74	0,285	0,365	0,596	0,287	0,365	0,687	0,265	0,359	0,784	0,238	0,353	0,723
Col												
A-F	0,222	0,311	0,588	0,234	0,307	0,566	0,215	0,303	0,636	0,220	0,307	1,218
F-A	0,210	0,305	0,692	0,205	0,297	0,645	0,213	0,295	0,631	0,197	0,289	0,988
F-F	0,129	0,177	0,297	0,128	0,177	0,390	0,125	0,173	0,341	0,129	0,168	0,377
F-L	0,127	0,175	0,322	0,127	0,175	0,402	0,121	0,171	0,338	0,128	0,167	0,353
Vehic /obst	0,129	0,175	0,275	0,132	0,177	0,366	0,127	0,175	0,373	0,131	0,170	0,298
Year												
1995-1999	0,114	0,153	0,229	0,117	0,153	0,341	0,108	0,151	0,269	0,115	0,147	0,279
2000-2004	0,137	0,186	0,293	0,143	0,186	0,401	0,132	0,183	0,315	0,141	0,179	0,355
2005-2010	0,199	0,270	0,421	0,202	0,273	0,589	0,185	0,268	0,470	0,206	0,263	0,584
Local												
In agglo	0,091	0,124	0,209	0,090	0,124	0,330	0,088	0,120	0,198	0,087	0,117	0,256
Gender												
F	0,094	0,131	0,238	0,096	0,130	0,291	0,094	0,128	0,214	0,092	0,124	0,290

Table C.24: M2: Standard errors of fixed effects for n=3000 and 100 replications

	Laplace			AGH15			AGH30			PQL		
Parameter	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	0,573	1,202	3,465	0,526	1,208	3,167	0,506	1,226	2,951	0,559	1,155	2,320
Age												
16-24	0,413	0,610	1,815	0,424	0,607	1,214	0,418	0,616	1,584	0,397	0,585	1,384
25-34	0,415	0,617	1,820	0,425	0,614	1,194	0,427	0,622	1,585	0,403	0,591	1,381
35-44	0,418	0,620	1,791	0,429	0,619	1,215	0,427	0,625	1,588	0,409	0,594	1,388
45-54	0,439	0,648	1,871	0,448	0,646	1,212	0,451	0,653	1,605	0,417	0,622	1,392
55-64	0,492	0,709	1,958	0,487	0,709	1,446	0,481	0,716	1,640	0,494	0,676	1,410
65-74	0,490	0,717	1,900	0,485	0,710	1,372	0,498	0,724	1,963	0,467	0,685	1,422
Col												
A-F	0,406	0,618	1,750	0,395	0,618	1,563	0,387	0,620	1,660	0,377	0,583	1,220
F-A	0,369	0,601	2,310	0,390	0,600	1,329	0,368	0,591	1,711	0,373	0,557	1,213
F-F	0,240	0,342	0,738	0,249	0,342	0,903	0,243	0,343	0,828	0,233	0,324	0,608
F-L	0,237	0,335	0,722	0,238	0,339	0,924	0,240	0,339	0,720	0,236	0,321	0,614
Vehic /obst	0,241	0,341	0,767	0,245	0,342	0,742	0,243	0,344	0,740	0,240	0,327	0,625
Year												
1995-1999	0,212	0,296	0,659	0,219	0,297	0,574	0,208	0,297	0,662	0,213	0,282	0,494
2000-2004	0,264	0,362	0,809	0,266	0,362	0,713	0,253	0,363	0,935	0,256	0,344	0,697
2005-2010	0,365	0,538	2,534	0,361	0,536	1,197	0,376	0,540	1,385	0,360	0,510	1,631
Local												
In agglo	0,172	0,238	0,470	0,173	0,239	0,549	0,169	0,240	0,655	0,171	0,227	0,488
Gender												
F	0,182	0,253	0,581	0,185	0,253	0,542	0,181	0,254	0,602	0,178	0,239	0,465

Table C.25: M2: Standard errors of fixed effects for n=800 and 500 replications

	Laplace			AGH15			AGH30			PQL		
Parameter	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	0,359	1,046	2,236	0,387	1,076	2,561	0,439	1,077	2,740	0,384	1,027	1,960
Age												
16-24	0,257	0,379	0,902	0,266	0,377	1,014	0,267	0,378	0,798	0,201	0,361	0,674
25-34	0,267	0,384	0,899	0,280	0,382	1,013	0,271	0,382	0,793	0,201	0,366	0,671
35-44	0,264	0,386	0,902	0,274	0,384	1,019	0,271	0,385	0,796	0,205	0,367	0,676
45-54	0,284	0,403	0,917	0,289	0,402	1,024	0,289	0,402	0,812	0,212	0,385	0,726
55-64	0,311	0,439	0,930	0,321	0,440	1,028	0,318	0,440	0,810	0,266	0,421	0,760
65-74	0,315	0,444	0,984	0,319	0,442	1,088	0,327	0,444	0,887	0,238	0,424	0,769
Col												
A-F	0,257	0,380	1,151	0,258	0,383	0,836	0,261	0,382	0,815	0,219	0,359	0,689
F-A	0,238	0,363	1,045	0,244	0,370	0,823	0,243	0,371	1,078	0,225	0,354	1,012
F-F	0,159	0,213	0,454	0,152	0,214	0,441	0,155	0,214	0,415	0,133	0,205	0,432
F-L	0,155	0,212	0,468	0,153	0,213	0,499	0,155	0,213	0,449	0,134	0,203	0,416
Vehic /obst	0,157	0,215	0,464	0,158	0,215	0,424	0,156	0,215	0,447	0,125	0,205	0,421
Year												
1995-1999	0,142	0,185	0,345	0,136	0,186	0,377	0,138	0,186	0,362	0,110	0,178	0,338
2000-2004	0,171	0,226	0,420	0,168	0,227	0,448	0,169	0,227	0,469	0,134	0,217	0,436
2005-2010	0,239	0,332	1,083	0,238	0,334	0,615	0,234	0,332	0,720	0,202	0,318	0,620
Local												
In agglo	0,110	0,149	0,322	0,108	0,150	0,295	0,108	0,150	0,300	0,097	0,144	0,283
Gender												
F	0,116	0,158	0,355	0,116	0,159	0,368	0,115	0,158	0,310	0,101	0,152	0,293

Table C.26: M2: Standard errors of fixed effects for n=2000 and 500 replications

	Laplace			AGH15			AGH30			PQL		
Parameter	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**	**min**	**mean**	**max**
Intercept	0,316	1,018	2,540	0,377	1,028	2,469	0,331	1,028	2,031	0,352	0,996	2,109
Age												
16-24	0,209	0,303	0,538	0,219	0,306	0,581	0,219	0,304	0,685	0,213	0,299	0,577
25-34	0,218	0,307	0,544	0,223	0,309	0,580	0,225	0,308	0,681	0,218	0,302	0,578
35-44	0,216	0,308	0,535	0,224	0,311	0,582	0,223	0,309	0,682	0,217	0,304	0,581
45-54	0,230	0,323	0,573	0,239	0,325	0,599	0,237	0,323	0,708	0,229	0,317	0,606
55-64	0,266	0,354	0,614	0,264	0,354	0,623	0,260	0,357	1,055	0,264	0,348	0,879
65-74	0,265	0,356	0,630	0,265	0,359	0,657	0,264	0,357	0,750	0,257	0,350	0,666
Col												
A-F	0,206	0,305	0,606	0,211	0,306	0,894	0,210	0,307	1,037	0,214	0,300	0,719
F-A	0,200	0,299	0,740	0,201	0,293	0,589	0,194	0,304	1,032	0,195	0,292	1,134
F-F	0,127	0,173	0,323	0,126	0,172	0,333	0,127	0,175	0,538	0,123	0,169	0,423
F-L	0,124	0,171	0,335	0,127	0,171	0,326	0,124	0,173	0,454	0,123	0,168	0,460
Vehic /obst	0,129	0,173	0,363	0,129	0,174	0,336	0,127	0,174	0,342	0,128	0,170	0,308
Year												
1995-1999	0,113	0,150	0,265	0,112	0,150	0,262	0,112	0,151	0,331	0,111	0,147	0,282
2000-2004	0,140	0,183	0,327	0,138	0,183	0,316	0,137	0,184	0,458	0,136	0,179	0,356
2005-2010	0,192	0,267	0,486	0,199	0,267	0,531	0,191	0,271	0,592	0,196	0,263	0,831
Local												
In agglo	0,090	0,121	0,220	0,088	0,120	0,210	0,089	0,123	0,320	0,088	0,118	0,259
Gender												
F	0,094	0,128	0,222	0,095	0,127	0,227	0,092	0,130	0,347	0,093	0,125	0,287

Table C.27: M2: Standard errors of fixed effects for n=3000 and 500 replications

	Laplace			AGH15			AGH30			PQL		
Parameter	**800**	**2000**	**3000**	**800**	**2000**	**3000**	**800**	**2000**	**3000**	**800**	**2000**	**3000**
Intercept	0,010	-0,042	-0,006	0,096	-0,014	0,083	0,131	-0,030	0,056	0,216	-0,105	0,031
Age												
16-24	-0,069	0,015	-0,004	-0,148	0,045	0,012	-0,041	-0,055	0,043	-0,118	-0,005	0,044
25-34	-0,065	-0,024	-0,019	-0,150	0,063	0,031	-0,070	-0,001	0,027	-0,131	0,026	0,042
35-44	-0,070	0,007	-0,015	-0,164	0,035	0,005	-0,014	-0,011	0,046	-0,095	-0,015	0,040
45-54	-0,036	0,016	-0,060	-0,192	0,065	-0,008	-0,007	-0,001	0,079	-0,152	-0,004	0,040
55-64	-0,074	-0,001	-0,067	-0,213	0,084	0,007	-0,090	-0,014	0,071	-0,162	-0,009	0,035
65-74	-0,020	-0,018	-0,017	-0,199	0,056	-0,017	-0,017	-0,016	0,004	-0,138	-0,021	0,059
Col												
A-F	-0,007	-0,141	0,000	-0,004	0,008	0,004	-0,069	-0,012	0,029	0,002	-0,027	-0,002
F-A	0,004	-0,057	-0,075	-0,083	-0,018	-0,002	-0,113	-0,033	-0,014	-0,176	-0,013	0,014
F-F	-0,061	-0,046	-0,031	-0,102	-0,030	-0,024	-0,041	-0,013	-0,009	-0,077	-0,007	-0,009
F-L	-0,096	-0,006	-0,024	-0,107	-0,040	-0,049	-0,080	0,019	-0,004	-0,073	-0,019	-0,004
V-O	-0,043	-0,031	-0,004	-0,080	-0,032	-0,029	-0,056	0,007	-0,003	-0,030	-0,031	-0,014
Year												
1995-1999	-0,041	0,028	0,012	-0,032	-0,048	0,007	-0,046	0,005	-0,013	-0,037	0,018	-0,007
2000-2004	-0,067	0,026	0,014	0,001	-0,069	0,001	-0,038	-0,025	-0,012	-0,059	0,033	-0,011
2005-2010	0,031	-0,002	0,021	-0,075	-0,019	0,028	-0,046	0,038	0,006	-0,096	-0,001	-0,033
Local												
Inag	-0,021	-0,012	0,003	-0,087	0,005	-0,007	-0,059	0,001	0,026	-0,009	0,025	-0,004
Gender												
F	-0,032	-0,015	0,000	-0,015	-0,027	-0,024	-0,010	-0,003	-0,022	-0,042	-0,028	-0,009
SD	-0,252	-0,352	-0,260	-0,161	-0,380	-0,210	-0,158	-0,309	-0,217	-0,190	-0,262	-0,477

SD: Standard Deviation

Table C.28: M2: Bias of mean estimates for 100 replications

	Laplace			AGH15			AGH30			PQL		
Parameter	**800**	**2000**	**3000**	**800**	**2000**	**3000**	**800**	**2000**	**3000**	**800**	**2000**	**3000**
Intercept	0,081	0,040	-0,066	0,046	-0,019	0,071	0,097	0,043	-0,012	0,079	-0,004	0,038
Age												
16-24	-0,029	-0,001	0,003	-0,005	-0,022	-0,012	-0,063	-0,064	-0,012	-0,002	0,004	-0,025
25-34	-0,044	-0,011	-0,023	-0,023	-0,025	-0,017	-0,085	-0,060	-0,008	-0,034	-0,019	-0,028
35-44	-0,023	-0,022	-0,011	-0,048	-0,018	-0,015	-0,082	-0,071	-0,010	-0,028	-0,006	-0,018
45-54	-0,001	-0,012	-0,007	-0,025	-0,039	0,001	-0,084	-0,068	-0,033	-0,025	-0,016	-0,017
55-64	-0,044	-0,007	-0,006	-0,071	-0,040	-0,018	-0,127	-0,064	-0,022	-0,047	-0,027	-0,028
65-74	-0,050	0,008	-0,011	-0,017	-0,016	-0,005	-0,042	-0,061	-0,014	0,037	-0,023	-0,029
Col												
A-F	-0,042	-0,044	0,005	-0,089	-0,020	-0,023	-0,018	0,014	-0,002	0,012	0,011	-0,020
F-A	-0,128	-0,012	-0,012	-0,114	-0,036	-0,031	-0,090	-0,037	-0,016	-0,070	-0,024	-0,028
F-F	-0,082	-0,016	-0,015	-0,070	-0,023	-0,033	-0,084	-0,019	-0,005	-0,040	-0,016	-0,014
F-L	-0,044	-0,037	-0,017	-0,073	-0,028	-0,024	-0,048	-0,031	-0,009	-0,043	-0,020	-0,018
Vehic /obst	-0,019	-0,006	0,000	-0,040	-0,011	-0,014	-0,010	0,000	0,004	-0,005	-0,002	-0,004
Year												
1995-1999	-0,035	-0,008	-0,009	-0,020	-0,002	-0,013	-0,004	-0,013	-0,014	0,017	0,006	-0,015
2000-2004	-0,058	-0,011	-0,010	-0,034	0,004	-0,006	-0,019	-0,022	-0,009	0,002	-0,002	-0,016
2005-2010	-0,026	0,011	0,011	-0,054	0,007	0,006	-0,015	-0,009	-0,021	0,007	0,008	-0,027
Local												
In agglo	-0,021	-0,011	-0,011	-0,005	-0,007	-0,013	-0,022	-0,006	-0,004	-0,012	-0,010	0,001
Gender												
F	-0,032	-0,005	-0,005	-0,027	-0,013	-0,011	-0,036	-0,012	-0,003	-0,028	-0,010	0,003
SD	-0,244	-0,302	-0,291	-0,223	-0,222	-0,264	-0,178	-0,223	-0,263	-0,329	-0,332	-0,340

SD: Standard Deviation

Table C.29: M2: Bias of mean estimates for 500 replications

		Laplace				AGH15				AGH30				PQL		
nR	**param.**	**800**	**2000**	**5000**		**800**	**2000**	**5000**		**800**	**2000**	**5000**		**800**	**2000**	**5000**
	all	0,006	0,009	0,005	0,015	0,010	0,003	0,005	0,006	0,004	0,014	0,005	0,013			
100	fixed	0,003	0,002	0,001	0,015	0,002	0,001	0,004	0,000	0,001	0,013	0,001	0,001			
	SD	0,063	0,124	0,068	0,026	0,144	0,044	0,025	0,095	0,047	0,036	0,068	0,227			
	all	0,006	0,005	0,005	0,005	0,003	0,004	0,006	0,004	0,004	0,007	0,006	0,007			
500	fixed	0,003	0,000	0,000	0,003	0,001	0,001	0,004	0,002	0,000	0,001	0,000	0,000			
	SD	0,060	0,091	0,085	0,050	0,049	0,070	0,032	0,050	0,069	0,108	0,110	0,116			

Table C.30: M2: MSE of mean estimates

Appendix D

Codes R

```
library (lme4)

#Importation fichier de données
new_maisv<-read.csv2("NEW_MAISV.csv")

#Les modalités de référence
new_maisv$age_cdctr<-relevel(new_maisv$age_cdctr, ref="75+")
new_maisv$localisation3<-relevel(new_maisv$localisation3,
ref="Hors agglo (N,D)/Lieu-dit")
new_maisv$annee1<-relevel(new_maisv$annee1, ref="1991-1994")
new_maisv$sexe_cdctr<-relevel(new_maisv$sexe_cdctr, ref="M")
new_maisv$col<-relevel(new_maisv$col, ref="L-F")
new_maisv$EES<-relevel(new_maisv$EES, ref="60-90")

#Aperçu et résumé des données
head(new_maisv)
summary(new_maisv)


#Modèle sans effet aléatoire (glm)
glm_new<- glm(mais3plus_veh2~ EES+age_cdctr+col+annee1+localisation3
+sexe_cdctr, family = binomial,data=new_maisv)


#Test presence effet aléatoire
#Function provided by B. Bolker in
#http://glmm.wdfiles.com/local--files/examples/glmmfuns.R
#(http://glmm.wikidot.com)
```

```
varprof <- function(mm,lower=0,upper=20,n=101) {
  sg <- seq(lower, upper, len = n)
  orig.sd <- attr(VarCorr(mm)[[1]],"stddev")
  dev <- mm@deviance
  nc <- length(dev)
  nms <- names(dev)
  vals <- matrix(0, nrow = length(sg), ncol = nc, dimnames = list(NULL, nms))
  update_dev <- function(sd) {
    .Call("mer_ST_setPars", mm, sd, PACKAGE = "lme4")
    .Call("mer_update_L", mm, PACKAGE = "lme4")
    res <- try(.Call("mer_update_RX", mm, PACKAGE = "lme4"), silent = TRUE)
    if (inherits(res, "try-error")) {
      val <- NA
    } else {
      .Call("mer_update_ranef", mm, PACKAGE = "lme4")
      .Call("mer_update_dev", mm, PACKAGE = "lme4") ## added for glmmML
      val <- mm@deviance
    }
    val
  }
  for (i in seq(along = sg)) {
      vals[i,] <- update_dev(sg[i])
    }
  update_dev(orig.sd) ## hack! to restore original sd
  data.frame(sd=sg,vals)
}

Test_eff_al<-function(modele)
{
dev0 <- varprof(modele, lower = 0, upper = 0, n = 1)[["ML"]] #
obsdev <- dev0 - deviance(modele)
p1<-pchisq(obsdev, df = 1, lower.tail = FALSE) # p-value
p2<-0.5 * pchisq(obsdev, df = 1, lower.tail = FALSE)
res=c(p1,p2)
return(res)
}

#Test de la présence d'effet aléatoire sur les différents modèles
```

```
#Les différents glmm possibles
glmm_new_annee1<- glmer(mais3plus_veh2~EES+age_cdctr+col+localisation3
+sexe_cdctr+(1|annee1), family = binomial,data=new_maisv)

glmm_new_col<- glmer(mais3plus_veh2~EES+age_cdctr+annee1+localisation3
+sexe_cdctr+(1|col), family = binomial,data=new_maisv)

glmm_new_col2<- glmer(mais3plus_veh2~REESES+age_cdctr+annee1+localisation3
+sexe_cdctr+(1|col), family = binomial,data=new_maisv)

 glmm_new_EES<- glmer(mais3plus_veh2~age_cdctr+col+annee1+localisation3
+sexe_cdctr+(1|EES), family = binomial,data=new_maisv)


glmm_new_age<- glmer(mais3plus_veh2~EES+col+annee1+localisation3+sexe_cdctr
+(1|age_cdctr), family = binomial,data=new_maisv)

glmm_new_sexe<- glmer(mais3plus_veh2~EES+age_cdctr+col+annee1+localisation3
+(1|sexe_cdctr), family = binomial,data=new_maisv)

glmm_new_localisation<- glmer(mais3plus_veh2~EES+age_cdctr+col+annee1
+sexe_cdctr+(1|localisation3), family = binomial,data=new_maisv)


#Test sur chaque glmm
Test_eff_al(glmm_new_col)
Test_eff_al(glmm_new_EES)
Test_eff_al(glmm_new_age)
Test_eff_al(glmm_new_sexe)
Test_eff_al(glmm_new_localisation)



#Les 3 modèles admettant un effet aléatoire avec les différentes méthodes
# et le temps d'exécution

#Laplace
system.time(glmm_new_annee1<- glmer(mais3plus_veh2~EES+age_cdctr+col
+localisation3+sexe_cdctr+(1|annee1), family = binomial,
data=new_maisv,verbose=TRUE))
```

```
system.time(glmm_new_col<- glmer(mais3plus_veh2~EES+age_cdctr+annee1
+localisation3+sexe_cdctr+(1|col), family = binomial,
data=new_maisv,verbose=TRUE))

system.time(glmm_new_EES<- glmer(mais3plus_veh2~age_cdctr+annee1+col
+localisation3+sexe_cdctr+(1|EES), family = binomial,
data=new_maisv,verbose=TRUE))

#Gauss-Hermite 4
system.time(Glmm_new_annee1<- glmer(mais3plus_veh2~EES+age_cdctr+col
+localisation3+sexe_cdctr+(1|annee1), family = binomial,data=new_maisv,
verbose=TRUE, nAGQ=4))

system.time(Glmm_new_col<- glmer(mais3plus_veh2~EES+age_cdctr+annee1
+localisation3+sexe_cdctr+(1|col), family = binomial,data=new_maisv,
verbose=TRUE, nAGQ=4))

system.time(Glmm_new_EES<- glmer(mais3plus_veh2~age_cdctr+annee1+col
+localisation3+sexe_cdctr+(1|EES), family = binomial,data=new_maisv,
verbose=TRUE, nAGQ=4))

#Gauss-Hermite 15
system.time(Glmm15_new_annee1<- glmer(mais3plus_veh2~EES+age_cdctr+col
+localisation3+sexe_cdctr+(1|annee1), family = binomial,data=new_maisv,
verbose=TRUE,nAGQ=15))

system.time(Glmm15_new_col<- glmer(mais3plus_veh2~EES+age_cdctr+annee1
+localisation3+sexe_cdctr+(1|col), family = binomial,data=new_maisv,
verbose=TRUE,nAGQ=15))

system.time(Glmm15_new_EES<- glmer(mais3plus_veh2~age_cdctr+annee1+col
+localisation3+sexe_cdctr+(1|EES), family = binomial,data=new_maisv,
verbose=TRUE, nAGQ=15))


#Gauss-Hermite 30
system.time(Glmm30_new_annee1<- glmer(mais3plus_veh2~EES+age_cdctr+col
+localisation3+sexe_cdctr+(1|annee1), family = binomial,data=new_maisv,
verbose=TRUE, nAGQ=30))
```

```
system.time(Glmm30_new_col<- glmer(mais3plus_veh2~EES+age_cdctr+annee1
+localisation3+sexe_cdctr+(1|col), family = binomial,data=new_maisv,
verbose=TRUE, nAGQ=30))

system.time(Glmm30_new_EES<- glmer(mais3plus_veh2~age_cdctr+annee1
+col+localisation3+sexe_cdctr+(1|EES), family = binomial,data=new_maisv,
verbose=TRUE, nAGQ=30))


#Gauss-Hermite 60
system.time(Glmm60_new_annee1<- glmer(mais3plus_veh2~EES+age_cdctr+col
+localisation3+sexe_cdctr+(1|annee1), family = binomial,data=new_maisv,
verbose=TRUE, nAGQ=60))

system.time(Glmm60_new_col<- glmer(mais3plus_veh2~EES+age_cdctr+annee1
+localisation3+sexe_cdctr+(1|col), family = binomial,data=new_maisv,
verbose=TRUE, nAGQ=60))

system.time(Glmm60_new_EES<- glmer(mais3plus_veh2~age_cdctr+annee1+col
+localisation3+sexe_cdctr+(1|EES), family = binomial,data=new_maisv,
verbose=TRUE, nAGQ=60))



#Quasi-vraisemblance
system.time(glmmPQ_new_annee1<- glmmPQL(mais3plus_veh2~EES+age_cdctr+col
+localisation3+sexe_cdctr, random= ~ 1 | annee1, family = binomial,
data=new_maisv, verbose=TRUE))

system.time(glmmPQ_new_col<- glmmPQL(mais3plus_veh2~EES+age_cdctr+annee1
+localisation3+sexe_cdctr, random= ~ 1 | col, family = binomial,
data=new_maisv, verbose=TRUE))

system.time(glmmPQ_new_EES<- glmmPQL(mais3plus_veh2~age_cdctr+col+annee1
+localisation3+sexe_cdctr, random= ~  1 | EES, family = binomial,
data=new_maisv,verbose=TRUE))


#OR et intervalles de confiance
```

```
#Pour glm
or_glm <- function(x, alpha=0.05){
          resu <- cbind(coef(x), confint(x,level=1-alpha))
         exp (resu)
          }

#Pour glmm
or_glmm <- function(x, alpha = 0.05){
          ## récup les coefs du modèles (effets fixes)
          b <- fixef(x)
          ## matrice de var-cov des effets fixes
          v <- as.matrix(vcov(x))
          ## écarts-types de b
          se <- sqrt(diag(v))
          ## quantile loi normale
          q <- qnorm(1 - alpha/2)
          ## résultat
          resu <- cbind(or = b, lower = b - q * se, upper = b + q * se)
          exp(resu)
          }


#AUC par validation croisée

#AUC par validation croisée pour glm
donnees<-new_maisv
n<-nrow(donnees)
K<-8 #nombre de validation croisée
taille<-n%/%K
alea<-runif(n)
rang<-rank(alea)
bloc<-(rang-1)%/%taille+1
bloc<-as.factor(bloc)

VC_auc_glm1<-function(formule,donnees,K)
{
all.auc=matrix(nr=K,nc=6)

for (k in 1:K) {
```

```
glm0<-glm(formule,family=binomial,data=donnees[bloc!=k,])

pred<- predict(glm0, newdata = donnees [bloc==k,], type= "response")

tab<-data.frame(vmais3plus=donnees$vmais3plus[bloc==k],pi=pred)

auc_tab<- with(tab, AUC(neg = pi[vmais3plus == 0], pos = pi[vmais3plus == 1]))
ICauc_tab<-ciAUC(auc_tab, alpha = 0.05)

all.auc[k,]<-c(auc_tab$auc, ICauc_tab$ci ,auc_tab$se, auc_tab$n0, auc_tab$n1)

}
apply(all.auc,2,mean)
}
```

```
#AUC par valdation croisée pour glmm
donnees<-new_maisv
n<-nrow(donnees)
K<-8
taille<-n%/%K
set.seed(5)
alea<-runif(n)
rang<-rank(alea)
bloc<-(rang-1)%/%taille+1
bloc<-as.factor(bloc)



VC_auc_fit<-function(formule,donnees,K)
{

all.auc=matrix(nr=K,nc=6)

for (k in 1:K) {
glmm0<-glmer(formule,family=binomial,data=donnees[bloc!=k,])
```

```
glmm<-glmer(formule,family=binomial,data=donnees[bloc==k,])


x=model.matrix(glmm)

Z <- getME(glmm, "Z")


#1 effet aléatoire
#eff_al<- ranef(glmm0)$col[,]
#eff_al<- ranef(glmm0)$annee1[,]
#eff_al<- ranef(glmm0)$EES[,]



#3 effets aléatoires
#eff_al<-c(ranef(glmm0)$ EES[,], ranef(glmm0)$col[,], ranef(glmm0)$annee1[,])

#2 effets aléatoires
eff_al<-c(ranef(glmm0)$ col[,], ranef(glmm0)$annee1[,])
#eff_al<-c(ranef(glmm0)$ EES[,], ranef(glmm0)$annee1[,])
#eff_al<-c(ranef(glmm0)$ EES[,], ranef(glmm0)$col[,])



pred<-1/(1+exp(-1*(x %*% fixef(glmm0)+Z%*% eff_al)))
#Remplacer à chaque fois col[,] par le bon effet aléatoire en fonction du modèle

tab<-structure(list(vmais3plus=new_maisv$vmais3plus[bloc==k],pi=pred),
class="data.frame", .Names=c("vmais3plus","pi"), row.names=c(NA,136L))


auc_tab<- with(tab, AUC(neg = pi[vmais3plus == 0], pos = pi[vmais3plus == 1]))
ICauc_tab<-ciAUC(auc_tab, alpha = 0.05)

all.auc[k,]<-c(auc_tab$auc, ICauc_tab$ci ,auc_tab$se, auc_tab$n0, auc_tab$n1)

}
apply(all.auc,2,mean)
```

```
}



#Matrice de confusion pour glm par  validation croisée
confusion_glm_vc <-function(formule,p)
{

conf=matrix(nr=8,nc=4)

for (k in 1:K) {

glm0<-glm(formule,family=binomial,data=new_maisv[bloc!=k,])
pred<- predict(glm0, newdata = new_maisv [bloc==k,], link= "response")


tab<-structure(list(vmais3plus=new_maisv$vmais3plus[bloc==k],pi=pred),
class="data.frame", .Names=c("vmais3plus","pi"), row.names=c(NA,103L))


tab$mais3 [tab$pi>=p]<-1
tab$mais3 [tab$pi<p]<-0

tab$vv [tab$vmais3plus==1 & tab$mais3==1]<-1
tab$ff [tab$vmais3plus==0 & tab$mais3==0]<-1
tab$vf [tab$vmais3plus==1 & tab$mais3==0]<-1
tab$fv [tab$vmais3plus==0 & tab$mais3==1]<-1
tab$vv [is.na(tab$vv)]<-0
tab$ff [is.na(tab$ff)]<-0
tab$vf [is.na(tab$vf)]<-0
tab$fv [is.na(tab$fv)]<-0

conf[k,]<-c(sum(tab$vv), sum(tab$fv), sum(tab$vf),sum(tab$ff))

}
m<-matrix(c(apply(conf,2,mean)),2,2)
}


#Matrice de confusion pour glmm par  VC
```

```
#Pour les modèles à 1 effet aléatoire
confusion_glmm_fitvc1 <-function(formule,p)
{


conf=matrix(nr=8,nc=4)

for (k in 1:K) {
glmm1<-glmer(formule,family=binomial,data=new_maisv[bloc!=k,])
glmm2<-glmer(formule,family=binomial,data=new_maisv[bloc==k,])


x=model.matrix(glmm2)
Z <- getME(glmm2, "Z")
eff_al<-ranef(glmm1)[[1]][,]


pred<-1/(1+exp(-1*(x %*% fixef(glmm1) +Z%*%eff_al)))

pred<-pred[,1]

tab<-structure(list(vmais3plus=new_maisv$vmais3plus[bloc==k],pi=pred),
class="data.frame", .Names=c("vmais3plus","pi"), row.names=c(NA,103L))

tab$mais3 [tab$pi>=p]<-1
tab$mais3 [tab$pi<p]<-0


tab$vv [tab$vmais3plus==1 & tab$mais3==1]<-1
tab$ff [tab$vmais3plus==0 & tab$mais3==0]<-1
tab$vf[tab$vmais3plus==1 & tab$mais3==0]<-1
tab$fv [tab$vmais3plus==0 & tab$mais3==1]<-1
tab$vv [is.na(tab$vv)]<-0
tab$ff [is.na(tab$ff)]<-0
tab$vf [is.na(tab$vf)]<-0
tab$fv [is.na(tab$fv)]<-0


conf[k,]<-c(sum(tab$vv), sum(tab$fv), sum(tab$vf),sum(tab$ff))
```

```
}
m<-matrix(c(apply(conf,2,mean)),2,2)
}


#Pour les modèles à 2 effets aléatoires
confusion_glmm_fitvc2<-function(formule,p)
{

conf=matrix(nr=8,nc=4)

for (k in 1:K) {
glmm1<-glmer(formule,family=binomial,data=new_maisv[bloc!=k,])

glmm2<-glmer(formule,family=binomial,data=new_maisv[bloc==k,])


x=model.matrix(glmm2)
Z <- getME(glmm2, "Z")
eff_al<-c(ranef(glmm1)[[1]][,],ranef(glmm1)[[2]][,])


pred<-1/(1+exp(-1*(x %*% fixef(glmm1) +Z%*%eff_al)))

pred<-pred[,1]

tab<-structure(list(vmais3plus=new_maisv$vmais3plus[bloc==k],pi=pred),
class="data.frame", .Names=c("vmais3plus","pi"), row.names=c(NA,103L))

tab$mais3 [tab$pi>=p]<-1
tab$mais3 [tab$pi<p]<-0


tab$vv [tab$vmais3plus==1 & tab$mais3==1]<-1
tab$ff [tab$vmais3plus==0 & tab$mais3==0]<-1
tab$vf[tab$vmais3plus==1 & tab$mais3==0]<-1
tab$fv [tab$vmais3plus==0 & tab$mais3==1]<-1
tab$vv [is.na(tab$vv)]<-0
tab$ff [is.na(tab$ff)]<-0
```

```
tab$vf [is.na(tab$vf)]<-0
tab$fv [is.na(tab$fv)]<-0


conf[k,]<-c(sum(tab$vv), sum(tab$fv), sum(tab$vf),sum(tab$ff))


}
m<-matrix(c(apply(conf,2,mean)),2,2)
}



#Pour les modèles à 3 effets aléatoires
confusion_glmm_fitvc3<-function(formule,p)
{

conf=matrix(nr=8,nc=4)

for (k in 1:K) {
glmm1<-glmer(formule,family=binomial,data=new_maisv[bloc!=k,])

glmm2<-glmer(formule,family=binomial,data=new_maisv[bloc==k,])


x=model.matrix(glmm2)
Z <- getME(glmm2, "Z")
eff_al<-c(ranef(glmm1)[[1]][,],ranef(glmm1)[[2]][,], ranef(glmm1)[[3]][,])


pred<-1/(1+exp(-1*(x %*% fixef(glmm1) +Z%*%eff_al)))

pred<-pred[,1]

tab<-structure(list(vmais3plus=new_maisv$vmais3plus[bloc==k],pi=pred),
class="data.frame", .Names=c("vmais3plus","pi"), row.names=c(NA,103L))

tab$mais3 [tab$pi>=p]<-1
tab$mais3 [tab$pi<p]<-0
```

```
tab$vv [tab$vmais3plus==1 & tab$mais3==1]<-1
tab$ff [tab$vmais3plus==0 & tab$mais3==0]<-1
tab$vf[tab$vmais3plus==1 & tab$mais3==0]<-1
tab$fv [tab$vmais3plus==0 & tab$mais3==1]<-1
tab$vv [is.na(tab$vv)]<-0
tab$ff [is.na(tab$ff)]<-0
tab$vf [is.na(tab$vf)]<-0
tab$fv [is.na(tab$fv)]<-0


conf[k,]<-c(sum(tab$vv), sum(tab$fv), sum(tab$vf),sum(tab$ff))


}
m<-matrix(c(apply(conf,2,mean)),2,2)
}



#Pour récupérer la partie fixe de chaque modèle
f_col<-mais3plus_veh2 ~ EES + age_cdctr + annee1 + localisation3
+ sexe_cdctr
f_EES<-mais3plus_veh2 ~ age_cdctr + col + annee1 + localisation3
+ sexe_cdctr
f_annee1<- mais3plus_veh2 ~ EES + age_cdctr + col + localisation3
+ sexe_cdctr

#Fonction qui simule les données et calcule mean, min et max des estimations,
#des biais et des erreurs standards
estim_sim2<-function(glmm, f_al, Q, r, n)
{
i<-1
p<-length (fixef(glmm))
q<- length(ranef(glmm)[[1]][,])
eff_al<-matrix(data=NA,r,q)
sd_eff_al<-rep(NA,r)
eff_fix<-matrix(data=NA,r,p)
se<-matrix(data=NA,r,p)
indic<-matrix(data=NA,r,4)
```

```
temps<- matrix(data=NA,r,3)


for (i in 1: r)

#while(i<=r)

{

#On simule les données selon notre base de données
#REESES_s = rnorm(n,0,sd=15.79)
EES_s<-rmultinom(n,1, c(0.047, 0.325, 0.237, 0.179, 0.136, 0.076))
localisation_s<- rbinom(n,1,prob=0.372)
sexe_cdctr_s<- rbinom(n,1,prob=0.293)
 age_cdctr_s<-rmultinom(n,1, c(0.039, 0.232, 0.243, 0.212, 0.125, 0.085, 0.064))
col_s<-rmultinom(n,1, c(0.237, 0.039, 0.055, 0.248, 0.247, 0.174))
annee1_s<-rmultinom(n,1, c(0.189, 0.561, 0.192, 0.058))
#juste pr récupérer la matrice de design x
mais3plus_s<-rbinom(n,1,prob=0.5)


Donnees_sim<-data.frame(mais3plus=as.factor(mais3plus_s),
localisation= as.factor(localisation_s), sexe_cdct= as.factor(sexe_cdctr_s),
 age_cdctr=t(age_cdctr_s), col=t(col_s), EES=t(EES_s), annee1=t(annee1_s))

Donnees_sim$mais3plus_veh2[Donnees_sim$mais3plus==1]<-"oui"
Donnees_sim$mais3plus_veh2[Donnees_sim$mais3plus==0]<-"non"

Donnees_sim$localisation3[Donnees_sim$localisation==0]<-"Hors agglo"
Donnees_sim$localisation3[Donnees_sim$localisation==1]<-"En agglo"

Donnees_sim$sexe_cdctr [Donnees_sim$sexe_cdct==0]<-"M"
Donnees_sim$sexe_cdctr [Donnees_sim$sexe_cdct==1]<-"F"

Donnees_sim$age_cdctr[Donnees_sim$age_cdctr.1==1]<-"75+"
Donnees_sim$age_cdctr[Donnees_sim$age_cdctr.2==1]<-"16-24"
Donnees_sim$age_cdctr[Donnees_sim$age_cdctr.3==1]<-"25-34"
Donnees_sim$age_cdctr[Donnees_sim$age_cdctr.4==1]<-"35-44"
Donnees_sim$age_cdctr[Donnees_sim$age_cdctr.5==1]<-"45-54"
Donnees_sim$age_cdctr[Donnees_sim$age_cdctr.6==1]<-"55-64"
```

```
Donnees_sim$age_cdctr[Donnees_sim$age_cdctr.7==1]<-"65-74"

Donnees_sim$col[Donnees_sim$col.1==1]<-"L-F"
Donnees_sim$col[Donnees_sim$col.2==1]<-"A-F"
Donnees_sim$col[Donnees_sim$col.3==1]<-"F-A"
Donnees_sim$col[Donnees_sim$col.4==1]<-"F-F"
Donnees_sim$col[Donnees_sim$col.5==1]<-"F-L"
Donnees_sim$col[Donnees_sim$col.6==1]<-"Véhic seul"

Donnees_sim$EES[Donnees_sim$EES.1==1]<-"60-90"
Donnees_sim$EES[Donnees_sim$EES.2==1]<-"0-20"
Donnees_sim$EES[Donnees_sim$EES.3==1]<-"20-30"
Donnees_sim$EES[Donnees_sim$EES.4==1]<-"30-40"
Donnees_sim$EES[Donnees_sim$EES.5==1]<-"40-50"
Donnees_sim$EES[Donnees_sim$EES.6==1]<-"50-60"

Donnees_sim$annee1[Donnees_sim$annee1.1==1]<-"1991-1994"
Donnees_sim$annee1[Donnees_sim$annee1.2==1]<-"1995-1999"
Donnees_sim$annee1[Donnees_sim$annee1.3==1]<-"2000-2004"
Donnees_sim$annee1[Donnees_sim$annee1.4==1]<-"2005-2010"

Donnees_sim$mais3plus_veh2<-as.factor(Donnees_sim$mais3plus_veh2)
Donnees_sim$localisation3<-as.factor(Donnees_sim$localisation3)
Donnees_sim$ sexe_cdctr <-as.factor(Donnees_sim$sexe_cdctr)
Donnees_sim$ age_cdctr <-as.factor(Donnees_sim$age_cdctr)
Donnees_sim$col<-as.factor(Donnees_sim$col)
Donnees_sim$annee1<-as.factor(Donnees_sim$annee1)
Donnees_sim$EES<-as.factor(Donnees_sim$EES)

Donnees_sim$age_cdctr<-relevel(Donnees_sim$age_cdctr, ref="75+")
Donnees_sim $localisation3<-relevel(Donnees_sim$localisation3, ref="Hors agglo")
Donnees_sim$annee1<-relevel(Donnees_sim$annee1, ref="1991-1994")
Donnees_sim$sexe_cdctr<-relevel(Donnees_sim$sexe_cdctr, ref="M")
Donnees_sim$col<-relevel(Donnees_sim$col, ref="L-F")
Donnees_sim$EES<-relevel(Donnees_sim$EES, ref="60-90")



#Pr récupérer la matrice design x
```

```
x<-model.matrix(f_al,Donnees_sim)


#Pour récupérer les paramètres selon lesquels on va simuler

#effets fixes
beta<-fixef(glmm)

#effets aléatoires
v<-names(ranef(glmm))
al<- as.integer(Donnees_sim[,v])
sigma<- sqrt(VarCorr(glmm)[[1]][,])
b_eff = rnorm(q,0,sd=sigma)

#On simule la partie linéaire
eta<- x%*%beta+b_eff[al]

#simulation de  la var explicative grâce au modèle et aux données simulées
Donnees_sim$vmais3plus_veh2<-rbinom(n,1,prob=plogis(eta))
Donnees_sim$mais3plus_veh2<-as.factor(Donnees_sim$vmais3plus_veh2)

#On ajuste les données simulées par un glmm et la méthode à tester
f<-formula(glmm)

st<-system.time(glmm0<- glmer(f, family = binomial,  data=Donnees_sim, nAGQ=Q))

ev1<-vcov(glmm0)[1,2]/sqrt(vcov(glmm0)[1,1]*vcov(glmm0)[2,2])
ev2<-vcov(glmm0)[1,3]/sqrt(vcov(glmm0)[1,1]*vcov(glmm0)[3,3])
ev3<-vcov(glmm0)[1,4]/sqrt(vcov(glmm0)[1,1]*vcov(glmm0)[3,4])
ev4<-vcov(glmm0)[1,5]/sqrt(vcov(glmm0)[1,1]*vcov(glmm0)[3,5])
ev5<-vcov(glmm0)[1,6]/sqrt(vcov(glmm0)[1,1]*vcov(glmm0)[3,6])

if(abs(ev1) < 0.9 & abs(ev2) < 0.9 & abs(ev3) < 0.9 & abs(ev4) < 0.9
& abs(ev5) < 0.9 & abs(ev1) >0.01  & abs(ev2) >0.01 & abs(ev3) >0.01
& abs(ev4) >0.01 & abs(ev5) >0.01)
{
fix<-fixef(glmm0)
```

```
## récup les temps d'éxécution
temps[i,]<-st[1:3]

## récup les effets aléatoires du modèle
eff_al[i,]<-ranef(glmm0)[[1]][,]

## récup l'écart-type des effets aléatoires du modèle
sd_eff_al[i]<-sqrt(VarCorr(glmm0)[[1]][,])

## récup la variance des effets aléatoires du modèle
## - variance  des eff al modele0
#var_eff_al_m[i]<- (VarCorr(glmm0)[[1]][,] - VarCorr(glmm)[[1]][,])^2


## récup les coefs du modèle (effets fixes)
          eff_fix[i,]<- fixef(glmm0)


 ## matrice de var-cov des effets fixes
          v <- as.matrix(vcov(glmm0))

 ## écarts-types des estimations des effets fixes
          se [i,]<- sqrt(diag(v))

indic[i,]<-c(AIC(glmm0), BIC(glmm0), logLik(glmm0), deviance(glmm0))

}

}


#eff al
mean_eff_al<-apply(na.omit (eff_al),2,mean)
biais_eff_al<- mean_eff_al- ranef(glmm)[[1]][,]
var_eff_al<- apply(na.omit (eff_al),2,var)
mse_eff_al<- (biais_eff_al)^2+ var_eff_al
max_eff_al<-apply(na.omit (eff_al),2,max)
min_eff_al<-apply(na.omit (eff_al),2,min)

#sd eff al
```

```
mean_sd_eff_al<- mean(na.omit(sd_eff_al))
biais_sd_eff_al<- mean_sd_eff_al- sqrt(VarCorr(glmm)[[1]][,])
var_sd_eff_al<- var(na.omit(sd_eff_al))
mse_sd_eff_al<- (biais_sd_eff_al)^2+ var_sd_eff_al
max_sd_eff_al<-max(na.omit(sd_eff_al))
min_sd_eff_al<-min(na.omit(sd_eff_al))


#eff fix
mean_eff_fix<- apply(na.omit(eff_fix),2,mean)
biais_eff_fix<- mean_eff_fix- fixef(glmm)
var_eff_fix<- apply(na.omit(eff_fix),2,var)
mse_eff_fix<- (biais_eff_fix)^2+ var_eff_fix
max_eff_fix<-apply(na.omit(eff_fix),2,max)
min_eff_fix<-apply(na.omit(eff_fix),2,min)


#se
mean_se <- apply(na.omit(se),2,mean)
biais_se<- mean_se- sqrt(diag(as.matrix(vcov(glmm))))
var_se<- apply(na.omit(se),2,var)
mse_se<- (biais_se)^2+ var_se
max_se<-apply(na.omit(se),2,max)
min_se<-apply(na.omit(se),2,min)


#indic
mean_indic<- apply(na.omit(indic),2,mean)
biais_indic<- mean_indic- c(AIC(glmm), BIC(glmm), logLik(glmm),deviance(glmm))
var_indic<- apply(na.omit(indic),2,var)
mse_indic<- (biais_indic)^2+ var_indic
max_indic<-apply(na.omit(indic),2,max)
min_indic<-apply(na.omit(indic),2,min)

#temps
mean_temps <- apply(na.omit(temps),2,mean)
max_temps<-apply(na.omit(temps),2,max)
min_temps<-apply(na.omit(temps),2,min)
```

```
l<-list(



 #eff al
eff_al=eff_al,
mean_eff_al= mean_eff_al,
biais_eff_al= biais_eff_al,
var_eff_al=var_eff_al,
mse_eff_al= mse_eff_al,
max_eff_al=max_eff_al,
min_eff_al=min_eff_al,


#Var eff al
sd_eff_al=sd_eff_al,
mean_sd_eff_al= mean_sd_eff_al,
biais_sd_eff_al= biais_sd_eff_al,
var_sd_eff_al= var_sd_eff_al,
mse_sd_eff_al= mse_sd_eff_al,
max_sd_eff_al=max_sd_eff_al,
min_sd_eff_al=min_sd_eff_al,



#eff fix
eff_fix=eff_fix,
mean_eff_fix= mean_eff_fix,
biais_eff_fix = biais_eff_fix,
var_eff_fix= var_eff_fix,
mse_eff_fix= mse_eff_fix,
max_eff_fix=max_eff_fix,
min_eff_fix=min_eff_fix,


#se
```

```
se=se,
mean_se = mean_se,
biais_se= biais_se,
var_se= var_se,
mse_se= mse_se,
max_se=max_se,
min_se=min_se,


#indic
indic=indic,
mean_AIC= mean_indic[1],
biais_AIC= biais_indic[1],
var_AIC= var_indic[1],
mse_AIC= mse_indic[1],
max_AIC=max_indic[1],
min_AIC=min_indic[1],


mean_BIC= mean_indic[2],
biais_BIC= biais_indic[2],
var_BIC= var_indic[2],
mse_BIC= mse_indic[2],
max_BIC=max_indic[2],
min_BIC=min_indic[2],


mean_logV= mean_indic[3],
biais_logV= biais_indic[3],
var_logV= var_indic[3],
mse_logV= mse_indic[3],
max_logV=max_indic[3],
min_logV=min_indic[3],


mean_deviance= mean_indic[4],
biais_deviance= biais_indic[4],
var_deviance= var_indic[4],
mse_deviance= mse_indic[4],
max_deviance=max_indic[4],
```

```
min_deviance=min_indic[4],

#temps
temps=temps,
mean_temps=mean_temps,
max_temps=max_temps,
min_temps=min_temps

)
}


#mse pour effets fixes ou effets fixes + sd effets aléatoires
mse<-function(biais, eff, model,sd)
{
biais_carre<-(biais)^2
mse_1<-mean(biais_carre)
c<-ncol(eff)
l<- nrow(eff)
b_carre<-matrix(ncol=c, nrow=l)
for( i in 1 : l)
{
if(sd==0){f<- fixef(model)}
else {f<-c(fixef(model), sqrt(VarCorr(model)[[1]][,]))}
b_carre [i,] <-(eff[i,]-f)^2
}
s<-apply (b_carre,1,mean)
mse_2<-mean (na.omit(s))
l<-list(mse_1,mse_2)

}

#En integrant la sd des eff al, on met en argument
biais<-c(sim100_annee1_800$biais_eff_fix, sim100_annee1_800$biais_sd_eff_al)
eff<- cbind(sim100_annee1_800$eff_fix, sim100_annee1_800$sd_eff_al)

#pour sd effets aléatoires seulement
library(lme4)
mse_sd<-function(biais, sd, model)
{
```

```
biais_carre<-(biais)^2
mse_1<-mean(biais_carre)
l<-length(sd)
b_carre<-rep(NA,l)
for( i in 1 : l)
{
b_carre [i] <-(sd[i]- sqrt(VarCorr(model)[[1]][,]))^2
}
mse_2<-mean (na.omit(b_carre))
l<-list(mse_1,mse_2)

}

#Intervalle de confiance pour beta
IC_beta <- function(beta, se,alpha = 0.05)
{
       ## quantile loi normale
       q <- qnorm(1 - alpha/2)
       ## résultat
      res <- cbind(beta=beta , lower = beta - q * se, upper = beta + q * se)
}

#Probabilité de couverture pour beta
couverture<-function(mean_beta,mean_se,alpha = 0.05,beta)
{
ic<- IC_beta (mean_beta, mean_se,alpha = 0.05)
np<-nrow(ic)
r<-nrow(beta)
cv<-rep(NA,np)

for (i in 1:np)
{
b<-na.omit(beta[,i])
b_inf<-b[b<=ic[i,3]]
b_sup<-b_inf[b_inf>=ic[i,2]]

cv[i]<-length(b_sup)
}
cv<-cv*100/length(na.omit(beta[,1]))
}
```

Appendix E

Article

Analyzing the maximum abbreviated injury scale in vehicle crashes using a logistic normal model

12/11/2013

Length of Paper: 5233 words, 4 tables, Total: 6233words + 35 references

Corresponding author:
Fatima Meguellati
Laboratoire Paul Painlevé - UMR CNRS 8524, Université Lille 1
Avenue Paul Langevin, - 59655 Villeneuve d'Ascq Cedex, France
Laboratoire d'Accidentologie, de Biomécanique et d'études du comportement humain
PSA Peugeot Citroën - Renault (LAB) - 132 rue des Suisses 92000 Nanterre, France
Phone: 0033678255681
fatima.meguellati@yahoo.fr

Assi N'Guessan
Laboratoire Paul Painlevé - UMR CNRS 8524, Université Lille 1
Avenue Paul Langevin, - 59655 Villeneuve d'Ascq Cedex, France
Phone: 0033328767457
assi.nguessan@polytech-lille.fr

Thierry Hermitte
Laboratoire d'Accidentologie, de Biomécanique et d'études du comportement humain
PSA Peugeot Citroën - Renault (LAB) - 132 rue des Suisses 92000 Nanterre, France
Phone: 0033176873513
thierry.hermitte@lab.com

ABSTRACT

This paper carries out a modelling centered on the estimation of Maximum Abbreviated Injury Scale-based injury predicting models using a logistic normal model with random effects. This study identifies Energy Equivalent Speed, collision type, crash year, location, driver's age and gender for the analysis of the gravity in vehicle crashes. The study shows that the collision type variable is best modeling by random effects than fixed effects and, that Energy Equivalent Speed, crash year, location, driver's age and gender are contributing factors with fixed effects to the injury severity. So we take advantage of the mixed logit model ability to account for unobserved effects that are difficult to quantify and may affect the model estimation. Crashes from the data base of detailed studies of personal accidents occurring from 1991 to 2010 in France were used. The estimation of the parameters (fixed and random effects) was done by several approximation methods. It was found that Energy Equivalent Speed is by far, the most significant variable. If we consider drivers aged 75 and over as a reference, drivers in the 25-34 age bracket contribute significantly followed by the 35-44 age bracket and then by the 55-65 age bracket. It was also observed that the vehicles crashed between 1995 and 1999 are more significant as compared to the reference one 1991-1994. To assess the performance of the model proposed, a binary logit model was compared with the mixed logit by the means of cross validation. The obtained results reveal that, our mixed logistic-normal model is preferred because it detects more vehicle crashes with serious injuries than the classic logistic regression.

Key Words: Crash data, Injury Scale, Binary response, Mixed effects logistic regression, Laplace's method, Adaptive Gauss-Hermite, Simulation, R.

INTRODUCTION

To accompany its goal of halving fatalities between 2011 and 2020, the European Union (EU) wants to set a target for reducing serious injuries resulting from road accidents (*1*). This new aim implies the study of seriousness criteria. In this paper we use the AIS (Abbreviated Injury Scale) index developed by the American Association for the Advancement of Automotive Medicine). This index is one of the most common anatomic scales for traumatic injuries used by the scientific community. It comes from a consensus founded on an anatomic mark and is based on the life threatening risk of the injury (*2*). AIS' coding rests upon a system of injury severity classification ranging from 1 to 6 (1: minor, 2: moderate, 3: serious, 4: severe, 5: critical, 6: beyond all medical care). The maximum AIS (MAIS) noted on an injured person suffering from multiple injuries enables to define the general level of injury seriousness of one person. So by its definition, the MAIS is a better estimation of severity outcome than French common classification based on hospitalization duration and defined by the ONISR, Observatoire Nationale Interministriel de la Sécurité Routière, (dead: dead at the time or in 30 days following the crash, severely injured: hospitalized more than 24 hours, slightly injured: hospitalized less than 24 hours, and no injury) and it allows comparisons with other countries. Moreover, the EU decides to use this criteria as a common definition to characterize the severity level (*1*). Thus, in this study we use the MAIS to characterize the general level of injury seriousness. We could consider the six levels of MAIS but unfortunately this leads to very poor estimation with huge confidence intervals. Furthermore, we know that an AIS value, for a road user injury, superior or equal to 3 corresponds to a critical life-threatening state (*2*). So, the dependent measure for this study is the dichotomous variable, MAIS3+, with a value of 1 if there is at least, in the crashed vehicle, one person with injury seriousness corresponding to MAIS value superior or equal to 3 and 0 in the other cases.

The main object of this paper is to build statistical models to explain and forecast this general level of injury severity according to certain factors which may have contributed to the accident occurrence. The more and more used approach consists in binary or multinomial logistic models which restrict the effects of explanatory variables to be the same across observations (*3*) (*4*) (*5*) (see (*6*) for more references). However, for a few years and taking into account the complexity of accident factors, injury analysis has more and more been made with fixed and random effect models (*7*) (*8*) (*9*) (see (*6*) for more references). Allowing the effects of some explicative factors to vary across the observations, those latter models can account for a certain non-observed heterogeneity due to unobserved factors like physical health and strength of injured person, vehicle type, safety technologies, driver's reaction etc. that may influence an injury outcome. This approach avoids potential bias and erroneous statistical inferences (*10*). This paper is part of this trend and offers a logistic normal regression model with fixed and random effects to analyze the MAIS3+ for a sample of vehicles, identifying among all the available seriousness factors those able to bring a certain non-observed heterogeneity on the general level of injury severity. Although this model has been already applied to analyze crash-injury severities, this research extends the current literature and introduces a novel effect random variable.

The paper is organized in the following way. Section 2 presents the data and information having motivated this work. Section 3 presents the methodological framework of the modeling and parameter estimation principle. Section 4 compares the performance of the used estimation methods. Section 5 entirely deals with the analysis of real data coming from the crashed vehicle sample having motivated this work. We end with section 6 presenting comments under the form of perspectives.

DATA

The data analyzed in this paper come from the base of detailed studies of personal accidents occurring from 1991 to 2010 and dealt with by the Biomechanical and Human Behavior Accidentology Laboratory (LAB) in France. This concerns 1722 crashed vehicles. A large part of crash data is collected on the accident scene in real time (discussions with the concerned persons, scene and vehicle photographs etc.).Thus each crashed vehicles in the base is described through several factors (or variables). Some variables give information about general items (place, time, weather, light, etc.), others on vehicles (drivers and vehicle characteristics, etc.) and about some road users (age, sex, seat in the vehicle, type of injury, etc.). Speed assessing or reconstruction variables are also noted. A large part of the variables are excluded of the study because they have a high rate of unknown values or present incoherence in the coding. As regards the variable selection, a first step is done by studying association between each of the remaining variables with the dependent variable, the MAIS3+. Only variables which the p-value of Chi-Square test is inferior to 0.05 and Cramer's V coefficient is superior to 0.1 are been kept. Then we apply a classical logistic model with these variables and keep only ones that are found statistically significantly different from zero at the 0.05 level significance (Wald test). So, each crashed vehicle is described through six variables (or factors) which are: Energy Equivalent Speed (EES), collision type (COL), the accident year (YEAR), the place (LOC), the driver's age (AGE) and the driver's sex (GENDER). Finally, keeping only the observations concerning light vehicle and omitting observations with missing data in variables we have obtained a final sample of 826 crashed vehicles observations.

The EES is estimated through expertise by visually comparing the damages on the crashed vehicle and those on a similar vehicle tested in the same type of collision during crash tests. This variable measures the deformation energy absorbed by the vehicle in a collision. Consequently, the EES is an important variable to describe and explain the injury seriousness (*11*). It is recoded in six speed brackets: 0-20 kmph, 20-30 kmph, 30-40 kmph, 40-50 kmph, 50-60 kmph and 60-90 kmph. The COL variable takes into account the impact on the concerned vehicle but also the one of the opposite vehicle if it is a two-vehicle accident. This variable is ranked in six levels: the concerned vehicle bumps head into the rear of another vehicle (F-A), the concerned vehicle is bumped into from behind by the front of another vehicle (A-F), both vehicles bump into each other's front (F-F), the concerned vehicle bumps head into the side of another vehicle (F-L), the concerned vehicle is bumped into its side by the front of another vehicle (L-F), the concerned vehicle bumps into a fixed obstacle (V-O). The year of the accident is coded into four categories: '1991-1994', '1995-1999', '2000-2004' and '2005-2010'. The variable LOC is coded into two levels: in built-up areas (Inag) or out of built-up areas (Hoag). The variable AGE is divided into 7 age brackets: '16-24 ', '25-34', '35-44 ', '45-54', '55-64 ', '65-74' and '75+'. Finally the variable GENDER has two modes: Male (M) and Female (F). Table 1 presents the data cross-tabulated to give a more detailed picture of the distributions of the variables.

METHODOLOGICAL APPROACH

Problem Formulation

Traditionally, a variable such as the MAIS3+ is modelled through a classic binary logistic regression, taking all the six factors described above as covariates. However, the accident phenomenon is of a complex nature and the causes are varied, some explanatory factors may then have heterogeneous non-observed effects on the accident and injury mechanisms. So, while

TABLE 1 Distribution of the Variables

	Sample		**MAIS3+**	
	N	%	N	%
Total	826	100.00	162	19.61
Variables	N	%	N	%
EES				
0-20	268	32.5	4	1.5
20-30	196	23.7	17	8.7
30-40	148	17.9	25	16.9
40-50	112	13.6	41	36.6
50-60	63	7.6	37	58.7
60-90	39	4.7	38	97.4
Col				
A-F	32	3.9	3	9.4
F-A	45	5.5	2	4.4
F-F	205	24.8	58	28.3
F-L	204	24.7	15	7.4
L-F	196	23.7	46	23.5
V-O	144	17.4	38	26.4
Loc				
Inag	307	37.2	23	7.5
Hoag	519	62.8	139	26.8
Age				
16-24	192	23.2	48	25.0
25-34	201	24.3	40	19.9
35-44	175	21.2	31	17.7
45-54	103	12.5	17	16.5
55-64	70	8.5	5	7.1
65-74	53	6.4	12	22.6
75+	32	3.9	9	28.1
Year				
1991-1994	156	18.9	57	36.5
1995-1999	463	56.1	75	16.2
2000-2004	159	19.3	21	13.2
2005-2010	48	5.8	9	18.8
Gender				
F	242	29.3	37	15.3
M	584	70.7	125	21.4

constructing the model, one should be able to define the explanatory variables which should be included among the fixed effect components and those which should be included among the

random effect components. It is therefore difficult to know how to decide which of the predictors have coefficients which significantly vary across the observations.

A great deal of papers about the random effect parameter selection exists (*12*), (*13*). However, in the everyday practice of injury seriousness assessment, a simple, pragmatic and easy-to-use selection criterion is needed. So, in this paper, we have opted for likelihood ratio parametric test (*14*) to assess whether the variability of each explanatory variable across the vehicles is significantly different from zero. Under the null hypothesis $\sigma^2 = 0$ (the alternative hypothesis: $\sigma^2 > 0$) with σ^2 the variance of the tested variable, the likelihood ratio test statistics is distributed as a mixture of two χ^2-distributions with 0 and 1 degrees of freedom; more precisely $0.5\chi_0^2 + 0.5\chi_1^2$. If the null hypothesis is rejected so the variable is taken as having random effects if not fixed effects. Thus, we found that the variability of the explanatory collision factor COL is significantly different from zero with a risk at 5% with a p-value equal to 4.34×10^{-8}. This result can be explained by the fact that there are unobserved or not incorporated factors (due to high rate of missing data) in the model that may influence the collision effects on the injuries severity. Indeed, in the model we don't have the characteristics of the person with injury severity MAIS3+ like the age, the health, the place in vehicle, the use of safety-belt etc. Also, the crash year from 1991 to 2010 include a large swing in design vehicle and safety technologies (*15*). So, a frontal-frontal collision, for example, which happened in 2005 has not the same effects on injuries severity for vehicles designed in 1990 as for vehicles designed in 2002, considering all the others parameters equal. In the rest of this paper, we suggest a mixed logistic-normal approach to model the general injury severity with collision type variable as random effect.

The Proposed Model and Parameter Estimation Principle

For the crash data analyzed here, the categorical variable collision (COL) with six levels is taken as a factor associated with random effect. So, we suppose that we have one random effect associated with q (q = 6) levels. In the following, we suppose that each random effect levels $COL_i \sim \mathrm{N}(0, \sigma^2), i = 1, \dots, 6$. Therefore, observations on each vehicle consist of a univariate binary response variable the MAIS3+ denoted y_{ij} (0 or 1), ($i = 1, \dots, 6; j = 1, \dots, n_i$), together with a 5×1 vector of covariates EES, AGE, YEAR, LOC, GENDER associated with fixed effects, and by a vector of covariate with 6 levels associated with a random effect. In this case, the logistic-normal mixed model is the common choice for the binary response variable MAIS3+.
Denote $Y = (Y_1^T, \dots, Y_6^T)^T$ the 826×1 total observation vector of responses where $Y_i = (y_{i1}, \dots, y_{in_i})^T$ is a $n_i \times 1$ vector response so that $\sum_{i=1}^{6} n_i = 826$. Conditioning on the random effect, the observations y_{ij} are assumed to be independent. So the log-likelihood at parameters (β, σ) for the logistic-normal mixed model is

$$logLik(Y|\beta,\sigma) = \sum_i \log \int_{\mathbb{R}} \prod_{j=1}^{n_i} \frac{e^{y_{ij}(X_{ij}^T\beta + \sigma U_i)}}{1+e^{(X_{ij}^T\beta+\sigma U_i)}} \Phi(U_i)\, dU_i \quad (1)$$

where U_i is the i^{th} level of the random effect COL, $\Phi(.)$ is the standard normal probability density function, $X_{ij}^T = (1, X_{ij1}^T, X_{ij2}^T, X_{ij3}^T, X_{ij4}^T, X_{ij5}^T)$, the covariates vector associated with fixed effects $X_{ij1}, X_{ij2}, X_{ij3}, X_{ij4}$ and X_{ij5} respectively represent $X_{EES}, X_{AGE}, X_{YEAR}, X_{LOC}$ and X_{GENDER} , and $\beta = (\beta_0, \beta_1^T, \beta_2^T, \beta_3^T, \beta_4^T, \beta_5^T)^T$ the $p \times 1$ vector of fixed effect coefficients, β_0 is a scalar, the β_m are vectors of dimension $(p_m - 1)$ where p_m is the number of modalities of the m^{th} fixed effect

variable X_{ijm} so that $p = 1 + \sum_{m=1}^{5}(p_m - 1)$. For example, β_1 is a vector of dimension 5 because a modality (the last one) of variable X_{EES} is taken as reference. Therefore, the $p \times 1$ fixed effect coefficients β, and the variance σ^2 parameterize the model. One knows that the logistic-normal mixed model is a special case of the generalized linear mixed models (GLMM). Many applications and statistical issues used this model and are reviewed in some books; see for instance (*16*).

In some important cases, the likelihood can be evaluated analytically and maximization proceeds by standard methods such as the EM algorithm (*17*), Fisher scoring (*18*) or Newton-Raphson (*19*). Here, we have non-normal likelihood and nonidentity link. In this case, estimation and inference are often hampered by the intractable integrations involved in evaluation of the log-likelihood (2) and have been well developed in the past two decades. Statisticians have proposed various ways to approximate the likelihood (1) including penalized quasi-likelihood (PQL) (*20*), Laplace approximations (*21*) and Adaptive Gauss-Hermite quadrature (AGH) (*22*), as well as Markov chain Monte Carlo (MCMC) algorithms (*23*).

The objective is to use a fast, accurate method but simple and well implemented in software method. Fast and accurate because it is important to quickly provide accurate critical information to emergency trauma centers to facilitate appropriate preparations for receipt of transported seriously injured occupants. Simple and well implemented in software to allow to non specialist users to use the methods. So in this paper, we use the Laplace approximation for parameter estimation and compare the results to those of AGH method with 15 quadrature points (AGH15) to see what compromise to do between speed and accuracy. We use here function glmer of the packages lme4 of the software R (*24*).

SIMULATION STUDY

Simulation Procedure

To compare the accuracy and time of the estimation methods, we have carried on simulations. 100 simulated datasets of 800 vehicles are generated and analyzed by Laplace and AGH15 method. We want to see how close the estimates are to the true values and how long the method estimation takes. The structure of the simulated dataset is similar to the accident database which has motivated this paper. It implies five fixed effect variables: the EES (X_{EES}) with six categories, the driver's age (X_{AGE}) with seven categories, the year of the accident (X_{YEAR}) with four categories, the place (X_{LOC}) with two categories, the sex (X_{GENDER}) with two categories and a random effect collision variable (X_{COL}) with six types of collision. The five fixed effect variables are generated through multinomial distributions $\mathcal{M}\left(n;\ \pi_1^{(0)}, \dots, \pi_r^{(0)}\right)$ of a dimension equal to the number of categories and where the probabilities of class $\pi_l^{(0)}$ $(l = 1, \dots, r)$ are fixed to proportions close to those observed in the analyzed database (see table 1).

X_{COL} is generated from a multivariate normal distribution $U \sim \mathrm{N}(0_6, \sigma^2 \times I_6)$ where 0_6 is the vector of dimension 6 composed of zero and I_6 is identity matrix of dimension 6 x 6. The dependent variable y_{ij} is binary (1 if there is at least, in the crashed vehicle, one person with injury seriousness corresponding to MAIS value superior or equal to 3 and 0 in the other cases) and is generated from a Bernoulli

$$y_{ij} \sim B\left(logit^{-1}\left(\beta_0 + \sum_{m=1}^{5} X_{ijm}^T \beta_m + U_i\right)\right), \quad (i = 1, \dots, 6; j = 1, \dots, n_i)$$

where the fixed effect variables $X_{ij1}, X_{ij2}, X_{ij3}, X_{ij4}$ and X_{ij5} respectively represent $X_{EES}, X_{AGE}, X_{YEAR}, X_{LOC}$ and X_{GENDER} In the simulations, the true values of the fixed effect parameters are set to
$\beta_0 = 6.440$, $\beta_1^T = (-8.448, -6.681, -5.625, -4.372, -3.338)$,
$\beta_2^T = (-1.220, -1.912, -1.659, -1.338, -2.233, -1.073)$
$\beta_3^T = (-0.988, -0.865, -0.757)$, $\beta_4 = -0.701$, $\beta_5 = -0.561$ and the standard deviation of the random effect is set to $\sigma = 0.684$

Proposed Methods Comparison

Table 2 presents the mean estimates, the mean biases, the standard errors values of the parameters and the mean squared errors for each method. The values in brackets near the methods show the percentage of converging replications. The estimations of the parameters are evaluated in relation to the number of converging replications. We see that Laplace has converged 73 times (73%) and AGH15 58 times (58%).

In relation to the mean estimates, we observe that underestimation rate (comparison of mean and true values) of fixed and random effects remains high for both methods. The used estimation methods give high but similar underestimation rates and mean solutions close to the true parameter values, particularly as far as the fixed effects are concerned.

For each fixed estimated parameter, we have added an estimation of its standard error by taking the square root of the reciprocal number's main diagonal of the Hessian Matrix inverse at the converging point. We have found that the approximate size of standard errors remains similar and in most cases under one. The mean range (mean difference between the maximum and minimum values) associated to standard errors by scenario and by method shows that the Laplace method gives estimations less dispersed from the means of the standard errors of beta than the AGH15 method. For example extra calculations from table 2 show that the mean range for standard errors is about 0.3777 for Laplace and 0.4616 for AGH15. On the whole, both methods supply almost identical simulated standard errors although they tend to be a bit smaller for AGH15 than Laplace.

Concurrently to the estimated parameter values and to their standard errors, we have assessed the approximation methods with the bias and mean square error calculation. The bias is obtained by the difference between the mean value and the true value of the parameter, and the mean square error specific to the fixed effect parameter is defined by:

$$MSE = \frac{1}{J}\sum_{j=1}^{J}(\hat{\beta}_j^{(mean)} - \beta_j^{(0)})^2$$

where $\hat{\beta}_{n,j}^{(mean)}$ is the estimated mean for each fixed effect parameter, calculated in relation to the converging replication number, *J* is the number of really estimated fixed effect parameter number and $\beta_j^{(0)}$the associated true value. It is a global empirical measure which enables to analyze the closeness between the solutions given by the approximation methods and the true values. In relation to the biases, we observe that both methods are almost identical although biases tend to be a bit smaller for AGH15 than Laplace. Further, both methods give mean estimates parameters very close to the true parameters values, particularly as far as the fixed effects are concerned. Furthermore, the results show that AGH15 method tends to have a bit smaller MSE than Laplace method. We reach the same conclusions for all the parameters' MSE (fixed and random effect together) as well as for the random effect only.

TABLE 2 Estimates, Biases, Standard Errors and Mean Squared Errors

		Laplace (73)					AGH15 (58)				
		Estim.	Bias	Standard Errors			Estim.	Bias	Standards Errors		
Parameter	**TV**	**mean**	**mean**	**min**	**mean**	**max**	**mean**	**mean**	**min**	**mean**	**max**
Intercept	6,440	6,359	-0,081	0,863	1,126	1,425	6,269	-0,172	0,852	1,150	1,455
EES											
0-20	-8,448	-8,390	0,058	0,741	1,084	1,508	-8,575	-0,127	0,739	1,108	1,540
20-30	-6,681	-6,529	0,152	0,670	0,931	1,184	-6,701	-0,020	0,631	0,945	1,192
30-40	-5,625	-5,415	0,210	0,622	0,904	1,186	-5,547	0,078	0,582	0,907	1,159
40-50	-4,372	-4,113	0,260	0,613	0,888	1,143	-4,226	0,146	0,553	0,890	1,137
50-60	-3,338	-3,070	0,268	0,621	0,903	1,141	-3,167	0,171	0,576	0,902	1,160
Age											
16-24	-1,220	-1,390	-0,170	0,492	0,603	0,781	-1,108	0,112	0,497	0,618	0,888
25-34	-1,912	-2,044	-0,132	0,505	0,618	0,791	-1,856	0,056	0,509	0,635	0,911
35-44	-1,659	-1,802	-0,143	0,505	0,618	0,811	-1,740	-0,081	0,514	0,641	0,887
45-54	-1,338	-1,493	-0,155	0,549	0,650	0,844	-1,243	0,095	0,523	0,666	0,911
55-64	-2,233	-2,373	-0,140	0,598	0,754	1,018	-2,262	-0,029	0,620	0,780	1,116
65-74	-1,073	-1,225	-0,152	0,575	0,727	0,944	-0,970	0,103	0,592	0,741	0,988
Year											
1995-1999	-0,988	-1,000	-0,012	0,274	0,318	0,376	-1,065	-0,077	0,279	0,324	0,425
2000-2004	-0,865	-0,900	-0,035	0,333	0,397	0,482	-0,944	-0,079	0,340	0,402	0,477
2005-2010	-0,757	-0,800	-0,043	0,470	0,603	0,972	-0,847	-0,090	0,491	0,622	0,926
Loc											
Inag	-0,701	-0,714	-0,013	0,228	0,276	0,332	-0,725	-0,025	0,227	0,280	0,347
Gender											
F	-0,561	-0,586	-0,025	0,245	0,294	0,385	-0,500	0,061	0,249	0,296	0,368
SD	0,684	0,539	-0,145				0,636	-0,049			
		MSE: 2,08E-02					MSE: 9,57E-03				

TV: True Values; Estim.: Estimate; SD: Standard Deviation; MSE: Mean Squared Error

To conclude, globally both methods give a very good accuracy of estimations but with a slight superiority to AGH15. However, as regard the average run times (14s for Laplace and 43 for AGH15), Laplace methods seems to do a much better compromise between fast and accuracy than AGH15.

REAL DATA ANALYSIS

Estimates of the Parameters of the Real Data

In this section we focus on a mixed logistic model estimation for the observed injury seriousness in a vehicle using a sample of 826 vehicles crashed in serious accidents. So we estimate the logistic-normal mixed model using the Laplace approximation method proposed in the previous section (the AGH15 method gives the same estimates).

The results are recorded in Table 3. P-values enable to measure the fixed effect significance. Given the modality reference of each covariate, we can see that the estimations have the same sign (negative) and on the whole are very significant (P-value < 0.05) with low risks whatever the approximation method. The EES is, by far, the most significant variable with risks associated to each modality being smaller than 10^{-2} possibly 10^{-3} if we take the 60-90 modality as

a reference. If we set a risk at 5% and if we consider drivers aged 75 and over as a reference, then drivers in the 25-34 age bracket very significantly contribute the modeling, followed by the 35-44 age bracket, and then the 55-65 age bracket. The youngest drivers come last. The vehicles crashed between 1995 and 1999 are more significant as compared to the reference one 1991-1994. On the contrary, the estimated fixed effect coefficient of vehicles crashed between 2005 and 2010 is not significant as compared to those observed in 1991-1994. As compared to the vehicles crashed out of built-up areas (the reference), the number of vehicles crashed in built-up areas is significant with a risk under 5%. The same is true for vehicles driven by women if we take male drivers as a reference.

Results Analysis and Interpretation

From the Table 3, we can see that the lower the EES is, the lower the odds ratio of a serious injury is. For example, considering all the other invariant parameters of the model, a vehicle with deformation energy between 0 and 20 kmph has 5000 ($1/[2 \times 10^{-4}]$) times less chance to contain someone with maximum injury seriousness than a vehicle with an EES between 60 and 90 kmph. On the contrary, a vehicle with EES between 50 and 60 kmph has 28 (1/0.0355) times less chance to contain someone with maximum injury seriousness than a vehicle with an EES between 60 and 90 kmph. The associated confidence intervals clearly show that these odds ratios are statistically significant at the 5% threshold.

Concerning driver's age, we can see that the odds ratio are less dispersed than the EES ones and are not all significant at the 5% threshold, vehicles driven by the 65-74 age bracket having a P-value superior to 0.05. Vehicles driven by the 16-24 age bracket have 0.295 more chance to contain people with serious injuries as compared to those driven by over-75s if we suppose that other fixed effect factors remain invariant. According to the estimations carried out in this survey, we can assume the fact that vehicles driven by young people (under 24) or by the elderly (over 65 or even 75) have more chance to be involved in accidents with serious injuries than the other age brackets. Even though they are based on only one case study, these results are consistent with the fact that, on the one hand, young drivers take more risks particularly in trespassing traffic laws and are less aware of the incurred risk (*25*) (*26*) and, on the other hand, the oldest drivers suffer from perceptual-motor and cognitive control decline (*27*). So, they have fewer reflexes and have a longer reaction time (*28*).

As for the year of the accident, we note that the vehicles crashed between 1995 and 1999 had 2,685 more chance not to contain seriously injured people than those crashed between 1991 and 1994 if we assume that the other model factors are invariant. The ratio is 2.375 for vehicles crashed between 2000 and 2004, and 2.132 between 2005 and 2010. But this last ratio is not significant at the 5% threshold. So, we can see that, in comparison with 1991-1994, the maximum injury seriousness in crashed vehicles decreases. However, the estimates show a slight increase of the odds to be involved in accidents with serious injuries from 1995-1999 to 2000-2004, also from 2000-2004 à 2005-2010. These results can be explained by the fact that time from 1991 to 2010 includes a large swing in car design and deployed safety technologies leading to confounding effects. However, globally the results reflect a decreasing trend.

As concerns the place, vehicles crashed in built-up areas have about 2.016 times more chance not to contain seriously injured people than those out of built-up areas. This result confirms that there are generally fewer deadly accidents in built-up areas than in rural areas (*29*) though built-up areas count more accidents but of lesser severity (*30*).

TABLE 3 Parameters and Odds Ratio Estimates Based on Real Data Using Laplace Method

						95% Confidence Interval	
Variables	**Estimates**	**SE**	**z**	**P-value**	**Odds Ratio**	**Lower**	**Upper**
Intercept	6.4404	1.2454	5.1710	0.0000			
EES							
0-20	-8.4481	1.1903	-7.0980	0.0000	0.0002	0.000	0.002
20-30	-6.6809	1.1039	-6.0520	0.0000	0.001	0.000	0.011
30-40	-5.6250	1.0825	-5.1960	0.0000	0.004	0.000	0.030
40-50	-4.3721	1.0706	-4.0840	0.0000	0.013	0.002	0.103
50-60	-3.3377	1.0769	-3.0990	0.0019	0.036	0.004	0.293
Age							
16-24	-1.2202	0.5397	-2.2610	0.0238	0.295	0.102	0.850
25-34	-1.9116	0.5625	-3.3990	0.0007	0.148	0.049	0.445
35-44	-1.6594	0.5567	-2.9810	0.0029	0.190	0.064	0.567
45-54	-1.3380	0.5919	-2.2610	0.0238	0.262	0.082	0.837
55-64	-2.2331	0.7311	-3.0540	0.0023	0.107	0.026	0.449
65-74	-1.0729	0.6420	-1.6710	0.0947	0.342	0.097	1.204
Year							
2005-2010	-0.7571	0.5343	-1.4170	0.1565	0.469	0.165	1.337
2000-2004	-0.8651	0.4102	-2.1090	0.0349	0.421	0.188	0.941
1995-1999	-0.9879	0.2915	-3.3880	0.0007	0.372	0.210	0.659
Loc							
Inag	-0.7009	0.3146	-2.2280	0.0259	0.496	0.268	0.919
Gender							
F	-0.5611	0.2827	-1.9850	0.0472	0.571	0.328	0.993
SD	0.684						

SD (Standard deviation of random effect); Pv (P-value)

Concerning driver gender, and taking male drivers as the reference modality, the negative sign of the estimated coefficient associated to female drivers shows that women are less dangerous drivers if the other factors do not vary a lot. The odds ratio for women is such that the vehicles they drive have about 1.752 (1/0.5706) times more chance not to contain seriously injured people than those driven by men. These results are consistent with a certain number of studies showing that male drivers have a more aggressive way to drive than women and that they tend to take more risks, especially young drivers (*31*) (*32*).

Prediction Performance

We want to know if the proposed model has a better prediction power than the classic logistic regression.

A first simple evaluation criterion is the error rate. It consists in comparing the real observed values of the dependent variable with those predicted by the model and counting the number of misclassifications. However, this criterion does not enable to report the error structure that is the way of making a mistake. In most applications, false positive and false negative errors are not equally important. Indeed, in the point of view of public health, to predict a serious crash as not serious is more problematic than the opposite. So, it is more interesting to use criteria as sensitivity (proportion of positives correctly classified by the model), accuracy (proportion of true

positives in classified positives), specificity (proportion of negatives correctly classified by the model) and the false positives rate (proportion of negatives misclassified by the model).
Although the model produces a continuous value for probability of injury, it is necessary to choose a cutpoint to decide when the crashed vehicle will be classified as containing seriously injured persons. In the absence of any particular reason, we generally take this threshold as equal to 0.5

Table 4 presents error rate, sensitivity and specificity estimated by cross-validation for the mixed logistic model and the classic logistic model. As regards the errors rate, we can see that the classic model has the smallest errors rate (13.47%). So, this model is the one which makes the least of misclassifications but the mixed model is very close with an error rate of 13.59%. Concerning the sensitivity, the mixed model better identifies the vehicles with seriously injured person than the classic model: it identifies 50.62% against 42.86%, which is 7.76% more. As regards the identification of vehicles with not seriously injured occupants (specificity), the classic model identifies them better than the mixed model: it identifies 97.59% against 95.17%, which is 2.42% more. So, we can see that the mixed model is more powerful than the classic model in the identification of vehicles with seriously injured occupants but less powerful in the identification of vehicles with not seriously injured occupants. Adopting a point of view of public health and thus prioritizing the sensitivity criterion, we consider that the mixed model is the most powerful for our problematic.

However, making no prioritization and making the cutpoint vary between 0 and 1, we can say that both these models have an excellent discriminating powerful in view of the estimated area under the receiver operator characteristic (ROC) curve close to 1.

TABLE 4 Performance Measures

Model	Error rate	Sensitivity	Specificity	AUC	95% Confidence intervals	
					Lower	Upper
Classic	13.47%	42.86%	97.59%	0.899	0.810	0.990
Mixed	13.59%	50.62%	95.17%	0.900	0.809	0.989

CONCLUDING REMARKS

In this paper, we modeled a road accident seriousness index by a dichotomous variable with the help of a mixed logistic-normal model and we found that the collision type is best modeling by random factor than fixed factor. Then, we wanted to choose the fastest, most accurate and simplest estimation method. Penalized Quasi-Likelihood being biased (*33*) and Monte Carlo simulations being complicated for non-specialists (*34*), Laplace approximation and Adaptive Gauss-Hermite method with 15 quadrature points have been compared through simulations. Globally both methods give a very good accuracy of estimations but with a slight superiority to AGH15. However, as regard the average run times, Laplace method seems to do a much better compromise between speed and accuracy than AGH15, so this method is preferred to estimate the parameters of the model established.

We also showed that our model has a discriminating power similar to the classic binary logistic model where all the explanatory factors are considered as fixed effect variables. Beyond the similitude of these two models' discriminating powers, the most important question in this study is to know the most performing model, i.e. the one with the greatest ability to make the difference between the seriously crashed vehicles and the not seriously crashed ones. To that effect, we have assessed the performance of these models by the means of cross validation. The

obtained results reveal that, on the average, our mixed logistic-normal model detects 7.76% more seriousness than the classic model without random effect. On the contrary, the classic model, on the average, detects 2.42% more non seriousness than the mixed model. Given that, in this study, the seriousness of body injuries is focused on, we prefer to use the mixed model with the collision type as random effect factor.

However the accident phenomenon being of a random nature, other factors than the collision type may be thought of as having a random effect. So, the predictive performance criterion of the model should be coupled with probabilistic criteria such as the BIC, the CAIC (*35*) for a better selection process of more adapted models.

Finally, in this paper, we have modeled the maximum body injury seriousness with a dichotomous variable with a value equal to 1 if there is body injury seriousness superior or equal to 3 in the crashed vehicles and a value equal to 0 in other cases. As the AIS index is injury seriousness scale going from 1 to 6 (possibly more), our current aim is to extend the results of this paper to a multinomial approach with a bigger dataset in order to better take into account the multivariate structure of the dependent variable MAIS and to suggest a model by body area.

ACKNOWLEDGEMENTS

This work was supported by the "Groupe d'Intérêt Economique" (GIE) of PSA-RENAULT of France under Agreement No. USTL-GIE PSA-RENAULT 248 02 515. The authors are grateful to the four anonymous reviewers for their comments and critical reading of the manuscript.

REFERENCES

1. European Comission. Mobility and transport: Road Safety. Serious injury. [Online]. http:ec.europa.eu/transport/road_safety/topics/serious_injuries/index_en.htm.

2. Ceesar, Inrets, Uclb, and Ivs. *Description et gravité des lésions traumatiques selon les classifications AIS 1998 et IIS 1994. Traduit de l'anglais. (Octobre 2004). The Abbreviated Injury Scale AIS Version 1998, The Injury Impairment Scale IIS Version 1994*, ISBN : 2-11-094954-6, Ed. Maison-Alfort, 2004.

3. Ulfarssoon, G. F., and F. L. Mannering. Difference in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car crashes. *Accident Analysis and Prevention*, Vol. 36, no. 2, 2004, pp. 135-147.

4. Kockelman, K. M., and Y. J. Kweon. Driver injury severity: An application of ordered probit models. *Accident Analysis and Prevention*, Vol. 34, 2002, pp. 313-321.

5. Al-Ghamdi, A. S. Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis and Prevention*, Vol. 34, 2002, pp. 729-741.

6. Lord, D., and F. Mannering. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transport Research Part A*, Vol. 44, 2010, pp. 291-305.

7. Anastasopoulosa, P. C., and F. L. Mannering. An empirical assessment of fixed and random parameter logit models using crash- and non-crash-specific injury data. *Accident Analysis and Prevention*, Vol. 43, 2011, pp. 1140-1147.

8. Kim, J. K., G. F. Ulfarsson, V. N. Shankar, and F. L. Mannering. A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. *Accident Analysis and Prevention*, Vol. 42, no. 6, 2010, pp. 1751-1758.

9. Milton, J. C., V. N. Shankar, and F. L. Maneering. Highway accident severities and the mixed logit model : an exploratory empirical analysis. *Accident Analysis and Prevention*, Vol. 40, 2008, pp. 260-266.

10. Mc Fadden, D., and K. Train. Mixed MNL models for discrete response. *Journal of Applied Econometrics*, Vol. 15, no. 5, 2000, pp. 447-470.

11. Zeidler, F., H. H. Schreir, and R. Stadelmann. Accident research and accident reconstruction by the EES-Accident Reconstruction Method. *Society of Automotive Engineers*, Vol. 94, no. 2, 1985, pp. 2399-2413.

12. Commenges, D., and H. Jacqmin-Gadda. Generalized score test of homogeneity based on correlated random effects models. *Journal of The Royal Statistical Society Series B-statistical Methodology*, Vol. 59, no. 1, 1997, pp. 157-171.

13. Yang, M. Bayesian variable selection for logistic mixed model with nonparametric random effects. *Computational Statistics and Data Analysis*, Vol. 56, no. 9, 2012, pp. 2663-2674.

14. Gao, S. Combining binomial data using the logistic normal model. *Journal of Statistical Computational and Simulation*, Vol. 74, no. 4, 2004, pp. 293-306.

15. Labrousse, M., T. Hermitte, V. Hervé, N. Bertolon, and A. Guillaume. Evolution of front car occupants injuries in frontal impacts considering the improvements of passive safety technologies. in *EAEC20110030 - EAE2011_E13*, 2011.

16. Longford, N. *Random Coeffcient Models*. Oxford: Clarendon Press., 1993.

17. Dempster, A. P., D. B. Rubin, and R. K. Tsutakawa. Estimation in Covariance Components Models. *Journal of American Statistical Association*, Vol. 76, 1981, pp. 341-353.

18. Longford, N. A Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models With Nested Random Effects. *Biometrika*, Vol. 74, 1987, pp. 817-827.

19. Lindstrom, M. J., and D. M. Bates. Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for repeated-Measures Data. *Journal of the American Statistical Association*, Vol. 83, 1988, pp. 1014-1022.

20. Breslow, N. E., and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, Vol. 88, 1993, pp. 9-25.

21. Raudenbush, W., M. Yang, and Y. M. Maximum Likelihood for Generalized Linear Models With Nested Random Effects via High-Order, Multivariate Laplace Approximation. *Journal of Computational and Graphical Statistics*, Vol. 9, no. 1, 2000, pp. 141-157.

22. Pinheiro, J. C., and E. C. Chao. Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics*, Vol. 15, 2006, pp. 58-81.

23. Booth, J. G.a.H.J.H. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society. Series B*, Vol. 62, 1999, pp. 265-285.

24. Bates, D. M., M. Mächler, and B. Bolker. (2011) LME4: linear mixed-effects models using S4 classes. R Package Version 0.999375-39. [Online]. http://CRAN.R-project.org/package=lme4.

25. Gregersen, N. P., and H.-Y. Berg. Lifestyle and accidents among young drivers. *Accident Analysis and Prevention*, Vol. 26, no. 3, 1994, pp. 297-303.

26. Blockley, P., and L. Hartley. Aberrant driving behavior: Errors and violations. *Ergonomics*, Vol. 38, no. 9, 1995, pp. 1759-1771.

27. Brouwer, W. H., and R. W. Ponds. Driving competence in older persons. *Disability \& Rehabilitation*, Vol. 16, no. 3, 1994, pp. 149-161.

28. Islam, S., and F. Mannering. Driver aging and its effect on male and female single-vehicle accident injuries: Some additional evidence. *Journal of Safety Research*, Vol. 37, 2006, pp. 267-276.

29. Jones, A. P., R. Haynes, V. Kennedy, I. M. Harvey, T. Jewell, and D. Lea. Geographical variations in mortality and morbidity from road traffic accidents in England and Wales. *Health and Place*, Vol. 14, 2007, pp. 519-535.

30. Eiksund, S. A geographical perspective on driving attitudes and behaviour among young adults in urban and rural Norway. *Safety Science*, Vol. 47, 2009, pp. 529-536.

31. Yagil, D. Gender and age-related differences in attitudes toward traffic laws and traffic violations. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 1, no. 2, 1998, pp. 123-135.

32. Iversen, H. H., and T. Rundmo. Attitudes towards traffic safety, driving behaviour and accident involvement in the Norwegian public. *Ergonomics*, Vol. 47, 2004, pp. 555-572.

33. Goldstein, H., and J. Rasbash. Improved Approximations for Multilevel Models with Binary responses. *Journal of the Royal Statistical Society*, Vol. Series B, no. 159, 1996, pp. 505-513.

34. Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H.H. Stevens, and J.-S. S. White. Generalized linear mixed models: a pratical guide for ecology and evolution. *Trends in Ecology and Evolution*, Vol. 24, no. 3, 2008, pp. 127-135.

35. Yu, D., and K. K.W. Yau. Conditional Akaike information criterion for generalized linear mixed models. *Computational Statistics and Data Analysis*, Vol. 56, 2012, pp. 629-644.

Bibliography

Abramowitz, M. and Stegun, I. (1974). *Handbook of mathematical functions*. New York: Dover.

Agresti, A. (2002). *Categorical data analysis*. 2nd Edition. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics.

Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis and Prevention*, 34(6):729–741.

Anastasopoulosa, P. C. and Mannering, F. L. (2011). An empirical assessment of fixed and random parameter logit models using crash- and non-crash-specific injury data. *Accident Analysis and Prevention*, 43(3):1140–1147.

Anderson, D. A. and Aitkin, M. (1985). Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society*, Serie B(47):204–210.

Anderson, R. L. and Bancroft, T. A. (1952). *Statistical Theory in Research*. McGraw-Hill Book Company.

Baker, S. P., O'Neill, B., Haddon, W., and Long, W. B. (1974). The injury severity score : a method for describing patients with multiple injuries and evaluating emergency care. *J. Trauma*, 14:187–196.

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R*. Springer.

Bates, D. M., Mächler, M., and Bolker, B. (2011). Lme4: linear mixed-effects models using s4 classes. r package version 0.999375-39. http://CRAN.R-project.org/package=lme4.

Ben Ahmed, W. (2004). *SAFE-NEXT : Une Approche Systémique Pour L'Extraction De Connaissances De Données. Application A La Construction Et A L'Interprétation De Scénarios D'Accidents De La Route*. PhD thesis, Laboratoire Génie Industriel, Ecole Centrale Paris, Châtenay Malabry.

Binghama, C. R., Shopea, J. T., and Zhub, J. (2008). Substance-involved driving: Predicting driving after using alcohol, marijuana, and other drugs. *Traffic Injury Prevention*, 9(6).

Blockley, P. and Hartley, L. (1995). Aberrant driving behavior: Errors and violations. *Ergonomics*, 38(9):1759–1771.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed effects models. *Biometrics*, 66(4):1069–1077.

Booth, J. G. and Hobert, J. H. (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society*, 62:265–285.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25.

Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models whith a single component of dispersion. *Biometrika*, 82:81–91.

Brouwer, W. H. and Ponds, R. W. (1994). Driving competence in older persons. *Disability & Rehabilitation*, 16(3):149–161.

Bull, J. P. (1975). The injury severity score of road traffic casualties in relation to mortality, time of death, hospital treatment time and disability. *Accident Analalysis & Prevention*, 7:249–255.

Carroll, R., Ruppert, D., and Gelfand, A. (1988). *Transformation and weighting in regression*. Chapman and Hall, London.

Ceesar, L., Inrets, L., Uclb, L., and Ivs, L. (2004). *Description et gravitÃ© des lÃ©sions traumatiques selon les classifications AIS 1998 et IIS 1994. Traduit de l'anglais. (Octobre 2004). The Abbreviated Injury Scale AIS Version 1998, The Injury Impairment Scale IIS Version 1994*. Ceesar, and Inrets, and Uclb, and Ivs.

Chen, M. H., Ibrahim, J. G., Shao, Q. M., and Weiss, R. E. (2003). Prior elicitation for model selection and estimation in generalized linear mixed models. *Journal of Statistical Planning and Inference*, 111:56–57.

Clarkson, D. B. and Zhan, Y. (2002). Using spherical-radial quadruature to fit generalized linear mixed effects models. *Journal of Computational and Graphical Statistics*, 11:639–659.

Commenges, D. and Jacqmin-Gadda, H. (1997). Generalized score test of homogeneity based on correlated random effects models. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 59(1):157–171.

Dejoy, D. M. (1992). An examination of gender differences in traffic accident risk perception. *Accident Analysis and Prevention*, 24:237–246.

Dempster, A. P., Rubin, D. B., , and Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of American Statistical Association*, 76:341–353.

Diaz, R. E. (2007). Comparison of pql and laplace 6 estimates of hierarchical linear models when comparing groups of small incident rates in cluster randomized trials. *Computational Statistics and Data Analysis*, 51:2871–2888.

Donohue, M. C., Overholser, R., Xu, R., and Vaida, F. (2011). Conditional akaike information under generalized linear and proportional hazards mixed models. *Biometrika*, 98(3):685–700.

Doran, H., Bates, D., Bliese, P., and Dowling, M. (2007). Estimating the multilevel rasch model: With the lme4 package. *Journal of Statistical Software*, 20(2):1–18.

Droesbeke, J.-J., Lejeune, M., and Saporta, G. (2005). *Modèles statistiques pour données qualitatives*. Technip, Paris.

Eiksund, S. (2009). A geographical perspective on driving attitudes and behaviour among young adults in urban and rural norway. *Safety Science*, 47:529–536.

Elvik, R. and Vaa, T. (2004). *The Handbook of Road Safety Measures*. Elsevier Ltd.

Engel, B. and Buist, W. (1998). Bias reduction of approximate maximum likelihood estimates for heritability in treshol models. *Biometrics*, 54:1155–1164.

Engel, B. and Keen, A. (1994). A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica*, 48(1):1–22.

Ferrandez, F. (1995). *L'étude détaillée d'accidents orientée vers la sécurité primaire: méthodologie de recueil et de pré-analyse*. Presses de l'Ecole Nationale des Pont et Chaussées.

Foulley, J. L., Delmas, C., and Robert-Granié, C. (2002). Méthodes du maximum de vraisemblance en modéle linéaire mixte. *Journal de la Société Francaise de Statistique*, 143(1-2):5–52.

Gao, S. (2004). Combining binomial data using the logistic normal model. *Journal of Statistical Computational and Simulation*, 74(4):293–306.

Goldman, N. and Whelan, S. (2000). Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Mol. Biol. Evol.*, 17:975–978.

Goldstein, H. (1986). Multilevel mixed linear analysis using iterative generalized least squares. *Biometrika*, 73:43–56.

Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society*, Series B(159):505–513.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., and Healy, M. (1998). *A user's guide to MLwiN*. Institute of Education.

Golub, G. H. and Welsch, J. H. (1969). Calculation of gauss quadrature rules. *Mathematics of Computation*, 23:221–230.

González, J., Tuerlinckx, F., De Boeck, P., and Cools, R. (2006). Numerical integration in logistic-normal models. *Computational Statistics and Data Analysis*, 51:1535–1548.

Green, P. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Society, Ser. B*, 46:149–192.

Green, P. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, 55(3):245–259.

Gregersen, N. P. and Berg, H.-Y. (1994). Lifestyle and accidents among young drivers. *Accident Analysis and Prevention*, 26(3):297–303.

Haleema, K. and Gana, A. (2011). Identifying traditional and nontraditional predictors of crash injury severity on major urban roadways. *Traffic Injury Prevention*, 12(3).

Hartley, H. O. and Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54:93–108.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61:383–385.

Harville, D. A. (1977). Maximum-likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72:320–340.

Haviotte, C. (2006-2007). Comparatif des bases de données d'accidentologie. Rapport interne, Ceesar et Lab.

Hedeker, D., , and Gibbons, R. D. (1994). A random effects ordinal regression model for multilevel analysis. *Biometrics*, 50:933–944.

Henderson, C. R., Kempthorne, O., Searle, S. R., and von Krosigk C. M. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15:192–218.

Hollnagel, E. (2004). *Barriers and accident prevention.* Ashgate Pub Ltd.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, 62(2):495–503.

Islam, S. and Mannering, F. (2006). Driver aging and its effect on male and female single-vehicle accident injuries: Some additional evidence. *Journal of Safety Research*, 37:267–276.

Iversen, H. H. and Rundmo, T. (2004). Attitudes towards traffic safety, driving behaviour and accident involvement in the norwegian public. *Ergonomics*, 47:555–572.

Joe, H. (2008). Accuracy of laplace approximation for discrete response mixed models. *Computational Statistics and Data Analysis*, 52:5066–5074.

Jones, A. P., Haynes, R., Kennedy, V., Harvey, I. M., Jewell, T., and Lea, D. (2007). Geographical variations in mortality and morbidity from road traffic accidents in england and wales. *Health and Place*, 14:519–535.

Kass, R., Tierney, L., and Kadane, J. (1990). *The validity of posterior expansions based on Laplace's methods.* Essays in Honor of George Bernard, eds. S. Geisser, J.S. Hodges, S.J. Press, and A. Zellner.

Kim, J. K., Ulfarsson, G. F., Shankar, V. N., and Kim, S. (2008). Age and pedestrian injury severity in motor-vehicles crashes: a heterodastic logit analysis. *Accident Analysis and Prevention*, 40(5):1695–1702.

Kim, J. K., Ulfarsson, G. F., Shankar, V. N., and Mannering, F. L. (2010). A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. *Accident Analysis and Prevention*, 42(6):1751–1758.

Kinney, S. K. and Dunson, D. B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics*, 63(3):690–698.

Kockelman, K. M. and Kweon, Y. J. (2002). Driver injury severity: An application of ordered probit models. *Accident Analysis and Prevention*, 34(4):313–321.

Kong, C. and Yang, J. (2010). Logistic regression analysis of pedestrian casualty risk in passenger vehicle collisions in china. *Accident Analysis and Prevention*, 42(4):987–993.

Kuk, A. Y. C. (1995). Assymptotically unbiaised estimation in generalized linear mixed models with random effects. *Journal of the Royal Statistical Society B - Methodological*, 57:395–407.

Labrousse, M., Hermitte, T., Hervé, V., Bertolon, N., and Guillaume, A. (2011). Evolution of front car occupants injuries in frontal impacts considering the improvements of passive safety technologies. In *EAEC*.

Le Coz, J.-Y. and Page, Y. (2003). La démarche accidentologique au service de l'évolution de la sécurité des véhicules. *Revue de la gendarmerie Nationale*, 207.

Leblanc, D., Lollivier, S., Marpsat, M., and Verger, D. (2000). *L'Econometrie et l'étude des comportements*. INSEE.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of The Royal Statistical Society Series B-statistical Methodology*, 58(4):619–656.

Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random effects model: an example. *Journal of the Royal Statistical Society*, 50(3):325–335.

Lesko, M. M., Woodford, M., White, L., O'Brien, S. J., Childs, C., and Lecky, F. E. (2010). Using abbreviated injury scale (ais) codes to classify computed tomography (ct) features in the marshall system. *BMC Medical Research Methodology*, 10(72).

Lin, X. and Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91:1007–1016.

Lindley, D. (1980). Approximate bayesian methods. In *Bayesian Statistics*, pages 223–245. Smith Valencia: Valencia University Press.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83:1014–1022.

Liu, Q. and Pierce, D. A. (1993). Heterogeneity in mantel-haenszel-type models. *Biometrika*, 80(3):543–556.

Liu, Q. and Pierce, D. A. (1994). A note on gauss-hermite quadrature. *Biometrika*, 81(3):624–629.

Longford, N. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74:817–827.

Magnus, J. R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New-York: Wiley.

Malyshkina, N. and Maneering, F. L. (2010). Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accident. *Accident Analysis and Prevention*, 42(1):131–139.

Martinez, M.-J. (2006). *Modéles linéaires généralisés á effets aléatoires : Contributions au choix de modÃ¨le au modÃ¨le de mÃ©lange*. PhD thesis, Universite Montpellier II.

Mc Fadden, D. and Train, K. (2000). Mixed mnl models for discrete response. *Journal of Applied Econometrics*, 15(5):447–470.

McCullagh, C. E. and Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall, London.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92:162–170.

McCulloch, C. E., Searle, S. R., and M., N. J. (2008). *Generalized, Linear, and Mixed models*. 2nd Edition, John Wiley & Sons.

McGilchrist, C. A. (1994). Estimation in generalized mixed models. *Journal of the Royal Stattistical Society, Series B-Methodological*, 56:61–69.

Meguellati, F., N'Guessan, A., and Hermite, T. (2014). Analyzing the maximum abbreviated injury scale in vehicle crashes using a logistic normal model. In *Transportation Research Board Compendium*.

Miltner, E. and Salwender, H. J. (1971). Influencing factors on the injury severity of restrained front seat occupants in car-to-car head-on collisions. *Accident Analysis and Prevention*, 27(2):143–150.

Milton, J. C., Shankar, V. N., and Maneering, F. L. (2008). Highway accident severities and the mixed logit model : an exploratory empirical analysis. *Accident Analysis and Prevention*, 40(1):260–266.

Naylor, J. C. and Smith, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, 31:214–225.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series B*, 56(1):61–69.

Nordfjaerna, T., Jorgensenb, S. H., and Rundmoa, T. (2010). An investigation of driver attitudes and behaviour in rural and urban areas in norway. *Safety Science*, 48(3):348–356.

O'Donnell, C. J. and Connor, D. H. (1996). Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. *Accident Analysis and Prevention*, 28(6):739–753.

O'Keefe, G. and Jurkovich, G. J. (2001). Measurement of injury severity and co-morbidity. In *In Injury Control*. Cambridge University Press.

Paleti, R., Eluru, N., and Bhat, C. R. (2010). Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes. *Accident Analysis and Prevention*, 42(6):1839–1854.

Pan, J. and Thompson, R. (2003). Gauss-hermite quadrature approximation for estimation in generalized linear mixed models. *Computational Statistics*, 18:57–78.

Patterson, H. D. and Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika*, 58:545–554.

Peitzman, A. B., Rhodes, M., Schwab, C. W., Yealy, D. M., and Fabian, T. C. (2007). *The Trauma Manual: Trauma and Acute Care Surgery*. Lippincott Williams & Wilkins.

Perron, T. and Bocquet, J. (1997). *Méthodologie d'analyse de sécurité primaire automobile pour la spécification fonctionnelle et l'évaluation prévisionnelle d'efficacité de systèmes d'évitement d'accidents*. PhD thesis, Laboratoire Génie Industriel, Ecole Centrale Paris, Châtenay Malabry.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the non-linear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4:12–35.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-plus*. Statistics and Computing Series. Springer-Verlag.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Raudenbush, W., Yang, M., , and M., Y. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. *Journal of Computational and Graphical Statistics*, 9(1):141–157.

Rodriguez, G. and Goldman, N. (2001). Improved estimation procedures for multilevel models with binary response: A case study. *Journal of the Royal Statistical Society A - General*, 164:339–355.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78:719–727.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance components*. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. John Wiley & Sons.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82:605–610.

Shun, Z. (1995). Another look at the salamander approach: A modified laplace approximation approach. *Journal of the American Statistical Association*, 92:341–349.

Shun, Z. and Mc Cullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society. Serie B*, 57:749–760.

Sivak, M., Soler, J., Trankle, U., and Spagnhol, J. M. (1989). Cross-cultural differences in driver risk-perception. *Accident Analysis and Prevention*, 21:355–362.

Solomon, P. and Cox, D. (1992). Non linear components of variance models. *Biometrika*, 79:1–11.

Stelmach, G. E. and Nahom, A. (1992). Cognitive-motor abilities of the elderly driver. *Human Factors*, 34(1):53–65.

Stiratelli, R., Laird, N., and Ware, J. (1984). Random effects models for serial observations with binary responses. *Biometrics*, 40:961–971.

Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal fixed effects model. *Biometrics*, 50:1171–1177.

Thompson, W. A. (1962). The problem of negative estimates of variance components. *Annals of Mathematical Statistics*, 33:273–289.

Tierny, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81:82–86.

Trottier, C. (1998). *Estimation dans les modéles linéaires généralisés á effets aléatoires*. PhD thesis, Institut National Polytechnique de Grenoble.

Ulfarssoon, G. F. and Mannering, F. L. (2004). Difference in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car crashes. *Accident Analysis and Prevention*, 36(2):135–147.

Url, A. (2013a). http://www.aaam.org/. American Advancement Automobile Medicine.

Url, C. www.ukccis.org. Co-operative Crash Injury Study.

Url, G. www.gidas.org. German In-Depth Accident Study.

Url, I. Reporting on serious road traffic casualtiess. http://internationaltransportforum.org/irtadpublic/index.html. International Traffic Safety Data and Analysis Group.

Url, I. (2013b). http://ec.europa.eu/transport/road_safety/topics/serious_injuries/index_en.htm. European Union.

Url, S. (2013c). Newsletter numéro 11 sécurité routière. ec.europa.eu/transport/road/safety/pdf/news/nl11_fr.pdf. European Union.

Venables, W.N.and Ripley, B. (2002). Modern applied statistics with s, 4th ed. springer, new york, isbn: 0-387-95457-0. http://www.stats.ox.ac.uk/pub/MASS4.

Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer Series in Statistics.

Wolfinger, R. (1993). Generalized linear mixed models : a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48:233–243.

Yagil, D. (1998). Gender and age-related differences in attitudes toward traffic laws and traffic violations. *Transportation Research Part F: Traffic Psychology and Behaviour*, 1(2):123–135.

Yang, M. (1998). *Increasing the Efficiency in Estimating Multilevel Bernouilli Models*. PhD thesis, Michigan State University, Department of Counseling, Educational Psychology and Special Education. Unpublished Ph.D. dissertation.

Yang, M. (2012). Bayesian variable selection for logistic mixed model with nonparametric random effects. *Computational Statistics and Data Analysis*, 56(9):2663–2674.

Yau, K. (2004). Risk factors affecting the severity of single vehicle traffic accidents in hong kong. *Accident Analysis and Prevention*, 36(3):333–340.

Yoonsang, K., Young-Ku, C., and Sherry, E. (2013). Logistic regression with multiple random effects: A simulation study of estimation methods and statistical packages. *The American Statistician*, 67(3):171–182.

Yu, D. and Yau, K. K. W. (2012). Conditional akaike information criterion for generalized linear mixed models. *Computational Statistics and Data Analysis*, 56:629–644.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects : a gibbs sampling approach. *Journal of the American Statistical Association*, 86:79–86.

Zeidler, F., Schreir, H. H., and Stadelmann, R. (1985). Accident research and accident reconstruction by the ees-accident reconstruction method. *Society of Automotive Engineers*, 94(2):2399–2413.

Zhang, D. and Lin, X. (2008). Random effect and latent variable model selection. *Journal of the American Statistical Association*, 192(2):19–36.

Zhou, X.-H., Perkins, A. J., and Hui, S. (1999). Comparisons of software packages for generalized linear multilevel models. *The American Statistician*, 53:282–290.