

Diverses alternatives pour déterminer les facteurs significatifs de la fréquence d'accidents dans l'assurance automobile

Maria del Carmen Melgar Hiraldo, José Antonio Ordaz Sanz and Flor Maria Guerrero Casas

Volume 73, Number 1, 2005

URI: <https://id.erudit.org/iderudit/1107005ar>

DOI: <https://doi.org/10.7202/1107005ar>

[See table of contents](#)

Publisher(s)

Faculté des sciences de l'administration, Université Laval

ISSN

1705-7299 (print)

2371-4913 (digital)

[Explore this journal](#)

Cite this article

del Carmen Melgar Hiraldo, M., Ordaz Sanz, J. & Guerrero Casas, F. (2005). Diverses alternatives pour déterminer les facteurs significatifs de la fréquence d'accidents dans l'assurance automobile. *Assurances et gestion des risques / Insurance and Risk Management*, 73(1), 31–54. <https://doi.org/10.7202/1107005ar>

Article abstract

Traffic accidents are one of the major problems in the developed societies nowadays. The main purpose of this paper is to analyze the most significant factors that explain the number of this type of accidents. This aspect represents a very useful question for the automobile insurance sector.

The literature points out that count data models are the most used ones on this important topic of the whole of the insurance industry. Traditionally, the Poisson regression and the negative binomial model have been the most provided specifications of this type of methodology we can find. However, zero-inflated models could be more appropriated to explain the existing underlying processes in accident situations.

In the present study, we use a database that has been kindly provided by a Spanish private insurance company. After carrying out a thorough descriptive analysis with these data, we compare the results we have obtained from the application of the three referred types of count models to them. Finally, we discuss about the differences pointing up the general advantages of the zero-inflated models. Nevertheless, all results are quite similar in a qualitative way, suggesting the existence of adverse selection and moral hazard processes in the automobile insurance market.

Assurances et gestion des risques, vol. 73(1), avril 2005, 31-54

Insurance and Risk Management, vol. 73(1), April 2005, 31-54

Diverses alternatives pour déterminer les facteurs significatifs de la fréquence d'accidents dans l'assurance automobile

**par María del Carmen Melgar Hiraldo,
José Antonio Ordaz Sanz et
Flor María Guerrero Casas**

RÉSUMÉ

Les accidents de circulation sont un des plus graves problèmes que l'on trouve actuellement dans les pays industrialisés. Le but de cet article est d'analyser les éléments qui contribuent à expliquer le nombre d'accidents de ce genre. Cette question est très utile pour le secteur de l'assurance automobile.

La littérature spécialisée signale que les modèles de comptage sont les plus utilisés dans cet important domaine de l'industrie de l'assurance. Traditionnellement, la régression de Poisson et le modèle binomial négatif ont été les méthodes employées le plus fréquemment. Cependant, les modèles à expansion de zéros pourraient être plus appropriés pour expliquer les processus sous-jacents aux distributions d'accidents automobiles.

Dans cette étude, nous utilisons une base de données fournie par une compagnie d'assurance privée espagnole. Après avoir fait une analyse descriptive complète de ces données, nous comparerons les résultats que l'on obtient de l'application des trois types de modèles de comptage développés dans le papier. Finalement, nous discuterons des différences en soulignant les avantages généraux des modèles à expansion de zéros. Néanmoins, les résultats sont qualitativement très ressemblants, et ils suggèrent en plus l'existence de problèmes d'aléa moral et d'antisélection dans le marché de l'assurance automobile.

Mots clés : Assurance automobile, taux de sinistralité, modèles de comptage.

Les auteurs:

María del Carmen Melgar Hiraldo, José Antonio Ordaz Sanz et Flor María Guerrero Casas enseignent à l'Université Pablo de Olavide, Carretera de Utrera Km.1, 41013 Séville, Espagne.

Courriel : mcmelhir@upo.es; jaordsan@upo.es; fguecas@upo.es.

Traffic accidents are one of the major problems in the developed societies nowadays. The main purpose of this paper is to analyze the most significant factors that explain the number of this type of accidents. This aspect represents a very useful question for the automobile insurance sector.

The literature points out that count data models are the most used ones on this important topic of the whole of the insurance industry. Traditionally, the Poisson regression and the negative binomial model have been the most provided specifications of this type of methodology we can find. However, zero-inflated models could be more appropriated to explain the existing underlying processes in accident situations.

In the present study, we use a database that has been kindly provided by a Spanish private insurance company. After carrying out a thorough descriptive analysis with these data, we compare the results we have obtained from the application of the three referred types of count models to them. Finally, we discuss about the differences pointing up the general advantages of the zero-inflated models. Nevertheless, all results are quite similar in a qualitative way, suggesting the existence of adverse selection and moral hazard processes in the automobile insurance market.

Keywords: Automobile insurance, accident rates, count data models.

I. INTRODUCTION

Depuis quelques années, les pays industrialisés ont une grande inquiétude à cause du grand nombre d'accidents qui se produisent sur les routes et les victimes mortelles que cela entraîne. Les conduites à grande vitesse et sous l'effet de l'alcool sont deux des causes les plus fréquentes d'accident. Il faut remarquer aussi qu'il y a toujours plus d'automobiles qui circulent et c'est sans doute une raison du nombre croissant de sinistres. Les autorités compétentes cherchent des solutions pour réduire la fréquence des infractions au code de la route en général, avec l'espérance que cela se répercutera sur la survenance d'accidents. L'entrée en vigueur du permis de conduire à points et le durcissement des punitions sont de plus en plus étendus dans tous les pays. L'assurance automobile pourrait jouer aussi un rôle important à ce sujet. La « menace » d'une prime élevée peut inciter à la prudence.

En Espagne, le grand nombre de véhicules immatriculés, avec une assurance obligatoire de responsabilité civile, fait de ce secteur la branche la plus notable sur l'ensemble de l'assurance non-vie et une des principales sur le marché de l'assurance en général.

Dans l'assurance automobile, l'analyse du nombre d'accidents que les conducteurs subissent est d'une importance extrême pour la détermination de la prime à payer. L'asymétrie d'information fait que l'assureur ne sait pas *a priori* si son client est plus ou moins sujet à risque. Il faut alors qu'il utilise l'information *a posteriori* et les statistiques sur les accidents. Alors que les études théoriques sur les contrats avec information asymétrique ont débuté dans les années 70, les modèles d'estimation empirique sont beaucoup plus récents.¹

Nous proposons dans ce travail plusieurs tests pour déterminer quelles sont les variables qui contribuent à des sinistralités plus hautes et plus basses. Les modèles économétriques qui s'utilisent normalement à cet effet sont des modèles de comptage, surtout la régression de Poisson et le modèle binomial négatif. Mais les modèles à expansion de zéros pourraient peut-être être plus convenables pour nos données. Pour le vérifier, nous appliquerons le modèle binomial négatif à expansion de zéros, en plus des modèles plus traditionnels.

Ce papier est organisé en 5 sections, cette brève introduction étant la première d'entre elles. La section 2 est dédiée à l'assurance automobile en Espagne. Nous y décrivons l'évolution du marché pendant la période 1999-2003 et quelques aspects de son fonctionnement. Ensuite, dans la section 3, nous signalons les principales caractéristiques des modèles de comptage que l'on utilisera pour faire les estimations. La section 4 inclut les résultats empiriques : une analyse descriptive de la base de données fournie par une compagnie d'assurance privée espagnole et les résultats des modélisations selon les trois modèles de comptage décrits en section 3. La dernière section résume les conclusions du travail. Pour terminer, nous incluons les références bibliographiques et une annexe où sont définies les variables utilisées.

2. L'ASSURANCE AUTOMOBILE EN ESPAGNE

Le grand nombre de véhicules qui circulent sur les routes espagnoles et la nécessité de se protéger des conséquences économiques adverses suite aux dégâts qui se produisent lors d'accidents, font de l'assurance automobile l'une des branches les plus importantes de l'assurance. Dans la dernière décennie, l'immatriculation de véhicules a suivi une tendance à la hausse ; en 1999, elle a atteint son maximum, avec presque 2 millions d'automobiles. À partir de là,

les chiffres ont été stables, à l'exception de l'année 2002 où il y a eu seulement 1 750 000 nouvelles immatriculations.²

Il y a en ce moment près de 20 millions de polices d'assurance automobile émises, ce qui fait de cette branche la première de l'assurance non-vie, bien que très légèrement à la baisse depuis l'année 2000. En 2003, presque 45% du montant des primes non-vie provenait de l'assurance automobile. Tel que montré au Tableau 1, on constate que, entre 1999 et 2002, environ 20% des primes perçues par le secteur de l'assurance provenaient de l'assurance automobile. En 2003, ce pourcentage était au-delà de 25%. On voit ainsi que, malgré la légère baisse observée sur le total du secteur non-vie, l'assurance automobile est plus forte globalement en 2003 que dans les années précédentes. Cela est dû au fait que l'assurance vie a perdu un peu d'importance par rapport au secteur non-vie. La Figure 1 indique le montant des primes obtenues dans chaque secteur.

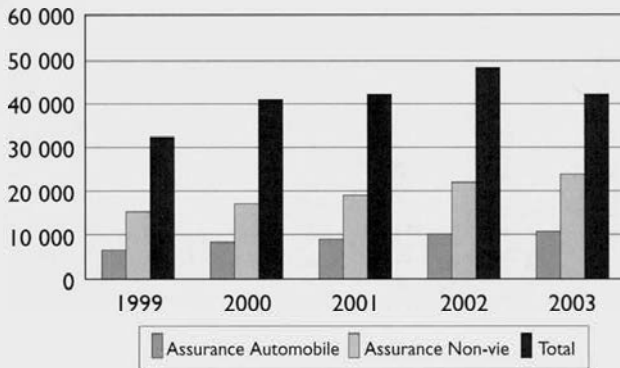
En Espagne, les propriétaires des véhicules à moteur sont obligés de s'assurer contre la responsabilité civile occasionnée par les dommages corporels ou matériels causés à d'autres personnes ou véhicules. En plus, ils peuvent souscrire des garanties facultatives contre d'autres risques. Les plus habituelles sont les suivantes : la garantie des dommages corporels subis par le conducteur, qui est obligatoire pour la plupart des compagnies d'assurance et grâce à laquelle le conducteur peut recevoir une indemnisation en cas de mort, blessures ou invalidité en raison d'un accident dont il porte la responsabilité ou pour lequel personne n'est responsable ; la garantie bris de glaces ; la

TABLEAU I
PRIMES AUTOMOBILE

Année	Primes automobile/ primes non-vie	Primes automobile/ primes totales
1999	43,50 %	20,39 %
2000	46,84 %	19,57 %
2001	46,58 %	21,08 %
2002	45,35 %	20,31 %
2003	44,50 %	25,52 %

Élaboration des auteurs à partir des données de ICEA (2000 à 2004).

FIGURE I
VALEUR DES PRIMES RECOUVRÉES
(EN MILLIONS D'EUROS)



Élaboration des auteurs à partir des données de ICEA (2000 à 2004).

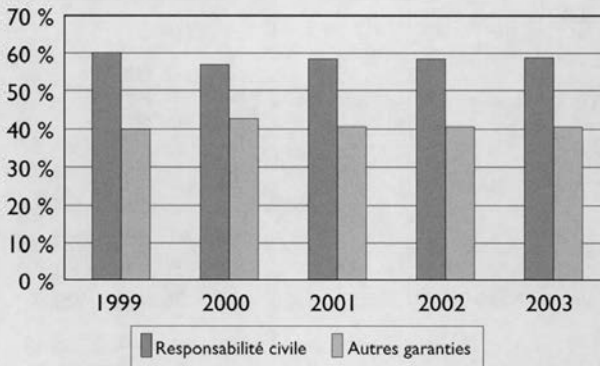
garantie vol ; la garantie incendie ; et la garantie pour les dégâts au véhicule assuré, qui peuvent être couverts à partir d'une limite fixée et d'une franchise, se restreindre aux cas où le tiers est identifié, à la perte totale du véhicule, ou, finalement, être illimités.

Le caractère obligatoire de la couverture de responsabilité civile fait que, sur le total des primes recouvrées, la partie correspondante à la responsabilité civile est plus importante que celle qui se rapporte aux garanties facultatives. La Figure 2 montre l'évolution de ces composantes sur le total de l'assurance automobile, entre 1999 et 2003. Tout au long de cette période, la responsabilité civile représente environ 60% des primes.

En Espagne, comme dans beaucoup d'autres pays, le grand nombre d'accidents routiers, surtout ceux qui font des victimes, est une grande source de préoccupation. Selon la *Dirección General de Tráfico* (1999 à 2003), un maximum a été atteint en 2000, avec presque 102 000 accidents, il est descendu les deux années suivantes jusqu'à 98 433 accidents, et ce maximum est finalement remonté aux alentours de 100 000 en 2003 (voir la Figure 3).

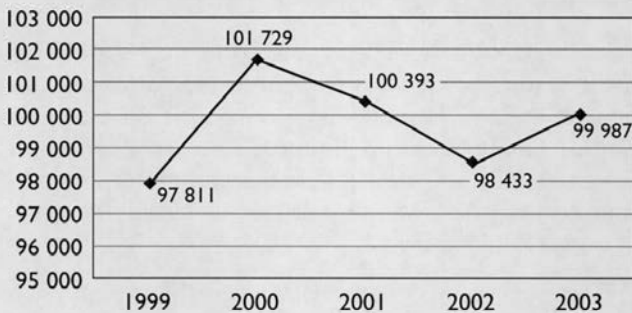
Pour l'assurance automobile, l'estimation du nombre de sinistres pouvant être subis, causant ou non des victimes, devient fondamentale pour ajuster le plus correctement possible les primes. Les compagnies d'assurance utilisent pour cela des statistiques qui leur permettent d'identifier *a priori* le risque d'accident d'un assuré selon des caractéristiques telles que le type de véhicule, son utilisation, son

FIGURE 2
DÉCOMPOSITION DES PRIMES DE L'ASSURANCE
AUTOMOBILE



Élaboration des auteurs à partir des données de ICEA (2000 à 2004).

FIGURE 3
NOMBRE D'ACCIDENTS AVEC DES VICTIMES



Élaboration des auteurs à partir des données de DGT (2000 à 2004).

âge et son sexe, son expérience comme conducteur ou la zone dans laquelle il circule normalement. La connaissance du comportement de l'assuré pendant les années précédentes est importante aussi. En ce qui concerne cette question, on a mis en place un fichier historique des accidents sur les cinq dernières années de tous les conducteurs espagnols assurés dans les compagnies participantes (il contient 85% des polices d'assurance) et qui permet de transmettre fidèlement l'information suite à un changement de compagnie pour un assuré.

L'objectif du travail que nous développons ici est de déterminer lesquelles de ces caractéristiques interviennent vraiment dans la sinistralité. Nous exposons ci-après la méthodologie que nous allons employer et les résultats de l'analyse.

3. LES MÉTHODES

La variable endogène que nous considérerons dans cette étude est le nombre d'accidents que les conducteurs subissent pendant la période prise en compte et qu'ils déclarent à leur compagnie d'assurance. Il s'agit donc d'une variable discrète qui répond aux exigences d'application des modèles de comptage pour son estimation. Parmi ceux-ci, la régression de Poisson et le modèle binomial négatif sont les types de modèles que l'on trouve le plus fréquemment dans les analyses de cette sorte. Cependant, les modèles à expansion de zéros, plus complexes, tiennent en compte certains aspects de la variable dépendante que les modèles plus traditionnels ne distinguent pas. Leur utilisation est convenable quand la quantité de valeurs nulles de la variable endogène est très élevée et, en plus, quand celles-ci peuvent avoir des interprétations différentes et cacher des caractéristiques de comportement de l'assuré qu'*a priori* on ne pourrait pas connaître. Nous exposons à la suite succinctement les modèles de comptage ou *count data models* que nous allons utiliser.³

Le modèle de régression de Poisson

La probabilité qu'une variable aléatoire Y prenne la valeur y_i pour l'individu i est égale à :

$$P(Y = y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \quad (1)$$

où $y_i = 0, 1, 2, \dots$ et λ_i est tel que $\ln \lambda_i = X_i \beta$, ou de façon équivalente, $\lambda_i = e^{X_i \beta}$, en notant X_i le vecteur des régresseurs et β celui des coefficients que l'on veut estimer.

La caractéristique fondamentale de la régression de Poisson est qu'il est nécessaire qu'il y ait de l'équidispersion pour pouvoir l'appliquer correctement. En effet, à partir de la formulation (1) on en déduit que $E(Y) = \lambda_i$ et $Var(Y) = \lambda_i$, c'est-à-dire $Var(Y) = E(Y)$. Cette lourde restriction ne se tient pas normalement quand le nombre de zéros de la variable endogène est élevé. Il y a dans ce cas de

la surdispersion et il faut employer une distribution de probabilité plus flexible. L'alternative choisie est souvent le modèle binomial négatif.

Le modèle binomial négatif

On suppose dans ce modèle que la variable Y suit une distribution de Poisson avec un paramètre μ_i tel que $\mu_i = \ln \lambda_i + \varepsilon_i$ où $\lambda_i = e^{X_i \beta}$, comme avant, et e^{ε_i} suit une distribution gamma avec paramètre ν . De cette façon, la probabilité que la variable Y prenne la valeur Y_i pour l'individu i peut se calculer selon l'expression suivante :⁴

$$P(Y = y_i) = \frac{\Gamma(y_i + \nu)}{\Gamma(y_i + 1) \cdot \Gamma(\nu)} \cdot \left(\frac{\nu}{\nu + \lambda_i} \right)^\nu \left(\frac{\lambda_i}{\nu + \lambda_i} \right)^{y_i} \quad (2)$$

où $\Gamma(\cdot)$ est la fonction gamma et $\nu > 0$.

En introduisant le paramètre de précision⁵ $\alpha > 0$, défini comme $\alpha = 1/\nu$, on peut démontrer que $E(Y) = \lambda_i$ et $Var(Y) = \lambda_i + \alpha \lambda_i^2$. La variance ne coïncide pas maintenant avec la moyenne de la variable, mais c'est une fonction quadratique de celle-ci. En effet, $Var(Y) = E(Y) + \alpha E(Y)^2$. Dans le cas où $\alpha = 0$, la moyenne et la variance de la variable dépendante seraient égales et on obtiendrait alors le modèle de Poisson. Donc, on peut considérer α comme une mesure du degré de surdispersion des données. Si l'on teste la nullité de α , on pourra déterminer la validité du modèle de Poisson ou, au contraire, la nécessité de choisir le modèle binomial négatif.

Cependant, ni le modèle de régression de Poisson ni le binomial négatif ne font de traitement spécial aux valeurs nulles de la variable dépendante. Il conviendrait de le faire parfois, parce que cette valeur peut enfermer des aspects nouveaux. Les modèles à expansion de zéros sont plus complets dans ce sens, car ils tiennent en compte ces possibles aspects différentiels.

Le modèle binomial négatif à expansion de zéros

Les modèles à expansion de zéros sont les développements les plus récents des modèles *count data*. Par rapport aux modèles standards, ils supposent que la variable Y est le produit d'une loi binaire et d'une loi de Poisson ou binomiale négative : $Y = ZY^*$ de telle sorte que Z est un processus modélisé comme un logit avec les valeurs $z_i \in \{0,1\}$ et Y^* détermine la valeur de Y sous la condition $z_i = 1$. L'idée à la base de ces modèles plus complexes est le fait d'une décision

initiale de chaque individu de participer ou pas, représentée par la variable Z . S'il n'y a pas de participation ($z_i = 0$), Y vaut zéro ; quand on participe ($z_i = 1$), Y se conduit selon la loi choisie, pouvant prendre en particulier la valeur nulle aussi. Ainsi :

$$P(Y = 0) = P(Z = 0) + P(Z = 1, Y = 0) = q_i + (1 - q_i)f(0)$$

$$P(Y = y_i > 0) = P(Z = 1, Y = y_i) = (1 - q_i)f(y_i)$$

où q_i est la probabilité de non participation, $q_i = P(z_i = 0)$, et $f(y_i)$ est la loi de Poisson, selon l'équation (1), ou la loi binomiale négative, selon l'équation (2).

Dans le premier cas, nous obtenons le modèle de Poisson à expansion de zéros (*zero-inflated Poisson* ou ZIP) et dans le deuxième le modèle binomial négatif à expansion de zéros (*zero-inflated negative binomial* ou ZINB). Évidemment, si tous les individus participaient ($q_i = 0$), on reviendrait au modèle traditionnel.

Pour le cas du modèle ZINB, la distribution de probabilités s'exprimerait comme :

$$P(Y = y_i) = q_i(1 - \min\{y_i, 1\}) + (1 - q_i) \frac{\Gamma(y_i + \nu)}{\Gamma(y_i + 1) \cdot \Gamma(\nu)} \cdot \left(\frac{\nu}{\nu + \lambda_i} \right)^\nu \left(\frac{\lambda_i}{\nu + \lambda_i} \right)^{y_i} \quad (3)$$

où q_i suit une distribution logistique: $q_i = \Lambda(\tau X_i \beta)$.

Cette formulation nous permet de démontrer que $E(Y) = (1 - q_i)\lambda_i$ et $Var(Y) = (1 - q_i)(1 + \alpha\lambda_i + q_i\lambda_i)\lambda_i$. Si $q_i = 0$, ces expressions coïncident avec celles du modèle binomial négatif. Et si en plus $\alpha = 0$, on aurait le modèle de régression de Poisson.

On trouve peu d'applications des modèles à expansion de zéros dans l'assurance automobile. Dans les dernières années, Lee *et al.* (2002) ont utilisé le modèle ZIP pour étudier la sinistralité des jeunes conducteurs pendant leur première année de conduite, avec des données provenant d'Australie. En Espagne, il n'y a pas de contribution dans ce sens. Nous pensons cependant que le modèle ZINB est le plus approprié pour déterminer, comme nous allons le faire, les variables significatives de la sinistralité pour deux raisons : la prépondérance des observations « 0 » pour la variable « nombre d'accidents » dans l'échantillon (77%, comme on verra plus tard) et le fait que cette valeur zéro peut indiquer que le conducteur assuré n'a pas eu de sinistre pendant la période considérée ou bien que, même s'il en a subi, il ne l'ait pas déclaré à sa compagnie d'assurance. Ainsi,

l'assuré a la possibilité de décider s'il va faire ou non la déclaration d'accident, décision qui peut se modéliser selon une loi logit.

4. LES RÉSULTATS EMPIRIQUES

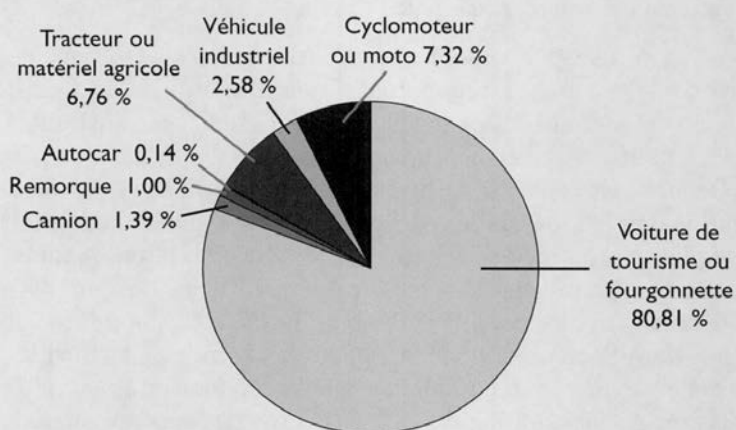
Analyse descriptive des données

La base de données dont nous disposons contient 60 000 assurés d'une importante compagnie d'assurance privée espagnole. L'information fait référence à diverses caractéristiques que nous avons regroupées, comme le font par exemple Dionne, Gouriéroux et Vanasse (1999) et Cohen (2002), dans quatre catégories différentes : caractéristiques du véhicule assuré, caractéristiques du conducteur, caractéristiques du contrat entre l'assuré et la compagnie, et finalement, caractéristiques des sinistres.

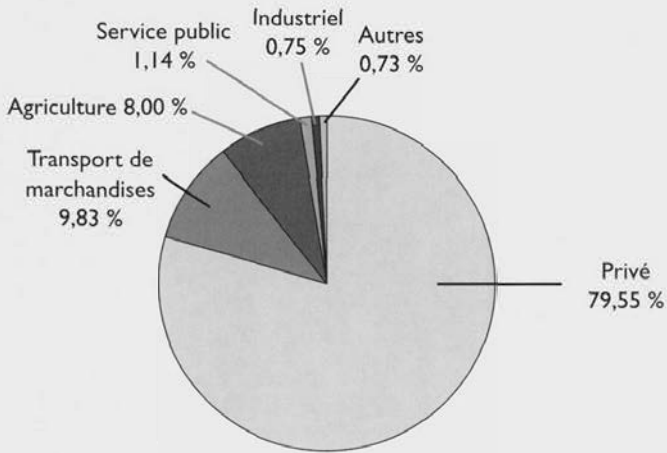
Les Figures 4 à 13 montrent les principaux résultats descriptifs pour les variables considérées dans l'étude et qui sont incluses dans l'annexe finale.⁶

La plupart des véhicules assurés sont des voitures de tourisme ou des fourgonnettes et ils s'utilisent principalement pour le service des particuliers. La moyenne d'âge des conducteurs est de 48 ans ; très peu ont moins de 25 ans. La majorité sont des hommes et il

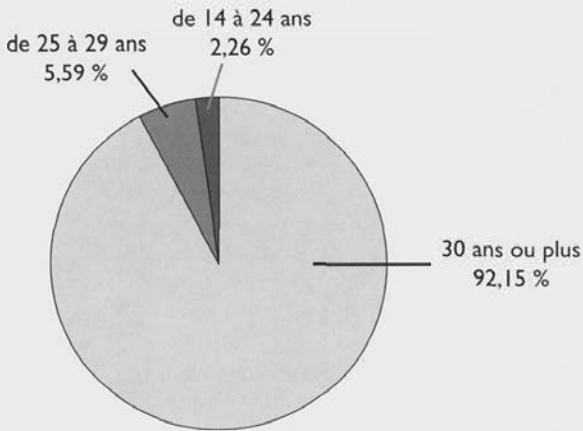
FIGURE 4
TYPE DU VÉHICULE ASSURÉ



**FIGURE 5
USAGE DU VÉHICULE ASSURÉ**



**FIGURE 6
ÂGE DU CONDUCTEUR**



s'agit d'une population de conducteurs ayant une longue expérience de conduite (23 ans en moyenne). La zone où la propension à avoir un accident est moyenne est la plus fréquentée suivie de celle avec la plus basse probabilité d'accident.⁷ En général, les primes diminuent au fur et à mesure que leur montant s'agrandit. Quand le degré de couverture augmente, la proportion d'assurés descend. Le niveau le plus bas correspond au minimum légal obligatoire (assurance aux

FIGURE 7
SEXE DU CONDUCTEUR

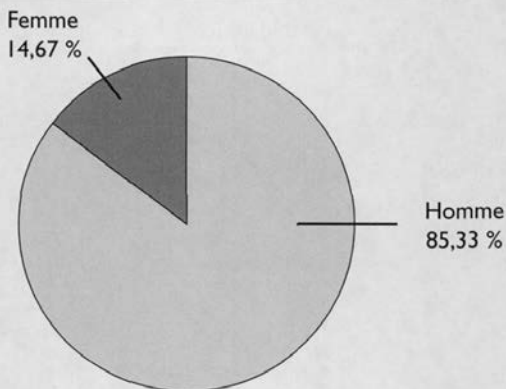
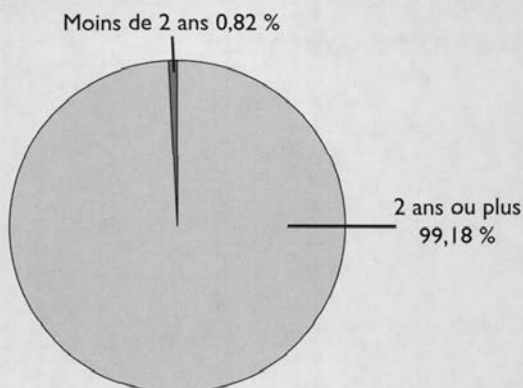


FIGURE 8
EXPÉRIENCE DU CONDUCTEUR



tiers), le plus haut couvre toutes les possibles éventualités (assurance tous risques) et les deux niveaux intermédiaires comprennent, en plus du minimum, d'autres garanties comme le vol de la voiture, l'incendie ou l'assurance contre la suspension du permis de conduire, par exemple. Enfin, en ce qui concerne les sinistres subis et dont la compagnie d'assurance a connaissance, nous remarquons que 13 909 assurés ont eu et déclaré en tout 19 841 accidents, ce qui fait une moyenne de 1,43 par conducteur accidenté. Parmi ceux-ci, la plus grande partie n'a eu qu'un accident (16,33% sur les 60 000 assurés).

FIGURE 9
ZONE HABITUELLE DE CIRCULATION

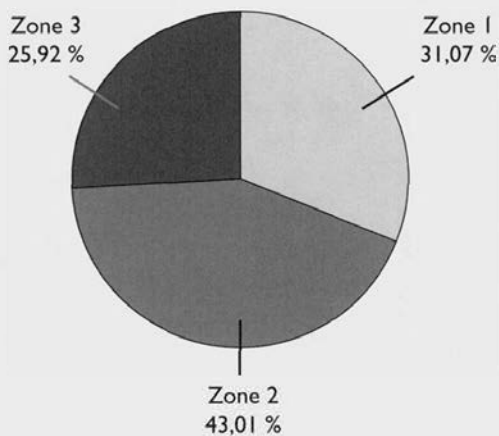


FIGURE 10
PRIME ANNUELLE PAYÉE PAR L'ASSURÉ

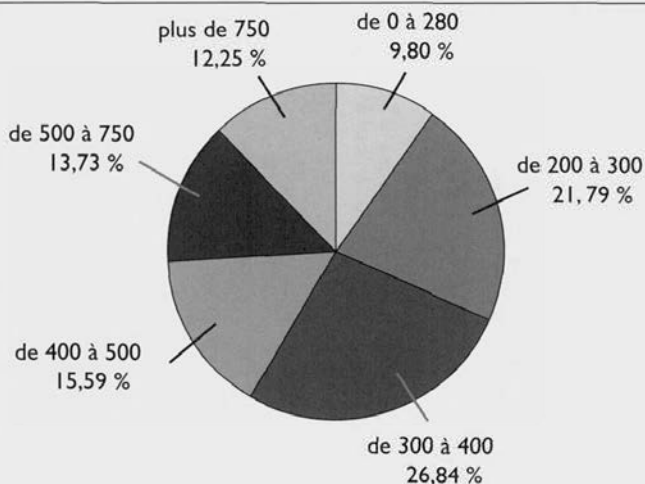


FIGURE 11
NIVEAU DE COUVERTURE

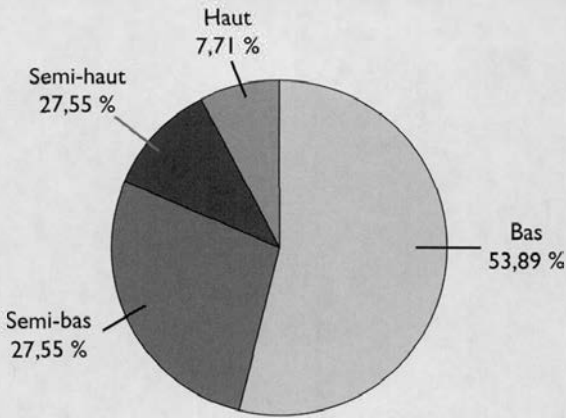
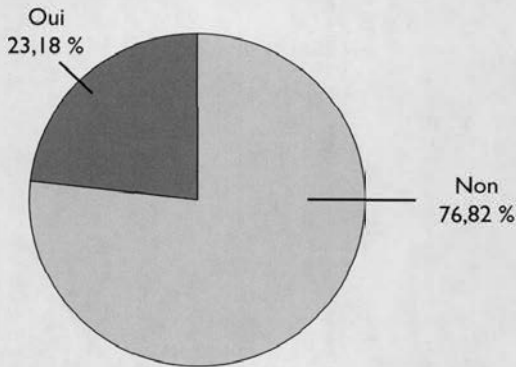


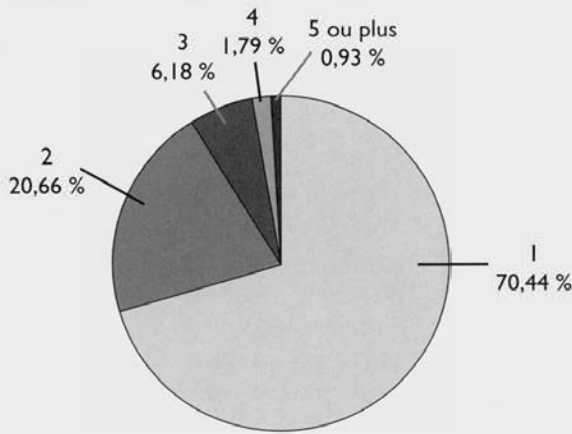
FIGURE 12
SURVENANCE D'ACCIDENTS



Estimation des modèles

Comme nous l'avons indiqué auparavant, l'objectif de l'étude que nous menons est de prévoir les variables qui contribuent le plus à la sinistralité dans l'assurance automobile. Nous utilisons pour cela un échantillon aléatoire de 15 000 assurés à partir duquel, et en utilisant le logiciel économétrique *Limdep v.7.0.*, nous faisons des estimations du nombre d'accidents subis et déclarés par les assurés en appliquant divers modèles *count data* : la régression de Poisson, le modèle binomial négatif et le modèle ZINB.

FIGURE 13
NOMBRE D'ACCIDENTS SUBIS ET DÉCLARÉS



Les Tableaux 2, 3 et 4 montrent les variables qui sont significatives, à un niveau de significativité de 5%, dans chacune des estimations effectuées.

Nous remarquons un premier point en commun dans les trois modélisations : l'absence de significativité des variables relatives au sexe du conducteur et à la zone de circulation. Parmi les variables significatives communes, il y a les catégories de véhicules, l'âge, l'expérience de conduite, ainsi que tous les niveaux de primes et de couverture. Par contre, les principales différences entre estimations se trouvent dans les diverses catégories de la variable « usage du véhicule » qui sont significatives et dans les niveaux de couverture. Nous analysons plus grandement ces différences.

Les usages « transport de marchandises » et « transport scolaire » apparaissent seulement dans certaines estimations. Le premier usage se trouve pertinent uniquement avec la régression de Poisson et c'est la variable la moins significative de toutes celles qui sont déterminantes. Cela pourrait expliquer peut-être la perte de significativité quand on applique les deux autres modèles de comptage. Cette explication n'est pourtant pas acceptable pour l'usage « transport scolaire », dont la significativité est élevée avec le modèle de Poisson et, malgré cela, il s'agit d'une variable qui n'est plus déterminante pour le modèle binomial négatif. On peut l'observer cependant de nouveau comme significatif quand on utilise le modèle ZINB.

TABLEAU 2
ESTIMATION DU NOMBRE D'ACCIDENTS SELON
LE MODÈLE DE RÉGRESSION DE POISSON

Poisson Regression			
Maximum Likelihood Estimate		Dependent variable	NOMSIN
Weighting variable	ONE	Number of observations	15000
Iterations completed	10	Log likelihood function	-10973.33
Chi-squared	1779.015	Restricted log likelihood	-11862.83
Degrees of freedom	17	Significance level	0.0000000
Chi-squared	21001.48021	G-squared	14234.82135
RsqP	0.1164	RsqD	0.1111

Variable	Coefficient	Std. Error	b/St.Er.	P[Z >z]
Constante	-2.331274360	0.17527834	-13.300	0.0000
CAMION	-0.499494085	0.13726493	-3.639	0.0003
REMORQUE	-1.806385611	0.45029189	-4.012	0.0001
CICL_MOT	-1.199356149	0.10862541	-11.041	0.0000
US_INDUS	-0.836173771	0.20595820	-4.060	0.0000
US_TMARC	-0.104964458	0.05219151	-2.011	0.0443
US_TSCOL	0.606691792	0.22316673	2.719	0.0066
US_AGRIC	-1.245518619	0.20275007	-6.143	0.0000
AGE	-0.002670004	0.00116219	-2.297	0.0216
ANCIE_2A	0.297772081	0.14289907	2.084	0.0372
P200_300	0.984455455	0.16797545	5.861	0.0000
P300_400	1.259722409	0.16617719	7.581	0.0000
P400_500	1.435261946	0.16726723	8.581	0.0000
P500_750	1.651876780	0.16717462	9.881	0.0000
P750_	1.827625188	0.17037010	10.727	0.0000
DG_SBAS	0.209422558	0.03613824	5.795	0.0000
DG_SHAUT	0.191822331	0.04773766	4.018	0.0001
DG_HAUT	0.250438747	0.05889440	4.252	0.0000

TABLEAU 3
ESTIMATION DU NOMBRE D'ACCIDENTS SELON
LE MODÈLE BINOMIAL NÉGATIF

Negative Binomial Model				
Maximum Likelihood Estimate		Dependent variable	NOMSIN	
Weighting variable	ONE	Number of observations	15000	
Iterations completed	22	Log likelihood function	-10567.96	
Chi-squared	821.3055	Restricted log likelihood	-10978.61	
Degrees of freedom	1	Significance level	0.0000000	
Variable	Coefficient	Std. Error	b/St.Er.	P[Z >z]
Constante	-2.316729434	0.17996220	-12.873	0.0000
CAMION	-0.593730999	0.15757478	-3.768	0.0002
REMORQUE	-1.876563748	0.48343084	-3.882	0.0001
CICL_MOT	-1.218441300	0.11669946	-10.441	0.0000
US_INDUS	-0.811924444	0.26764260	-3.034	0.0024
US_AGRIC	-1.242707708	0.20903781	-5.945	0.0000
AGE	-0.002968410	0.00145526	-2.040	0.0414
ANCIE_2A	0.472073480	0.19904058	2.372	0.0177
P200_300	0.970897871	0.16782230	5.785	0.0000
P300_400	1.248782853	0.16600098	7.523	0.0000
P400_500	1.419439441	0.16894228	8.402	0.0000
P500_750	1.631494652	0.16835337	9.691	0.0000
P750_	1.814438397	0.17459892	10.392	0.0000
DG_SBAS	0.213097334	0.04307454	4.947	0.0000
DG_SHAUT	0.211052137	0.05716204	3.692	0.0002
DG_HAUT	0.261571012	0.07709611	3.393	0.0007
Overdispersion parameter for negative binomial model				
Alpha	1.189956205	0.065644457	18.127	0.0000

TABLEAU 4 – ESTIMATION DU NOMBRE D'ACCIDENTS SELON LE MODÈLE BINOMIAL NÉGATIF À EXPANSION DE ZÉROS

Zero Altered Neg.Binomial Regression Model

Logistic distribution used for splitting model

Comparison of estimated models

	Pr [0 means]	Number of zeros		Log- likelihood
Poisson	0.77621	Act.= 11558	Prd.=11643.2	-10975.40389
Neg. Bin.	0.80153	Act.= 11558	Prd.=12023.0	-10566.32717
Z.I.Neg\ Bin	0.80914	Act.= 11558	Prd.=12137.1	-10602.78682

Note, the ZIP log-likelihood is not directly comparable

Variable	Coefficient	Std. Error	b/St.Er.	P[Z >z]
Constante	-1.28440599	0.17655471	-7.275	0.0000
CAMION	-0.44590356	0.12996288	-3.431	0.0006
REMORQUE	-1.44735847	0.40273782	-3.594	0.0003
CICL_MOT	-0.96146994	0.12097544	-7.948	0.0000
US_INDUS	-0.62899615	0.23138352	-2.718	0.0066
US_TSCOL	0.54887892	0.21092137	2.602	0.0093
US_AGRIC	-0.96731043	0.19198964	-5.038	0.0000
AGE	-0.00234278	0.00118028	-1.985	0.0472
ANCIE_2A	0.34755214	0.15420227	2.254	0.0242
P200_300	0.75314203	0.15211197	4.951	0.0000
P300_400	0.97798477	0.15747760	6.210	0.0000
P400_500	1.10971377	0.16368736	6.779	0.0000
P500_750	1.28133158	0.16793417	7.630	0.0000
P750_	1.40035194	0.17378581	8.058	0.0000
DG_SBAS	0.16411052	0.03634207	4.516	0.0000
DG_SHAUT	0.17065633	0.04626868	3.688	0.0002
DG_HAUT	0.21886105	0.06141913	3.563	0.0004
Overdispersion parameter				
Alpha	0.05179639	0.02204607	2.349	0.0188
Zero inflation model				
Tau	-0.44048315	0.14566054	-3.024	0.0025

Tous les niveaux de couverture sont significatifs dans les trois modèles. Il y a cependant des différences si l'on compare les niveaux entre eux. Effectivement, dans le modèle ZINB (Tableau 4), plus la couverture choisie est importante, plus la sinistralité est grande, par rapport à celle des assurés qui sont couverts à l'égard des tiers. Ces résultats concordent avec ceux de Dionne, Gouriéroux et Vanasse (1999) et Cohen (2002) mais non pas avec ceux de Chiappori et Salanié (1997, 2000). La base de donnée avec laquelle ces derniers auteurs ont fait leur analyse ne contenait que des conducteurs avec peu d'expérience; c'est peut-être l'explication des différences que l'on trouve. Quand on applique les deux autres modèles de comptage, la relation est différente. Le degré de couverture le plus élevé se distingue toujours sur les autres et présente une sinistralité plus grande qu'eux, mais dans les niveaux intermédiaires la sinistralité est un peu plus forte pour ceux qui ont la couverture semi-basse, bien que les différences ne soient pas trop importantes (voir Tableaux 2 et 3).

5. CONCLUSION

Dans la plupart des pays développés, l'assurance automobile est la branche la plus importante de l'assurance non-vie et une des principales aussi dans le marché de l'assurance en général. En Espagne, plus du 25% des primes de l'année 2003 proviennent de ce secteur. Le grand nombre de véhicules immatriculés chaque année et l'obligation de s'assurer contre la responsabilité civile que l'on peut occasionner aux tiers constituent sans doute les causes principales de ce développement.

En plus de l'inquiétude généralisée de la société face au grand nombre de victimes que les accidents de circulation entraînent, les assureurs ont une préoccupation additionnelle. Une bonne estimation du nombre d'accidents de tous types qui peuvent survenir permettra un meilleur ajustement des primes. Dans l'étude que nous avons présentée, nous avons fait plusieurs tests pour déterminer les facteurs qui font que la fréquence d'accident soit plus ou moins élevée. En plus des deux modèles de comptage qui s'utilisent normalement dans les analyses similaires à celle que nous faisons (la régression de Poisson et le modèle binomial négatif), nous avons appliqué aussi le modèle binomial négatif à expansion de zéros. Ce modèle permet de distinguer les deux interprétations de la valeur « 0 accidents ». D'un côté, elle peut indiquer simplement que l'assuré n'a pas eu d'acci-

dent pendant l'intervalle de temps considéré. Mais, d'un autre côté, elle peut vouloir dire que le conducteur n'a rien déclaré à sa compagnie d'assurance, même s'il a eu un ou plusieurs accidents.

Du point de vue théorique, le modèle binomial négatif à expansion de zéros paraît donc être le plus approprié, parce qu'en plus de recueillir la double interprétation de la valeur nulle de la variable endogène, il y a de la surdispersion dans nos données, à cause du grand pourcentage de conducteurs qui n'ont pas eu ou n'ont pas déclaré d'accidents (77%).

Du point de vue empirique, bien que les résultats qualitatifs des trois modélisations faites ne sont pas très différents, la significativité des coefficients de surdispersion et du modèle à expansion de zéros (Tableau 4) confirment que le modèle ZINB est celui qui fournit la meilleure information sur les variables qui déterminent la fréquence des accidents subis et déclarés.⁸

Finalement, nous signalons quelques aspects remarquables de la comparaison entre les estimations. Les principales différences concernent l'usage du véhicule et le niveau de couverture choisi par l'assuré. Sur ce dernier sujet, quand on applique le modèle ZINB, on observe que les prédictions théoriques en présence d'asymétrie d'information se confirment : plus le niveau de couverture est complet, plus haute est la fréquence d'accident estimée. Il peut y avoir là une composante d'antisélection, puisque les conducteurs qui savent qu'ils sont plus sujets à risques essaient de se protéger avec le plus de garanties possibles. D'un autre côté, on ne peut négliger la composante d'aléa moral: le conducteur qui est assuré contre la plupart des contingences sera moins prudent que celui qui n'est assuré qu'à l'égard des tiers. Quant au reste des variables, à l'exception du sexe et de la zone de circulation habituelle, qui ne paraissent pas influencer sur la sinistralité, elles restent significatives dans les trois modèles.

Références

- Boyer, M. et Dionne, G., 1989, « An Empirical Analysis of Moral Hazard and Experience Rating », *Review of Economics and Statistics*, LXXI (1), 128-134.
- Cameron, A. et Trivedi, P., 1986, « Econometric Models Based on Count Data: Comparison and Applications of Some Estimators and Tests », *Journal of Applied Econometrics*, 1, 29-54.
- Cameron, A. et Trivedi, P., 1998, «Regression Analysis of Count Data», Cambridge University Press.
- Chiappori, P.A. et Salanié, B., 1997, « Empirical Contract Theory: The Case of Insurance Data », *European Economic Review*, 41, 943-950.

- Chiappori, P.A. et Salanié, B., 2000, « Testing for Asymmetric Information in Insurance Markets », *Journal of Political Economy*, 108, 1, 56-78.
- Cohen, A., 2002, « Asymmetric Information and Learning: Evidence from the Automobile Insurance Market », *Discussion Paper* n°371, Harvard Law School, Cambridge.
- DGT, 2000, Anuario Estadístico de Accidentes 1999. Dirección General de Tráfico, Ministerio del Interior, Madrid.
- DGT, 2001, Anuario Estadístico de Accidentes 2000. Dirección General de Tráfico, Ministerio del Interior, Madrid.
- DGT, 2002, Anuario Estadístico de Accidentes 2001. Dirección General de Tráfico, Ministerio del Interior, Madrid.
- DGT, 2003, Anuario Estadístico de Accidentes 2002. Dirección General de Tráfico, Ministerio del Interior, Madrid.
- DGT, 2004a, Anuario Estadístico de Accidentes 2003. Dirección General de Tráfico, Ministerio del Interior, Madrid.
- DGT, 2004b, Anuario Estadístico General 2003. Dirección General de Tráfico, Ministerio del Interior, Madrid.
- Dionne, G., Gouriéroux, C. et Vanasse, C., 1999, « Evidence of Adverse Selection in Automobile Insurance Markets » dans *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation*, Dionne, G. et C. Laberge-Nadeau (eds.), 13-46.
- Greene, W.H., 1995, *Limdep Version 7.0: User's Manual*. Bellport, NY: Econometric Software.
- ICEA, 2000, Evolución del Mercado Asegurador en 1999. Informe n°800. ICEA (Investigación Cooperativa entre Entidades Aseguradoras y Fondos de Pensiones), Madrid.
- ICEA, 2001, Evolución del Mercado Asegurador en 2000. Informe n°828. ICEA, Madrid.
- ICEA, 2002, Evolución del Mercado Asegurador. Estadística Año 2001. Informe n°866. ICEA, Madrid.
- ICEA, 2003, Evolución del Mercado Asegurador. Estadística Año 2002. Informe n°890. ICEA, Madrid.
- ICEA, 2004, Evolución del Mercado Asegurador. Estadística Año 2003. Informe n°917. ICEA, Madrid.
- Jones, A.M., 2001, « Applied Econometrics for Health Economists – A Practical Guide », Office of Health Economics. London.
- Lee, A.H, Stevenson, M.R., Wang, K. et Yau, K.K.W., 2002, « Modeling Young Driver Motor Vehicle Crashes: Data with Extra Zeros », *Accident Analysis and Prevention*, 34, 515-521.
- Melgar, M.C., 2004, « Análisis de las Componentes de la Demanda de Seguro », Aplicación al Seguro del Automóvil. Tesis Doctoral. Universidad de Sevilla.
- Puelz, R. et Snow, A., 1994, « Evidence on Adverse Selection: Equilibrium Signalling and Cross-Subsidization in the Insurance Market », *Journal of Political Economy*, 102, n° 2, 236-257.
- Winkelmann, R., 2003, « Econometric Analysis of Count Data », Springer.

Notes

1. Boyer et Dionne (1989) ont testé pour la première fois l'hypothèse d'asymétrie d'information sur des données de l'assurance automobile. Ensuite, depuis Puelz et Snow (1994), les tests se sont multipliés.

2. Ces données ont été obtenues de DGT (2004b).

3. Pour avoir plus de détails sur les modèles *count data*, voir Cameron et Trivedi (1986, 1998) ou Winkelmann (2003).

4. Selon la terminologie de Cameron et Trivedi (1986), ce modèle s'appellerait binomial négatif II et c'est un cas particulier (avec $k = 0$) du modèle plus général que l'on obtient

en prenant $v_i = \frac{1}{\alpha} (e^{X_i\beta})^k$. Si $k = 1$, on aurait le modèle binomial négatif I, où la moyenne et la variance de Y sont proportionnelles.

5. Voir Jones (2001).

6. Une analyse descriptive plus exhaustive peut se trouver dans Melgar (2004).

7. Le climat de la zone, l'état des routes, l'âge des conducteurs, etc., sont des aspects qui peuvent faire que, dans certaines zones d'Espagne, il y ait une probabilité d'accident plus élevée que dans d'autres.

8. Voir Greene (1995).

ANNEXE – LISTE DES VARIABLES

Le type du véhicule assuré

- TOU_FOUR = 1 si le véhicule assuré est une voiture de tourisme ou une fourgonnette, 0 sinon (référence)
- CAMION = 1 si le véhicule assuré est un camion, 0 sinon
- REMORQUE = 1 si le véhicule assuré est une remorque, 0 sinon
- AUTOCAR = 1 si le véhicule assuré est un autocar ; 0 sinon
- TRAC_MA = 1 si le véhicule assuré est un tracteur ou un matériel agricole, 0 sinon
- VEH_IND = 1 si le véhicule assuré est un véhicule industriel, 0 sinon
- CICL_MOT = 1 si le véhicule assuré est un cyclomoteur ou une moto, 0 sinon

L'usage du véhicule assuré

- US_PRIVÉ = 1 si l'usage du véhicule assuré est l'usage particulier, 0 sinon (référence)
- US_SP = 1 si l'usage du véhicule assuré est le service public, 0 sinon
- US_LOCAT = 1 si l'usage du véhicule assuré est la location, 0 sinon
- US_ECOL = 1 si l'usage du véhicule assuré est l'autoécole, 0 sinon
- US_ACVN = 1 si l'usage du véhicule assuré est l'achat et vente, 0 sinon
- US_INDUS = 1 si l'usage du véhicule assuré est l'industriel, 0 sinon
- US_TMAR = 1 si l'usage du véhicule assuré est le transport de marchandises, 0 sinon
- US_TSCOL = 1 si l'usage du véhicule assuré est le transport scolaire, 0 sinon
- US_TGV = 1 si l'usage du véhicule assuré est le transport général de voyageurs, 0 sinon
- US_AGRIC = 1 si l'usage du véhicule assuré est agricole, 0 sinon
- US_SPC = 1 si l'usage du véhicule assuré est la suspension du permis de conduire, 0 sinon

L'âge de l'assuré

- AGE = âge de l'assuré le 15 décembre 2002

L'expérience de l'assuré dans la conduite

- ANCIE_2A = 1 si l'assuré a le permis de conduire depuis moins de 2 ans (jusqu'au 15 décembre 2002), 0 sinon

Le sexe de l'assuré

- FEMME = 1 si l'assuré est une femme, 0 sinon

ANNEXE – LISTE DES VARIABLES (SUITE)

La zone où l'assuré circule habituellement

- ZONE_1 = 1 si l'assuré circule habituellement dans la zone où la probabilité d'accident est la plus basse, 0 sinon (référence)
- ZONE_2 = 1 si l'assuré circule habituellement dans la zone où la probabilité d'accident est moyenne, 0 sinon
- ZONE_3 = 1 si l'assuré circule habituellement dans la zone où la probabilité d'accident est la plus élevée, 0 sinon

La prime annuelle payée par l'assuré

- P0_200 = 1 si la prime annuelle payée par l'assuré ne dépasse pas 200 €, 0 sinon (référence)
- P200_300 = 1 si la prime annuelle payée par l'assuré est de plus de 200 € et ne dépasse pas 300 €, 0 sinon
- P300_400 = 1 si la prime annuelle payée par l'assuré est de plus de 300 € et ne dépasse pas 400 €, 0 sinon
- P400_500 = 1 si la prime annuelle payée par l'assuré est de plus de 400 € et ne dépasse pas 500 €, 0 sinon
- P500_750 = 1 si la prime annuelle payée par l'assuré est de plus de 500 € et ne dépasse pas 750 €, 0 sinon
- P750_ = 1 si la prime annuelle payée par l'assuré est de plus de 750 €, 0 sinon

La couverture de l'assurance

- DG_BAS = 1 si l'assuré bénéficie du degré de couverture le plus bas, 0 sinon (référence)
- DG_SBAS = 1 si l'assuré bénéficie d'un degré de couverture semi bas, 0 sinon
- DG_SHAUT = 1 si l'assuré bénéficie d'un degré de couverture semi haut, 0 sinon
- DG_HAUT = 1 si l'assuré bénéficie du degré de couverture le plus haut, 0 sinon

Le nombre de sinistres subis et déclarés par l'assuré

- NOMSIN = nombre d'accidents soufferts par l'assuré et déclarés à la compagnie d'assurance entre le 16 juin 2002 et le 15 juin 2003