risks

MDPI

Article

Modelling Unobserved Heterogeneity in Claim Counts Using Finite Mixture Models

Lluís Bermúdez [1,*], Dimitris Karlis [2] and Isabel Morillo [1]

[1] Departament de Matemàtica Econòmica, Financera i Actuarial, Universitat de Barcelona, Diagonal 690, 08034-Barcelona, Spain; imorillo@ub.edu

[2] Department of Statistics, Athens University of Economics and Business, 10434 Athens, Greece; karlis@aueb.gr

* Correspondence: lbermudez@ub.edu; Tel.: +34-93-403-4853; Fax: +34-93-403-4892

Received: 20 December 2019; Accepted: 22 January 2020; Published: 29 January 2020

Abstract: When modelling insurance claim count data, the actuary often observes overdispersion and an excess of zeros that may be caused by unobserved heterogeneity. A common approach to accounting for overdispersion is to consider models with some overdispersed distribution as opposed to Poisson models. Zero-inflated, hurdle and compound frequency models are typically applied to insurance data to account for such a feature of the data. However, a natural way to deal with unobserved heterogeneity is to consider mixtures of a simpler models. In this paper, we consider *k*-finite mixtures of some typical regression models. This approach has interesting features: first, it allows for overdispersion and the zero-inflated model represents a special case, and second, it allows for an elegant interpretation based on the typical clustering application of finite mixture models. *k*-finite mixture models are applied to a car insurance claim dataset in order to analyse whether the problem of unobserved heterogeneity requires a richer structure for risk classification. Our results show that the data consist of two subpopulations for which the regression structure is different.

Keywords: zero-inflation; overdispersion; automobile insurance; risk classification; risk selection

JEL Classification: C51

1. Introduction and Aims

In insurance datasets, for the purposes of modelling claim counts, there is a problem of unobserved heterogeneity caused by differences in driving habits and behaviour among policyholders that cannot be observed or measured by the actuary (for example, driving ability, driving aggressiveness or the degree of obeying traffic regulations). This often leads to overdispersion and a relatively large number of zeros, which cannot be fully remedied by Poisson regression models. Many attempts have been made in the actuarial literature to account for such features of the data (for example, compound frequency models, also known as mixture models, and their zero-inflated or hurdle versions). This paper aims to explore whether the problem of unobserved heterogeneity requires a richer structure, though the use of finite mixtures of regression models, than the previous models have.

In a competitive market, insurance companies need to use a pricing structure that ensures that the exact weight of each risk is fairly distributed within the portfolio. If an insurance company does not achieve at least the same success with respect to this goal as their competitors, the policyholders with lower risk will be tempted to move to another company that offers better rates for them. Such an adverse selection process would lead the less unsuccessful company to lose its financial equilibrium, with insufficient income from premiums to pay for the claims reported by the remaining policyholders with higher risk.

In most developed countries, the car insurance market is a highly competitive market. Therefore, to avoid such an adverse selection process, a particularly complex pricing structure is designed by actuaries.

A thorough review of the modelling of claim counts for car insurance can be found in (Denuit et al. 2007). In general terms, to handle this problem, the actuary segments the portfolio into homogeneous classes so that all the insured parties belonging to a particular class pay the same premium. This procedure is referred to as risk classification, tariff segmentation or a priori ratemaking.

In short, the classification or segmentation of risks involves establishing different classes of risk according to the nature of claims and probability of their occurrence. To this end, factors are determined to classify each risk, and its influence on the observed number of claims is estimated. To achieve this, risk analysis based on generalized linear models (GLMs) is widely accepted. Focusing on claim frequency, a regression component is included in the claim count distribution to take individual characteristics into account.

A very common GLM used for these purposes is the Poisson regression model and its generalisations. Introduced by Dionne and Vanasse (1989) in the context of car insurance, the model can be applied if a series of classification variables, referred to as a priori variables, plus the number of claims for each individual policy are known. However, the Poisson regression model is usually rejected because of the presence of overdispersion and an excess of zeros. This rejection may be interpreted as a sign that the portfolio is still heterogeneous: not all factors influencing risk can be identified, measured and introduced into the a priori modelling. This phenomena is known as the problem of unobserved heterogeneity.

In parallel, another way to account for unobserved heterogeneity is to consider that the claims record for each insured party reveals the differences in driving habits and behaviour among policyholders that cannot be observed or measured via the a priori variables. Therefore, the idea of considering individual differences in policies within the same a priori class by using an a posteriori mechanism has emerged, i.e., tailoring an individual premium based on the claims record for each insured party. This concept has received the name of a posteriori ratemaking, experience rating or the bonus-malus system (see Denuit et al. (2007)).

One way to deal with overdispersion is to consider compound frequency models (mixture models) with some overdispersed distribution. This is best achieved by moving from the simple Poisson model to the negative binomial model (Dionne and Vanasse (1992)) or to the Poisson-inverse Gaussian model (Dean et al. (1989)). To account for the excess of zeros, some generalizations of the Poisson model have been considered. Lambert (1992) introduced the zero-inflated Poisson regression model and, since then, there has been a considerable increase in the number of applications of zero-inflated regression models based on several different distributions. A comprehensive discussion of these applications can be found in Winkelmann (2008). Similarly, hurdle models are also widely applied to insurance claim count data. A common assumption in all these models is that all policyholders behave in the same way with regard to a priori variables, and thus they all have the same regression structure.

In this paper, we examine whether this assumption is realistic. The models proposed in this paper account for unobserved heterogeneity by choosing a finite number of subpopulations. To account for overdispersion and an excess of zeros, we consider a k-finite mixture of Poisson and negative binomial regression models. As Park and Lord (2009) show for vehicle crash data analysis, a finite mixture of Poisson or negative binomial regression models is especially useful where count data are drawn from heterogeneous populations. For modelling claim counts, the idea behind this is that the data consist of subpopulations of policyholders, "caused" by the unobserved heterogeneity, for which the regression structure, used to account for the observed or a priori variables, is different. These models allow each component in the discrete mixture to have its own score, i.e., for there to be different behaviour for each group of policyholders, whereas classical claim frequency models use a single score.

To sum up, this paper aims to explore whether resolving the problem of unobserved heterogeneity requires a richer structure than that which is present in typical compound frequency models and their zero-inflated or hurdle versions. By applying finite mixtures of regression models, we will examine whether unobserved risk factors that are not considered in the a priori tariff, such as a driver's reflexes, aggressiveness, or knowledge of the Highway Code, establish the existence of subpopulations of policyholders with different a priori behaviour. To achieve this goal, the proposed models are fitted to a set of car insurance claims data to compare their goodness of fit with the traditional claim frequency

models and to assess if we need to account for this extra heterogeneity. Finally, we discuss whether the proposed models help to search for better alternatives to account for unobserved heterogeneity.

In the next section, the models and computational details used are defined. In Section 3, we summarize the database obtained from a Spanish insurance company and the results from fitting the models to it. Finally, we offer some concluding remarks in Section 4.

2. Finite Mixture of Regression Models

The central idea for a finite mixture of regression models is that we assume that the entire population can be split into k subpopulations (also called clusters, components or segments). Assuming a discrete-valued response y_i for the i-th individual, we then assume that

$$P(y_i) = P(Y_i = y_i) = \sum_{j=1}^{k} \pi_j P(y_i|\theta_{ij}), \quad \theta_{ij} > 0, \;\; y_i = 0, 1, \ldots,$$

where $0 < \pi_j < 1$ with $\sum_{j=1}^{k} \pi_j = 1$ are the mixing proportions indicating the probability that a randomly selected observation belongs to the j-th subpopulation and $P(y|\theta)$ is some discrete distribution indexed by some parameter vector θ. In our case presented below, $P(y|\cdot)$ will be assumed to belong to one of the Poisson or negative binomial families. Note that we assume that for each individual we have a set of parameters θ_{ij} that depend on each component and they may depend on some covariate information for the i-th individual.

We further assume that the mean of the j-th component can be modelled by a vector of covariates containing information on the i-th individual, denoted by $\mathbf{x}_i$. In the general setting, this covariate vector that characterises the i-th individual can be different for different components, and therefore we should use also a subscript j. As, in our model, we use the same covariates for all components, we drop this second index. Assuming, without loss of generality, that $\theta = (\mu, \phi)$, where μ is the mean of the distribution (this can be easily obtained with a reparameterisation) and ϕ some parameter related to overdispersion (set equal to 1 for the Poisson distribution), we further assume that:

$$\log \mu_{ij} = \mathbf{x}_i' \beta_j$$

where now β_j is a component-specific vector of coefficients.

Note that the above formulation can be seen in the context of a GLM. However, we prefer to describe the model in a more general setting since some of the families we may use instead of Poisson and negative binomial models do not belong to the exponential family or therefore to the general GLM setting.

The above generic formulation can be expanded by allowing additional covariates to the rest of the parameters for each component, as well as to the vector of mixing proportions. A well-known model of this type is the finite mixture of Poisson regressions in Wang et al. (1996) (see also Grun and Leisch 2007, 2008). Finite mixtures of regression models have been widely used in different settings, see Hennig (2000) for a thorough discussion.

This type of modelling has some interesting features: first, the zero-inflated model is a special case; second, it allows for overdispersion; and third, it allows for a neat interpretation based on the typical clustering application of finite mixture models.

It is useful to show that if we denote by μ_j and σ_j^2 the mean and the variance of the j-th component, then the mean μ and the variance σ^2 of the mixture are given by

$$\mu = \sum_{j=1}^{k} \pi_j \mu_j, \;\; \text{and} \;\; \sigma^2 = \sum_{j=1}^{k} \pi_j (\mu_j^2 + \sigma_j^2) - \mu^2 \tag{1}$$

These formulas will be useful later on for our calculations.

2.1. Finite Mixture of Poisson Regressions

The case of a finite mixture of Poisson regressions is by far the best known and most commonly applied in practice. It dates back to Wang et al. (1996) and assumes that

$$P(y_i|\mu_{ij}) = \sum_{j=1}^{k} \pi_j \frac{\exp(-\mu_{ij})\mu_{ij}^{y_i}}{y_i!}, \quad y_i = 0, 1, \ldots$$

with $\mu_{ij} = \exp(\mathbf{x_i}'\beta_j)$. The zero-inflated Poisson regression is a special case. The model allows for overdispersion with respect to the simple Poisson regression model. For more details see Grun and Leisch (2007); Wang et al. (1996).

2.2. Finite Mixture of Negative Binomial Regressions

For the negative binomial model, we assume

$$P(y_i|\mu_{ij}, \phi_j) = \frac{\Gamma(\phi_j + y_i)}{\Gamma(\phi_j) y_i!} \left(\frac{\mu_{ij}}{\phi_j + \mu_{ij}}\right)^{y_i} \left(\frac{\phi_j}{\phi_j + \mu_{ij}}\right)^{\phi_j}, \quad \phi_j > 0, \quad y_i = 0, 1, \ldots$$

and $\mu_{ij} = \exp(\mathbf{x_i}'\beta_j)$, i.e., the probability function of a negative binomial with mean μ_{ij} and variance $\mu_{ij} + \frac{\mu_{ij}^2}{\phi_j}$.

Note that we assume a separate overdispersion parameter ϕ_j for each component. Such a model has been fitted by Byung-Jung et al. (2014) and Zou et al. (2013). With respect to the finite mixture of Poisson regressions, the model has an extra overdispersion parameter and therefore allows for more flexible distributions in terms of components.

It is evident that the negative binomial model also contains the simple Poisson model as a special case ($\phi_j \to \infty$).

2.3. Other Models

Although in this paper we focus on the two families of models introduced above, there are other models that fit into this context for which we do not present results. They relate to Poisson-inverse Gaussian regression models (Dean et al. (1989)) and finite mixtures of them; some nonparametric random effects Poisson regression models (see Aitkin (1999)), i.e., the model assumes some random effect on the intercept of the Poisson regression and thus actually fits a finite mixture of Poisson regression model where the estimated coefficients (apart from the intercept) are the same for all components; and hurdle-type models (a hurdle model is a modified count model in which the two processes generating the zeros and the positives are not constrained to be the same, see Mullahy (1986)). As mentioned above, we do not formulate zero-inflated models as we treat them as special cases of finite mixture models.

2.4. Estimation via EM Algorithm

Under the umbrella of a finite mixture, estimation for this particular family of models is rather simple. We follow the standard approach of combining the observed data $(Y_i, \mathbf{X}_i)$ with unobserved latent vectors $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ik})$ with $Z_{ij} = 1$ if the i-th observation belongs to the j-th component, and 0 otherwise. As is typical, the EM algorithm consists of estimating the Z's by their conditional expectation and then fitting a standard regression model to the response, Y, using a weighted likelihood, based on the weights derived during the E-step. A formal description of a generic algorithm is given in what follows.

E-step: Using the current estimates, $\hat{\pi}_j$ and $\hat{\theta}_{ij}$, $i = 1, \ldots, n$ and $j = 1, \ldots, k$, calculate

$$w_{ij} = E(Z_{ij}) = \frac{\hat{\pi}_j P(y_i|\hat{\theta}_{ij})}{\sum_{j=1}^{k} \hat{\pi}_j P(y_i|\hat{\theta}_{ij})}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, k \tag{2}$$

and then

M-step:

M1 Update the mixing proportions using

$$\hat{\pi}_j = \frac{\sum_{i=1}^{n} w_{ij}}{n}, \quad j = 1, \ldots, k$$

M2 Update the regression coefficients and the component-specific parameters by fitting a single regression model for the j-th component with response y_i, covariates $\mathbf{x}_i$ using a weighted likelihood approach with weights w_{ij}.

It is clear that the M-step is not in a closed form. Also, note that actually we fit k models with the same data but different weights. This can be run in parallel to speed up the process. All the pros and cons of the EM algorithm for finite mixtures apply. Also, standard procedures for finite mixtures are applicable, such as for example model selection. We will discuss some computational details later.

Finally, we need to emphasize the issue of identifiability. Conditions for identifiability for such finite mixtures of regression models are given in Hennig (2000). For such a finite mixture of regression models for count data, problems may occur if the covariates are categorical and they can have a small number of different combinations. In our case, we have seven binary variables leading to 2^7 combinations. Not all of them appear in the data but we still have quite a large number of distinct combinations for the model matrix. In general, it is hard to show that identifiability exists since the conditions are hard to evaluate. We believe that in our case no particular problem exists. From a practical point of view we have worked with several initial values to examine whether we became trapped with different solutions. This did not happen, adding to our belief that our model is identifiable.

2.5. Computational Details

An important aspect for the successful application of the EM algorithm is that appropriate initial values need to be selected, as otherwise one may be trapped in local rather than global maxima. We selected our initial values as follows.

We started by fitting a simple Poisson regression model. This also gave sufficient initial values for the simple negative binomial regression. Initial values for the overdispersion parameter in these two models were set equal to the observed overdispersion (as proposed in Breslow (1984)).

From here on we describe the approach for each model. Therefore, when we refer to “model", we imply either the mixture of Poisson models or the mixture of negative binomial models. Initial values for $k = 2$ were selected by perturbing the simple ($k = 1$) regression model. Specifically, we fitted a single regression and keeping the fitted values, we split them into two components with mixing probabilities of 0.5 each, and means equal to 1.2 and 0.8 of the fitted values. Then, to fit a model with $k + 1$ components, we used the solution with k components and a new component at the centre (that of a single one-component regression), with mixing probability 0.05. The other mixing probabilities were rescaled to sum to 1. Extensive simulation has shown that this approach works well to locate the maximum. Other approaches can be found in Papastamoulis et al. (2016).

All our computations were made in `R`. We used our own code, while some of the models can be fitted using the `gamlss`, `VGAM` and `flexmix` packages in `R`. However, we found some convergence problems and less flexibility while using the standard packages.

Convergence was detected when the relative change between two successive iterations was smaller than 10^{-8}. For fitting the separate regression models we used the standard GLM approach (IRLS algorithm) for Poisson and negative binomial regression.

3. Data and Results

3.1. Data Description

The original database is a random sample of the car portfolio of a major insurance company operating in Spain in 1996. Only cars categorized as being for private use were considered. The data contain information from 80,994 policyholders. Seven exogenous variables plus the annual number of accidents recorded were considered here. For each policy, the information at the beginning of the period and the total number of claims from policyholders were reported. The definition and some descriptive statistics of the variables are presented in Table 1. This dataset has previously been used in Pinquet et al. (2001), Brouhns et al. (2003), Bolancé et al. (2003, 2008), Boucher et al. (2007, 2009), Boucher and Denuit (2008), Bermúdez (2009) and Bermúdez and Karlis (2012).

The meaning of the variables that refer to the Spanish market should also be clarified. The variable *ZON* distinguishes between driving zones of greatest risk (Madrid, Catalonia and central northern Spain) and the rest. Regarding the type of coverage provided by the of policies (variable *COV*), the classification adopted here responds to the most common types of car insurance policy available on the Spanish market. The simplest policy only includes third-party liability. This simplest type of policy makes up the baseline group, while variable *COV* equals 1 denotes policies which, apart from the guarantees contained in the simplest policies, also include comprehensive and collision coverage.

Table 1. Dependent and explanatory variables used in the models.

Variable	Definition	Mean	St. dev.
N	total number of claims reported by policyholders (0: 71,087; 1: 6,744; 2: 2,067; 3: 690; 4: 248; 5: 95; 6: 34; >6: 29)	0.1833	0.5873
GEN	equals 1 for women and 0 for men	0.1600	0.3666
URB	equals 1 when driving in urban area, 0 otherwise	0.6690	0.4706
ZON	equals 1 when driving in Madrid, Catalonia or northern Spain, 0 otherwise	0.4326	0.4954
LIC	equals 1 if the driving license is 4 or more years old, 0 otherwise	0.9766	0.1511
LOY	equals 1 if the client is in the company for more than 5 years, 0 otherwise	0.1441	0.3512
COV	equals 1 if includes comprehensive and collision coverage, 0 otherwise	0.5087	0.4999
POW	equals 1 if horsepower is greater than or equal to 5500cc, 0 otherwise	0.8058	0.3955

3.2. Fitted Models

We fitted models of increasing complexity to this dataset, starting from a simple Poisson regression model. We used AIC and BIC to select the best among a series of candidate models. All models were run in R. Table 2 compares the fitted models for Poisson and negative binomial distributions, resulting in the best fit being obtained with a 2-finite mixture of negative binomial regression models (2FMNB). Finite mixture models with $k > 2$ were also fitted, but no improvement in terms of AIC or BIC was achieved. This result gives rise to the conclusion that this portfolio is comprised of two groups of policyholders.

Table 2. Information criteria for selecting the best model for the data.

Model	Log-Likelihood	Parameters	AIC	BIC
Poisson	−42,585.08	8	85,186.15	85,260.57
Negative binomial	−38,453.13	9	76,924.27	77,007.98
Zero-inflated Poisson	−38,836.59	9	77,691.19	77,774.91
Zero-inflated negative binomial	−38,453.13	10	76,926.27	77,019.28
2-Finite Poisson mixture	−38,449.61	17	76,933.21	77,091.36
2-Finite negative binomial mixture	−38,347.81	19	76,733.62	76,910.36

As expected, a large improvement is obtained by moving from a simple Poisson model to a compound frequency model with some overdispersed distributions. The best fit is achieved by the negative binomial model. Zero-inflated models, while providing an improvement on the basic Poisson model, allowing for overdispersion in this case, were not helpful for the negative binomial model. It seems that the problem is not extra zeros but the existence of another group of policyholders. Therefore, assuming that we have two distinct subpopulations, we may move towards a finite mixture model.

In this case, using a 2-finite mixture of regression models, a large improvement was obtained by moving from one component Poisson to a 2-finite mixture of Poisson regression models. Note that this improvement is better than that obtained with the zero-inflated Poisson model. However, as the best fit is obtained by the 2FMNB, it seems that there is still some extra overdispersion which needs to be modelled appropriately, assuming within each component an overdispersed distribution like a negative binomial.

Figure 1 shows boxplots for the fitted mean values per component for both mixture models. We can observe that the group separation is not the same for the two models. Different models can have similar likelihoods but very different properties and potential. Comparing Poisson and negative binomial mixtures we see that they model different aspects and therefore, as they are close in likelihood terms, can focus on separate things. The first component for the Poisson model is more concentrated towards 0. The opposite is true for the second component. Clearly, 2FMNB fits the data better than the 2-finite mixture of Poisson regression models.

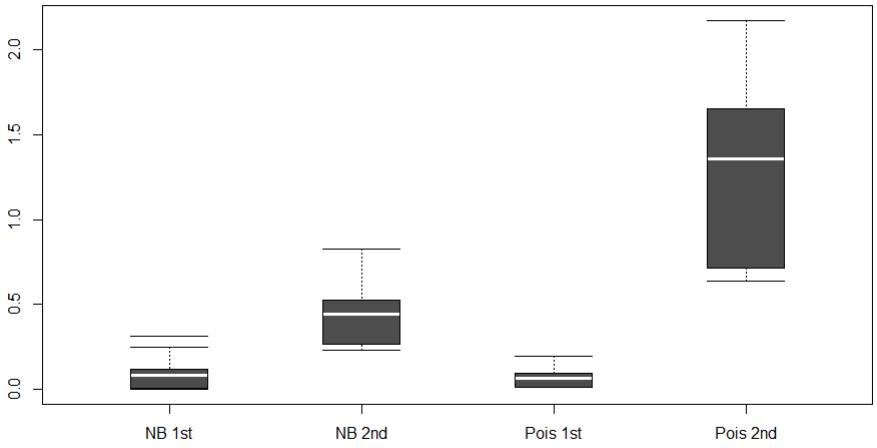

Figure 1. Boxplots of the fitted means for each of the two components for both models.

Figure 1 shows the distinct characteristics of the two assumed distributions. For the case of the Poisson distribution, as the variance is determined from the mean, we see that the two components are

further away as an attempt to model the excess of variance. Recall that the total variance is in fact the sum of the between variance (how much the components differ) and the within variance (inside each component). In contrast, in the negative binomial case, the two components are closer since the extra overdispersion parameter regulates the variability. This is also a warning that use of the Poisson model can lead to an erroneous inference of the mean for each component.

From Figure 1, we can also observe that the group separation is characterised by a low mean for the first component and a high mean with higher variance for the second. One may assume that this group separation is revealed by driving characteristics, such as driving ability, aggressiveness or degree of obeying traffic regulations, that are the source of the unobserved heterogeneity. In this case, we can consider those policyholders who belong to the first component to be "good" drivers, whereas policyholders in the second component can be considered "bad" drivers.

Table 3 summaries the results for the 2FMNB and the case with $k = 1$, i.e., no mixture. For the 2FMNB, we report the estimated regression coefficients for each component and p-values for testing the hypothesis that the variable is statistically significant. For the simple negative binomial model, we report the coefficient and standard p-values based on the Wald test.

Table 3. The fitted models for both the negative binomial and the 2FMNB. The p-value for the 2FMNB refers to that of LRT when the variables is removed from both components, whereas for the simple negative binomial it refers to the Wald test.

	2FMNB				**Negative Binomial**	
	1st comp.	2nd comp.	p-value		Estimate	p-value
Intercept	−6.1420	−1.1364	< 0.0001	Intercept	−2.4144	< 0.0001
GEN	0.2633	0.0086	0.0124	GEN	0.0774	0.0103
URB	0.3407	−0.0762	0.0017	URB	0.0165	0.4870
ZON	0.3745	0.0564	< 0.0001	ZON	0.1324	< 0.0001
LIC	0.2413	−0.2423	0.0448	LIC	−0.1610	0.0230
LOY	0.3707	0.1289	< 0.0001	LOY	0.2019	< 0.0001
COV	3.1438	0.6373	< 0.0001	COV	1.0024	< 0.0001
POW	0.2502	0.1148	< 0.0001	POW	0.1440	< 0.0001
ϕ	0.2321	0.6051		ϕ	0.2527	
π	0.6686	0.3314				

For the 2FMNB model, to assess the significance of the variables, we calculated a Likelihood Ratio Test (LRT) statistic. Note that this, as a variable selection problem, is not standard, as each covariate appears in both components. Also note that even since standard errors can be derived through the Hessian, such a procedure can be very unstable. Also bootstrap based standard errors can be very time-consuming. Therefore, to see the importance of the covariates, we maximised the log-likelihood with and without each variables and we obtained the LRT, compared with a χ^2 distribution with 2 degrees of freedom. Also, note that covariates that perhaps were not significant for a simple model (no mixture) can be significant in the mixture model, as the two components can allow for separate effects, which are lost when combining to one model.

Comparing the models from Table 3, we can see some interesting points. First, coefficient estimates of the negative binomial model are in some way a linear combination of coefficient estimates of each component of the 2FMNB. Second, in the negative binomial model, only *URB* is not significant at a level of 95%, whereas in the 2FMNB all covariates are significant at that level: the *URB* variable that is deemed significant for the mixture is not significant in the simple model. The reason is that they have opposite signs in the mixture, and therefore when we estimate one coefficient for the simple case, the effect is cancelled out: we estimate some average effect which is close to zero and has large variance. This implies that the mixture can more clearly reflect the importance of the variables and the existence of two groups

of policyholders that behave in different ways with regard to these a priori covariates. Finally, focusing on the dispersion parameters of the negative binomial distribution for each component, we can conclude that the second component presents larger dispersion than the first.

To summarize, we may assume two groups of policyholders with different regression structures. This is is particularly noticeable for the variable *URB*. For policyholders considered to be "good" drivers, driving in an urban area increases the probability of making a claim; whereas it decreases for "bad" drivers. This is reasonable: "good" drivers who make a claim are more likely to make it when driving in an urban area, and it will probably just be a small claim. In contrast, "bad" drivers are less likely to make a claim in an urban area since with their driving behaviour (more aggressive and ignoring traffic rules) they are more likely to make a claim when driving outside the urban areas.

3.3. Usage of FM Models for Actuarial Purposes

In this section, we aim to show the advantages and limitations of using a finite mixture of regression models with respect to other models, such as compound frequency models and their zero-inflated or hurdle versions.

First, with respect to compound frequency models, 2-finite mixture models account for unobserved heterogeneity more effectively, providing a better fit. 2FMNB separates the policyholders in two groups allowing for better classification and thus providing a better picture for managerial matters. Assuming that this group separation is caused by their driving capabilities or behaviour, we may consider that we have a group of "good" drivers and another of "bad" drivers.

Second, with respect to zero-inflated and hurdle models, the problem of unobserved heterogeneity is addressed in a more parsimonious way, trying to fix two issues at the same time: overdispersion and an excess of zeros. Zero-inflated and hurdle models focus on the excess of zeros and only implicitly correct for overdispersion; while finite mixture models do both explicitly. Also, note that the interpretation offered by finite mixtures is more reasonable: zero-inflation implies that some drivers will never have an accident, whereas finite mixture models say that there is still some small probability that good drivers will have an accident, this sounds more reasonable in practice from the actuarial point of view (see the discussion in Lord et al. (2007) on the usage of zero-inflated models for car accidents).

Third, the regression structure for each component provided by the 2-finite mixture models is very different from the single score given by compound frequency models and their zero-inflated or hurdle versions. This supports the aforementioned idea that the data consist of two subpopulations, "caused" by, or as a result of, the unobserved factors, for which the regression structure, used to account for the observed factors, is different. The 2-finite mixture models produce a wider picture of the portfolio and therefore offer better chances for accurate risk analysis. As mentioned above, the 2-finite mixture models enable us to see the importance of the variables more clearly. Significant variables according to the LRT test used here, with opposite signs in the mixture, may not be significant for the simple models as they only estimate one coefficient and the effect is cancelled out, estimating some average effect which is close to zero.

However, finite mixture models presents a limitation that impedes their effective use for ratemaking purposes. Although the 2-finite mixture models proposed here separate the policyholders into two types of drivers, they do not allow us to know the type of driver a particular new policyholder is. In other words, for a new customer, although we can estimate different premiums for each component, i.e., for "good" drivers and for "bad" drivers, we cannot find out in which category the new driver belongs, unless we have already observed the number of claims they have made, which is useless. This is because the model is a regression-type model and one needs to observe both the response (number of claims) together with the covariates in order to calculate the posterior probability. Also note that the mixing proportions, π_j, do not offer information on this since they refer to a randomly selected client without taking into account their characteristics. One solution might be to move some of the covariates to the mixing proportions. Therefore, for each new driver, we can have an estimate on the component that they belong to and use this to calculate their premium.

To evaluate the usefulness of finite mixture models, the differences between the 2FMNB and its respective regression model with one component (negative binomial) are analysed through the mean (a priori pure premium) and the variance (necessary for a priori loaded premium) of the number of claims per year for some profiles of the insured parties. Five different, yet representative, profiles were selected from the portfolio and classified according to their risk level. The profiles can be seen in Table 4. We selected the profiles so as to have different increasing means. The first can be classified as the best profile since it presents the lowest mean score. The second was chosen from among the profiles considered as good drivers, with a lower mean value than the mean of the portfolio. The third profile was chosen with a mean score lying very close to the mean of the portfolio. Finally, a profile considered as being for a bad driver (with a mean score above the mean of the portfolio) and the worst driver profile were selected.

Table 4. The 5 profiles used for the comparisons.

Profile Name	GEN	URB	ZON	LIC	LOY	COV	POW
Best	0	1	0	1	0	0	0
Good	1	1	0	0	0	0	1
Average	0	0	0	1	0	1	0
Bad	1	1	0	0	0	1	1
Worst	1	1	1	0	1	1	1

Table 5 shows the results for the five profiles for the two models with respect to the mean and the variance. For the finite mixture model, we have used the same mixing proportion ($\pi = (0.6686, 0.3314)$) for all profiles when we calculate the total mean (2FMNB) from the mean for each component (2FMNB-1 and 2FMNB-2). With respect to the mean, one can see that 2FMNB coincides to a great extend with the negative binomial model. However, we observe larger differences between the means for each component. The group of "good" drivers is far below the group of "bad" drivers. From a practical point of view, the means for each component can be considered as a lower bound and upper bound of the negative binomial means.

Meanwhile, the variance for 2FMNB is greater than for the negative binomial model for all the profiles. As we have commented, finite mixture models allow for unobserved heterogeneity more efficiently. In the same way as mentioned above for means, we see major differences between the variances for each component. Thus, "bad" drivers present greater dispersion than "good" drivers.

Table 5. The mean and the variance derived from the simple negative binomial model (NB) and the 2FMNB.

Profile	Mean				Variance			
	NB	2FMNB	2FMNB-1	2FMNB-2	NB	2FMNB	2FMNB-1	2FMNB-2
Best	0.077	0.080	0.004	0.233	0.101	0.183	0.004	0.323
Good	0.113	0.115	0.005	0.336	0.164	0.279	0.005	0.524
Average	0.207	0.200	0.063	0.476	0.378	0.496	0.081	0.852
Bad	0.309	0.289	0.117	0.636	0.688	0.756	0.176	1.306
Worst	0.432	0.419	0.247	0.766	1.170	1.159	0.509	1.735

Following the traditional two-step methodology, finite mixture models may open up the opportunity to evaluate the extent of a posteriori ratemaking. Bonus-malus systems are usually applied to account for the unobserved heterogeneity. In a posteriori ratemaking, actuaries consider the past claims record of each policyholder in order to update their a priori premiums, assuming that the number of claims

reported by policyholders reveals unobservable risk characteristics. In this context, the mean for the first component (2FMNB-1) can be seen as the limit of the a posteriori premium with bonuses. In this assumption, we consider the group of "good" drivers as the policyholders that do not report a claim in many years. In contrast, the mean for the second component would be the limit of the a posteriori premiums with maluses.

In summary, on the basis of the 2FMNB outcome, we can conclude that the use, for ratemaking purposes, of a negative binomial model, together with a bonus-malus system to account for the unobserved heterogeneity, has at least two limitations. First, after an a priori premium is obtained with a negative binomial model, we need to take many years with no claims to reach the level of 2FMNB means for the group of "good" drivers. Second, in the mean time, we may fail to account for the effect of the a priori variables because we assume that all drivers, "good" and "bad", behave in the same way with respect to these a priori variables.

4. Conclusions

In this paper, we propose the use of a 2-finite mixture of Poisson and negative binomial regression models to allow for the overdispersion and the excess of zeros usually detected in a car insurance dataset and commonly explained by the presence of unobserved heterogeneity. Assuming the existence of two types of clients, described separately by each component in the mixture, improves the modelling of the dataset. The idea is that the data consist of two subpopulations for which the regression structures are different.

These models are applied to a car insurance claims dataset in order to analyse whether the problem of unobserved heterogeneity requires richer structure for risk classification compared with the classical models used to allow for such a feature of the data, i.e., compound frequency models and their zero-inflated versions. From this application, we conclude the following.

First, our results show that this portfolio is comprised of two groups of policyholders or drivers. According to their driving habits or behaviour, such as driving ability, aggressiveness and degree of or obeying traffic regulations, the first group, characterised by a very low mean, can be considered the group of policyholders who are "good" drivers. In contrast, the second group, defined by a high mean with higher variance, can be considered as the group of policyholders who are "bad" drivers.

Second, the two groups of policyholders exhibits different regression structures, i.e., they behave in different ways with regard to the a priori factors. This is is highlighted particularly for the variable related to driving in an urban area or not: for policyholders considered "good" drivers, driving in a urban area increases the probability of having a claim, whereas it decreases for "bad" drivers. Furthermore, simpler models, such as a negative binomial model, fail to reflect the importance of the variables, and therefore lead to an inadequate risk classification.

Third, the two groups of policyholders have very different expected claim frequency values. When using the usual two-step ratemaking procedure, to prevent an adverse selection process, and assuming that the number of claims reported by policyholders reveals their unobservable risk characteristics, a bonus-malus system is considered to update the a priori premiums obtained with a compound frequency model. However, in this case, we would need many years without observing claims from a certain policyholder to reach the premium level provided by the 2FMNB for the group of "good" drivers.

To avoid the aforementioned limitations, we highly recommend the use of telematics devices for ratemaking purposes (see Guillén et al. (2019)). Vehicle telematics allows driving habit information to be collected that will dramatically reduce the unobserved heterogeneity caused by driving habits behavioural variation. Combining traditional a priori rating factors with the new information obtained telemetrically would make it unnecessary to use a time-consuming bonus-malus system and, simultaneously, it will lead to a more efficient risk classification. In other words, including this new information in the a priori ratemaking would allow us to differentiate between "good" and "bad" drivers from the beginning, without the need of a posteriori adjustment and taking into account the importance of all the rating factors more clearly.

Finally, although the 2-finite mixture models proposed here separate the policyholders into two types of drivers, they do not allow us to know the type of driver a particular policyholder is. This could be achieved in different ways, i.e. taking into account the past claim record of each individual or introducing covariates into the mixing probabilities of the mixtures. This may be the goal for future research.

Author Contributions: Conceptualisation, L.B., D.K. and I.M.; methodology, L.B and D.K.; software, D.K.; validation, L.B., D.K. and I.M.; formal analysis, L.B and D.K.; investigation, L.B and D.K.; resources, I.M.; data curation, L.B and I.M.; writing–original draft preparation, L.B., D.K. and I.M.; writing–review and editing,L.B. and I.M.; visualisation, D.K. and I.M.; supervision, L.B and D.K.; project administration, L.B.; funding acquisition, L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Spanish Ministry of Economy [grant number ECO2015-66314-R].

Acknowledgments: The authors wish to acknowledge discussions with researchers at Riskcenter at the University of Barcelona and the constructive comments made by two anonymous referees, which helped to improve the quality of the paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Aitkin, Murray. 1999. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* 55: 117–28.

Bermúdez, Lluís. 2009. A priori ratemaking using bivariate Poisson regression models. *Insurance: Mathematics and Economics* 44: 135–41.

Bermúdez, Lluís, and Dimitris Karlis. 2012. A finite mixture of bivariate Poisson regression models with an application to insurance ratemaking. *Computational Statistics & Data Analysis* 56: 3988–99.

Bolancé, Catalina, Montserrat Guillén, and Jean Pinquet. 2003. Time-varying credibility for frequency risk models: Estimation and tests for autoregressive specifications on the random effects. *Insurance: Mathematics and Economics* 33: 273–82.

Bolancé, Catalina, Montserrat Guillén, and Jean Pinquet. 2008. On the link between credibility and frequency premium. *Insurance: Mathematics and Economics* 43: 209–13.

Boucher, Jean-Philippe, and Michel Denuit. 2008. Credibility premiums for the zero inflated Poisson model and new hunger for bonus interpretation. *Insurance: Mathematics and Economics* 42: 727–35.

Boucher, Jean-Philippe, Michel Denuit, and Montserrat Guillén. 2007. Risk classification for claim counts: A comparative analysis of various zero-inflated mixed Poisson and hurdle models. *North American Actuarial Journal* 11: 110–31.

Boucher, Jean-Philippe, Michel Denuit, and Montserrat Guillén. 2009. Number of accidents or number of claims? an approach with zero-inflated Poisson models for panel data. *Journal of Risk and Insurance* 76: 821–46.

Breslow, Norman E. 1984. Extra-Poisson variation in log-linear models. *Applied Statistics* 33: 38–44.

Brouhns, Natacha, Montserrat Guillén, Michael Denuit, and Jean Pinquet. 2003. Bonus-malus scales in segmented tariffs with stochastic migration between segments. *Journal of Risk and Insurance* 70: 577–99.

Byung-Jung, Park, Dominique Lord, and Chungwon Lee. 2014. Finite mixture modeling for vehicle crash data with application to hotspot identification. *Accident Analysis & Prevention* 71: 319–26.

Dean, Charmaine, Jerald Lawless, and Gordon Willmot. 1989. A mixed Poisson-inverse-gaussian regression model. *Canadian Journal of Statistics* 17: 171–81.

Denuit, Michael, Xavier Marechal, Sandra Pitrebois, and Jean-François Walhin. 2007. *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems.* New York: Wiley.

Dionne, George, and Charles Vanasse. 1989. A generalization of actuarial automobile insurance rating models: the negative binomial distribution with a regression component. *ASTIN Bulletin* 19: 199–212.

Dionne, George, and Charles Vanasse. 1992. Automobile insurance ratemaking in the presence of asymmetrical information. *Journal of Applied Econometrics* 7: 149–65.

Grun, Bettina, and Friedrich Leisch. 2007. Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics and Data Analysis* 51: 5247–52.

Grun, Bettina, and Friedrich Leisch. 2008. Flexmix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software* 28:1–35. doi:10.18637/jss.v028.i04.

Guillén, Montserrat, Jens Perch Nielsen, Mercedes Ayuso, and Ana Pérez-Marín. 2019. The use of telematics devices to improve automobile insurance rates. *Risk Analysis* 39: 662–72.

Hennig, Christian. 2000. Identifiablity of models for clusterwise linear regression. *Journal of Classification* 17: 273–96.

Lambert, Diane. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34: 1–14.

Lord, Dominique, Simon Washington, and John N. Ivan. 2007. Further notes on the application of zero-inflated models in highway safety. *Accident Analysis & Prevention* 39: 53–57.

Mullahy, John. 1986. Specification and testing of some modified count data models. *Journal of Econometrics* 33: 341–65.

Papastamoulis, Panagiotis, Marie-Laure Martin-Magniette, and Cathy Maugis-Rabusseau. 2016. On the estimation of mixtures of Poisson regression models with large number of components. *Computational Statistics & Data Analysis* 93: 97–106.

Park, Byung-Jung, and Dominique Lord. 2009. Application of finite mixture models for vehicle crash data analysis. *Accident Analysis and Prevention* 41: 683–91.

Pinquet, Jean, Montserrat Guillén, and Catalina Bolancé. 2001. Long-range contagion in automobile insurance data: Estimation and implications for experience rating. *ASTIN Bulletin* 31: 337–48.

Wang, Peiming, Martin L. Puterman, Iain Cockburn, and Nhu Le. 1996. Mixed Poisson regression models with covariate dependent rates. *Biometrics* 52: 381–400.

Winkelmann, Rainer. 2008. *Econometric Analysis of Count Data, 4th edition*. New York: Springer.

Zou, Yajie, Yunlong Zhang, and Dominique Lord. 2013. Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis. *Accident Analysis & Prevention* 50: 1042–51.