FIXED VERSUS RANDOM EFFECTS IN POISSON REGRESSION MODELS FOR CLAIM COUNTS: A CASE STUDY WITH MOTOR INSURANCE

BY

JEAN-PHILIPPE BOUCHER AND MICHEL DENUIT

ABSTRACT

This paper examines the validity of some stylized statements that can be found in the actuarial literature about random effects models. Specifically, the actual meaning of the estimated parameters and the nature of the residual heterogeneity are discussed. A numerical illustration performed on a Belgian motor third party liability portfolio supports this discussion.

KEYWORDS

Poisson mixture, risk classification, claim frequencies, random effects, fixed effects.

1. INTRODUCTION

1.1. Motivation

Risk classification techniques for claim counts have been the topic of many papers appeared in the actuarial literature. DIONNE & VANASSE (1989, 1992) used a Negative Binomial regression model, while DEAN, LAWLESS & WILLMOT (1989) used a Poisson-Inverse Gaussian distribution to fit the number of claims. Recently, GOURIEROUX & JASIAK (2004) introduced the integer valued auto-regressive (INAR) model with unobserved heterogeneity. DENUIT & LANG (2004) used Generalised Additive Models. YIP & YAU (2005) presented several parametric Zero-Inflated count distributions, and BOUCHER, DENUIT & GUIL-LEN (2006) resorted to Hurdle Models.

The mixed Poisson distribution is often used to account for unkwown characteristics of the driver, influencing the number of accidents reported to the company. When panel data are available, these hidden features can alternatively be captured by an individual heterogeneity term that is constant over time (the standard reference for panel data is HSIAO (2003); the particular case of count variables is treated in CAMERON & TRIVEDI (1998)). This paper aims to confront the two approaches with emphasis on the actual meaning of the estimated

parameters in a mixed Poisson regression when random effects and covariates are correlated. In such a case, parameters estimates should be seen has the apparent effects of the covariates on the frequency. Keeping this in mind allows for a better understanding of the resulting price list.

1.2. Agenda

Two kinds of models can be used with longitudinal data: the fixed effects model and the random effects model. Both are briefly described in Section 2. In random effects models, three kinds of heterogeneity distributions are considered in this paper: Gamma, Inverse-Gaussian and Log-Normal. In Section 3, following the work of MUNDLAK (1978), we link the fixed effects model to the random effects model by a regression on the individual heterogeneity terms. We show that the combination of the fixed effects model with this regression, gives approximately the same results as the random effects model. The final Section 4 concludes.

1.3. Description of the Data

In this paper, we work with a Belgian motor third party liability insurance portfolio comprising 9,894 policies followed for a period of 3 consecutive years (from 1997 to 1999). Thus, we work with 29,682 observations. For each contract, we have informations about the annual number of claims together with some characteristics of the insured: sex of the driver (man or woman), age of the driver (divided in 3 classes: 17-22, 23-30 and more than 30), power of the vehicle (less than 66kW or more than 66kW) and the size of the city where the insured was living (big, medium or small, based on the number of residents). Figure 1.1 describes the observed annual claim frequency and the distribution of the policyholders according to their characteristics.

2. PANEL DATA MODELS

2.1. Presentation

Our portfolio is composed of $N = 9,884$ policyholders. Each policyholder i is observed during $T = 3$ periods. Let $N_{i,t}$ be the number of reported claims for insured i during year t. Such data are called longitudinal data (or panel data). They consist of repeated observations of individual units that are followed over time. Each individual is assumed to be independent of the others but correlation between observations relating to the same individual is permitted. Here, we assume that the number of claims per year obeys to a Poisson distribution with a parameter specific to each policyholder. Specifically, $N_{i,t}$ is assumed to be Poisson distributed with mean $\theta_i \lambda_{i,t}$, $i = 1, ..., N$, $t = 1, ..., T$. The

Distribution of the portfolio by characteristics of the insured

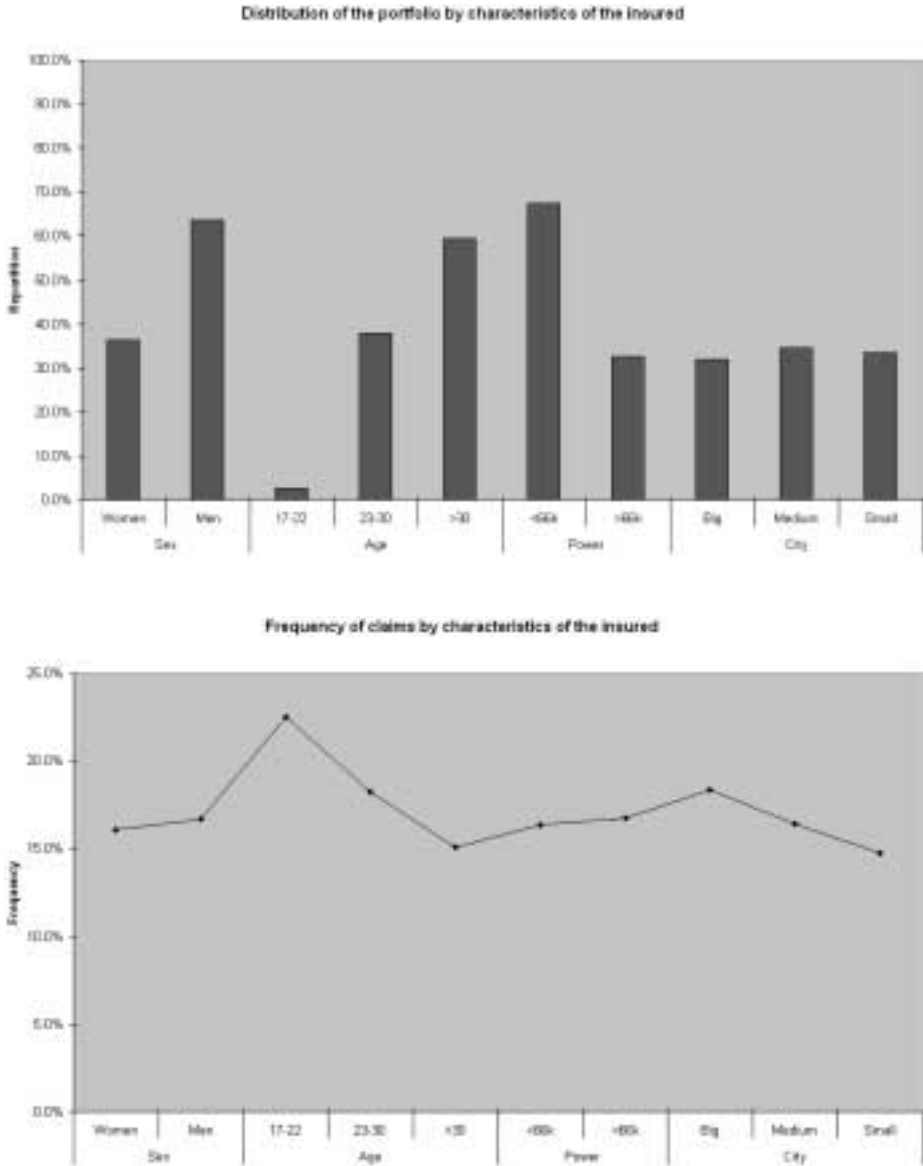Frequency of claims by characteristics of the insured

Figure 1.1. Description of the motor third party liability portfolio.

expected annual claim frequency is a product $\theta_i \lambda_{i,t}$ of a static factor θ_i times a dynamic factor $\lambda_{i,t}$. The former accounts for the dependence between observations relating to the same insured. The latter introduces the observable characteristics (that are allowed to vary in time). In general, $\ln \lambda_{i,t}$ is expressed as a linear combination of the observable characteristics, that is $\lambda_{i,t} = \exp(\beta_0 + \boldsymbol{\beta}' \boldsymbol{x}_{i,t})$,

where β_0 is the intercept, $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_p)$ is a vector of regression parameters for explanatory variables $\boldsymbol{x}_{i,t} = (x_{i,t,1}, \ldots, x_{i,t,p})'$.

There are two standard ways of dealing with panel data. In the random effects model, the heterogeneity parameter is treated as a random variable θ_i^{RE} with unit mean. At the portfolio level, the θ_i^{RE}'s are assumed to be independent and identically distributed. In the fixed effects model, the heterogeneity parameter θ_i^{FE} is treated as a parameter to be estimated for each individual. In this case, no intercept enters the model (to ensure identifiability). Conceptually, these two models are quite different. While the fixed effects model makes inferences conditional on the effects present in the sample, the random effects model draws conclusion for the population.

A major difference between the two approaches is that the fixed effects model provides only estimates for the parameters of time varying characteristics, since all the other parameters can been seen as part of the individual term θ_i^{FE}. Further explanations concerning differences between these two models are discussed in Section 3. A standard reference for linear models of panel data is HSIAO (2003), and a good review for count data is provided in CAMERON & TRIVEDI (1998).

2.2. Random Effects Model

2.2.1. Description

In the random effects model (henceforth, quantities relating to the random effects model will be indicated by the superscript RE), θ_i^{RE} is considered as a positive random variable, with probability density function $g(\cdot)$. Given θ_i^{RE}, the annual claim numbers $N_{i,1}, N_{i,2}, \ldots, N_{i,T}$ are independent. The joint probability function of $N_{i,1}, \ldots, N_{i,T}$ is thus given by

$$\Pr\left[N_{i,1} = n_{i,1}, \ldots, N_{i,T} = n_{i,T}\right]$$
$$= \int_0^\infty \left(\prod_{t=1}^{T} \exp\left(-\theta\lambda_{i,t}^{RE}\right) \frac{\left(\theta\lambda_{i,t}^{RE}\right)^{n_{i,t}}}{n_{i,t}!}\right) g(\theta)\, d\theta. \tag{2.1}$$

Estimation of parameters is performed using maximum likelihood estimators or moment techniques. The GEE method of LIANG & ZEGER (1986) can be a solution to account for the dependence between each observation of the same insured as shown in DENUIT, PITREBOIS & WALHIN (2003). However, for this paper, we will restrict ourselves to maximum likelihood estimates for all models analysed.

2.2.2. Poisson-Gamma Model

If θ_i^{RE} follows the Gamma distribution with mean 1 and variance $\frac{1}{\nu}$, the joint probability function of $N_{i,1}, \ldots, N_{i,T}$ writes

$$\Pr\left[N_{i,1} = n_{i,1}, ..., N_{i,T} = n_{i,T} \right]$$

$$= \left(\prod_{t=1}^{T} \frac{\left(\lambda_{i,t}^{RE} \right)^{n_{i,t}}}{n_{i,t}!} \right) \frac{\Gamma(n_{i,\cdot} + v)}{\Gamma(v)} \left(\frac{v}{\sum_{i=1}^{T} \lambda_{i,t}^{RE} + v} \right)^{v} \left(\sum_{i=1}^{T} \lambda_{i,t}^{RE} + v \right)^{-n_{i,\cdot}}, \qquad (2.2)$$

where $n_{i,\cdot} = \sum_{t=1}^{T} n_{i,t}$. In this case,

$$\mathbb{E}\left[N_{i,t} \right] = \lambda_{i,t}^{RE} < \mathbb{V}\left[N_{i,t} \right] = \lambda_{i,t}^{RE} + \left(\lambda_{i,t}^{RE} \right)^{2} / v.$$

Maximum likelihood estimations of parameters and variances can be obtained as follows. The first order conditions for parameters β_0^{RE}, $\boldsymbol{\beta}^{RE}$ and v lead to the system

$$\sum_{i=1}^{n} \sum_{t=1}^{T} n_{i,t} - \lambda_{i,t} \frac{\sum_t n_{i,t} + v}{\sum_t \lambda_{i,t} + v} = 0 \qquad (2.3)$$

$$\sum_{i=1}^{n} \sum_{t=1}^{T} x_{i,t} \left(n_{i,t} - \lambda_{i,t} \frac{\sum_t n_{i,t} + v}{\sum_t \lambda_{i,t} + v} \right) = \mathbf{0} \qquad (2.4)$$

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{n_{i,\cdot}-1} \frac{1}{j+v} \right) - \log\left(1 + \frac{\sum_t \lambda_{i,t}}{v} \right) + \sum_t \frac{\lambda_{i,t} + n_{i,t}}{\sum_t \lambda_{i,t} + v} = 0. \qquad (2.5)$$

Numerical procedures can then be used to solve these equations.

2.2.3. *Poisson-Inverse Gaussian Model*

The Inverse Gaussian distribution is another good candidate to model the heterogeneity parameter (WILLMOT (1987), DEAN, LAWLESS & WILLMOT (1989) and TREMBLAY (1992) for an application with insurance data). SHOUKRI ET AL. (2004) showed that the Poisson with Inverse-Gaussian heterogeneity of mean and variance equal to 1 and τ respectively, has the joint probability function

$$\Pr\left[N_{i,1} = n_{i,1}, ..., N_{i,T} = n_{i,T} \right]$$

$$= \left(\prod_{t=1}^{T} \frac{\left(\lambda_{i,t}^{RE} \right)^{n_{i,t}}}{n_{i,t}!} \right) \left(\frac{2}{\pi\tau} \right)^{0.5} e^{1/\tau} \left(1 + 2\tau \sum_{t=1}^{T} \lambda_{i,t}^{RE} \right)^{-s_i/2} K_{s_i}(z_i) \qquad (2.6)$$

where $K_j(.)$ is the modified Bessel function of the second kind, $s_i = n_{i,\cdot} - 0.5$ and

$$z_i = \frac{1}{\tau} \sqrt{1 + 2\tau \sum_{t=1}^{T} \lambda_{i,t}^{RE}}.$$

The modified Bessel function of the second kind has some useful properties that can be used to find the maximum likelihood estimators (for more details, see SHOUKRI ET AL. (2004)).

Now, $\mathbb{V}[N_{i,t}] = \lambda_{i,t}^{RE} + \tau(\lambda_{i,t}^{RE})^2$. The maximum likelihood estimators of β_0^{RE} and $\boldsymbol{\beta}^{RE}$ solve

$$\sum_{i=1}^{n} \sum_{t=1}^{T} n_{i,t} - \lambda_{i,t} M(n_{i,\cdot}) = 0 \tag{2.7}$$

$$\sum_{i=1}^{n} \sum_{t=1}^{T} \boldsymbol{x}_{i,t} (n_{i,t} - \lambda_{i,t} M(n_{i,\cdot})) = \boldsymbol{0}, \tag{2.8}$$

where the function $M(.)$ can be expressed as the ratio of the modified Bessel function of the second kind

$$\frac{K_{n_{i,\cdot}+1/2}(\alpha)}{K_{n_{i,\cdot}-1/2}(\alpha)} = M(n_{i,\cdot}) \sqrt{1 + 2\tau \sum_{t=1}^{T} \lambda_{i,t}^{RE}}.$$

The estimation of the parameter τ can be found by solving

$$-\frac{1}{2\tau} - \frac{1}{\tau^2} - \frac{s_i \sum_t \lambda_{i,t}}{1 + 2\tau \sum \lambda_{i,t}} + \frac{\partial \log K_{s_i}(z_i)}{\partial \tau} = 0 \tag{2.9}$$

where the derivative of the function K is equal to:

$$\frac{\partial \log K_{s_i}(z_i)}{\partial z_i} = -M(n_{i,\cdot}) \sqrt{1 + 2\tau \sum_t \lambda_{i,t}} + \frac{(n_{i,\cdot} - \frac{1}{2})\tau}{\sqrt{1 + 2\tau \sum_t \lambda_{i,t}}} \tag{2.10}$$

$$\frac{\partial z_i}{\partial \tau} = \frac{\sum_t \lambda_{i,t}}{\tau \sqrt{1 + 2\tau \sum_t \lambda_{i,t}}} - \frac{\sqrt{1 + 2\tau \sum_t \lambda_{i,t}}}{\tau^2}. \tag{2.11}$$

Again, numerical procedures are needed to obtain the solutions.

2.2.4. *Poisson-Log Normal Model*

In biostatistical circles, the Poisson Log-Normal model is often used, after HINDE (1982). In this case, the error term can be expressed as $\theta_i^{RE} = \exp(\varepsilon_i)$ for some Gaussian noise ε_i. From this, the mean parameter has the form $\exp(\boldsymbol{x}_{i,t}' \beta^{RE} + \varepsilon_i) = \gamma_{i,t}$, with ε_i following a Gaussian distribution with mean $\mu = -\sigma^2/2$ (a simplifying normalization) and variance σ^2. In this case,

TABLE 2.1.

PARAMETER ESTIMATES $\hat{\beta}^{RE}$ OF THE POISSON RANDOM EFFECTS MODELS.

Variable	Parameter	Gamma	Inv. Gaussian	Log Normal
Intercept	–	–2.0059 (0.034)	–2.0063 (0.034)	–2.0064 (0.034)
Age	17-22 23-30 > 30	0.4009 (0.088) 0.1964 (0.034) 0	0.4030 (0.088) 0.1983 (0.034) 0	0.4034 (0.088) 0.1988 (0.034) 0
City	Big Medium Small	0.2337 (0.041) 0.1106 (0.041) 0	0.2345 (0.041) 0.1100 (0.041) 0	0.2346 (0.041) 0.1098 (0.041) 0
v τ σ^2	– – –	2.6638 (0.314) – –	– 0.3940 (0.048) –	– – 0.3363 (0.036)
Log-Likelihood	–	–12,937.60	–12,936.30	–12,936.02

$$\Pr\left[N_{i,1} = n_{i,1}, ..., N_{i,T} = n_{i,T} \right]$$
$$= \int_{-\infty}^{\infty} \left(\prod_{t=1}^{T} \frac{e^{-\gamma_{i,t}} (\gamma_{i,t})^{n_{i,t}}}{n_{i,t}!} \right) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2} \right) d\varepsilon_i. \tag{2.12}$$

Numerical procedures can be used to find maximum likelihood estimates. Routines are now available in standard statistical packages, such as SAS (with the NLMIXED procedure). Now, $\mathbb{E}\left[N_{i,t} \right] = \lambda_{i,t}^{RE}$ and $\mathbb{V}\left[N_{i,t} \right] = \lambda_{i,t}^{RE} + \left(e^{\sigma^2} - 1 \right)\left(\lambda_{i,t}^{RE} \right)^2$.

The Poisson-Log Normal model is interesting because it has a natural interpretation (see, e.g., WINKELMANN (2003)). The error term ε_i is often considered as a factor that captures the effects of hidden exogenous variables. If there are many hidden variables, and if these variables are independent, then central limit theorems can be invoked in order to establish the normality of ε_i.

2.2.5. *Numerical Example*

Estimations of the parameters for the 3 random effects models are displayed in Table 2.1. Sex of the driver and power of the car have been removed from all models since they were not statistically significant, with respective p-values of approximately 0.38 and 0.12 for all models (specifically, p-values of 0.3841 and 0.1163 in the Poisson-Gamma model, 0.3861 and 0.1205 in the Poisson-Inverse Gaussian model, 0.3843 and 0.1205 in the Poisson-LogNormal model).

However, other variables have considerable impact such as the age of the driver or the size of the city where the insured lives. From Table 2.1, we see that young drivers and policyholders living in big cities exhibit higher expected claim frequencies. All models seem to have approximately the same quality of fit since their log-likelihood are almost equal for the same number of parameters.

2.3. Fixed Effects Model

2.3.1. *Description*

In the fixed effects model, all characteristics that are not time-varying are captured by the individual heterogeneity term θ_i^{FE}. In our case, the intercept β_0 has to be removed (and is included in θ_i^{FE}). As sex of the driver has been removed from the model, the remaining explanatory variables do vary with time and enter the fixed effects model.

The Poisson fixed effects model has been proposed by PALMGREN (1981) and HAUSMAN ET AL. (1984). The standard way of evaluating the parameters of this model is the conditional maximum likelihood of ANDERSEN (1970). The idea of the conditional method is to obtain an estimator of $\boldsymbol{\beta}^{FE}$ without having to estimate each θ_i^{FE}.

As proved in CAMERON & TRIVEDI (1998), the maximum likelihood and conditional maximum likelihood estimation methods always yield identical estimates for covariates parameter $\boldsymbol{\beta}^{FE}$ in case of Poisson distribution. Specifically, the estimated parameters solve

$$\sum_{i=1}^{n} \sum_{t=1}^{T} \boldsymbol{x}_{i,t} \left(n_{i,t} - \lambda_{i,t}^{FE} \frac{\sum_t n_{i,t}}{\sum_t \lambda_{i,t}^{FE}} \right) = \boldsymbol{0}. \tag{2.13}$$

Note that only insureds with varying caracteristics (and at least one claim) are used in the estimation of $\boldsymbol{\beta}^{FE}$. Estimates of θ_i^{FE} can then be obtained from

$$\theta_i^{FE} = \frac{\sum_t n_{i,t}}{\sum_t \lambda_{i,t}^{FE}}. \tag{2.14}$$

Both estimates of $\boldsymbol{\beta}^{FE}$ and θ_i^{FE} are consistent when $T \to \infty$ and $N \to \infty$, but only the estimate of $\boldsymbol{\beta}^{FE}$ is consistent for fixed T and $N \to \infty$, as for insurance data.

2.3.2. *Numerical Example*

Application of the fixed effects Poisson model to the Belgian motor portfolio leads to the parameters estimates of $\boldsymbol{\beta}^{FE}$ displayed in Table 2.2. Some interesting conclusions can be drawn from the fixed effects model. The results are quite different from those obtained with the random effects models. Indeed, young

TABLE 2.2.

PARAMETER ESTIMATES $\hat{\boldsymbol{\beta}}^{FE}$ OF THE POISSON FIXED EFFECTS MODEL.

Variable	Parameter	Estimate (Std. err.)
Age	17-22	0.4754 (0.205)
	23-30	–0.1603 (0.126)
	> 30	0
City	Big	–0.7945 (0.523)
	Medium	–0.3418 (0.530)
	Small	0
Log-Likelihood		–7803.577

drivers can be seen as better drivers than older ones and insureds coming from big cities now seem to be better drivers. It is worth to stress that the estimated parameters in Table 2.2 have to be thought in the sense of fixed effects model, where all individual impacts are removed.

The true effect of young age and living in large cities is thus to decrease the annual expected claim frequency. The apparent higher risk that is often found to be associated with these characteristics in empirical studies then results from their association with dangerous individual characteristics.

The high values of the standard errors suggest that almost all parameter are not statistically different from zero. This is nevertheless not a problem here, since our aim is to show that a fixed effects model followed by a regression of the θ_i^{FE}'s on the observable characteristics produces almost the same results than a random effects model.

2.4. Fixed or Random Effects Model?

Comparaison between equations (2.13) and (2.4) or (2.8) leads to the conclusion that there are no differences between fixed or random effects when T is large enough. However, for fixed small T, parameters estimates for these two models can be significantly different (as it can be seen from Tables 2.1-2.2).

A reason for such a difference between parameters estimates comes from the construction of the random effects model. Indeed, in the development of equation (2.1), we have made the following crucial assumption: the random effects θ_i^{RE} are independent and identically distributed. This means that the conditional probability density function of θ_i^{RE} given $x_{i,t}$ equals $g(\cdot)$, that does not depend on $x_{i,t}$. If the distribution of θ_i^{RE} depends on the $x_{i,t}$'s, the parameters estimates $\hat{\boldsymbol{\beta}}^{RE}$ may be inconsistent and should not be used, as shown by MUNDLAK (1978).

Figure 2.1 shows us the distribution of the θ_i^{FE} (mean and 0 to 95th percentile) according to the characteristics of the insured. Clearly, we can see that the

distribution of the heterogeneity term varies with the characteristics of the
insured. The dispersion is much more important in the least dangerous classes,
like Age 17-22 and Big cities. On average, the θ_i^{FE}'s are larger there, which
explains the apparent riskiness in Table 2.1. The heterogeneity is not identically
distributed, which can cause an inconsistency in the evaluation of the para-
meters $\boldsymbol{\beta}^{RE}$ of the random effects model.

Application of Hausman test to our data leads, without surprise, to the
rejection of the null hypothesis of uncorrelation between regressors and random
effects (p-value of less than 0.01% for a chi-square distribution with 4 degrees
of freedom). This result means that the heterogeneity term is not identically
distributed accross insureds.

The fixed effects model is preferred in cases where conclusions have to be
made on the sample, while the interests of random effects model are on the
overall population. For insurance data, fixed effects model cannot be used
since annual premiums cannot be calculated for new policyholders. Expected
claim frequencies can be computed only for policyholders in the portfolio for
several years. Moreover, estimates for the θ_i^{FE}'s based on just a few observa-
tions must be considered with caution. Additionnaly, fixed effects models are
difficultly handled by insurance companies since too many individual effects
had to be considered. Random effects models should then be preferred, but,
as we mentioned, some theorical aspects prohibit its use since estimates of the
parameters are biased when heterogeneity is not independent from the regres-
sors. The purpose of the next section is to legitimate the use of the random
effects models for insurance ratemaking, provided parameters are estimated
and interpreted with care.

3. REGRESSION OF THE θ_i^{FE}'S ON THE OBSERVABLE CHARACTERISTICS

3.1. Residual Heterogeneity

As noted by PINQUET (2000), the random effects are often correlated with
covariates for insurance data. Therefore, they relate to some residual hetero-
geneity, that is, orthogonal to the observable variables. The aim of this section
is to demonstrate on a basis of the Belgian data set that a fixed effect Poisson
regression followed with a regression of the resulting θ_i^{FE}'s on the observable
characteristics gives almost the same values for the regression coefficients than
a random effect Poisson regression. This legitimates the use of random effects
techniques in actuarial science. We also consider mixed effects model with non
identically distributed θ_i^{RE}, and we reach a similar conclusion.

Specifically, let us now explain the $\hat{\theta}_i^{FE}$'s as a function of covariates. To this
end, we consider the $\hat{\theta}_i^{FE}$'s as realizations from some probability density function
$h(\mu_i, \tau)$ with mean $\mu_i = \exp(\hat{\boldsymbol{x}}_i' \boldsymbol{\delta})$ and variance τ. The term $\hat{\boldsymbol{x}}_i$ is an adaptation
of covariates $\boldsymbol{x}_{i,t}$ for $t = 1, ..., T$, since there is only one indivual heterogeneity
term for all periods t (here, we took $\hat{\boldsymbol{x}}_i = \boldsymbol{x}_{i,2}$). We will take for h the Gamma,

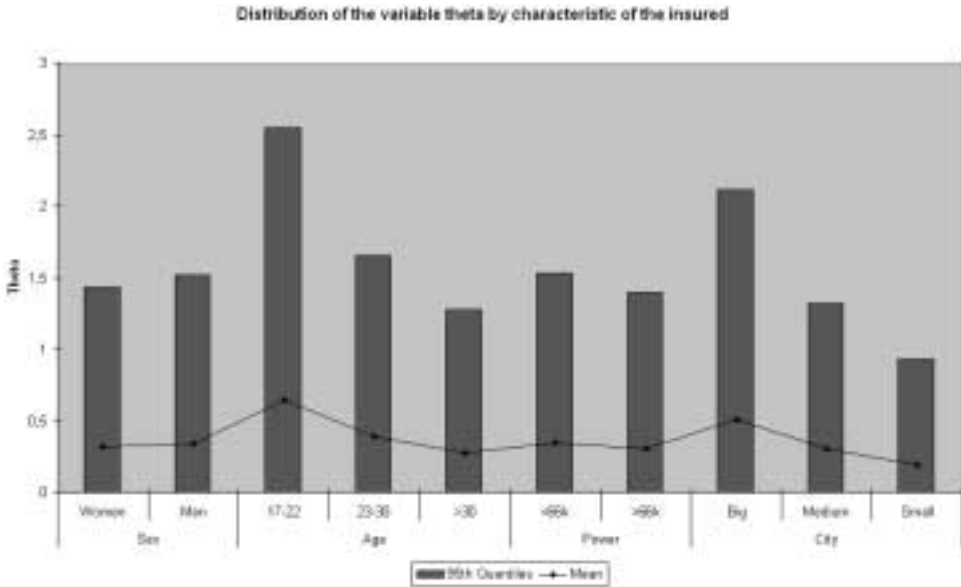Distribution of the variable theta by characteristic of the insured

Fig. 2.1. Distribution of the θ_i^{FE}'s according to the observable characteristics.

Inverse Gaussian and LogNormal densities (also considered in the random effects model).

We saw on Figure 2.1 that the distribution of the θ_i^{FE}'s was influenced by the observable characteristics. Another way of explaining the random effects model (second approximation) is to consider non-equally distributed hetero-geneity. In consequence, instead of assuming that the θ_i^{RE}'s are independent and identically distributed with mean 1 and variance τ, we use another model. The evaluation is done with known λ_i^{FE} and we assume that the θ_i^{FE}'s have mean $\mu_i = \exp(\hat{x}_i' \delta)$ and variance τ_i. Formally, we now consider that

$$
\Pr\left[N_{i,1} = n_{i,1}, ..., N_{i,T} = n_{i,T} \right]
$$
$$
= \int_0^\infty \left(\prod_{t=1}^T \frac{\left(\lambda_{i,t}^{FE} \theta\right)^{n_{i,t}} \exp\left(-\lambda_{i,t}^{FE} \theta\right)}{n_{i,t}!} \right) f(\theta \,|\, \mu_i, \tau_i) d\theta, \tag{3.1}
$$

where $f(\cdot \,|\, \mu_i, \tau_i)$ denotes some probability density function with mean $\mu_i = \exp(\hat{x}_i' \delta)$ and variance τ_i. Again, we consider for f the Gamma, Inverse Gaussian and LogNormal densities. If the two-step procedure based on the θ_i^{FE}'s, or the one-step procedure based on heterogeneous θ_i^{RE}, is coherent with the random effect model, we expect that

$$
\hat{\beta}^{RE} \approx \hat{\beta}^{FE} + \hat{\delta}. \tag{3.2}
$$

As proved by MUNDLAK (1978), the fact that there exists a correlation between θ_i^{RE} and the covariates causes a bias on the estimation of the $\boldsymbol{\beta}^{RE}$. However, despite the presence of this bias, it is still possible to use the parameter estimates for the premium calculation if the heterogeneity term is only considered as residual heterogeneity. In consequence, these estimates represent the apparent effects on the frequency of claims and not the real effect, since we must use $(\hat{\boldsymbol{\beta}}^{FE} + \hat{\boldsymbol{\delta}})$ instead of $\hat{\boldsymbol{\beta}}^{FE}$. For insurance ratemaking, this distinction does not really matter since apparent effect is the interest when some important classification variables are missing.

Remark 3.1. The equation used to construct a regression analysis on the individual fixed effects is based on equation (2.14). The individual fixed effects are expressed as the ratio of the sum of the number of claims in the T years on the sum of the $\lambda_{i,t}^{FE}$ for the same period. Consequently, there is a significant presence of zero value for the θ_i^{FE}, which causes a problem for a regression analysis. In consequence, weighted regression seems appropriate for modelling the individual fixed effects. However, weighted observations without claim need to be removed from the dataset since it is not possible to work with a zero-valued observation in a regression analysis of heterogeneity. Still because of the weighted regression, it becomes impossible to have the same dispersion parameter for random effects models and its first approximation, since some random variations are removed by using average values.

3.2. Gamma Heterogeneity

Let us begin with Gamma distributed θ_i^{FE}'s. Since the Gamma distribution is a member of the exponential family, the estimation of $\boldsymbol{\delta}$ is based on GLM (see MCCULLAGH & NELDER (1989)) and solves

$$\sum_{i=1}^{n} \frac{1}{\omega_i} \left(\theta_i^{FE} - \mu_i \right) \frac{\partial \mu_i}{\partial \boldsymbol{\delta}} = \sum_{i=1}^{n} \frac{v}{\mu_i} \left(\theta_i^{FE} - \mu_i \right) \hat{\boldsymbol{x}}_i$$

$$= \sum_{i=1}^{n} \frac{v}{\mu_i} \left(\frac{\sum_t n_{i,t}}{\sum_t \lambda_{i,t}^{FE}} - \mu_i \right) \hat{\boldsymbol{x}}_i \qquad (3.3)$$

$$= \sum_{i=1}^{n} \frac{v}{\sum_t \lambda_{i,t}^{FE} \mu_i} \left(\sum_t n_{i,t} - \sum_t \lambda_{i,t}^{FE} \mu_i \right) \hat{\boldsymbol{x}}_i = \boldsymbol{0}$$

where $\omega_i = \mu_i^2/v$ is the variance of the distribution and μ_i is the mean function that is expressed as $\exp(\hat{\boldsymbol{x}}_i' \boldsymbol{\delta})$. Since $N_{i,t}$ is a discrete variable, the resulting first order condition has the form of a Gamma distribution for $\sum_t n_{i,t}$, with mean $\sum_t \lambda_{i,t}^{FE} \mu_i$ and variance proportional to the square of the mean.

   If we allow for non identically distributed random effect, we get the following contribution for policyholder i to the likelihood in the random effects model:

TABLE 3.1.

PARAMETER ESTIMATES FOR GAMMA HETEROGENEITY

Variable	Parameter	RE-FE Difference	1st model	2nd model
Intercept	–	–2.0128 (0.034)	–1.9971 (0.044)	–2.0062 (0.034)
Age	17-22	0.8763 (0.038)	0.9016 (0.132)	0.8625 (0.085)
	23-30	0.3567 (0.018)	0.3339 (0.047)	0.3533 (0.034)
	> 30	0	0	0
City	Big	1.0282 (0.234)	1.0223 (0.056)	1.0369 (0.041)
	Medium	0.4524 (0.166)	0.425 (0.055)	0.4570 (0.041)
	Small	0	0	0
Dispersion	v	2.6638 (0.314)	0.1975 (0.065)	2.6437 (0.310)

$$\Pr\left[N_{i,1} = n_{i,1}, ..., N_{i,T} = n_{i,T}\right]$$

$$= \int_0^\infty \left(\prod_{t=1}^T \frac{\left(\lambda_{i,t}^{FE}\theta\right)^{n_{i,t}}\exp\left(-\lambda_{i,t}^{FE}\theta\right)}{n_{i,t}!}\right)\left(\frac{\theta v}{\mu_i}\right)^v (\theta)^{-1}\frac{\exp\left(\frac{-\theta v}{\mu_i}\right)}{\Gamma(v)}\,d\theta \qquad (3.4)$$

$$= \left(\prod_{t=1}^T \frac{\left(\lambda_{i,t}^{FE}\mu_i\right)^{n_{i,t}}}{n_{i,t}!}\right)\frac{v^v}{\Gamma(v)}\frac{\Gamma(n_{i,\cdot} + v)}{\left(\mu_i\sum_{t=1}^T \lambda_{i,t}^{FE} + v\right)^{n_{i,\cdot} + v}}.$$

Only two observations have been removed from the weighted regression. The parameters of the first model have been found by a two-step procedure. Firstly, the individual heterogeneity parameters θ_i^{FE} have been evaluated using equation (2.14). Afterwards, a weighted Gamma regression has been performed to find estimates of δ. For the second model, equation (3.4) has been applied with $\lambda_{i,t}^{FE}$ known from application of the fixed effects model.

We can see in Table 3.1 that the parameter estimates of δ for the first and second models are approximately equal to the difference between the random and the fixed effects model.

3.3. Inverse Gaussian Heterogeneity

Once again, instead of assuming that the heterogeneity term is independant of the covariates, we suppose that the heterogeneity term has Inverse Gaussian distribution with mean μ_i and a corresponding variance ω_i that is equal to $\mu_i^3\tau$. Inverse Gaussian distribution can also be estimated using first-order condition of the GLM (see MCCULLAGH & NELDER (1989)).

The estimator of $\boldsymbol{\delta}$ then solves

$$\sum_{i=1}^{n} \frac{1}{\omega_i} \left(\theta_i^{FE} - \mu_i\right) \frac{\partial \mu_i}{\partial \boldsymbol{\delta}} = \sum_{i=1}^{n} \frac{\tau}{\mu_i^2} \left(\theta_i^{FE} - \mu_i\right) \hat{\boldsymbol{x}}_i$$

$$= \sum_{i=1}^{n} \frac{\tau}{\mu_i^2} \left(\frac{\sum_t n_{i,t}}{\sum_t \lambda_{i,t}^{FE}} - \mu_i\right) \hat{\boldsymbol{x}}_i \qquad (3.5)$$

$$= \sum_{i=1}^{n} \frac{\tau}{\sum_t \lambda_{i,t}^{FE} \mu_i^2} \left(\sum_t n_{i,t} - \sum_t \lambda_{i,t}^{FE} \mu_i\right) \hat{\boldsymbol{x}}_i = \boldsymbol{0}.$$

If we allow for random effects with different means, the contribution of policy-holder i to the likelihood is

$$\Pr\left[N_{i,1} = n_{i,1}, \ldots, N_{i,T} = n_{i,T}\right]$$

$$= \int_0^\infty \left(\prod_{t=1}^{T} \frac{\left(\lambda_{i,t}^{FE} \theta\right)^{n_{i,t}} \exp\left(-\lambda_{i,t}^{FE} \theta\right)}{n_{i,t}!}\right) \left(2\pi\tau\theta^3\right)^{-1/2} \exp\left(\frac{-(\theta - \mu_i)^2}{2\mu_i^2 \theta\tau}\right) d\theta \qquad (3.6)$$

$$= \left(\prod_{t=1}^{T} \frac{\left(\lambda_{i,t}^{FE} \mu_i\right)^{n_{i,t}}}{n_{i,t}!}\right) \left(\frac{2}{\pi\tau\mu_i}\right)^{0.5} \exp\left(\frac{1}{\tau\mu_i}\right) \left(1 + 2\left(\tau\mu_i\right) \sum_{t=1}^{T} \lambda_{i,t} \mu_i\right)^{-s_i/2} K_{s_i}(z_i)$$

where $K_j(.)$ is the modified Bessel function of the second kind, $s_i = n_{i,\bullet} - 0.5$ and

$$z_i = \frac{1}{\tau} \sqrt{1 + 2\left(\tau\mu_i\right) \sum_{t=1}^{T} \lambda_{i,t}^{FE} \mu_i}.$$

Table 3.2 shows the results of these models on our insurance data. As expected, the results are quite the same as those obtained with the Gamma heterogeneity models. Even the intercept of the first approximation model has approximately the same value, due to the presence of high values in the data.

One difference between models is the parameter τ. This difference comes from the fact that $\tau_i^{RE_1} = \tau^{RE_2} \mu_i$, where RE_1 and RE_2 are the first and second approximation models. Since the value of μ_i is 0.3326, we can link the two estimated parameters τ by the following approximation : $0.3940 \approx 1.090 \times 0.3326 = 0.3625$.

3.4. Log-Normal Heterogeneity

Let us now assume that $\theta_i^{FE} = \exp(\varepsilon_i)$, where ε_i follows the Gaussian distribution with mean $\mu_i = \boldsymbol{x}_i'\boldsymbol{\delta} - \frac{\sigma^2}{2}$ and variance $\omega_i = \sigma^2$. Obviously, the Gaussian distribution can be solved using first-order condition of the GLM models (see McCullagh & Nelder (1989)). In consequence, the estimation of $\boldsymbol{\delta}$ solves

TABLE 3.2.

PARAMETER ESTIMATES FOR INVERSE GAUSSIAN HETEROGENEITY

Variable	Parameter	RE-FE Difference	1st model	2nd model
Intercept	–	–2.0127 (0.034)	–1.9888 (0.033)	–2.0098 (0.033)
Age	17-22	0.8784 (0.037)	0.8758 (0.169)	0.8431 (0.090)
	23-30	0.3585 (0.018)	0.3101 (0.045)	0.3376 (0.034)
	> 30	0	0	0
City	Big	1.0289 (0.234)	1.0097 (0.059)	1.0271 (0.041)
	Medium	0.4517 (0.166)	0.4456 (0.047)	0.4603 (0.040)
	Small	0	0	0
Dispersion	τ	0.3940 (0.048)	4.2234 (0.704)	1.090 (0.158)

$$\sum_{i=1}^{n} \frac{1}{\omega_i} \left(\log\left(\theta_i^{FE}\right) - \log(\mu_i) \right) \frac{\partial \mu_i}{\partial \boldsymbol{\delta}} = \sum_{i=1}^{n} \frac{1}{\sigma^2} \left(\log\left(\theta_i^{FE}\right) - \log(\mu_i) \right) \hat{\boldsymbol{x}}_i$$

$$= \sum_{i=1}^{n} \frac{1}{\sigma^2} \left(\log\left(\sum_t n_{i,t}\right) - \log\left(\sum_t \lambda_{i,t}^{FE}\right) - \log(\mu_i) \right) \hat{\boldsymbol{x}}_i \quad (3.7)$$

$$= \sum_{i=1}^{n} \frac{1}{\sigma^2} \left(\log\left(\sum_t n_{i,t}\right) - \log\left(\sum_t \lambda_{i,t}^{FE} \mu_i\right) \right) \hat{\boldsymbol{x}}_i = \boldsymbol{0}.$$

Allowing for non identically distributed random effects, the joint probability density function of the claim numbers for insured i is

$$\Pr\left[N_{i,1} = n_{i,1}, ..., N_{i,T} = n_{i,T} \right]$$

$$= \int_0^\infty \left(\prod_{t=1}^{T} \frac{e^{-\gamma_i} \gamma_i^{n_i}}{n_{i,t}!} \right) \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(\varepsilon_i - \mu_i)^2}{2\sigma^2} \right) d\varepsilon_i \quad (3.8)$$

with $\gamma_i = \exp(\boldsymbol{x}_i' \boldsymbol{\beta} + \varepsilon_i)$. Closed form of the Poisson-Log-Normal distribution is not possible. However, by numerical approximations or by the NLMIXED procedure in SAS, it is possible to find estimators of $\boldsymbol{\delta}$.

The numerical results are displayed in Table 3.3. Approximation of a weighted regression gives interesting results that are similar to those obtain with the Gamma and the Inverse-Gaussian distributions.

4. CONCLUDING REMARKS

The results obtained in this paper give us the legitimity to use random effects models even if there exists a correlation between the regressors and

TABLE 3.3.

PARAMETER ESTIMATES FOR LOG-NORMAL HETEROGENEITY

Variable	Parameter	RE-FE Difference	1st model	2nd model
Intercept	–	–2.0129 (0.034)	–2.0037 (0.044)	–2.0210 (0.034)
Age	17-22 23-30 > 30	0.8788 (0.037) 0.3591 (0.018) 0	0.8793 (0.132) 0.3295 (0.047) 0	0.8668 (0.086) 0.3560 (0.034) 0
City	Big Medium Small	1.0292 (0.234) 0.4515 (0.166) 0	1.0209 (0.056) 0.4539 (0.055) 0	1.0381 (0.041) 0.4562 (0.041) 0
Dispersion	σ^2	0.3363 (0.036)	2.2640 (0.3773)	0.3422 (0.036)

the heterogeneity. The parameter estimates do not identify the impact of these regressors on the premium but only the apparent effects. Since it is usually the interest of the actuary in ratemaking, there is no problem with this interpretation. However, such a correlation indicates clearly that a correction should be done to obtain a more accurate model. Especially, the apparent high risk of young drivers should deserve some attention. The analysis conducted in this paper shows that the fixed effects are very heterogeneous for these individuals. Instead of penalizing these insureds in the a priori ratemaking, an appropriate bonus-malus scheme could be designed. Merit rating systems improve the fairness of the tariff in that respect.

We have focused our attention on the Poisson distribution since it is commonly used in practice for risk classification. However, other count distributions, such as the Negative Binomial distribution (see HAUSMAN, HALL & GRILICHES (1984) or ALLISON & WATERMAN (2002) for an alternative fixed effects model) or the Zero-Inflated models also deserve consideration for the analysis of claim frequencies. A similar analysis could be performed for these regression models, too. Moreover, continuous probability models, such as the Gamma or the Log-Normal distributions are routinely used to model the amount of claims. Again, a study in the vein of the present one could be performed for claim severities.

References

ALLISON, P.D. and WATERMAN, R. (2002) Fixed-effects Negative Binomial regression models. *Sociological Methodology*, **32**, 247-265.

ANDERSEN, E.B. (1970) Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society B*, **32**, 283-301.

BOUCHER, J.-P., DENUIT, M. and GUILLEN, M. (2006) Risk classification for claim counts: A comparative analysis of various zero-inflated Mixed Poisson and Hurdle models. *Working Paper*, Institut des Sciences Actuarielles, Université Catholique de Louvain.

CAMERON, A.C. and TRIVEDI, P.K. (1998) *Regression Analysis of Count Data*. Cambridge University Press.

DEAN, C., LAWLESS, J.F. and WILLMOT, G.E. (1989) A mixed Poisson-Inverse Gaussian regression model. *Canadian Journal of Statistics*, **17**, 171-182.

DENUIT, M. and LANG, S. (2004) Nonlife ratemaking with Bayesian GAM's. *Insurance: Mathematics and Economics*, **35**, 627-647.

DENUIT, M., PITREBOIS, S. and WALHIN, J.-F. (2003) Tarification automobile sur données de panel. *Bulletin of the Swiss Association of Actuaries*, 51-81.

DIONNE, G. and VANASSE, C. (1989) A generalization of automobile insurance rating models: The Negative Binomial distribution with regression component. *ASTIN Bulletin*, **19**, 199-212.

DIONNE, G. and VANASSE, C. (1992) Automobile insurance ratemaking in the presence of asymetrical information. *Journal of Applied Econometrics*, **7**, 149-165.

GOURIEROUX, C. and JASIAK, J. (2004) Heterogeneous INAR(1) model with application to car insurance. *Insurance: Mathematics and Economics*, **34**, 177-192.

HAUSMAN, J.A., HALL, B.H. and GRILICHES, Z. (1984) Econometric models for count data with application to the Patents-R&D Relationship. *Econometrica*, **52**, 909-938.

HINDE, J. (1982) Compound Poisson regression models. In R. Gilchrist, ed., *GLIM 82: Proceeding of the International Conference on Generalised Linear Models*, New York, Springer-Verlag.

HSIAO, C. (2003) *Analysis of Panel Data*. Cambridge University Press, Cambridge.

LIANG, K.Y. and ZEGER, S.L. Longitutinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

MCCULLAGH, P. and NELDER, J.A. (1989) *Generalized Linear Models*. Chapman and Hall, London.

MUNDLAK, Y. (1978) On the pooling of time series and cross section data. *Econometrica*, **46**, 69-85.

PALMGREN, J. (1981) The Fisher information matrix for log-linear models arguing conditionnally in observed explanatory variables. *Biometrika*, **68**, 563-566.

PINQUET, J. (2000) Experience rating through heterogeneous models. In *Handbook of Insurance*, edited by G. Dionne. Kluwer Academic Publishers.

SHOUKRI, M.M., ASYALI, M.H., VANDORP, R. and KELTON, D. (2004) The Poisson Inverse Gaussian regression model in the analysis of clustered counts data. *Journal of Data Science*, **2**, 17-32.

TREMBLAY, L. (1992) Using the Poisson Inverse Gaussian in bonus-malus systems. *ASTIN Bulletin*, **22**, 97-106.

WILLMOT, G. (1987) The Poisson-Inverse Gaussian distribution as an alternative to the Negative Binomial. *Scandinavian Actuarial Journal*, 113-127.

WINKELMANN, R. (2003) *Econometric Analysis of Count Data*. Springer, New York.

YIP, K.C.H. and YAU, K.K.W. (2005) On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics*, **36**, 153-163.

JEAN-PHILIPPE BOUCHER AND MICHEL DENUIT
Institut des Sciences Actuarielles
Université Catholique de Louvain
6 rue des Wallons
B-1348 Louvain-la-Neuve
Belgium.