Fifth International Conference in Financial and Actuarial Mathematics

Fitting Generalised Linear Models to Car Claims data

Liberato Camilleri, Marianne Cassar
University of Malta,
Department of Statistics and Operations Research
liberato.camilleri@um.edu.mt

Abstract:
Generalised linear models (GLMs) overcome the limitations of Normal regression models since they can accommodate any distribution that is a member of the exponential family. These models allow transformation of the response variable through the canonical link function. This paper presents two GLMs to analyze a data set provided by a car insurance company. The first model is a lognormal regression model that relates claim size to a number of demographic, car-related and policy-related predictors and the second model is a Poisson regression model that relates the number of claims filed by a policy holder to these explanatory variables. An appropriate model that describes the aggregate claim amount in a portfolio of insurance contracts during a fixed period combines both claim size and number of claims through a compound Poisson distribution.

Key words:
Poisson and Lognormal distributions, Iterative reweighted least squares algorithm, Generalized Linear Models, Compound Poisson distribution

Introduction

One of the most far-reaching contributions in statistical modelling is the concept of generalized linear models introduced by John Nelder and Robert Wedderburn (1972). These models relate the response variable to the linear predictor (non-random component) through any invertible link function and accommodate any error distribution that is in the exponential family. Analyzing car claim data using traditional ordinary least squares regression models and ANOVA methods can be problematic. Firstly the distribution of car claim size is very often right skewed and do not follow a Normal distribution; secondly the number of claims made by a policyholder is a discrete variable and would be better accommodated by a discrete distribution. GLMs, on the other hand, provide an integrated conceptual and theoretical framework that can be used to analyze both continuous and categorical response data. Logistic and Probit regression models are appropriate to analyze Binomial response data; whereas Loglinear models are suitable to

analyze Multinomial and Poisson response data. The iteratively re-weighted least squares algorithm that maximizes the log-likelihood function in Generalized Linear models makes use of Fisher scoring. Although GLMs accommodate most of the assumptions of Regression models they still rely on the assumption that the responses are independent.

Estimation

The unity of several statistical methods to analyze response data that departs from the normality assumption was demonstrated by (Nelder and Wedderburn 1972) using the idea of a generalized linear model. This section provides an outline of the properties of GLMs as a comprehensive structure.

Consider a random variable Y_i whose probability mass function, if it is discrete, or probability density function, if it is continuous is assumed to follow the form of the exponential family of distributions.

$$P(Y_i = y_i) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right]$$

It is assumed that the responses Y_i are independent and identically distributed and the distribution of each Y_i is a member of the exponential family. Moreover the known values of the explanatory variables influence the distribution of Y_i through a single linear function or linear predictor η_i

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$$

It is also assumed that the mean $\mu_i = E(Y_i)$, and linear predictor η_i are related by a smooth invertible link function $g(\cdot)$.

$$\eta_i = g(\mu_i)$$

Considering the likelihood L as a function of $\boldsymbol{\beta}$, one can find the maximum of L by maximizing $\log L = \sum_{i=1}^{J} l_i$ where $l_i = \log P(Y_i = y_i)$. This is realized by solving the maximum likelihood equations $\frac{\partial l_i}{\partial \beta_j} = 0$ for $j = 1,..., p$

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

Since $l_i = \dfrac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)$, $\mu_i = E(Y_i) = b'(\theta_i)$ and $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$ then

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a_i(\phi)} = \frac{y_i - \mu_i}{a_i(\phi)}, \quad \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) \text{ and } \frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$$

Moreover, by setting $d_i = \dfrac{\partial \eta_i}{\partial \mu_i}$ then $\dfrac{\partial l_i}{\partial \beta_j} = \dfrac{y_i - \mu_i}{a_i(\phi)} \dfrac{1}{b''(\theta_i)} \dfrac{1}{d_i} x_{ij}$

Since $\sigma_i^2 = \text{var}(Y_i) = a_i(\phi) b''(\theta_i)$ then

$$\frac{\partial l_i}{\partial \beta_j} = \frac{y_i - \mu_i}{\sigma_i^2 d_i} x_{ij}$$

The log-likelihood function is maximized by solving the equations $\dfrac{\partial l}{\partial \beta_j} = 0$ for $j = 1, \ldots, p$. Let U_j be the scores with respect to parameters β_j such that

$$U_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{I} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{I} \frac{(y_i - \mu_i) x_{ij}}{\sigma_i^2 d_i}$$

In general the equations $U_j = 0$ for $j = 1, \ldots, p$ are non-linear and they have to be solved by numerical iteration. The Newton-Raphson approach to solving these equations would be to set up an iterative scheme for the vector $\boldsymbol{\beta}$. The m^{th} approximation is given by

$$\hat{\boldsymbol{\beta}}^{(m)} = \hat{\boldsymbol{\beta}}^{(m-1)} - \mathbf{H}^{-1}\mathbf{U}^{(m-1)}$$

where $\mathbf{H} = \left[\dfrac{\partial^2 l}{\partial \boldsymbol{\beta} \, \partial \boldsymbol{\beta}'} \right]$ is the Hessian matrix evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(m-1)}$ and the score vector $\mathbf{U}^{(m-1)}$ is also evaluated at the previous iteration. These second derivatives are often complicated to calculate. An alternative procedure, which is sometimes simpler than the Newton-Raphson method is called the Fisher scoring technique. It involves replacing the matrix of second derivatives by its matrix of expected values where

$$\Psi_{jk} = -E\left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right] = E\left[\frac{\partial l}{\partial \beta_j} \frac{\partial l}{\partial \beta_k} \right] = \sum_{i=1}^{I} \frac{x_{ij} x_{ik}}{\sigma_i^4 d_i^2} E\left(Y_i - \mu_i\right)^2$$

Since $E(Y_i - \mu_i)^2 = \text{var}(Y_i) = \sigma_i^2$ then $\Psi_{jk} = \sum_{i-1}^{I} \dfrac{x_{ij} x_{ik}}{\sigma_i^2 d_i^2}$

So $\boldsymbol{\Psi} = \left[\frac{\partial l}{\partial \boldsymbol{\beta}} \frac{\partial l}{\partial \boldsymbol{\beta}'}\right]$ can be expressed as $\mathbf{X}'\mathbf{W}\mathbf{X}$, where $\mathbf{W}$ is a diagonal matrix whose diagonal elements are $w_{ii} = 1/(\sigma_i^2 d_i^2)$ and $\hat{\boldsymbol{\beta}}^{(m)} = \hat{\boldsymbol{\beta}}^{(m-1)} + \left[\boldsymbol{\Psi}^{(m-1)}\right]^{-1}\mathbf{U}^{(m-1)}$. Multiplying throughout by $\boldsymbol{\Psi}^{(m-1)}$ we get $\boldsymbol{\Psi}^{(m-1)}\hat{\boldsymbol{\beta}}^{(m)} = \boldsymbol{\Psi}^{(m-1)}\hat{\boldsymbol{\beta}}^{(m-1)} + \mathbf{U}^{(m-1)}$. The right hand side of this iterative scheme can be written as:

$$\boldsymbol{\Psi}^{(m-1)}\hat{\boldsymbol{\beta}}^{(m-1)} + \mathbf{U}^{(m-1)} = \sum_i \frac{x_{ij}}{\sigma_i^2 d_i^2}\left[\sum_k x_{ik}\hat{\beta}_k^{(m-1)} + d_i(y_i - \mu_i)\right] = \mathbf{X}'\mathbf{W}\mathbf{z}$$

where $z_i = \sum_k x_{ik}\hat{\beta}_k^{(m-1)} + d_i(y_i - \mu_i) = \eta_i + d_i(y_i - \mu_i)$

$$(\mathbf{X}'\mathbf{W}\mathbf{X})\hat{\boldsymbol{\beta}}^{(m)} = \mathbf{X}'\mathbf{W}\mathbf{z} \text{ and } \hat{\boldsymbol{\beta}}^{(m)} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}$$

The generalized linear model maximum likelihood estimators are obtained by an iterative weighted least squares procedure.

Application

To implement the theory of GLMs we utilized a data set provided by a local insurance car company, to relate the number of claims filed annually by each policyholder and the claim size made by each claimant to a number of predictors. These explanatory variables included policy-related variables (cover subscription, premium paid annually by policyholder), car-related variables (number of owned cars, engine size) and individual covariates (age of policyholder). Premium paid annually and claim size are continuous variables; number of claims and number of owned cars are discrete variables; whereas cover subscription (Third party only, third party fire and theft, fully comprehensive), engine size (less than 1100, 1100-1500, more than 1500cc) and age of policy holder (18-30, 31-50, more than 50 years) are categorical variables.

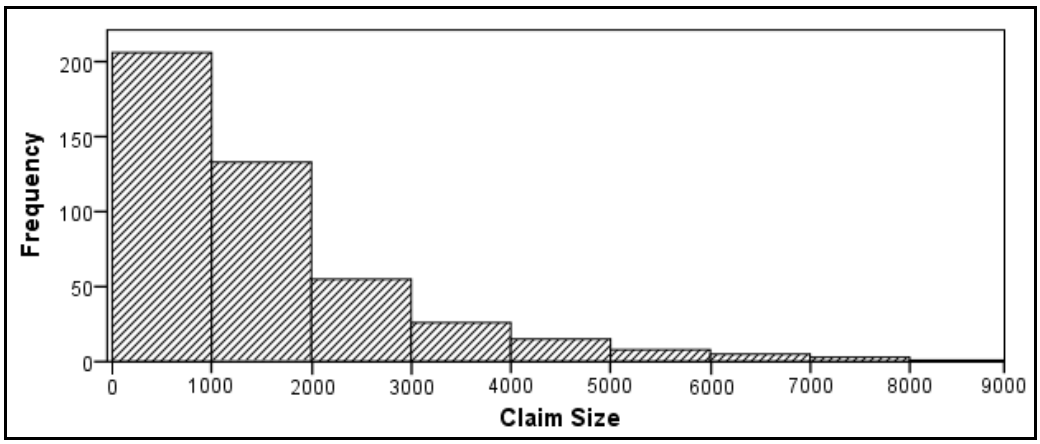

Figure 1: Frequency distribution of car claim size

The data comprised 9107 policyholders of which 497 made at least one claim. The frequency distribution of claim size, displayed in Figure 1, was considerably right skewed and fitting a Normal regression model to this data was not deemed appropriate. EasyFitXL was used to identify the best contender for the distribution of claim size using the Kolmogorov Smirnov, Anderson Darling and Chi squared criteria. Undoubtedly the Lognormal distribution is identified as the best fitting distribution for claim size.

Distribution	Kolmogorov Smirnov		Anderson Darling		Chi squared	
	Statistic	Rank	Statistic	Rank	Statistic	Rank
Lognormal	0.0231	1	0.2676	1	4.1677	2
Normal	0.1701	17	26.127	17	131.54	18

Table 1: Goodness of fit using Lognormal and Normal distributions

Figure 2 displays that 94.5% of policyholders made no car claims throughout a year, 5.2% made one claim and the remaining 0.3% made at least two claims. The Poisson distribution is appropriate for the number of claims made yearly by a policyholder because this random variable is discrete, does not have an obvious maximum and making a claim is considered to be a rare event.

Figure 2: Frequency distribution of number of car claims

Fitting a Lognormal regression model to car claim size

If the logarithm of the response variate $\log(y_i)$ has a Normal distribution with mean μ_i and variance σ^2 then y_i has a lognormal distribution. Assume that claim sizes y_i, $i = 1,...,n$ are independent and follow a lognormal distribution whose density function is:

$$f(y_i) = \frac{1}{y_i\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\log y_i - \mu_i)}{2\sigma^2}\right)^2 \quad \text{for } y_i > 0.$$

This can be expressed as:

$$\log f(y_i) = \left[\frac{\mu_i \log y_i - \frac{1}{2}\mu_i^2}{\varphi} - \left(\log(y_i) + \frac{1}{2}\log(2\pi\varphi) + \frac{(\log y_i)^2}{2\varphi} \right) \right]$$

The mean and variance of y_i are respectively

$$E(Y_i) = \exp\left(\mu_i + 0.5\sigma^2\right) \text{ and } \operatorname{Var}(Y_i) = \exp\left(2\mu_i + \sigma^2\right)\left[\exp\left(\sigma^2\right) - 1\right]$$

If an identity link function $(\eta_i = \mu_i)$ is assumed, implying that $d_i = \partial\eta_i / \partial\mu_i = 1$, one can compute the linear predictor η_i, the predicted values μ_i, the iterative weights $w_{ii} = 1/\sigma_i^2$ and the working variate $z_i = y_i$, at each iteration.

Predictor	Wald Chi-Square	df	P-value
Intercept	6957.628	1	0.000
Engine size	3.875	2	0.144
Age-Group	2.205	2	0.332
Cover	0.380	2	0.827
Premium paid annually	4.988	1	0.026

Table 2: Tests of model effects

The tests of model effects displayed in table 2 indicate that the premium paid annually (in thousands of Euro) is the best predictor of car claim size. This is followed by engine size, age of policyholder and cover subscription. This four-predictor model explain 15.7% of the total variance in the responses indicating that there are other important predictors that contribute significantly in explaining variation in car claim sizes.

Term	Parameter	Std. Error	95% Wald Confidence Interval	
			Lower	Upper
Intercept	6.898	0.113	6.676	7.120
Engine size (less than 1100cc)	-0.236	0.140	-0.510	0.038
Engine size (1100-1500cc)	0.033	0.099	-0.162	0.229
Engine size (More than 1500cc)	0	.	.	.
Age-Group (18-30 years)	0.254	0.171	-0.081	0.589
Age-Group (31-50 years)	0.047	0.097	-0.144	0.237
Age-Group (More than 50 years)	0	.	.	.
Cover (Third party only)	-0.064	0.110	-0.280	0.152
Cover (Third party fire and theft)	0.007	0.130	-0.247	0.262
Cover (Fully comprehensive	0	.	.	.
Premium	0.351	0.157	0.043	0.659
(Scale)	0.830	0.059	0.723	0.953

Table 3: Parameter estimates and corresponding 95% confidence intervals

The parameter estimates, displayed in table 3, reveal interesting contrasts between the levels of the main effects. Policyholders that pay large premiums tend to make bigger claims than other policyholders; young claimants tend to make bigger claims than older ones. Moreover, policyholders possessing small sized engine cars tend to make smaller claims than those possessing large sized engine cars and

claimants who insure their cars under a third party cover tend to make smaller claims than those who insure their cars under a third party fire and theft or a fully comprehensive cover. The residual plot displayed in Figure 3 exhibits no systematic patterns indicating no model misspecifications. Approximately 95% of all Pearson residuals lie between the ± 2 threshold values which conform to what is expected.

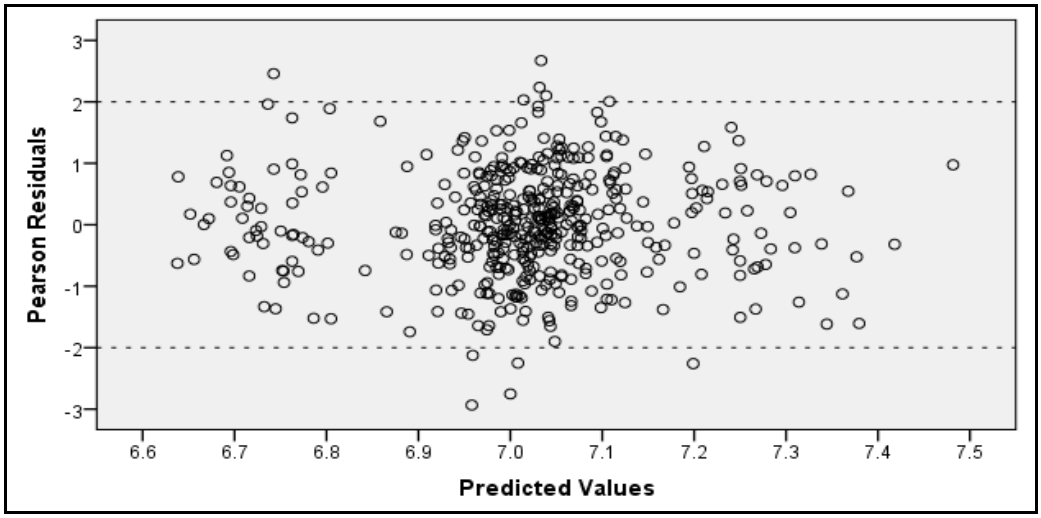

Figure 3: Residual plot

Fitting a Poisson regression model to number of claims

Suppose that the response variates $y_i,\ i = 1,...,n$ are independent and Poisson distributed $y_i \sim Poi(\mu_i)$ whose mass function is:

$$P(Y_i = y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$

The log link is the canonical link function for a random variable having a Poisson distribution $\eta_i = \log \mu_i$, implying that $d_i = \partial\eta_i/\partial\mu_i = 1/\mu_i$. Given the fact that the mean and variance of y_i are both μ_i, one can compute the linear predictor η_i, the predicted values μ_i, the working variate $z_i = \eta_i + (y_i - \mu_i)/\mu_i$ and the iterative weights $w_{ii} = \mu_i$ at each iteration.

Predictor	Wald Chi-Square	df	P-value
Intercept	2.050	1	0.152
Age-Group	3.094	2	0.213
Cover	47.627	2	0.000
Number of cars owned	15.198	1	0.000
Premium paid annually	6.558	1	0.010

Table 4: Tests of model effects

The tests of model effects displayed in table 4 indicate that cover subscription is the best predictor of number of car claims. This is followed by the number of cars owned by policyholder, the premium paid annually (in thousands of Euro) and the age of policyholder. This four-predictor model explain 29.6% of the total variance in the responses indicating that there are other important predictors that contribute significantly in explaining variation in the number of car claims made annually by a policyholder.

Term	Parameter	Std. Error	95% Wald Confidence Interval	
			Lower	Upper
Intercept	-0.486	0.5841	-1.631	0.659
Age-Group (18-30 years)	0.131	0.1657	-0.193	0.456
Age-Group (31-50 years)	0.169	0.0967	-0.020	0.358
Age-Group (More than 50 years)	0	.	.	.
Cover (Third party only)	-0.667	0.1321	-0.926	-0.408
Cover (Third party fire and theft)	-0.675	0.1129	-0.896	-0.454
Cover (Fully comprehensive	0	.	.	.
Number of cars owned	2.228	0.5716	1.108	3.349
Premium paid annually	0.001	0.0002	0.000	0.001
(Scale)	1			

Table 5: Parameter estimates and corresponding 95% confidence intervals

The parameter estimates, displayed in table 5, reveal interesting contrasts between the categories of the predictors. Policyholders that pay larger premiums tend to make bigger claims than other policyholders; old claimants tend to make fewer claims than younger ones. Moreover, policyholders who insure their cars under a Fully Comprehensive cover have a tendency to make more claims than others and the number of claims made annually increases with the number of cars owned by the policyholders. Approximately 95% of all Pearson residuals lie between the ± 2 threshold values which conform to what is expected.

The compound Poisson distribution

The decomposition of the aggregate claim amount *S* paid annually by the insurer allows consideration of the number of claims and corresponding claim amounts separately. A practical advantage of this is the factors affecting claim numbers and claim amounts may well be different. For instance, a prolonged spell of bad weather may have a significant effect on claim numbers but little or no effect on the distribution of individual claim amounts. On the other hand, inflation may have a significant effect on the cost of repairing cars, and hence on the distribution of individual claim amounts, but little or no effect on claim numbers.

If the random variable *N*, representing number of claims made by policyholders has a Poisson distribution with parameter λ and $X_1, X_2, ..., X_N$ are corresponding claim amounts which are assumed independent and identically distributed, then the total claim amount $S = \sum_{i=1}^{N} X_i$ has a compound Poisson distribution. This is

obtained by marginalizing the joint distribution of (S,N) over *N*, which in turn is attained by joining the marginal distribution of *N* with the conditional distribution $S|N$. If the number of claims *N* has a Poisson distribution with mean λ and the total claim amount *S* has a compound Poisson distribution with parameter λ then,

$$E[N]=Var[N]=\lambda$$

$$M_N(t)=E\left[e^{tN}\right]=\sum_{N=1}^{n}\frac{e^{tN}\lambda^N e^{-\lambda}}{N!}=e^{-\lambda}\sum_{N=1}^{n}\left\{\frac{\left(\lambda e^t\right)^N}{N!}\right\}=e^{-\lambda}e^{\lambda e^t}=e^{\lambda(e^t-1)}$$

The expectation of *S* is obtained by applying the identity $E[S]=E[E[S\,|\,N]]$

$$E[S\,|\,N=n]=\sum_{i=1}^{n}E[X_i]=nm_1 \text{ and } E[S\,|\,N]=Nm_1$$

$$E[S]=E[Nm_1]=E[N]m_1=\lambda m_1$$

The variance of *S* is obtained by using $Var[S]=E[Var[S\,|\,N]]+Var[E[S\,|\,N]]$

$$\text{Var}[S\,|\,N=n]=\sum_{i=1}^{n}Var[X_i]=n(m_2-m_1^2) \text{ and } Var[S\,|\,N]=N(m_2-m_1^2)$$

$$Var[S]=E[N(m_2-m_1^2)]+Var[Nm_1]=(m_2-m_1^2)E[N]+m_1^2Var[N]=\lambda m_2$$

The moment generating function of *S* is the moment generating function of *N* evaluated at $\log M_X(t)$.

$$M_S(t)=E\left[\left(M_X(t)\right)^N\right]=E\left[e^{\left(N\log M_X(t)\right)}\right]=M_N\left(\log M_X(t)\right) \text{ where } M_N(t)=e^{\lambda(e^t-1)}$$

$$M_S(t)=M_N\left(\log M_X(t)\right)=e^{\lambda\left(e^{\log M_X(t)}-1\right)}=e^{\lambda(M_X(t)-1)}$$

A very important property is that the sum of the independent compound Poisson random variables is itself a compound Poisson random variable. If $S_1,...,S_n$ are independent random variables each having a compound Poisson distribution with parameters λ_i and $F_i(x)$ for $i=1,...,n$ where $F_i(x)$ is the distribution of individual claim amounts then $A=\sum_{i=1}^{n}S_i$ also has a compound Poisson distribution with parameters $\Lambda=\sum_{i=1}^{n}\lambda_i$ $F(x)=\frac{1}{\Lambda}\sum_{i=1}^{n}\lambda_i F_i(x)$. Table 6 displays the distribution of claim size and the average number of car claims per age group and per cover.

Age Group	Distribution of claim size	Mean number of claims
18 – 30 years	$\ln N(1277.5,\ 0.829)$	$\lambda_1 = 0.05108$
31 – 50 years	$\ln N(1117.4,\ 0.963)$	$\lambda_2 = 0.06195$
More than 50 years	$\ln N(1039.1,\ 0.967)$	$\lambda_3 = 0.05334$

Cover subscription	Distribution of claim size	Mean number of claims
Third party only	$\ln N(1112.9,\ 0.895)$	$\lambda_1 = 0.03757$
Third party fire and theft	$\ln N(1123.6,\ 0.871)$	$\lambda_2 = 0.03907$
Fully comprehensive	$\ln N(1117.4,\ 0.942)$	$\lambda_3 = 0.08201$

Table 6: Distributions of claim sizes and mean number of claims

Conclusion

The GLMs identified one significant predictor (premium paid annually) for claim size and three significant predictors (cover subscription, number of cars owned, premium paid annually) for number of filed claims. One of the limitations of this study is that the explanatory variables explained a small portion of the variation in the response variables. Indeed other explanatory variables, for instance, speed of car before impact and driving behaviour would have improved predictions if they were recorded.

An alternative approach to address the data heterogeneity is by fitting Latent class models. These models assume that the observed data are actually composed of several homogeneous segments that are mixed together in unknown proportions. The segments are considered latent (unobserved) because the number of clusters and the number of individuals they comprise are unknown. The objective is to estimate the true number of segments and derive a prediction regression model for each segment using the expectation-maximization (EM) algorithm that maximizes the expected log-likelihood function. Indeed the main advantage of using these models over traditional clustering techniques is that estimation and segmentation are carried out simultaneously.

References:

1. M. Aitkin, D. Anderson, B. Francis, J. Hinde. Statistical Modelling in GLIM, Oxford Science Publications, 1994.
2. A. Dempster, N. Laird, D. Rubin. Maximum Likelihood from Incomplete Data via EM algorithm. Journal of the Royal Statistical Society, 39, 1-38, 1977.
3. P. McCullagh, J. Nelder. Generalized Linear Models, London: Chapman & Hall Publications, 1989.
4. J.A. Nelder, R.W.M. Wedderburn. Generalized Linear Models. Journal of the Royal Statistical Society, A, 135, 370-384, 1972.
5. A. Skrondal, S. Rabe-Hesketh. Generalized Latent Variable Modelling. Chapman & Hall/CRC, 2004.