In the name of God

E.C.O. College of Insurance
Allameh Tabatabaei University

Master Thesis

A Credibility Premium for the Zero-Inflated Poisson Models and New Hunger for Bonus Interpretation

In the Subject

Actuarial Science

By:

Batool Hasanzadeh

Supervisor: Dr. Amir Teimour Payandeh
Advisor: Dr. Reza Ofoghi

Tehran-Iran
August 2009

Abstract

The purpose of this thesis is to explore and compare the credibility premiums in generalized zero-inflated count models for panel data. Predictive premiums based on quadratic loss and exponential losses are derived. It is shown that the credibility premiums of the zero-inflated model allow for more flexibility in the prediction. Indeed, the future premiums not only depend on the number of past claims, but also on the number of insured periods with at least one claim.

The hunger for bonus is a well-known phenomenon in insurance, meaning that the insured does not report all of his accidents to save bonus on his next year's premium. However, actuaries and researchers continue to model the number of claims with standard count distributions, neglecting this phenomenon. The model also offers another way of analyzing the hunger for bonus phenomenon. The accident distribution is obtained from the zero-inflated distribution used to model the claims distribution, which can in turn be used to evaluate the impact of various credibility premiums on the reported accident distribution. This way of analyzing the claims data gives another point of view on the research conducted on the development of statistical models for predicting accidents. A numerical illustration supports this discussion and we consider the real claims data set of third party car insurance contracts of the Saman insurance company from 1384-1387 (Iranian Calendar System).

Key Words: Credibility; Count data; Quadratic loss; Exponential loss; Zero-inflated models; Number of accidents; Hunger for Bonus.

To My Family

Abstract

The purpose of this thesis is to explore and compare the credibility premiums in generalized zero-inflated count models for panel data. Predictive premiums based on quadratic loss and exponential losses are derived. It is shown that the credibility premiums of the zero-inflated model allow for more flexibility in the prediction. Indeed, the future premiums not only depend on the number of past claims, but also on the number of insured periods with at least one claim.

The hunger for bonus is a well-known phenomenon in insurance, meaning that the insured does not report all of his accidents to save bonus on his next year's premium. However, actuaries and researchers continue to model the number of claims with standard count distributions, neglecting this phenomenon. The model also offers another way of analyzing the hunger for bonus phenomenon. The accident distribution is obtained from the zero-inflated distribution used to model the claims distribution, which can in turn be used to evaluate the impact of various credibility premiums on the reported accident distribution. This way of analyzing the claims data gives another point of view on the research conducted on the development of statistical models for predicting accidents. A numerical illustration supports this discussion and we consider the real claims data set of third party car insurance contracts of the Saman insurance company from 1384-1387 (Iranian Calendar System).

Key Words: Credibility; Count data; Quadratic loss; Exponential loss; Zero-inflated models; Number of accidents; Hunger for Bonus.

Table of Contents

1. Introduction

1.1. Introduction

Because risk classification in insurance involves unobserved risk characteristics, Bayesian modelling offers an intellectually acceptable approach. Indeed, these characteristics are usually modelled by the introduction of a random effect in the classification process. Consequently, a posteriori analysis following claims experience is an interesting task because a Bayes revision of the heterogeneity component allows estimating more precisely these unobserved characteristics. At each insured period, the random effects can be updated for past claim experience, revealing some individual information.

Boucher et al. (2006) developed new models to fit the number of claims. Count data usually exhibit a great number of zeros than expected from the Poisson model. In this situation, the zero-inflated Poisson is commonly used. This is a mixture of Poisson and a degenerate distribution at zero. The models are generalizations of the zero-inflated Poisson distribution for panel data, where a random effect has been added to the model (for zero-inflated models applied to cross-section data, see Yip and Yau 2005). In this thesis, I explore the predictive premiums of one of these models. Predictive premiums are developed based on quadratic loss or on exponential loss (Ferreira, 1977; Denuit and Dhaene, 2001; Morillo and Bemùdez, 2003). I show that the credibility premiums of the zero-inflated model allow for more flexibility in the prediction. Indeed, the future premiums not only depend on the number of past claims, but also on the number of insured periods with at least one claim. I also use this zero-inflated model to give another way of analyzing the hunger for bonus situation, that is to say the possibility that the insured does not report all his accidents to save his bonus on his next year premium. I use the zero-inflated distribution to model the claims distribution, the accident distribution is deducted, which can in turn be used to evaluate the impact of various credibility premiums on the claims distribution. This way of analyzing the claims data gives another point of view on the research conducted on the development of statistical models for predicting accidents.

In this chapter, I address the relevance and importance of the thesis topic and will go through the literature review of the subject. In the second chapter, I will define in details different definition and concepts that I need for understanding the thesis. In the third chapter, I will introduce the zero-inflated Poisson models with random effects and derive useful results of credibility theory, more particularly the predictive premium for quadratic loss and exponential loss. Then, I will introduce a new way of analysing the hunger for bonus situation, where a distinction between the true accident frequency and the claims distribution is made. Finally, in the last chapter, chapter five, I will apply the results of the chapter three on experienced claims data of the Saman insurance company for the period 1384-1387. I will also present recommendations for further research. In Appendix A, I review Baysian Credibility Premium in order to make the readers familiar with this useful way of obtaining mean of predictive distribution. Also, Maximum Likelihood Estimators will appear in Appendix B. The equivalence of some words is given in appendix C.

1.2. Relevance and Importance of the Thesis Subject

Insurance companies need to know the actual number of accidents, but as philison (1960), Leimaire (1977) pointed out the hunger for bonus is a well known phenomenon that represents the fact that insured do not report all their accidents to save bonus on the following premium year’s. However, actuaries and researchers continue to model the number of claims with standard count distributions, neglecting the bonus hunger phenomenon. Hence, I assume that the number of accidents is based on a Poisson distribution but that number of claims is generated by censorship of this Poisson distributions.

Insurance data usually include a relatively large number of zeros (no claims). Deductible and no claims discounts (bonus) increase the proportion of zeros, since small claims are not reported by insured drivers. High number of zero values led to the idea that a distribution with excess zero can provide a good fit, such as zero inflated distribution.

1.3. Literature Review

Much of the literature on this thesis concentrated on the credibility premium and zero inflated models. Random effects have an important role in our work and our models consider data in a period of time, also I will use panel data.

1.3.1. Credibility Models

In actuarial parlance, the term credibility was originally attached to experience rating formulas that were convex combination (weighted average) of individual and class estimates of the individual risk premium. Credibility theory is the art of combining different collections of data to obtain an accurate overall estimate. It provides actuaries with techniques to determine insurance premiums for contracts that belongs to heterogeneous portfolio, where is limited or irregular claim experience for each contract but ample claim experience for the portfolio. Credibility theory can be seen one of quantitative tools that allows the insurers to perform experience rating, that is, to adjust future premiums based on past experiences. This technique has a long history in actuarial science, with fundamental contributions dating back to Mowbray (1914). Whitney (1918) introduced the intuitively appealing concept of using a weighted average of (1) average claims from the risk class and (2) average claims over all risk classes to predict future expected claims.

In part, credibility predictors succeed because they are known to be the best possible predictors in a broad variety of situations. Bühlmann (1967) described a fundamental model containing latent (unobserved) effects that are common to all claims from a risk class. Bühlmann called these “structure effects.” The “best” linear unbiased predictors that can be derived from this model turn out to be the same intuitively appealing linear credibility predictors described above. Bühlmann’s basic model formulation extends readily to encompass a large class of models for a review that is oriented towards linear regression and longitudinal data models (Frees, Young and Luo 1999).

To account for the entire distribution of claims, a common approach used in credibility is to adopt a Bayesian perspective. Keffer (1929) initially suggested using a Bayesian perspective for experience rating in the context of group life insurance. Subsequently, Bailey (1945, 1950) showed how to derive the linear credibility form from a Bayesian perspective as the mean of a predictive distribution.

In addition to the works cited above, I also note the work of Miller and Hickman (1975) and Pinquet (1997). Miller and Hickman (1975) examined credibility in the context of aggregate loss distributions. Pinquet (1997) was also interested in automobile claims; he considered collision claims arising from two lines, at fault and no fault coverage. Both of these papers assumed parametric distributions for the number of claims and amount distributions and used Bayesian techniques to develop estimators.

An excellent introduction to the credibility theory can be found, in Goovarerts and Hoogstad (1987), Herzog (1994), Dannenburg, Kass and Goovearts (1996), Klugman et al (2004, Chapter 16) and Bühlmann and Gisler (2005). See also Norberg (2004) for an overview with useful references and links to statistics and linear estimation. The underlying assumption of credibility theory which sets it apart from formulas based on the individual risk alone is that the risk parameter is regarded as a random variable. This naturally leads to a

Baysian approach to credibility theory. The book by Klugman (1992) provides an in-depth treatment of the question. See also the review papers by Makov et al (1996) and Makov (2002). The connection between credibility formulas and Mellin transform in the Poisson case has been established by Albrecht (1984). .

In a couple of seminal papers, Dionne and Vanasse (1989, 1992) proposed a credibility model which integrates a priori and a posteriori information on an individual basis. The unexplained heterogeneity was the modeled by the introduction of a latent variable representing the influence of hidden policy characteristics. Taking this random effect to be Gamma distributed yields the Negative Binomial model for the claim number. An excellent summary of the statistical models that may lead to experience rating in insurance can be found in Pinquet (2000). The nature of serial correlation (endogenous or exogenous) is discussed there.

There are many applications of credibility techniques to vary branches of insurance. Let us mention a nonstandard one, by Rejesus et al (2006). These authors examine the feasibility of implementing an experience-base premium rate discussed in crop insurance.

1.3.2. Claim Count Distribution

Other credibility models for claims counts can be found in the literature, going beyond the mixes Poisson model. The model suggested by Shengwang, Wei & Whitmore (1999) employs the Negative Binomial distribution of the annual claim numbers together with a Pareto structure function. Some credibility models are design for stratified portfolios. Desjardins, Dionne & Pinquet (2001) considered fleets of vehicles and used individual characteristics of both the vehicles and carriers. See also Angers, Desjardins, Dionne & Guertin (2006).

An interesting alternative to the Negative Binomial model can be obtained using the conditional specification technique introduced by Arnold et al (1999). The idea is to specify the joint distribution of (N_t, Θ) through its conditionals. More precisely, the conditional distribution of N_t given $\Theta = \theta$ is $Poi(\gamma(\theta))$ for some function $\gamma: R^+ \to R^+$, and the conditional distribution of Θ given $N_t = n_t$ is $gamma(\alpha(k), \beta(k))$ where $\alpha(\cdot)$ and $\beta(\cdot)$ are two functions mapping N to R^+. For an application of the model to experience rating, see Sarabia et al (2004).

1.3.3. Loss Functions

The quadratic loss function is the most widely used in practice. The results with the exponential loss function are taken from Bermudez et al (2000). Early references about the use of this kind of loss function include Ferrira (1997) and Lemaire (1979). Morillo and Bermudez (2003) used an exponential loss function in connection with the Poisson-Inverse Gaussian model.

Other loss function can be envisaged. Young (1998a) uses a loss function that is a linear combination of a squared error term and a second-derivative term. The squared-error term major the accuracy of the estimator, while the second derivative-term constrains the estimators two linear. See also Young and De Vylder (2000), where the loss function is a linear combination of a squared-error term and a term that encourages the estimator to be close to constant, especially in the tails of the distribution of claims, where Young (1997) noted the difficulty with her semi parametric approach. Young (2000) resorts to a loss function that can be decomposed into a squared-error term and a term that encourages the credibility premium to be constant. This author shows that by using this loss function, the problem of upward divergences noted in Young (1997) is reduced. See also Young (1998b).

Young (2000) also provides a simple routine for minimizing the loss function based on the discussion of De Vylder in Young (1998a).

Adopting the semi parametric model proposed in Young (1997, 2000) but considering that the piecewise linear function has better characteristics in simplicity and intuition than the kernel. Huang et al (2003) used the piecewise linear function as the estimate of the prior distribution and to obtain the estimates for credibility formula.

1.3.4. Zero-Inflated Poisson Model

In this thesis I use zero inflated Poisson (ZIP) model. The ZIP models can be considered as a mixture of a zero point mass and Poisson distribution and where first use to study soldering defects on print wiring boards (Lambert, 1992). To account for overdispersion in the Poisson part, generalizations of the model are possible and include the Zero-Inflated Negative Binomial (ZINB) distribution. It can show that many collected count data display variability bigger than the mean (Ridout, 1998). The extra variability could be due to the clustering or heterogeneity of the data. These data sets are found in diverse disciplines and they are known as overdispersed count data. The negative binomial distribution is a popular choice in modelling overdispersed count data because it is more flexible in accommodating overdispersion in comparison with the Poisson model (Lawless, 1987). In addition to overdispersion, count data may also exhibit a great number of zeros than expected from the Poisson model. The Zero Inflated Poisson model is commonly used in modelling data with excess zero. It is a mixture of Poisson and a degenerate distribution at zero. Lambert (1992) used the ZIP in modelling a manufacturing process. However, count data usually exhibit the joint presence of excess zero counts and overdispersion. In this event, the zero inflated negative binomials distribution provides a better fit. See Yip and Yau (2005) for an application to insurance claim count data. Yip and Yau presented the ZIP, ZINB, zero inflated generalized Poisson and zero inflated double Poisson (ZIDP) to accommodate the excess zero for insurance claim data. Gupta et al (1996) introduced zero adjusted generalized Poisson distribution.

In application of this mechanism, a reported claim implies an increase in the premium of the following years. This induces a “hunger for bonus” (Lemaire 1995): there is an incentive not to report all incurred claims since the increase of the future premiums can be higher than the insurance benefit.

1.3.5. Time Dependent Random Effect

Purcaru & Denuit (2003) studied the kind of dependence arising in these credibility models for claim counts. Albrecht (1985) studied such credibility models for claim counts, whereas Gerber & Jones (1975) and Sundt (1981, 1988) dealt with general random variables.

A fundamental difference between static A1–A3 and dynamic B1–B3 credibility models (I will discuss about them in chapter 2) is that the latter incorporate the age of the claims in the risk prediction, whereas the former neglect this information. Since I intuitively feel that the predictive ability of a claim should decrease with its age, dynamic specification seems more acceptable. As pointed out by Pinquet et al (2001), dynamic credibility models agree with economic analysis of multiperiod optimal insurance under moral hazard. In this optic, the stationary of the $\Theta_i's$ implies that the predictive ability of claims depends mainly on the lag between the date of prediction and the date of occurrence (because of time translation invariance of the marginal’s of the $\Theta_i's$).

Empirical studies performed on panel data, as in Pinquet et al (2001) and Bolancé et al (2003), support time-varying (or dynamic) random effects. An interesting feature of credibility premium derived from stationary random effects with a decreasing correlogram is that the age of the claims are taken into account in the *a posteriori* correction: a recent claim will be more penalized than an old one (whereas the age of the claim is not taken into account with static random effects).

This kind of *a posteriori* correction reconciles actuaries' and economists' approaches to experience rating. Henriet & Rochet (1986) distinguished two roles played by *a posteriori* corrections, showing that these two roles involve different rating structures. The first role deals with the problem of adverse selection, where the very aim is to evaluate as accurately as possible the true distribution of reported accidents. This is the classical actuarial perspective. The second role is linked to morale hazard and implies that the distribution of reported accidents over time must be taken into account to maintain incentives to drive carefully. This means that more weight must be given to recent information in order to maintain such incentives. This is the economic point of view. The credibility model B1–B3 with dynamic random effects, although theoretically more intricate, takes these two objectives into account.

1.3.6. Credibility and Panel Data Model

Frees et al (1999) developed links between credibility theory and longitudinal (or panel) data models. They demonstrated how longitudinal data models can be applied to the credibility ratemaking problem. As pointed out by this authors, by expressing credibility ratemaking applications in the frame work of longitudinal data models, actuaries can realize several benefits: (1) Longitudinal data models provide a wide variety of models from which of choose. (2) Standard statistical software makes analyzing data relatively easy. (3) Actuaries have another method for explaining the ratemaking process. (4) Actuaries can use graphical and diagnostic tools to select a model and assess its usefulness.

2. Statistical and Mathematical Concepts

2.1. Introduction

In chapter one, we elaborate on the elements of this thesis, however, there is an urgent need to get familiar with topics which have impacts on the research framework. Knowing the concepts and rules of these important issues has a vital role in understanding research steps.

The structure of this chapter is as follows: In second section, we give a brief discussion about zero-inflated models. An attempt is done to get familiar with credibility theory and also credibility premium. Also, we formally introduce predictive premium, random effect and present some loss functions that we need to consider. At the end of this chapter, we will introduce the Panel data.

2.2. Zero-Inflated Poisson Models

Insurance data usually include a relatively large number of zeros (no claims). Deductible and no claims discounts (bonus) increase the proportion of zeros, since small claims are not reported by insured drivers. High number of zero values led to the idea that a distribution with excess zero can provide a good fit, such as zero inflated distribution.

The uses of other distributions to model claim counts were motivated by the hunger for bonus situations that can occur in practice. The zero-inflated distributions applied to the number of claims can be used to model the behaviour of insured and model the probability of reporting an accident. Indeed, because the models linked to a reporting decision at the period level, and not at the accident level as with Lemaire's model, we can conceive that each year, a number of insured’s will not claim at all, whatever the case. However, in this situation, one might question why these insured’s procure insurance. Some explanations refer to their fear of insurance, their having minimal protection (mandatory insurance), or their being insured only for major (with probability close to 0).

Another way of interpreting this model has some close connections with Lemaire's model, which also assumes that the number of accidents is Poisson distributed. In addition, it considers the probability of each accident's being reported. However, unlike the Lemaire's model, our models assume that the insured’s do not really know how a bonus-malus system works and do not use any kind of algorithm when deciding whether to claim. More specifically, the first accident of each insured year indicates the way the insured will act for the rest of the year. Accordingly, if the first accident is reported, so will all the other accidents and if the first accident is not reported, nor will the other accidents. This is clearly an approximation, but seems realistic because insured’s think that once they have lost their bonus, the other claims do not have an impact. Those that will not claim at first, because they made an effort to financially support their decision, tend to defend the way they act and will consequently not claim other accidents. In some highly uncommon situations, where a major accident followed a non-reported accident, an insured would probably claim to his insurer.

However, if the vast majority of the insured reports less than two claims per year and given that major accidents are infrequent, this situation happens with a probability very closed to 0. Nevertheless, this approximation error should always be kept in mind and be considered when the accident distribution is analyzed. However, we also think that this non-optimal strategy of deciding to report or not their first claim, followed by the same reporting behaviour for every subsequent claim, can be applied to other jurisdictions. Indeed, these irrational behaviours of insured’s can simply be explained by the fact that many of them do not understand the way insurers set the premiums.

Using a reporting decision at the period level allows us to distinguish the underreporting from the driving behavior. Consequently, using zero-inflated distributions on the number of claims, the idea is too uncensored these zero-inflated distributions to obtain an approximation of the accident frequency distribution. By removing all the effects of reporting that we modeled by the censorship parameter, ∅ the accident process is Poisson distributed (as in Lemaire's model), which is simple and easy to be understand. It seems intuitive to model the accident process by some classic count distribution such as the Poisson distribution because its interpretation is direct, as a limit of a Binomial distribution with the number of tries going to infinity and the accident probability tending to 0. Additionally, note that the zero-inflated models allow us to approximate the accident distribution, even without a deep understanding of the knowledge of the bonus-malus system.

For modelling of claim counts we use two kinds of ZI models in this thesis: MZIP-Gamma and ZI-MVNB; the ZI-MVNB model further generalizes the MZIP- Gamma models in the decision to claim the accident or not. Indeed, in the standard approach and in Lemaire's model, the decision to claim or not is made at each accident, while for the MZIP- Gamma model, the decision is made only for the first accident of each insured period, other accidents being filed similarly. In contrast, for the ZI-MVNB model, the decision is done only for the first accident of the first insured period, other accidents being reported similarly. This model seems appealing in the modeling of basic bonus-malus systems (like in Canada), where the bonus is lost if accident has been claimed in the last 3 or 5 years.

Obviously, ideally, the ∅ parameter of the ZI-MVNB model should be modeled as dynamic, where it can decrease gradually each year, maybe because of the impact of a major accident. In short, we can interpret this model as a situation where some insured will not claim at all for all insured periods. This kind of insured buys insurance only to obey the law, meaning that they will not report accident because their coverage is minimal or because an increase would make their premium too expensive.