Mathematical Statistics
Stockholm University

Generalized Linear Models for Traffic Annuity Claims, with Application to Claims Reserving

Patricia Mera Benner

Examensarbete 2010:2

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Examensarbete **2010:2**,
http://www.math.su.se/matstat

Generalized Linear Models for Traffic Annuity Claims, with Application to Claims Reserving

Patricia Mera Benner*

April 2010

Abstract

Currently the standard/common reserving techniques used by the general insurance actuaries are based on the assumptions that future claims are going to behave with the same pattern/trend as they did in the past, no allowance is made for any individual claim information, any change in speed with which the claims are settled or for any factors that may change the pattern. The main objective of this thesis is to provide the reserving actuary with an alternative method than can be applied in cases when the common reserving methods fail to deliver a suitable result, but also to investigate the outcome of using an alternative model that allows for individual claims information. In this paper we will try to predict the loss reserve for personal injury claims that compensates for the loss of income a claimant will have as a consequence of traffic injury. Also known as traffic annuity claims, these types of claims are rather difficult to estimate; given that the victims personal circumstances will have a significant effect on how the compensation will turn out to be. We will apply a Generalized Linear Model to individual data, and test if this type of individual claims method can work as a support for the actuaries to improve the reserve estimation and at the same facilitate the understanding of which factors that are important for predicting the loss reserve for traffic annuity claims

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: patricia.benner@trygghansa.se. Supervisor: Ola Hössjer.

Generalized Linear Models for Traffic Annuity Claims, with Application to Claims Reserving

Patricia Mera Benner

April 2010

Abstract

Currently the standard/common reserving techniques used by the general insurance actuaries are based on the assumptions that future claims are going to behave with the same pattern/trend as they did in the past, no allowance is made for any individual claim information, any change in speed with which the claims are settled or for any factors that may change the pattern.

The main objective of this thesis is to provide the reserving actuary with an alternative method than can be applied in cases when the common reserving methods fail to deliver a suitable result, but also to investigate the outcome of using an alternative model that allows for individual claims information.

In this paper we will try to predict the loss reserve for personal injury claims that compensates for the loss of income a claimant will have as a consequence of traffic injury. Also known as traffic annuity claims, these types of claims are rather difficult to estimate; given that the victim's personal circumstances will have a significant effect on how the compensation will turn out to be.

We will apply a Generalized Linear Model to individual data, and test if this type of individual claims method can work as a support for the actuaries to improve the reserve estimation and at the same facilitate the understanding of which factors that are important for predicting the loss reserve for traffic annuity claims.

Foreword
Acknowledgement

I started to write this report in the late spring of 2006 at the Non-Life Insurance Company Trygg-Hansa, Nordic Actuarial division, Stockholm. In March 2008, I re-establish my thesis after working at Trygg-Hansa for almost three years. I decided to make some changes in the report given that I felt more mature in my role as an actuary. This report constitutes my master thesis and consists of 30 credits for a master degree of a total of 240 credits in mathematical statistics, from the Mathematics Economy program at the Stockholm University.

First of all I would like to thank my current manager and supervisor Lars Klingberg, Reserving Chief Actuary at Trygg-Hansa, for all his support, effort and understanding. Lars has constantly provided good ideas and discussions for what this thesis should include.

I will also like to thank my second supervisor Olov Dahlberg, Senior Actuary at Trygg Hansa, for his continuous support and guidance; Olov has played and plays an essential role in my continuous development as an actuary and my understanding of the insurance business.

I would as well want to express my gratitude to my supervisor at the University Ola Hössjer, for his valuable comments, guidance and supervision during the completion of this thesis.

Lots of thanks also to Trygg Hansa's reserving and capital modeling team for being such good working mates.

At last, but not least I would like to thank my mother, rest in peace, for all the encouragement she gave me through my life.

Tables of contents

1. Introduction

One of the most important things that an insurance company needs to be familiar with is how much money it needs to set aside to face future liabilities, claims incurred but not reported or not enough reported. The assessment of the reserve plays a vital role in the insurance company's management, pricing and financial state and it is therefore crucial that the company can recognize at any given time how much its ultimate responsibility is going to be.

It is also essential to understand the importance of setting the right claim cost. The reserve is classified as a liability in the company balance sheet and a reserve that is wrongly estimated either excessively or thinly could represent a false picture of the company's financial state and lead to considerable problems for the company. For example a reserve that is inadequate can lead to under-pricing of risk and as a result the premium rate will be based on an optimistic estimation of the company's current liabilities and will damage its competitive position in the market.

The main objective of reserving is to predict a "best estimate" of the ultimate cost of claims and the responsibility of the actuary is to choose a model/technique that produces an ultimate cost estimate that is as close to reality as possible. Unfortunately the "best estimate" of the claim cost is not always related to precision, as the standard deviation in the more common techniques is rather wide and the outcome of the methods is used as a distribution of possible outcomes rather than precise cost estimation.

2. Purpose

There exist a number of statistical methods available for actuaries for use to estimate claims reserves, each method requires different assumptions and achieves diverse level of prudence. The selection of model will depend on the actuary and on what type of claims we wish to estimate.

One of the most common use reserving techniques is the Chain Ladder. It is a relatively simple reserving method based on the assumption that claims on average are going to developed at the same rate as they developed in the past.

The main dilemma the actuaries face by using Chain Ladder is that it works rather poorly when the historical pattern is not stable, it doesn't allow for any individual claims information or for any change in the speed with which aims are settled, or for any other factors which may change. In many cases legal, environmental or society changes can disturb the development of future claims, indicating that the use of the past to predict the future may be a rather difficult task and will probably not work in practice as well as other methods.

One of the reason that Trygg-Hansa wants to investigate if a new alternative model for traffic annuity claims may give a better estimate of the future loss reserve is that the estimated average total cost using payment trend differs considerably from the one obtained when we trend the incurred cost (payments + expected payments). As a consequence of this the company feels an uncertainty with the models that they are currently using and wants to investigate the factors that are creating this uncertainty.

To this end we introduce a generalized linear model with11 covariates and fit this model to a data set of 778 observations.

We leave for future work to apply this model to claims reserving, i.e. summing predicted loss reserves for a number of individuals within the reserving class.

3. Background

To reserve accurately the actuary can not only rely on her/his mathematical/statistical skills, he/she also needs to spend great time understanding the business, pricing, claims handlers and the data. An actuary who ignores these factors would probably not be able to explain movements or deviations when they occur. How precise the reserve estimation will be may depend on what types of claims we are estimating and on the effort the actuary has made to understand the data and the business.

It is widely known that there are some business classes which are easier to predict than others, and the claims are usually classified into two groups, short and long tailed classes.

The short tailed classes are the claims reported and paid to the insurance company shortly after the accident occurs. Minor car accidents not resulting in any bodily injury damage are good examples of these types of claims. Generally all property damage type of policies are short tailed. These types of claims are normally well behaved and have a stable pattern over time making it ideal for the actuary to estimate.

On the contrary long tailed claims are those that take many years to pay, the time between when the claims are reported and settled can be as long as 40 years, creating a natural variation in the reserve estimation. Car accidents that result in bodily injury claims are usually long tailed, a claim where a child is severely injured will probably take many years to settle.

The complexity of the long tailed claims and the lack of precision in the standard models used make reserve estimation a significant challenge for the actuary.

In conclusion there are some business classes/groups of claims where the standard reserving techniques work better and can be applied with great confidence by the actuary and there are some classes where the standard techniques can be challenged and maybe should be challenged by the actuaries in the struggle to predict the "best estimate" of the future cost of the company.

The traffic annuity claims presented in this thesis are a perfect example of claims that are long tailed, very difficult to predict, with a great variation from claim to claim and where the standard reserving methods work rather poorly. To be able to understand the complexity of these types of claims we need to truly understand what it is that makes these types of claims so difficult to predict. I will therefore spend the rest of this section explaining the characteristics of the market before we get into mathematical details in the next section.

3.1 Personal Injury Claims Settlement

To understand the difficulty of estimating the total cost of traffic annuity claims, it is important to first of all understand how the Swedish Motor Insurance business works and how the legal entities work with it. The social reforms play a very important role on how much compensation the injured will receive and should therefore not be underestimated.

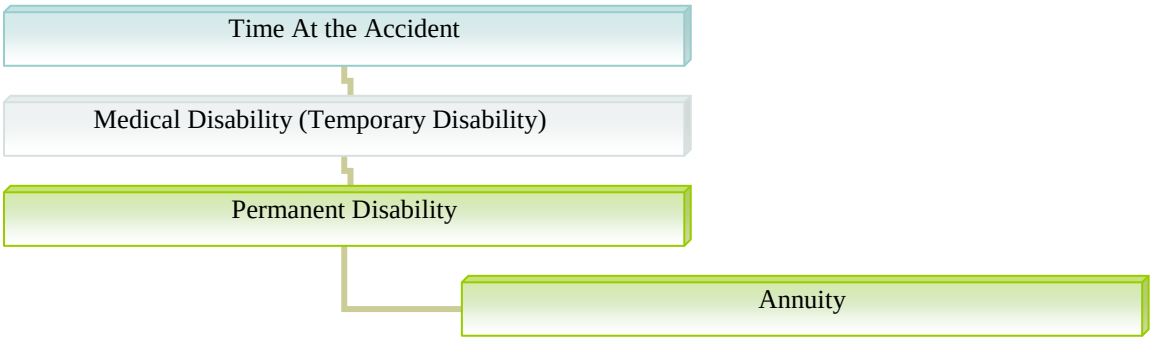
According to Swedish law, anyone that suffers from a personal injury claim as a result of a traffic accident is entitled basically to full compensation for their damages. The insurance company is accountable to cover for damages cause by the insured vehicle but also for liabilities to third parties.

Traffic accidents often occur in a fraction of a second but their consequences can last for days, months, years or in the worst case for the rest of life. A large number of road users involved in traffic accidents recover from their injuries, but some of them never recover fully and suffer from some kind of permanent disability.

When an individual gets injured as a consequence of a traffic accident, the person is entitled to compensation for medical emergency and for future loss of income the victim may have as result of the accident.

In Sweden the claims settlement for personal injury claims is often divided into two stages, medical emergency and permanent disability.

Figure 1.



Medical emergency represents the period from when the accident occurred until the time when the injurer's medical situation has become stable. The length of the period will generally depend on the complexity of injury.

To be able to categorize the degree of an injury, in Sweden the disability is often measured as a percentage (1-99%). In the medical emergency phase the disability percentage is more or less considered to be temporary, mostly due to the uncertainty of the injury. The percentage will only be considered definitive when the victim's residual adverse effect from the accident is more or less permanent.

The compensation during this period will embrace non-financial expenses such as pain and suffering, but also compensations for loss of income within the period.

Permanent disability represents the period when the disability percentage of the victim is considered permanent. From this stage the victim will claim compensation for the annual loss of income that is attributable to the injury.

The process to calculate the annual loss of income the victims have is done based on factors such as illness, unemployment and residual earning capacity as well as wider economic factors such as inflation and taxation. The insurance company's main rationale is to try to estimate what the work situation of the victim would have been if the accident never occurred, basing this on how the injured work and salary situation was prior to the accident but also on what the victim's carrier pattern would have been if the accident never occurred.

The assumptions made to calculate the amount paid for loss of income is calculated after withdrawing any social compensation or benefit (including workers compensation agreements). The amount paid will in most cases be paid as a monthly annuity during the victim's lifetime and then reduced by a certain amount.

3.2 Swedish Road Traffic Injury Commission

In Sweden, the social security system plays a key role in regulating the claim. For this reason, it is important that we understand how this entity works.

The Traffic Injury Commission TSN (Traffic Injury Commission, Swedish abbreviation) is an entity created by the Swedish government with the purpose of acting as a moderator between the Insurance Company and the claimant. The main objective is that the claimant gets his compensation as equally and uniformly as possible.

The insurance company is obligated to consult TSN for a reviewed opinion before reaching a final settlement in cases where the claimant and the insurance company disagree or when the medical disability of the victim is 10 percent or more.

In theory, the victim can appeal the decision taken by TSN; however in the vast majority of the cases the claimant accepts the compensation and the case is considered settled.

Is also worth mentioning that in Sweden the compensation agreed by the victim and the insurance company can be re-examined for up to ten years after the claim was settled. The reopening of the case is primarily due to deterioration in the victim's medical condition, as result, the injured work capacity has worsened and the compensation for loss of income needs to be recalculated.

3.3 Traffic Annuity Claims

A traffic annuity claim is a personal injury claim that emerges when a policy-holder is injured (or a third party injured) as a result of a traffic accident and find himself incapable to work as before. In Sweden this annuity amount covers the future loss of income that the claimant may have, the annuity is set, so that it can be paid out over the entire lifetime of the injured.

A traffic annuity usually deals with heavily disabled victims of accidents of various kinds, making it very difficult to even attempt to predict the future loss of income for the victim. The complexity of estimating the cost of the annuity makes these types of claims a great challenge for the claims handler and actuaries to predict.

An annuity in which the payment is connected to the life loss of income of one person is usually adjusted to an indexation mechanism. In Sweden the annuity indexation depends on the status of the annuity. If the annuity is open the monthly value of the annuity is indexed by a wage index and if closed it is indexed by the official CPI (Consumer price index, KPI, Swedish abbreviation)

When an annuity is calculated it is usually divided into two stages, pre retirement and post retirement. The first one is the stage where calculation is made for the loss of income the injured will get until he or she retires and the second represents the amount paid when the injured retires. It is important to separate these two stages, given that the calculation for each period will differ. For the pre retirement stage, the calculation is done based on parameters such as age, salary, gender, medical disability, worker compensation, occupation and CPI. The post retirement will be related to the amount calculated in stage 1, but will depend on how many years of compensation the injured will get prior to retirement.

The post retirement compensation also known as disability pension is payable in full, three quarters, two third, half of one quarter of the full rate basic pension and supplementary pension. For example, full disability pension will be paid for individuals that correspond to the correct full old age pension.

For simplicity in this paper we will only review the annuity amount calculated for pre retirement.

3.4 Swedish Social Insurance

According to Swedish law if you live and work in Sweden you are covered by Swedish social insurance (Försäkringskassan). The Swedish social insurance agency compensates individuals for a number of different benefits, such as sickness, activity, parental, child allowance and pension.

For this thesis it is important to understand how the Swedish social insurance works together with insurance companies, when they compensate victims injured as a consequence of a traffic accident. The compensation the insurance company pays out will to a large extent depend on how much compensation the victim receives from the social insurance. In other terms, what is not covered by the social insurance will be covered by the insurance company,

There are two benefit that are essential for victims' injured as a result of a traffic accident; Sickness and activity compensation. The injured will receive one or the other compensation depending on the circumstances.

The sickness compensation is a benefit for people age 30-64 that will almost certainly not be able to work full-time due to a disability, injury or illness. The compensation is paid out as full, three-quarter, half or a quarter depending on how your work capacity has been reduced. The amount of compensation received will be related to the victim's yearly salary, the social insurance entity will use this information to calculate an assumed income estimating how the victim's earning would have been if he/she had continued to work. If the victim has low or no income the compensation will be based on the injurer's age (guarantee benefit).

The Activity compensation is a benefit granted to young people age 19-29 that are not able to work due to illness, injury or disability. The benefit is similar to the sickness compensation, with the only difference that the injured can participate in recovery activities.

We will come back later in this paper and discuss the impact and importance the social insurance has in the GLM analysis done in this thesis,

4. Method

The main purpose of this thesis is to first of all, find a model that fits the data on an individual basis and then apply it to expected future annuities.

4.1 Claims Reserving

The insurance company receives premiums from its costumers when they purchase an insurance cover. As compensation the insurer will cover for all damages occurring during the insured period. Generally the company will receive the premiums long before the claim occurs, the company will then need to set aside a reserve that covers future payments. Most claims are reported to the company within a few days after the accident occurs but there are also claims that take months and sometimes years before they are reported to the insurance company.

Claims reserving can be described by the following key elements:

- Case Reserve, which is the amount set aside for claims that have been reported to the company but are not yet fully settled. This amount is usually set by claims handlers.

- Incurred But Not Yet Reported (IBNYR), which is the amount set aside by the actuary for claims that are not yet reported to the company.

- Incurred But Not Enough Reported (IBNER) is the allowance for changes in the claims handlers estimate.

Usually companies do not distinguish between IBNYR and IBNER, instead they report the reserve IBNYR+IBNER as IBNR (incurred but not reported).

4.2 The Chain Ladder Method

The Chain Ladder method is one of the oldest and most well-known reserving methods used to estimate the ultimate claim cost. It is based on a simple algorithm which basically builds on past experience. The Chain ladder technique is deterministic in nature, given that the outcome is a point estimate of the ultimate claims rather than a range of estimates.

For several years the Chain Ladder methodology has been explained from a deterministic algorithm which was not derived from a stochastic model. But in more recent years a variety of articles have been written trying to find the stochastic model that motivates the basic Chain Ladder algorithm, all in the search of quantifying the variability of the estimated ultimate claims amount, see for instance Mack (1994), Mack and Venter (2000) and Taylor (2000).

Generally the data is presented in form of claims triangles, where the rows represent accident periods and the columns represent development periods. The data in the triangle can be of different types such: Numbers of claims, incurred claims, paid claims, case reserve etc.

4.3 Generalized Linear Model (GLM)

Generalized linear models play a very important role in statistical inference. They represent a mathematical way of quantifying the relationship between a response variable and a set of independent variables, including a general class of statistical models.

The use of GLM provides an important advantage over the multiple linear regression model. One of the benefits is that we now go beyond the assumption of a normal distribution for response variables and can use any member of the exponential dispersion family of distributions. This includes Poisson, gamma, binomial and log- normal distributions. Another important advantage is that the relationship between the expected response variable $\mathrm{E}(\mathbf{Y})$ and the dependent variables does not longer need to be linear. Instead of

$$\mathrm{E}(\mathbf{Y}) = \boldsymbol{\beta}\mathbf{X} \tag{4.1}$$

we can allow for the coordinate wise transformation

$$g(\mathrm{E}(\mathbf{Y})) = \boldsymbol{\beta}\mathbf{X} \tag{4.2}$$

of $\mathrm{E}(\mathbf{Y})$. Here $\mathbf{X}, \boldsymbol{\beta}$, $\mathbf{Y}$ and $\boldsymbol{\mu} = \mathrm{E}(\mathbf{Y})$ are defined as

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ . & . & \cdots & . \\ 1 & X_{n1} & \cdots & X_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_1 \\ . \\ . \\ . \\ \mu_n \end{pmatrix},$$

where $\mathbf{X}$ is the $n \times (k+1)$ matrix of covariates, usually called the design matrix. It contains all the values of independent variables and one column of 1: s corresponding to the intercept. $\boldsymbol{\beta}$ is a $(k+1) \times 1$ matrix of parameters, containing k+1 regression parameters that need to be estimated.

Generally the function $g(\cdot)$ is referred as the link function, since it rationality is to "link" the expected value $\boldsymbol{\mu} = \mathrm{E}(\mathbf{Y})$ to the linear predictor $\mathbf{X}\boldsymbol{\beta}$ in a general, flexible way.

4.3.1 The Exponential dispersion family of distribution

The distribution of a random variable Y_i, for observation number i $(i = 1,2,....., n)$ belongs to the exponential family of distribution if its probability density function can be written in the form

$$f_{Y_i}(y_i;\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(y_i,\phi,\omega_i)\right\}. \tag{4.3}$$

We assume here that the dispersion parameter is a positive constant, $\phi > 0$, and that the weighs ω_i are also positive, $\omega_i > 0$, for all observation numbers $i = 1,2,....., n$.
Further, θ_i is the canonical parameter that depends on the observation number i, the functions $b(\cdot)$ and $c(\cdot)$ depend on the distribution and $y_i\theta_i$ is the multiplicative term between the response variable and the parameters.

The mean and variance of Y_i are specified by (cf. McCullagh and Nelder 1989)

$$\mathrm{E[Y_i]} = \mu_i = b'(\theta_i) \tag{4.4}$$

$$Var[\mathrm{Y_i}] = \sigma_i^2 = b''(\theta_i)\phi/\omega_i \tag{4.5}$$

where

$$V(\mu_i) = b''(\theta_i) \tag{4.6}$$

is known as the variance function. Combining (4.2) and (4.4) we find that θ_i is related to the regression parameters through

$$g(b'(\theta_i)) = \beta_0 + \sum_{j=0}^{k}\beta_j \mathrm{X}_{i,j} = \eta_i \tag{4.7}$$

where $g(\mu) = \theta^{-1}((b')^{-1}(\mu))$, assuming $\theta_i = \theta(\eta_i)$ for all i and some function $\theta(\cdot)$.

4.4 Maximum Likelihood Estimation of GLM

The maximum likelihood method is used to determine the parameters that maximize the probability of sampled data. It also provides an efficient method for quantifying uncertainty by means of confident intervals.
The method is considered to be flexible and it can be applied to most models and different types of data.
More about The Maximum Likelihood method in Appendix A.2

4.5 Goodness of Fit

Two statistics that are useful in assessing the goodness of fit are the scale deviance D defined as

$$D = 2(\ell_{sat} - \ell(\hat{\beta}_0, \ldots, \hat{\beta}_k, \hat{\phi})) \tag{4.19}$$

where ℓ_{sat} is the log likelihood obtained from the saturated model with $\hat{\mu}_i = y_i$ for all i.

$\hat{\beta}_o, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are maximum likelihood estimated of $\beta_o, \beta_1, \ldots, \beta_k$, and Pearson's chi-square χ^2 statistic defined in Appendix A.1 .

4.6 Actuarial use of GLM

Generalized linear models have proved to be a very useful tool in general insurance, see for instance Brockman and Wright (1992) and Ohlsson and Johansson (2004). It has been widely used by actuaries as a pricing instrument for many years, and in more recent years it has also become a powerful tool for other areas in General Insurance.

The benefits of generalized linear models are being recognized in many areas of actuarial practice, several actuarial papers have been written in the last 30 years describing the use of GLM within loss reserving, claim cost estimation, average claim cost and much more. A number of statistical software's have been developed for actuarial use and as actuaries continue to become more familiar with GLM we expect the number to increase.

5. Data Analysis

The data used in this paper was extracted from Trygg-Hansa's claims and annuity database. From this register we have only selected the claims that have an annuity monthly payment as consequences of liabilities in a traffic accident. The amount set as an annuity is only based on the claimant's inability to work in the future.

Due to the long tail nature of traffic personal injury claims it was quite a challenge to obtain a set of data that could be used in this thesis. The claims staging environment has changed on several occasions, making it very complicated to get historical data with the quality needed.

We were able to create a dataset with 902 independent observations, where we only included data with accident years between 1990 and 2005. We didn't include any data for more recent accident years due to the simple fact that it takes at least three years before an annuity can be settled. Claims older than 1990 were also excluded since development changed considerably from the 70's to the 90's.

The first step was to calculate the monthly payment of each observation as if the annuity was settled in 2009. We needed to do so due to the annuity indexation mechanism described in section 3.3.
We achieved this by calculating the monthly value backwards in time to the value it had when the accident occurred. By doing this we could then easily apply a wage index to calculate the monthly value as if the annuity was settled today.

The wage index used depended on what type of occupation the claimant had when the accident occurred, and was gathered from Statistics Sweden (SCB). Unfortunately we couldn't find any statistics prior to 1996. For those years we decided to use the government's official CPI.

5.1 Selecting a GLM

An important aspect of this thesis is model building. It is essential that we ensure that the selected model fits data well. For that reason, it is crucial that we begin with an exploratory data analysis in order to get a better picture of how the variables interact with each other. The purpose of this exploratory analysis is to select a distribution for the response variables, a link function for the GLM and a set of predictor variables (covariates).

We plot the response variable against each of the candidate predictor variables (properly divided into classes) and inspect visually whether or not there is any correlation.

5.2 Variables

5.2.1 Number Observations

The number of annuity claims in each accident year is a crucial statistic given that several of the claims that are open today are as old as 1973. It is therefore very important to understand the distribution of number of claims by accident year for our dataset.

Our first thought was to include all statistics available from the system. We were able to get data from accident years 1973 to 2006, but we could at an early stage determine that claims older than 1990 were in many cases not representative of the types of claims we believe will occur in the future.
Different circumstances, such as law, governmental and environment changes make it very complicated to rely on the past to predict the future. The number of observations per accident year during 1990-2006 is shown in Figure 2.

We can also observe from Figure 2 that a great number of the annuities are distributed between accident years 1990-2000. This is not surprising given that we know that annuities take a long time to settle. Annuities are often settled within 10-15 years, and as a result, annuities from 2001 may not be fully represented in the annuity system until 2015.
We can also observe that it takes at least three years before an annuity is settled see Figure 2 and the low values for years 2004-06, annuities that are settled quickly are usually cases where the injured medical situation is stable.

Figure 2: Numbers of Annuities per Accident Year

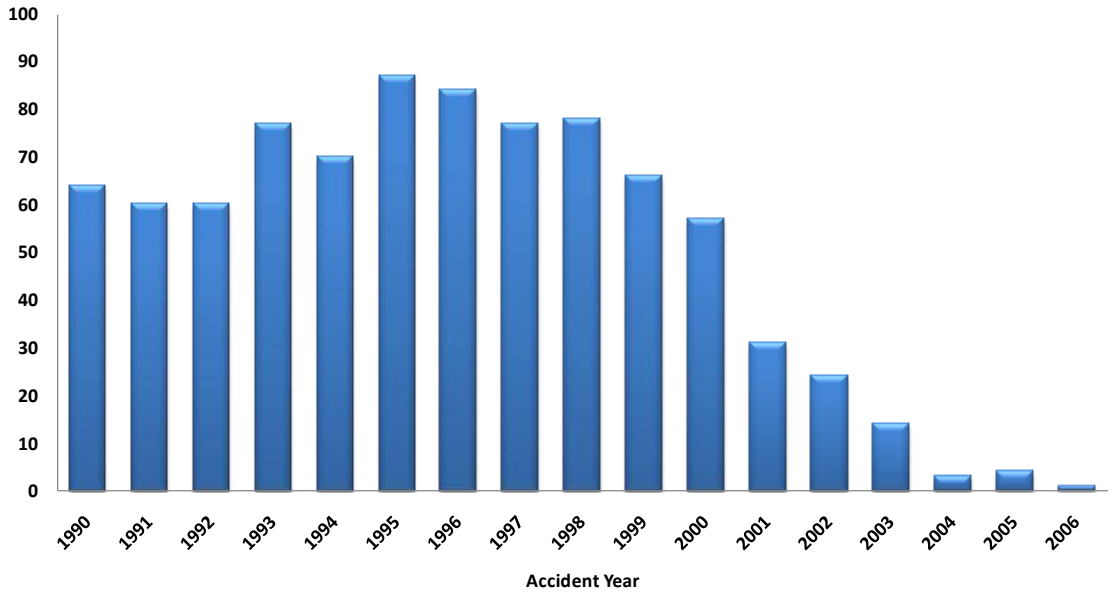
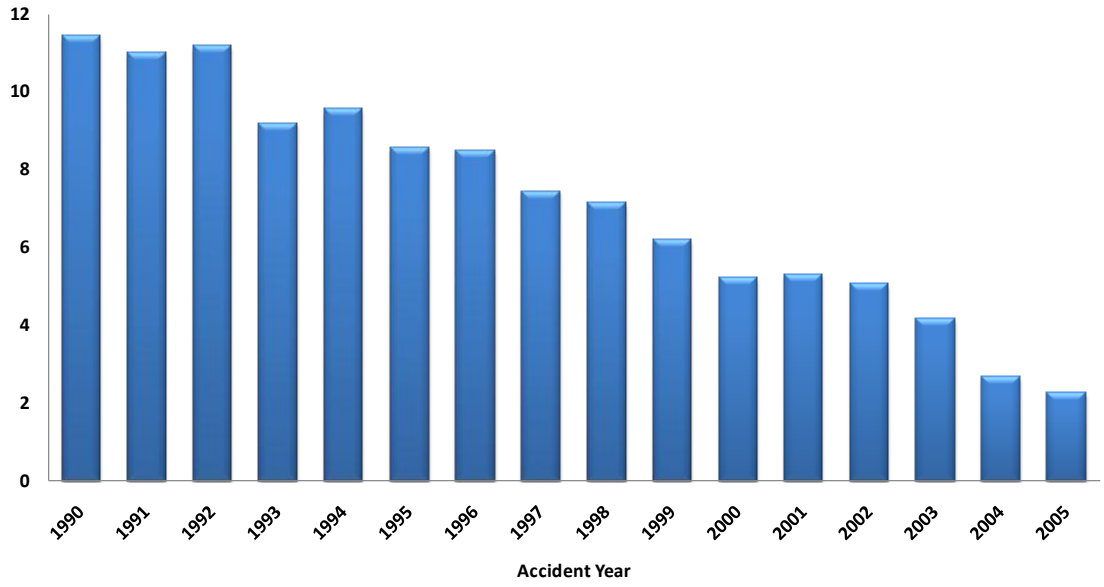


Figure 3 shows the average number of years it takes to settle an annuity claim, the average year is expect to be around 10-12 years depending on what type of injury the claimant may have. The average number of years decreases for more recent accident years. This is as expected as we expect the number of claims to increase considerably in future years when they remaining open claims are settled.

Figure 3: Average Time to Settled Annuities

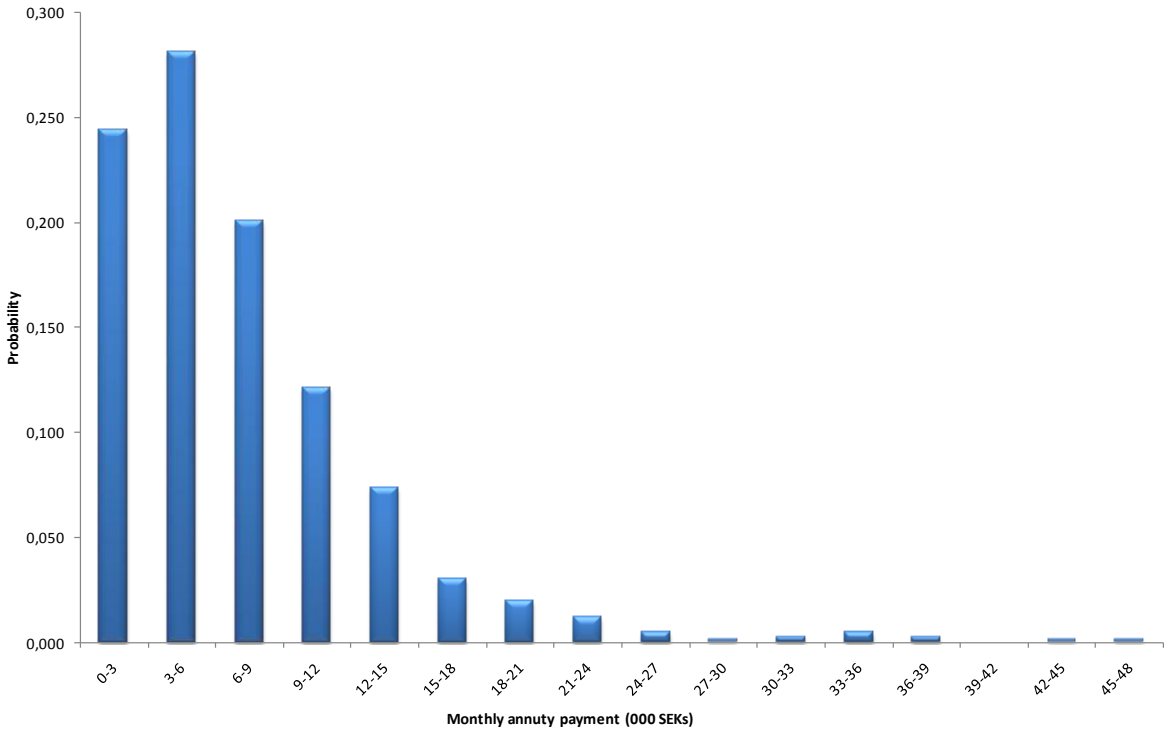


5.2.2 The response variable - Monthly annuity payment

One of the important characteristics of the generalized linear model is that we don't need to limit our response variable to be quantitative and normally distributed. We can use a variety of distributions from the exponential dispersion family to fit the model (as mentioned in Section 4.3). The response variable can be of different types such as binary, ordinal, continuous etc.

By examining data we can determine what type of response variable y_i we are dealing with. It is clear from the graphs below that the monthly annuity payments have a distribution where most of the compensation is less than 10000 SEK. We can also observe that the distribution is skewed to the right and that there are some outstanding values that we might consider excluding later on.

Figure 4: Observed Monthly Annuity Payments as at 2009



Is well-known that the Gamma distribution is a flexible model for a skewed distribution on the positive axis. We thus assume that the density of Y_i is gamma distributed, i.e.

$$f(y_i;\alpha_i,\beta_i) = \frac{1}{\Gamma(\alpha_i)\beta_i^{\alpha_i}} y_i^{\alpha_i-1} e^{y_i/\beta_i},\ i = 0,1,2,3\ldots$$

where

for $\alpha_i > 0$

is the Gamma function.

The mean and the variance are defined as

$$\mathrm{E}[y_i] = \alpha_i\beta_i$$

$$Var(y_i) = \alpha_i\beta_i^2$$

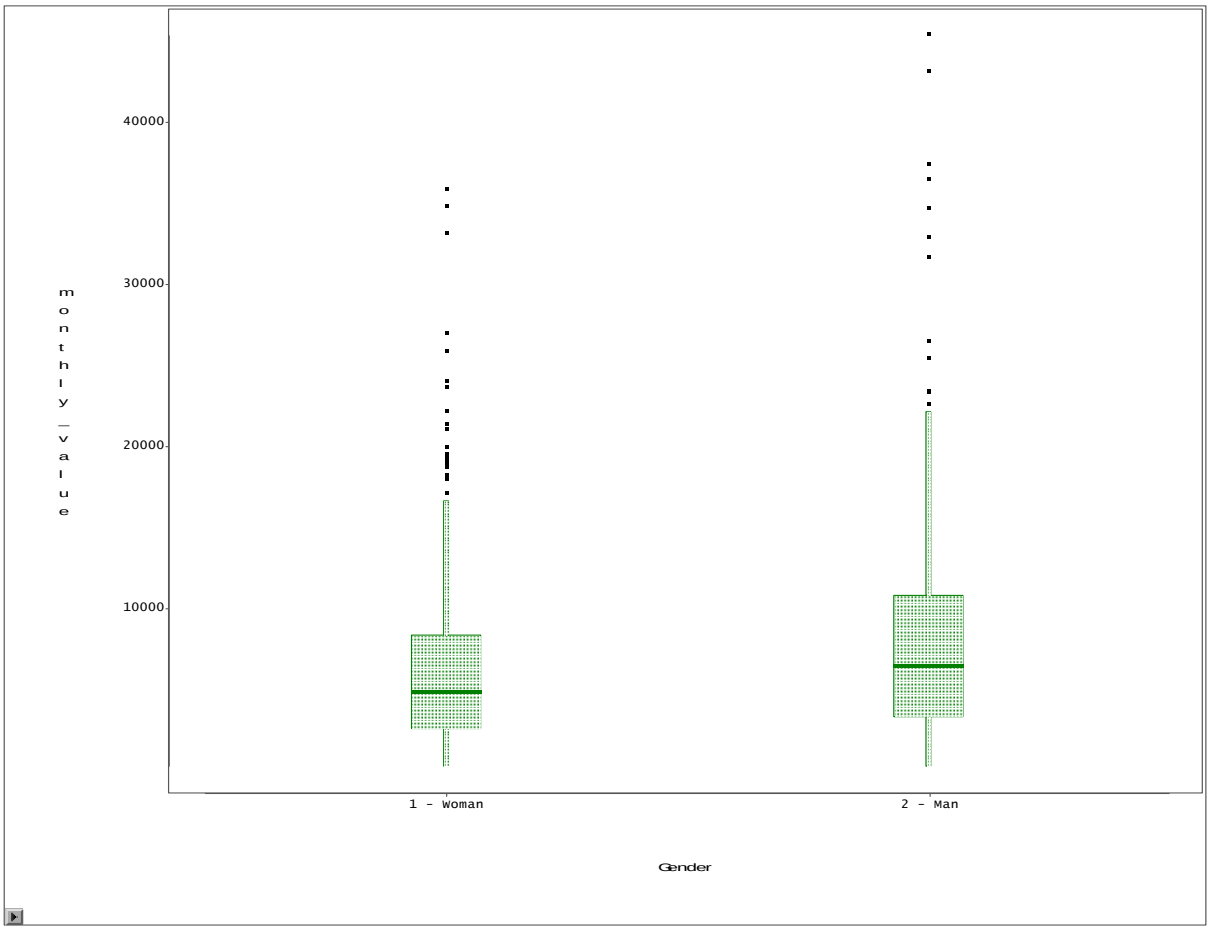
If $\alpha_i = \alpha$ is fixed and β_i varies, we thus get $\mu_i = \alpha\beta_i$, a quadratic variance function $V(\mu_i) = \mu_i^2$ and the dispersion parameter $\phi = \alpha^{-1}$. This implies that the standard deviation of an observation is proportional to its mean.

5.2.1 Gender

The gender variable stands for the classification between men and women.

In Figure 5 we plot the response variable against gender, to see if we could find any clear correlation between them. We can observe that the monthly annuity value is higher for men than for women. This is not a surprise given the differences in the social and economic status of men and women. One import factor in the calculation of the annuity is based on salary. Women frequently work with public related jobs that are usually less paid on average than jobs in the private sector, where a great number men work .

Figure 5: Box Plot of Loss of Income Compensation Split by Gender Class



We let gender correspond to the first covariate $X_{i,1}$ of the design matrix in section 4.3 with $X_{i,1} = 0$ for women and $X_{i,1} = 1$ for men. This is described in table 1.

Table 1: Covariate for Gender Class

Levels	Class	Values	Covariates
2	Gender	Woman, Man	$X_{i,1} = 0, X_{i,1} = 1,$

5.2.2 Age at the time of the accident

The process to calculate an annuity for future loss of income is very complex and a great challenge for the claims handlers. It can in many cases feel like a guessing game as an attempt must be made to take all contingencies into account. These include factors such as illness, unemployment and residual earning capacity as well as wider economic factors such as inflation and taxation. When the injured is a child, the uncertainty increases considerably given that it does not have a relevant pre-injury salary to use as a starting point to estimate future loss of income. In Sweden the child's background, his family members, profession and social status form the basis for calculating the future loss of income. In other words, the assumption made is that the child will follow a similar career path as his family members.

Thus the age of the victim plays a significant role in how the annuity for future loss of income is calculated. We decided to classify the variable into four groups: Children, Young Adult, Adult and Mature Adult (see Table 2).

Table 2: Age Group for Time of Accident

Age Group	Frequency	Percentage	Age Interval
Children	31	3%	0-19
Young Adult	120	14%	20-25
Adult	469	54%	26-45
Mature Adult	256	29%	>46
Total	901	100%	

In Figure 6 we plot the response variable aligned with each age group. From the graph we can notice that victims that were injured as a child have a higher compensation compared to other age groups. There is strong evidence that the age of the victim at the time of the accident plays an import role in what compensation the victim will get.

It is reasonable to assume that children will get higher compensation due to fact that it actually takes longer to settle these types of claims. Generally because we need to wait until the child is in a working age before we even can make a patent statement on of how much the compensation will be. Table 3 gives an idea of how many years on average it takes to settle claims. We believe that there is a clear relation between the age of the victim at the time of the accident and the time it takes to settle the claim.

Table 3: Average Age of Settlement of the Annuity for Different Age Groups

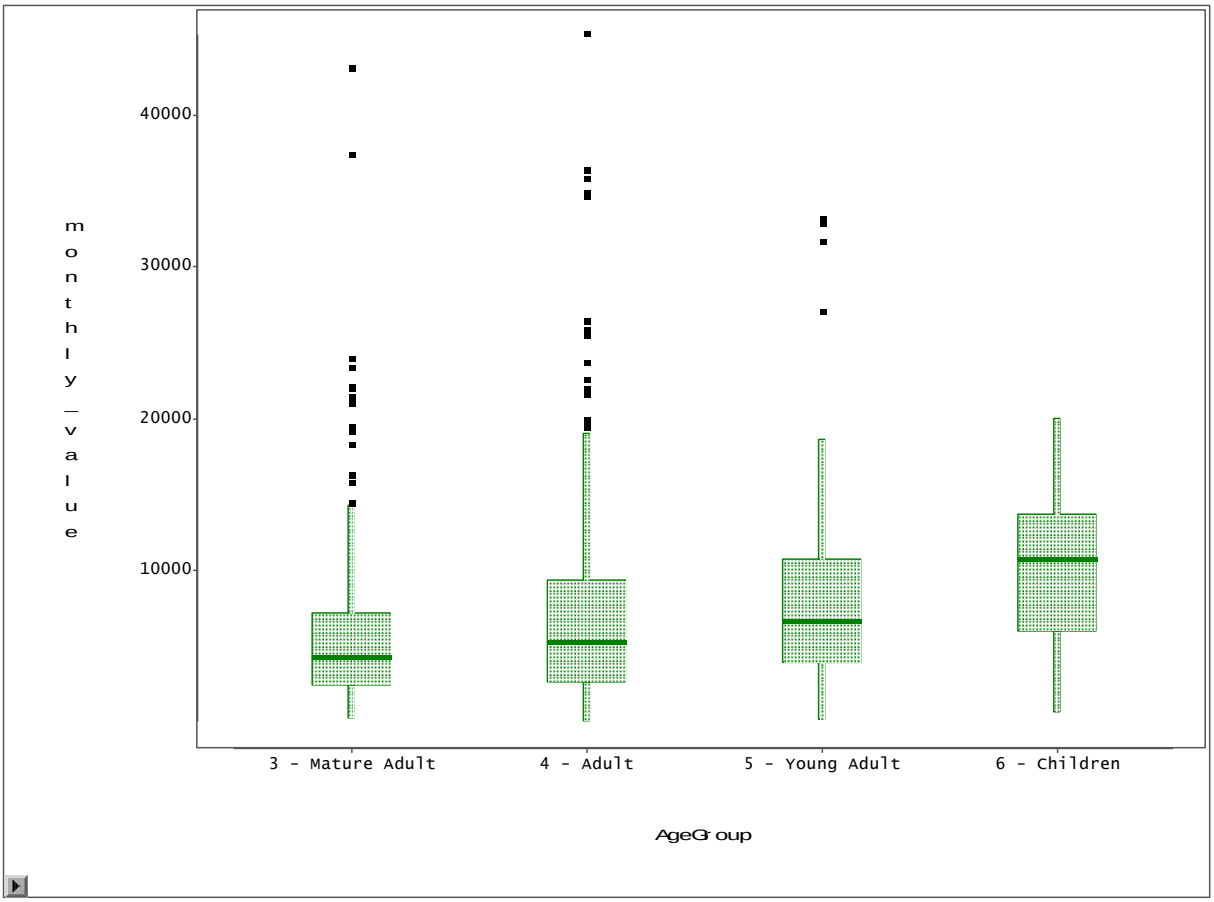
Children	Young Adult	Adult	Mature Adult
12	10	9	7

Children claims take on average at least eleven years to settle, eleven years is a rather optimistic number given that we excluded all claims occurred prior to 1990 (see Section 5.2.1). We believe that the average age it takes to settle children claims is more likely to be around 18 years.

Unfortunately, by not including claims occurred prior to 1990, the number of children in our data set is low (see Table 2).

Taking all these factors in account it might be reasonable to separate children annuities from the rest of the groups and do a separate analysis for this type of annuities and include data from earlier accident years. We will return to this discussion later in this paper.

Figure 6: Box Plot of Loss of income Compensation Split Age Group



In the design matrix in Section 4.3 we use the covariates $X_{i,2}$, $X_{i,3}$ and $X_{i,4}$ for the age at accident of observation i, with $(X_{i,2}, X_{i,3}, X_{i,4}) = (0,0,0)$ for mature adults, $(1,0,0)$ for Adults, $(0,1,0)$ for young adults and $(0,0,1)$ for children. This is also described in Table 4.

Table 4: Covariates for Age Class

Levels	Class	Values	Covariates
4	Age Group	Mature Adult,	$(X_{i,2}, X_{i,3}, X_{i,4}) = (0,0,0)$
		Adult	$(X_{i,2}, X_{i,3}, X_{i,4}) = (1,0,0)$
		Young Adult	$(X_{i,2}, X_{i,3}, X_{i,4}) = (0,1,0)$
		Children	$(X_{i,2}, X_{i,3}, X_{i,4}) = (0,0,1)$

5.2.3 Salary Income

Salary Income represents the yearly income the injured had at the time when a settlement of the victim's compensation was reached between the insurance company and the injured. This variable is extremely important given that the compensation the victim will get is to a great extent based on this.

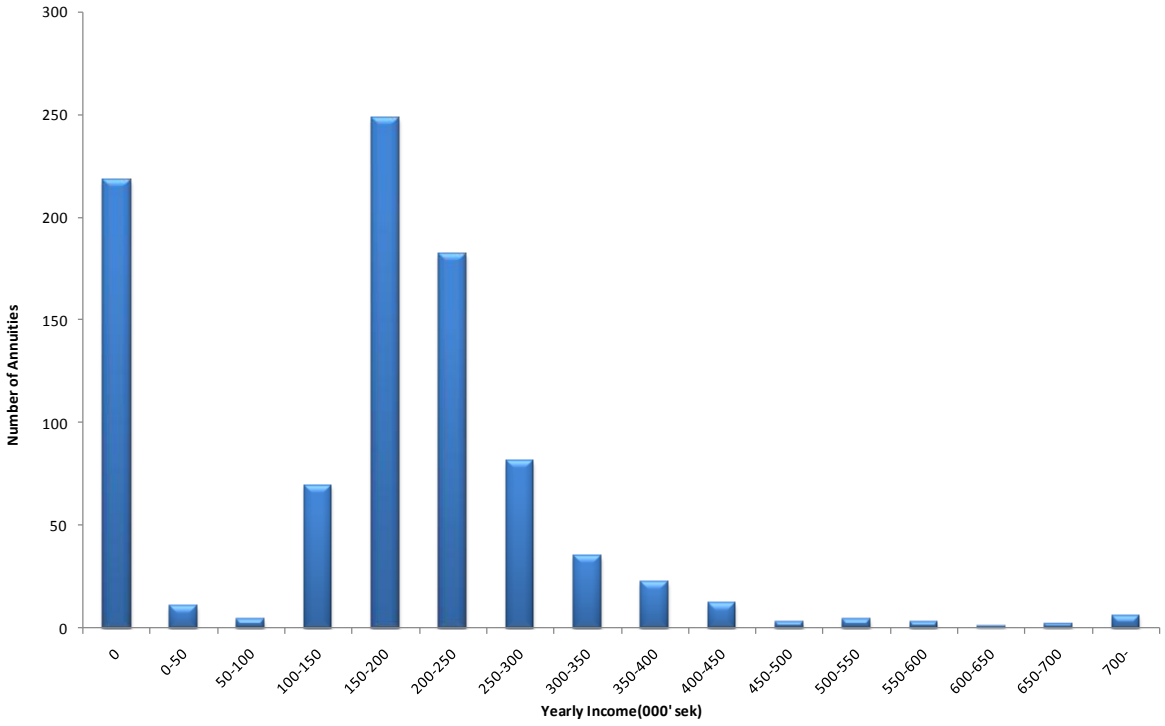
Personal Injury claims, particularly claims that become annuities are usually handled manually by claims handlers. On a case by case basis, the claims handler will follow the case and close it when a settlement between the injured and insurance company is reached. When the case is not settled the yearly income of the victim will be updated each year using the wage index for that particularly case. In many cases the settlement will take years, the salary of the individual would probably change, but given that the cases are settled manually, claims handlers may fail to send this information into the claims database.

For simplicity we assume that salary in our sample dataset is the salary income the individual had at the time when a settlement was reached between the insurance company and the victim. We use an average wage index to index the salary to the current value.

To group the variable into a categorical variable it was essential to analyze the observed salary values. In Figure 7 we can study the distribution of the yearly salary income for all annuities. Note that a great number of injuries have salary zero and that there are some extreme values in the tail. It is rather critical to understand why some observations have zero values. One of the reasons is that if a victim at the time of the accident was a child, a student or unemployed the yearly salary will probably be zero, but the annuity will still be calculated on the basis of a yearly salary.

We can also observe in Figure 7 that there are some outstanding values in the tail that we might consider excluding later.

Figure 7: Observed Yearly Salary



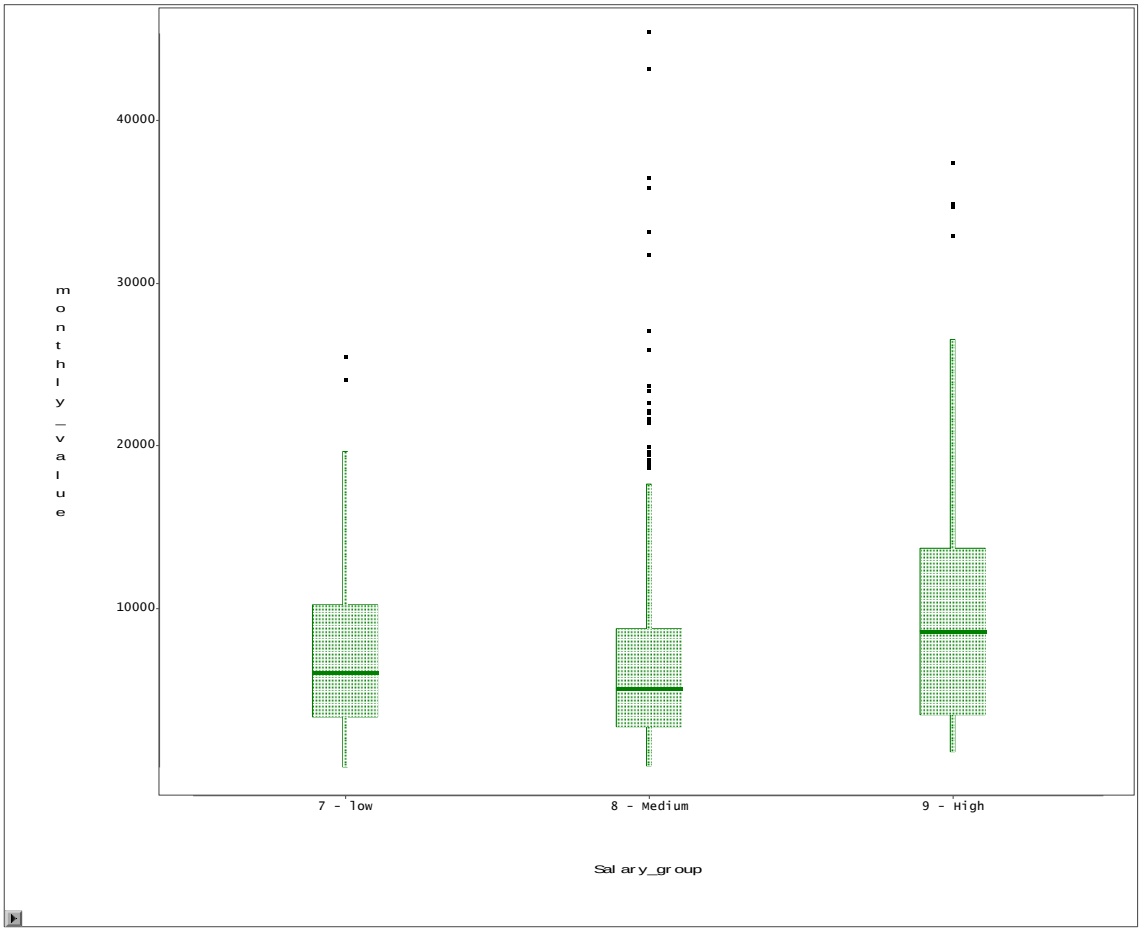
After examining the data we decided to group the variable into three groups; Low, Medium and High (also described in Table 5)

Table 5: Groups for Salary Income

Salary Group	Frequency	Percentage	Salary (000 SEKm)
Low	233	26%	0-100
Medium	615	68%	101-350
High	52	6%	351-
Total	876	100%	

In Figure 8 we can observe that the group with highest compensation is the group with highest yearly earnings. This is no surprise given that the monthly annuity compensation is calculated based on the victim's income. We can also observe that the victims with lowest income have on average higher compensations than the group with medium earnings. This may seem a bit contradictory, but it can easily be explained by two factors. The first factor is that there are a number of observations where the victim's yearly salary may not be correct updated in the system and as a result the data can be misleading. The other important factor is the impact of social insurance (see Section 3.4). This information is important to keep in mind when we choose our final model.

Figure 8: Box Plot of Loss of income Compensation Split by Salary Group



In the design matrix in Section 4.3 we use the covariates $X_{i,5}$, $X_{i,6}$ and $X_{i,7}$ to describe the salary income of observation i, with $(X_{i,5}, X_{i,6}) = (0,0)$ for low income, $(1,0)$ for medium income and $(0,1)$ for high income. This is also described in Table 6.

Table 6: Covariates for Salary Class

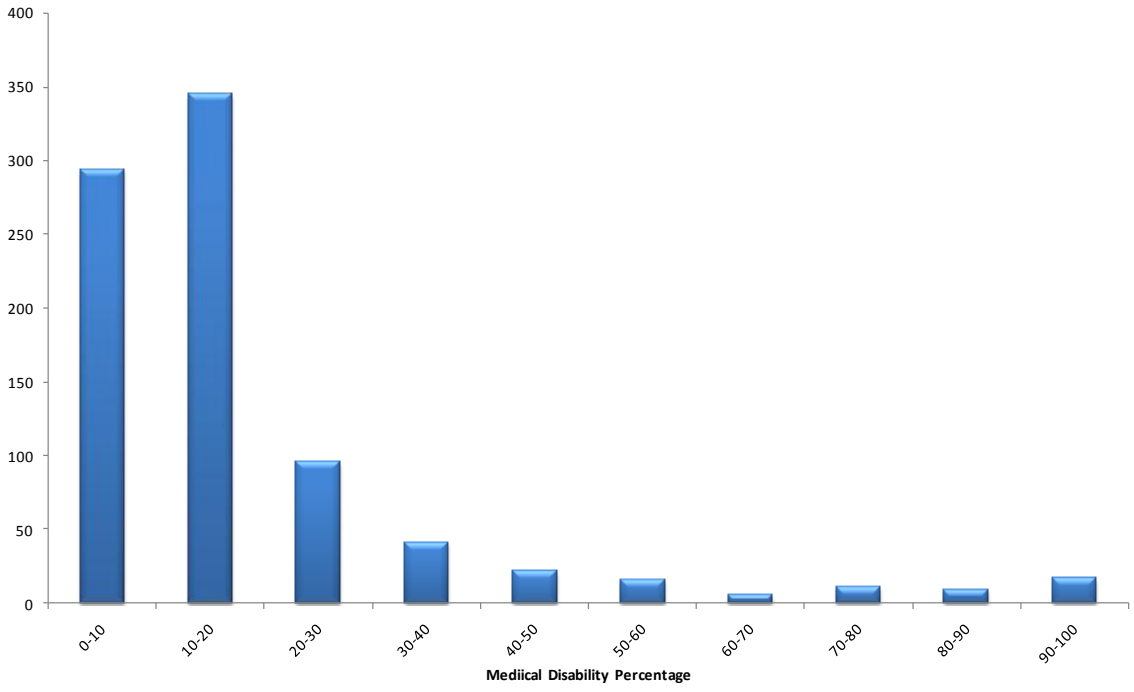
Levels	Class	Values	Covariates
3	Salary Group	Low	$(X_{i,5}, X_{i,6}) = (0,0)$
		Medium	$(X_{i,5}, X_{i,6}) = (1,0)$
		High	$(X_{i,5}, X_{i,6}) = (0,1)$

5.2.4 Medical Disability Percentage

Medical disability percentage represents the permanent medical assessment that the victim will have as a consequence of a traffic injury, stated as a percentage 1-99%. The assessment is known to be an important factor in determining the overall level of compensation for non-financial damage, but no clear relationship has been established for the compensation of loss of income that a person may get.

Figure 9 shows the medical disability distribution in our sample dataset. It is clear from the graph that most of the annuities have a medical disability percentage that is less than 50 % and that there are significant number of annuities that have a disability less than 20%.

Figure 9: Number of Annuities Split by Medical Disability Percentage



It is important to point out that a medical disability of 15 % does not necessary mean that the victim is 15 % incapable of working, in many cases the victim may be 100% compensated for loss of income, but the medical disability percentage may be less than 50%. Consequently it is crucial to keep in mind when we apply our model that using medical disability percentage to estimate future loss of income can be deceptive. It can give over-compensation as well as under-compensation in comparison to the actual loss.

We categorized the variable into 4 groups. Table 7 shows how the variable was classified and distributed within the chosen classes in our sample data.

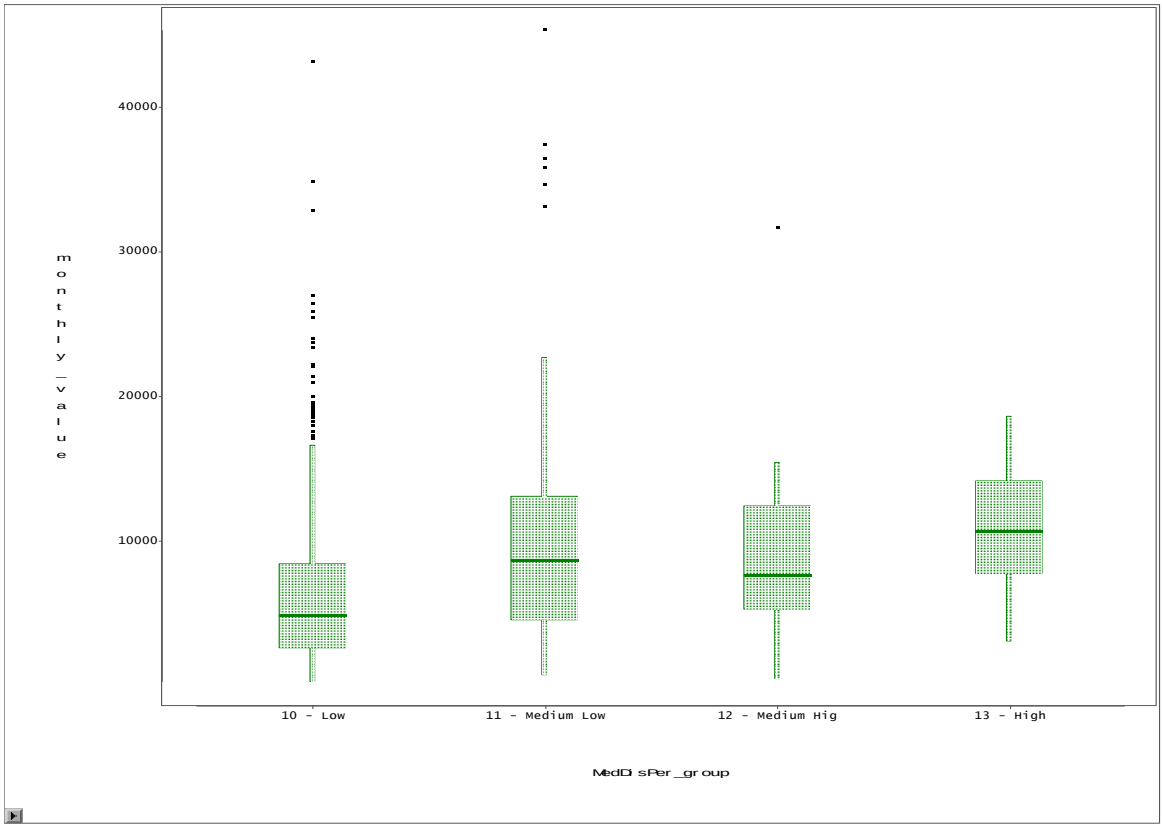
Table 7: Groups for Medical Disability

Medical Disability Group	Frequency	Percentage	Medical Disability (%)
Low	716	82%	0-25
Medium Low	100	11%	26-50
Medium High	28	3%	51-75
High	32	4%	76-99
Total	876	100%	

The second step was to examine if this variable had any statistical significance in determining the level of compensation for the financial loss that the claimant will get. We plot the variable against the response variable (see Figure 10) and found that there was no clear linear relationship between the two variables. However there seems to be a weak trend that higher medical disability leads to larger expected annuity payment.

It is worth commenting, that the group of individuals with the highest annuity compensation (with a medical disability interval of 0-75%) have medical disability between 26 – 50%. After discussing this with claims handlers, they confirm that the claims that are more expensive are not the ones where the victim is totally impaired, rather the cases where the victim is 25-50 % disabled. This may come as a surprise, but the fact that individuals that are for instance 70% disable will probably not be able to work for the rest of his life, and the compensation will be decided at an earlier stage. An individual that is partially disable is often more difficult to estimate given that the probability of working can be negotiated with the insurance company. As a result the insurance company would have to wait and see the development of the victim's injury before a final settlement can be agreed.

Figure 10: Box Plot of Loss of Income Compensation for Medical Disability Group



In the design matrix in Section 4.3 we use the covariates $X_{i,7}$, $X_{i,8}$ and $X_{i,9}$ to describe medical disability of Observation i, with $(X_{i,7}, X_{i,8}, X_{i,9}) = (0,0,0)$ for Group 1, $(1,0,0)$ for Group 2, $(0,1,0)$ for Group 3 and $(0,0,1)$ for Group 4. This is also described in Table 8.

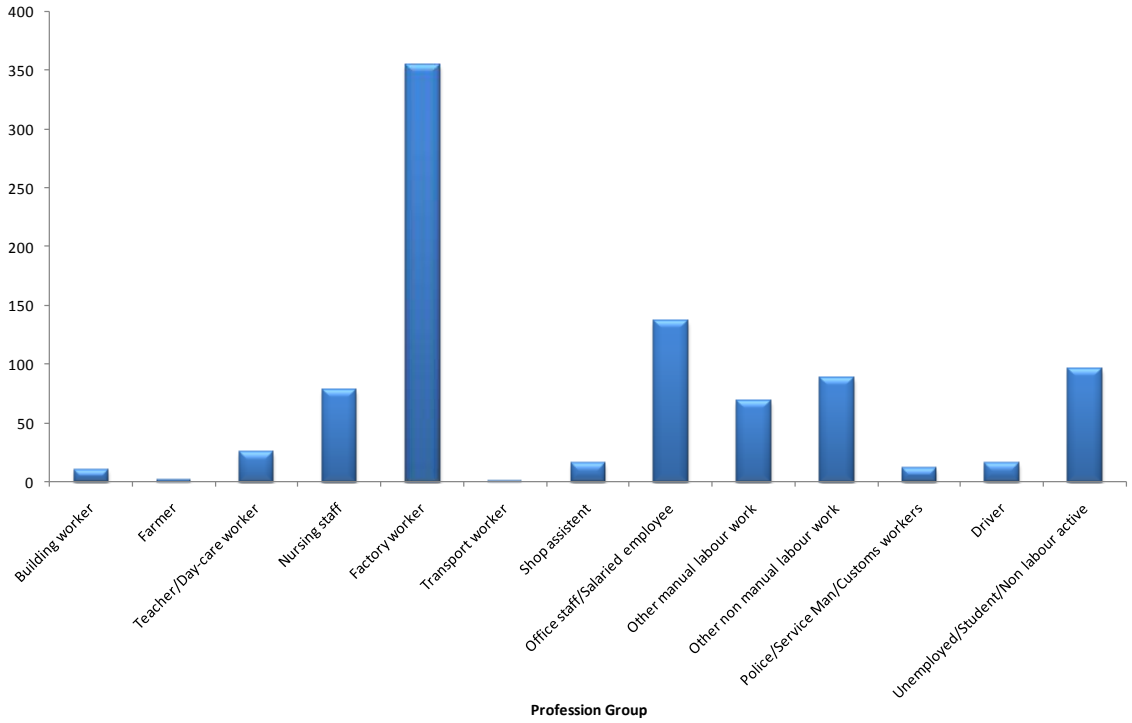
Table 8: Covariates for Medical Disability

Levels	Class	Values	Covariates
4	Medical Disability Percentage	Low	$(X_{i,7}, X_{i,8}, X_{i,9}) = (0,0,0)$
		Medium Low	$(X_{i,7}, X_{i,8}, X_{i,9}) = (1,0,0)$
		Medium High	$(X_{i,7}, X_{i,8}, X_{i,9}) = (0,1,0)$
		High	$(X_{i,7}, X_{i,8}, X_{i,9}) = (0,0,1)$

5.2.5 Variable Occupation/Profession

The variable represents the types of profession the injured had at the time of the accident. In Trygg-Hansa's database structure we could find thirteen different classifications of the victim's occupation (for more information see Appendix A2). The distribution of the number of annuities split by occupation is described in Figure 11.

Figure 11: Number of Annuities Split by Profession Group



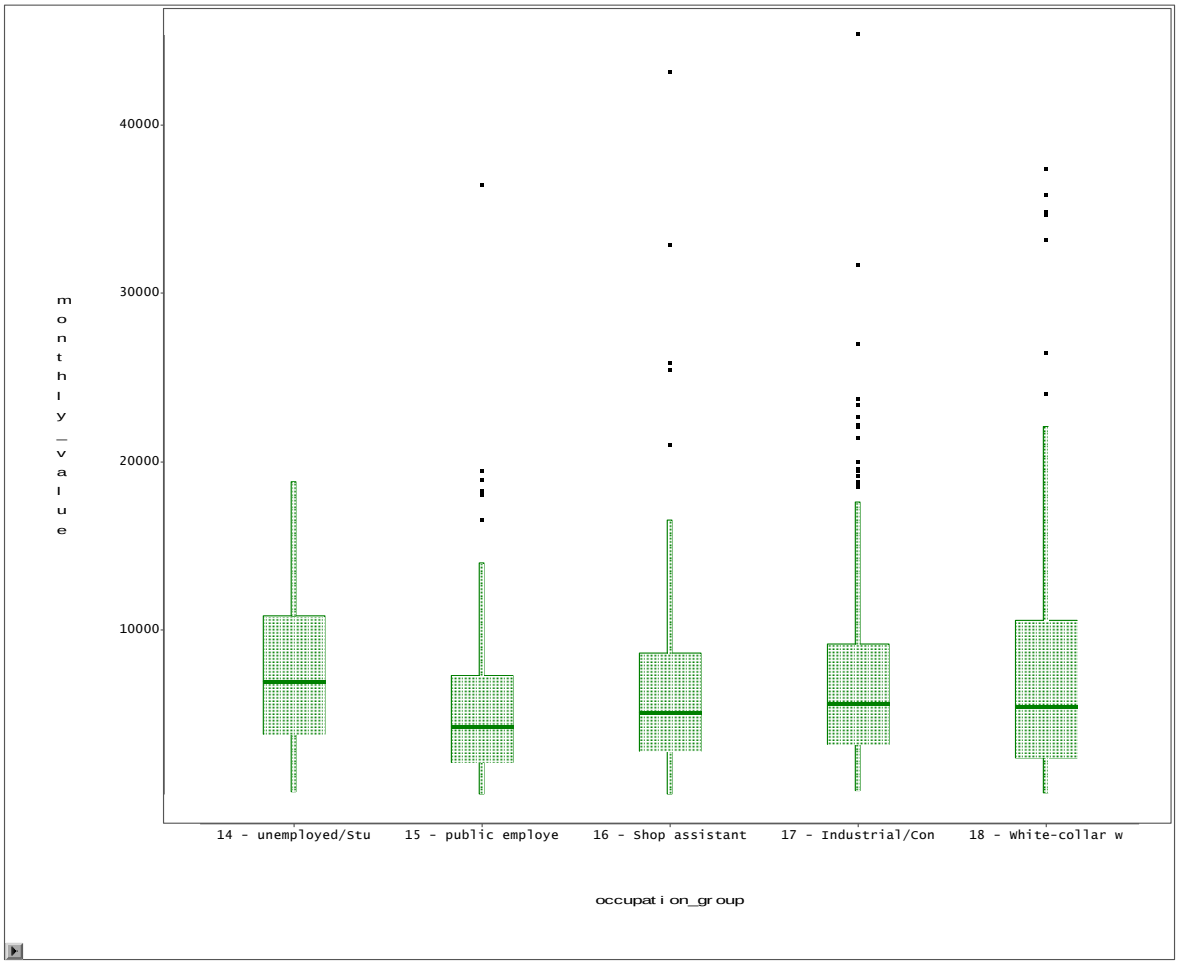
We classified the variable into five groups different groups, also described in Table 9. We believed that it was important to keep victims that were unemployed or not work active at the time of the accident as one group, the reason behind this is that we know that in cases where the victim is a child or a student the occupation or carrier pattern will change from what was initially set as their occupation. In this type of cases the profession group will probably not be as significant as for the other groups.

Table 9: Merging of Professions into Groups

Profession Group	Frequency	Percentage	Profession
Unemployed/Student/Not work active	90	10%	13
Public employee	100	11%	3 and 4
Shop assistant/Other manual workers	180	21%	7, 9 , 10 and 12
Industrial/Construction/Police	369	42%	1, 2, 5, 6 and 11
White-collar worker	137	16%	8
Total	876	100%	

In Figure 12 we plot the monthly annuity value to the different profession groups so that we could get a descriptive view of how the variable interacts with the financial compensation of the victim.

Figure 12: Box Plot of Loss of Income Compensation for Profession Group



In the design matrix of Section 4.3 we use the covariates $X_{i,10}$, $X_{i,11}$, $X_{i,12}$ and $X_{i,13}$ to describe profession group of observation i, with $(X_{i,10}, X_{i,11}, X_{i,12}, X_{i,13}) = (0,0,0,0)$ for Group 1, $(1,0,0,0)$ for Group 2, $(0,1,0,0)$ for Group 3, $(0,0,1,0)$ for Group 4 and $(0,0,0,1)$ for Group 5. This is also described in Table 8.

Table 8: Covariate for Profession Group

Levels	Class	Values	Covariates
5	Profession Group	1	$(X_{i,10}, X_{i,11}, X_{i,12}, X_{i,13}) = (0,0,0,0)$
		2	$(X_{i,10}, X_{i,11}, X_{i,12}, X_{i,13}) = (1,0,0,0)$
		3	$(X_{i,10}, X_{i,11}, X_{i,12}, X_{i,13}) = (0,1,0,0)$
		4	$(X_{i,10}, X_{i,11}, X_{i,12}, X_{i,13}) = (0,0,1,0)$
		5	$(X_{i,10}, X_{i,11}, X_{i,12}, X_{i,13}) = (0,0,0,1)$

5.3 Additional Variables Not Included In The Analysis

5.3.1 Income Disability Percentage

Income disability percentage is an important reference in determining the size of the monthly annuity payment. It represents the degree of reduced capacity to work after the victim is exposed to a traffic accident

The variable is generally linked to the medical disability percentage (see Section 5.2.4), but not in all cases. A 50% medical disability does not necessary imply a 50% income disability. In many cases the medical disability is only used as a guide in determining the income disability of the victim. The victim's working and living condition after the accident plays a more significant role.

Unfortunately this variable can't be found in our system and can only be collected manually. Given the lack of time the variable could not be included in the analysis. In future development of this thesis it would be important to include this variable.

5.3.2 Workmen's injury

According to Swedish law, workers' compensation is coordinated with the Swedish social security system. Injuries experienced at work are often qualified under the Work Injuries Insurance Act.

A traffic injury sustained as a result of an accident at work, or on the way to work is covered by the worker's compensation insurance. This means that the amount paid by the insurance company will only be the share that social insurance doesn't cover. As a result the size of the annuity will to a large extent depend on this variable. It is therefore important to adjust for this when creating the mathematical model

The main data issue is that we can't identify when an injury is covered by the Workers Injury Insurance Act or not. Unfortunately, this information can only be collected manually. To solve this problem we divided data into two different policy groups, Commercial and Private insured claims, based on what type of cover the claimant had at the time of the accident. Private claims are injuries that are only covered by the insurance company and will therefore not face this problem. On the contrary, a significant proportion of Commercial claims are covered by the workers compensation. Unfortunately, we are not able to classify when the compensation is paid as an integrated part of the social insurance and when it is paid by the insurance company. For simplicity, we will therefore not include these claims. In the future it is important to include them to get the overall estimate for traffic annuity claims.

5.3.3 Compensation from the Swedish Social Insurance Agency (Försäkringskassan)

The insurance company covers for the loss of income that the social insurance doesn't cover. As a result, what the insurance company pays out will to a large extend depend on this.

For example in cases where the victim's earning is very high the social insurance will only cover a top amount, the rest will be covered by the insurance company. These types of cases will have a large annuity reserve and will be very expensive for the insurance company.

Another important example is that if the victim was young or unemployed when the accident occurred, the social insurance will pay out a guarantee benefit that may differ from what not be what the victim claims to be entitled to. As a consequence the insurance company will pay a large amount to cover this.

To get a correct comparison between the amount the insurance company pays out for each individual we need to get full information of how much compensation the victim receives from The Swedish Social insurance agency. The problem is that we don't have this information in our database; unfortunately this can only be gathered manually.

We believe it is essential that in future development of this thesis, this variable is added to the dataset.

5.4 Link Function

The link function provides the relationship between the linear predictor and the mean of the response variable. There are many commonly used link functions, and their choice can be somewhat arbitrary. It can be convenient to match the domain of the link function to the range of the distribution function's mean.

The link function is defined in section 4.7 as

$$g(\mu_i) = \eta_i = \sum_j x_{i,j}\beta_j .$$

For continuous data with no heteroscedasticity it is appropriated to use an identity link: $\eta_i = \mu_i$.

If each covariate has a multiplicative effect one mean response, we express this by means of a log link, $\eta_i = \log(\mu_i)$ with its inverse $\mu_i = e^{\eta_i}$. This is the standard choice of link function in non-life insurance.

5.5 Model Selection and Parameter Estimation for Selected GLM

The first step in our model selection was to build a model strategy. This we did in Section 5.2 by including those covariates into the GLM that we a priori considered important and which also were seen to have an effect on the response variables from Box plots. We needed to build a model that was relatively simple and made sense and at the same time captured most of the information in the data. In the second step we examine whether the model provides a reasonable approximation of the data.

Our chosen model selection strategy is the backward selection method. This technique is based on including all variables in the model (significant or not), and then removing the variables step by step from the model until all remaining variables are significant.

5.5.1 Model Selection Including All Variables.

We argued in Section 5.1 that a Gamma response variable is reasonable. We thus start by applying a GLM with a Gamma distributed response variable using all the covariates described in Section 5.2.

Table 7: Class Level Information

Class	Levels	Values
Gender	2	Woman, Man
Age_group	4	Adult, Children, Mature Adult, Young Adult
Salary_group	3	low, Medium, High
MedDisPer_group	4	Low, Medium Low, Medium High, High
Occupation group	5	Unemployed/Student/Non labour active, Public employee, Shop assistant/Other manual Workers, Industrial/Construction/Police , White-collar worker

Model Information

Data Set	WORK.TEST2
Distribution	Gamma
Link Function	Log
Dependent Variable	monthly_value

Number of Observations Read	856
Number of Observations Used	856

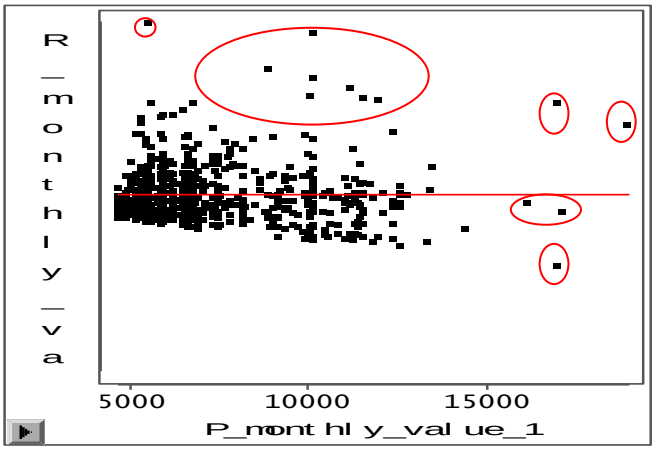
By analyzing the output (see Table 8) for this model we could conclude that the deviance of the Gamma distributed model was 477 on 842 degrees of freedom and that the scaled deviance and Pearson χ^2 values adjusted for DF were rather small indicating a good fit..

Table 8: Criteria for Assessing Goodness Of Fit Full Model

Criterion	DF	Value	Value/DF
Deviance	842	477.5239	0.5671
Scaled Deviance	842	927.6886	1.1018
Pearson Chi-Square	842	501.8923	0.5961
Scaled Pearson X2	842	975.0292	1.1580
Log Likelihood		-8323.9593	

We plot the residuals against the predicted values to see if the residuals were scattered randomly about 0 and to spot a trend. The residuals appear to behave randomly, and this suggests that the model fits data well. We also check if we have some possible outliers or errors in the data. We could observe that there are some observations (see Figure 13 marked in red) that may have an overly large impact on the parameter estimates. To quantify the influence of the four most extreme observations we observe the change in the deviance and in Pearson's χ^2.

Figure 13: Residuals vs Predicted Values



We found that if we excluded the four observations in red, the deviance and Pearson's χ^2 decreased and we got a better fit of the data (see Table 9).

Table 9: Criteria for Assessing Goodness of Fit Full Model Excluding Outliers

Criterion	DF	Value	Value/DF
Deviance	837	454.3861	0.5429
Scaled Deviance	837	919.5373	1.0986
Pearson Chi-Square	837	440.0780	0.5258
Scaled Pearson X2	837	890.5821	1.0640
Log Likelihood		-8247.8532	

The output of Table 10 contains an analysis of parameter estimates, illustrating the estimates of model parameters, their standard error, and a Wald test of each parameter. It also includes a Wald test pre adjusted, including the tests before the outliers where excluded. From this table we can observe that by excluding the potential outliers we got better parameter estimates for almost all covariates with the exception of the Gender parameter.

Tables 10: Analysis of Parameter Estimates of Full Model Excluding Outliers

Parameter		DF	Estimate	Error	Square	Pr>Chi	Pre Adjusted Pr>Chi
Intercept		1	10.0532	0.1784	3174.34	<.0001	<.0001
Gender	Woman	1	-0.0642	0.0533	1.45	0.2288	0.0613
Gender	Man	0	0.0000	0.0000	.	.	.
Age_group	Adult	1	-0.1539	0.0737	4.36	0.0368	0.0870
Age_group	Children	1	0.1650	0.1588	1.08	0.2987	0.3328
Age_group	Mature Adult	1	-0.2748	0.0818	11.29	0.0008	0.0038
Age_group	Young Adult	0	0.0000	0.0000	.	.	
Salary_group	low	1	-0.5072	0.1191	18.15	<.0001	<.0001
Salary_group	Medium	1	-0.6002	0.1130	28.21	<.0001	<.0001
Salary_group	High	0	0.0000	0.0000	.	.	.
MedDisPer_group	Low	1	-0.4402	0.1311	11.27	0.0008	0.0014
MedDisPer_group	Medium Low	1	-0.0924	0.1451	0.41	0.5240	0.9932
MedDisPer_group	Medium High	1	-0.1442	0.1844	0.61	0.4345	0.4874
MedDisPer_group	High	0	0.0000	0.0000	.	.	.
occupation_group	Unemployed/Student/Non Labour active	1	-0.2325	0.1116	4.34	0.0372	0.0514
occupation_group	Public employee	1	-0.2780	0.0971	8.19	0.0042	0.0145
occupation_group	Shop assistant/Other manual workers	1	-0.1890	0.0826	5.24	0.0221	0.0543
occupation_group	Industrial/Construction/Police	1	-0.0893	0.0749	1.42	0.2333	0.1704
occupation_group	White-collar workers	0	0.0000	0.0000	.	.	
Scale		1	2.0237	0.0912			

The type 3 analysis shown in Table 11 gives us a more general view of how the parameters fit the data. By using this information we can conclude that the variables Salary_group and MedDisPer_group are highly significant and that the gender variable is not significant and should probably be excluded from the model.

Table 11: LR Statistics for Type 3 Analysis

Source	DF	Chi-Square	Pre Adjusted Pr > ChiSq
Gender	1	1.45	0.0612
Age_group	3	15.69	0.0077
Salary_group	2	31.88	<.0001
MedDisPer_group	3	30.34	<.0001
occupation_group	4	11.22	0.0985

5.5.2 Model Selection Excluding Gender

A third Genmod analysis where we excluded the variable Gender gave the following output:

Table12: Criteria for Assessing Goodness of Fit without Gender in Full Model

Criterion	DF	Value	Value/DF
Deviance	838	455.1031	0.5431
Scaled Deviance	838	919.6340	1.0974
Pearson Chi-Square	838	439.6281	0.5246
Scaled Pearson X2	838	888.3635	1.0601
Log Likelihood		-8248.5781	

Table 13: LR Statistics for Type 3 Analysis without Gender in Full Model

Source	DF	Chi-Square	Pr > ChiSq
Age_group	3	16.31	0.0010
Salary_group	2	36.10	<.0001
MedDisPer_group	3	35.51	<.0001
occupation_group	4	12.42	0.0145

We found that by excluding the gender covariate (see Table 13), the variables Age_group and Occupation_group are now more significant than they were when we included Gender as a covariate. We could also conclude that the scaled Pearson χ^2 value/DF was reduced by 0,004, also indicating a better fit.

The parameter estimates for the different variables, shown in Table 14, give a more detailed analysis of all variables. Here we could notice that the variable Age group children, medical disability medium high, medical disability high and occupation group industrial/construction/police were not significant. In next section we will go into more detail of how to proceed.

Tables 14: Analysis of Parameter Estimates of Full Model without Gender

Parameter		DF	Estimate	Error	Square	Pr>Chi
Intercept		1	10.0593	0.1786	3173.05	<.0001
Age_group	Adult	1	-0.1567	0.0737	4.51	0.0336
Age_group	Children	1	0.1686	0.1589	1.13	0.2887
Age_group	Mature Adult	1	-0.2795	0.0818	11.69	0.0006
Age_group	Young Adult	0	0.0000	0.0000	.	.
Salary_group	low	1	-0.5188	0.1187	19.10	<.0001
Salary_group	Medium	1	-0.6235	0.1114	31.34	<.0001
Salary_group	High	0	0.0000	0.0000	.	.
MedDisPer_group	Low	1	-0.4658	0.1294	12.95	0.0003
MedDisPer_group	Medium Low	1	-0.1026	0.1449	0.50	0.4787
MedDisPer_group	Medium High	1	-0.1555	0.1843	0.71	0.3989
MedDisPer_group	High	0	0.0000	0.0000	.	.
occupation_group	Unemployed/Student/Non Labour active	1	-0.2342	0.1117	4.40	0.0359
occupation_group	Public employee	1	-0.2984	0.0958	9.71	0.0018
occupation_group	Shop assistant/Other manual workers	1	-0.1787	0.0822	4.72	0.0298
occupation_group	Industrial/Construction/Police	1	-0.0828	0.0748	1.23	0.2678
occupation_group	White-collar workers	0	0.0000	0.0000	.	.
Scale		1	2.0207	0.0910		

5.5.3 Model Selection Excluding Children

In Section 5.2.2 we discussed the impact of the variable child in the data. We know from this that the number of observation is small, 31, and that the data quality may not be the most optimal for a GLM model based on the variables mentioned in Section 5.2.

A traffic claim where the victim is a child is usually a very volatile and unstable claim. The development of the child's individual situation can change dramatically during the years, making it very complex to estimate the total cost of the annuity. For example the first medical diagnosis made by the doctor can change up to 100% from what was he/she initially estimated. As a result the variable medical disability percentage will not be optimal in the first stage of the claim, or maybe not until the child's healing progress has stabilized.

Another important dilemma is that children usually don't have a loss of income until they reach maturity. The child's initial income is usually 0, and the loss of income calculation is based on the child's close relatives' yearly income. To be able to use salary as a variable, the claims handler should update the claim information in the system so that the collected salary represents the yearly earnings that the loss of income calculation is based on. A similar dilemma is observed when using the variable occupation_group.

As a result we believe that Children should be analyzed separately, in a more detailed and cautious approach. We leave this for future development. See Appendix A.3 for more information of the GLM done for Children.

A fourth Genmod analysis assuming a Gamma distributed response variable excluding the parameter Children was fitted to 820 observations. The model gave the following output.

Table 15: Criteria for Assessing Goodness of Fit without Gender and Child Age Group

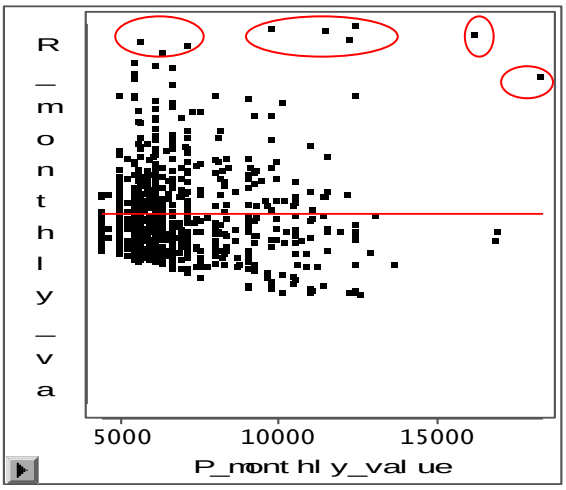
Criterion	DF	Value	Value/DF
Deviance	808	445.9322	0.5519
Scaled Deviance	808	887.1314	1.0979
Pearson Chi-Square	808	432.4913	0.5353
Scaled Pearson X2	808	860.3923	1.0648
Log Likelihood		-7937.7804	

Table 16: LR Statistics for Type 3 Analysis without Gender and Child Age Group

Source	DF	Chi-Square	Pr > ChiSq
Age_group	2	11.77	0.0028
Salary_group	2	36.02	<.0001
MedDisPer_group	3	36.17	<.0001
Occupation_group	4	12.39	0.0147

Table 15 shows a slight worsening of the fit (the scale deviance increase with 0, 0005), despite this we decided to use the model based on statistical consideration and on subject-matters discussed through this paper. The Type 3 output also indicates that all variables are significant, in particularly the groups Salary and MedDisPer. For a more detailed analysis of the parameters see Appendix A3, Table 17.
In Figure 14 we plot the residuals against predicted values to check for some possible outliers. We can observe that there are some observations (see Figure 14 marked in red) that may have an overly large impact on the parameter estimates. It is common that when we exclude parameters, some outliers that were not visible before are now visible. We also need to keep in mind that even if the observations may have a large influence it does not necessarily mean that we need to exclude them.

Figure 14: Residuals vs. Predicted Values



5.5.3.1 Model Excluding Children and Outliers

Our final approach was to fit a model that excluded the variable Gender, the parameter Children and some possible outliers and got the following output:

Table 18: Criteria for Assessing Goodness of Fit

Criterion	DF	Value	Value/DF
Deviance	778	421.4391	0.5417
Scaled Deviance	778	853.5737	1.0971
Pearson Chi-Square	778	407.9070	0.5243
Scaled Pearson X2	778	826.1662	1.0619
Log Likelihood		-7644.3697	

Table 19: LR Statistics for Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
Age_group	2	11.77	0.0028
Salary_group	2	40.16	<.0001
MedDisPer_group	3	33.91	<.0001
occupation_group	4	12.54	0.0138

Tables 20: Analysis of Parameter Estimates of Full Model without Gender

Parameter		DF	Estimate	Standard Error	Chi - Square	Pr>Chi
Intercept		1	10.1418	0.1880	2908.73	<.0001
Age_group	Adult	1	-0.1468	0.0740	3.93	0.0474
Age_group	Mature Adult	1	-0.2821	0.0837	11.37	0.0007
Age_group	Young Adult	0	0.0000	0.0000	.	.
Salary_group	low	1	-0.5541	0.1228	20.36	<.0001
Salary_group	Medium	1	-0.6759	0.1152	34.45	<.0001
Salary_group	High	0	0.0000	0.0000	.	.
MedDisPer_group	Low	1	-0.4880	0.1400	12.16	0.0005
MedDisPer_group	Medium Low	1	-0.1067	0.158	0.46	0.4994
MedDisPer_group	Medium High	1	-0.2497	0.2052	1.48	0.2236
MedDisPer_group	High	0	0.0000	0.0000	.	.
occupation_group	Unemployed/Student/Non Labour active	1	-0.2641	0.1154	5.24	0.0221
occupation_group	Public employee	1	-0.3013	0.0985	9.36	0.0022
occupation_group	Shop assistant/Other manual workers	1	-0.2100	0.0842	6.22	0.0126
occupation_group	Industrial/Construction/Police	1	-0.1114	0.0761	2.14	0.1432
occupation_group	White-collar workers	0	0.0000	0.0000	.	.
Scale		1	2.0254	0.0947		

The fit of the model is reasonable with a scaled deviance of 853 on 778 df, which gives a scale deviance/df = 1,0971. The effects of Salary and MedDisPer group are both highly significant. The main conclusion is that there is a clear relationship between how much the victim will be compensated each month, the victim's yearly earning and the disability percentage.

From the output we can also conclude that the parameter estimates for the MedDisPer group low is significant however medium high and high are not statistically significant. This is in line with what was discussed in Section 5.2.4.

In Figure 15 and 16 we plot the observed and the predicted monthly annuity value next to each other to see how good the predicted model fits the data in a more visual way. We can observe that the predicted values starts at 4200 SEK while the observe values has smaller values and that the distribution of the predicted values has a smaller dispersion.

Figure 15: Observe Monthly Annuity Value

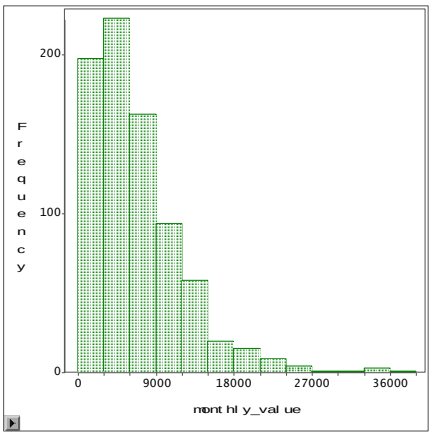
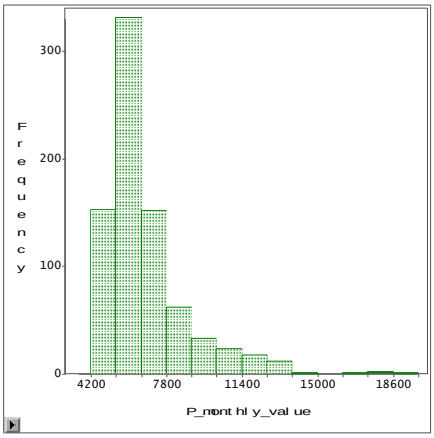


Figure 16: Predicted Monthly Annuity Value



Using the information in Table 21 we can write the model equation (4.2) as follow:

Log (E (Monthly Value)) = 10,1416 - 0,2821 * $\overline{X}_2$ - 0,1468 * $\overline{X}_3$ - 0,5541* $\overline{X}_5$ - 0,6759 * $\overline{X}_6$ - 0,4880 * $\overline{X}_7$ - 0,1067 * $\overline{X}_8$ - 0,2497 * $\overline{X}_9$ - 0,2641 * $\overline{X}_{10}$ - 0,3013 * $\overline{X}_{11}$ - 0,2100 * $\overline{X}_{12}$ - 0,1114 * $\overline{X}_{13}$

Tables 21: Parameter Information

Parameter	Variable	Included i Final Mod	Gender	Age_Group	Salary_G roup	MedDisPer_G roup	Occupation_Group
$\overline{X}_1$	Gender	No	Women/Men				
$\overline{X}_2$	Age_Group	Yes		Mature Adult			
$\overline{X}_3$	Age_Group	Yes		Adult			
$\overline{X}_4$	Age_Group	No		Children			
$\overline{X}_5$	Salary_Group	Yes			Low		
$\overline{X}_6$	Salary_Group	Yes			Medium		
$\overline{X}_7$	MedDisPer_ Group	Yes				Low	
$\overline{X}_8$	MedDisPer_ Group	Yes				Medium Low	
$\overline{X}_9$	MedDisPer_ Group	Yes				Medium High	
$\overline{X}_{10}$	Occupation_ Group	Yes					Unemployed/Student/ Not work active
$\overline{X}_{11}$	Occupation_ Group	Yes					Public employee
$\overline{X}_{12}$	Occupation_ Group	Yes					Shop assistant/Other manual workers
$\overline{X}_{13}$	Occupation_ Group	Yes					Industrial/Construction/Police

6. Discussion

6.1 Summary

One of the main purposes of this thesis was to try to provide the reader with a deeper understanding of the Swedish traffic annuity claims and the complexity of using existing reserving methods.

Swedish bodily injury claims are one of the most long tailed claims in the world. We estimate that in exceptional cases it can take up to 40 years to close a claim. This is a remarkable aspect of the Swedish motor business. A background on legal, security and society gives the reader a basic knowledge of the Swedish motor market. This facilitates understanding of the problems and interpretation of the GLM analysis with more awareness of how the variables interact with each other.

To conclude, we introduced a generalized linear model with gamma distributed response variable and 11 covariates and fit this model to a data set of 778 observations as defined in Section 5.2. All variables where examined in detail so that the selected model represented the best estimate of what we believed an individual with a number of characteristic should receive. In the future, we hope our GLM can be used for claims reserving.

6.2 Data Quality

Throughout this paper we have on several occasions mentioned the constant setback we have encountered each time we examined data. It has been quite a challenge to gather all information needed and some rough assumptions have been made so that we could continue the work. It may seem that this is an exceptional case, but the truth is that actuaries constantly struggle with data. Thus it is no surprise that one important concern in our model is data quality.

We know that claims handlers do a very good job updating claims information in the system, but unfortunately this does not always apply to claims that become annuities. The problem is that annuity claims are usually managed manually, As a result some information is saved in claims file documents not in the systems. The variable Salary is a clear example of this types of data errors (see Section 5.2.3).

In our sample data we could find numerous observations that seemed erroneous, cases where for example the yearly annuity compensation was higher than the yearly income of the victim or even cases where the yearly income was 0. Knowing that the victim's yearly salary is used as a starting point in calculating the victim's loss of income we could only assume that the data was wrong or not correctly updated in the system.

Our first thought was to exclude all observations that seemed to be incorrect, but excluding them from the model didn't give a better fit, primarily due to that other covariates also play a significant role. As a result we conclude that even if the some data may in some cases be erroneous, including them in the model gives more information than excluding them.

It is important to mention once more that the personal injury claims discussed in this paper are of a very long tailed nature. It is therefore even more important that the insurance companies are aware of the importance of possessing a data warehouse that holds loads of information (too much is never enough) that is regularly updated. Information about the victim's living conditions, career pattern and medical situation is essential if we want to succeed in using a GLM analysis based on individual information. Models' using individual information requires a minimum of data quality and history that are just not in place today. Data

staging environment has improved considerably in latest years, but there is a need to work closely with claims and IT so that the information required fit the actuary's necessities.

Finally we learned from this thesis that there is a basic need for actuaries to sit down with the claims - and IT departments and work on their differences, share information and agree on what information is required. We cannot spend our time regretting what was not done in the past; we can learn from it and add this knowledge to face new claims.

6.3 Children

As we have commented through this thesis the group that is most volatile and complicated to predict is Children. To really understand the dilemma that actuaries face when they try to estimate the cost for annuities when the injured is a child we can ask the following questions: How can we be able to estimate how much compensation a child will receive when the accident occurs? How can we expect stability in the analysis when the stability is based on the child's life development? Can we decide the career pattern the child is going to take? Can we assume that the child will become a dentist when he grows up? These questions describe in a simple way the difficult the actuaries and claims handlers have in setting the right cost.

Unfortunately we could not fit a GLM model to Children data due to the difficulties in using the variables mentioned in Section 5.2. To be able to do a proper GLM analysis for Children it is necessary that we gather information that currently is not available, as well as more data. Information regarding the child's background, such as the victim's family situation is essential, the victim's family yearly earnings plays a significant role in estimating the initial compensation for loss of income the child will receive (see Section 5.2.2). Another problem that we need to keep in mind is that we only have 31 observations for the group Children, mainly due to the fact that for this thesis decided to exclude all annuities that occurred prior to 1990 (see Section 5.2.2).

6.4 Conclusion and Outlook

In recent years the topic of claims reserving has involved a number of controversies, mostly related to which model provides the best reserve estimate. The answer to that question may never be fully settled, given that different models fit different problems or data sets. It is important to keep in mind that model selection is not an exact science, and that our purpose as actuaries is to search for the better ones. Claims reserving is a practical exercise and it is essential to understand and learn from the data before we select our model and not apply the same model assumptions in all situations.

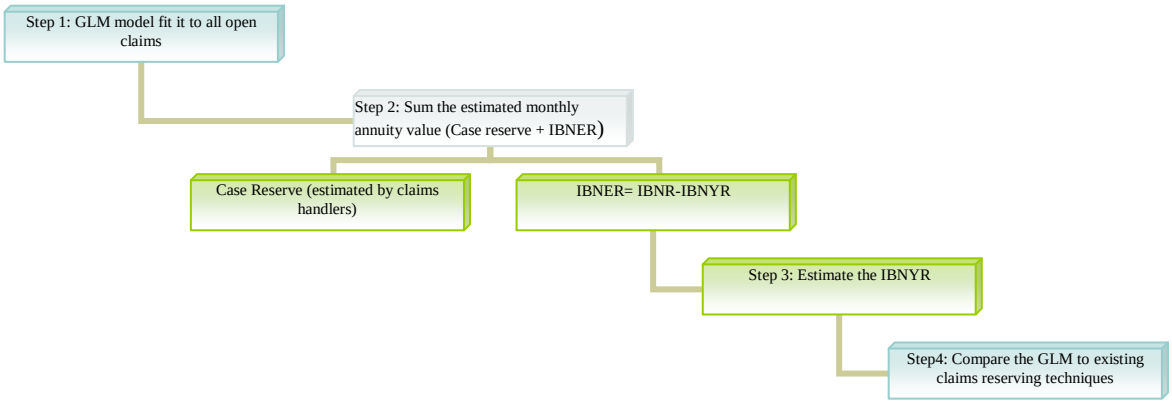
It is worth mentioning that even though several papers have been written in latest years about choosing an alternative model to Chain Ladder, there are still only used by a limited number of actuaries. The reasons are mostly based on understanding of such models but also data quality, data is simply not available neither in the level or the quality needed. However the main reason is that there is no general need from the insurance companies to use more complex models than chain ladder to estimate the company's loss reserves. But in recent years, requirements from Solvency II, has accelerated the discussion around the variability of the loss reserves and as a result the request for new alternative models has intensify.

6.5 Going Forward

When we started working with this paper our initial plan was to come up with a Generalized Linear Model based on the injured individual situation and applied it to claims that are open today. Unfortunately we underestimated the work and commitment this theme required and this could not be achieved in this paper. It is important to keep in mind that we have come a long way. The model presented in this thesis works well as a foundation for future developments of this subject as well as an introduction for the Swedish motor business.

The main purpose of this paper was to provide actuaries with an alternative method to calculate claims that are incurred but not reported (see Section 4.1), also known as IBNR. To achieve this there are some outstanding points, described in Figure 19 that needs to be completed.

Figure 19.



The first step is to calculate the monthly annuity cost for all open claims; we use the GLM analysis and applied to all open claims that are open today.

The second step is to sum up the predicted monthly annuity to get the estimated total cost for all open claims. This is a rather straight forward exercise where we estimate the life cycle of all individual using standard life insurance mathematics. The total cost should be calculated taking in account assumptions of future inflation and interest rates.

In Section 4.1 we got familiar with the term total reserve which is the sum of the Case Reserves plus IBNR. Case reserve is usually set by claims handlers and is considered to be a known item. In simple terms, the IBNR is the sum of two components: IBNYR+IBNER. The IBNER is defined as the estimate that allows for changes in the claims handlers estimate, and the IBNYR is the amount set aside for claims that are not yet reported to the company. The IBNER is given by the procedures in step 1-2.

The third step is to estimate IBNYR which can be done using several different reserving techniques. One method would be to use a frequency severity approach, where frequency can be estimated from the reporting pattern and severity could possibly be based on results from the generalized linear model.

As the procedure above is completed, we have an estimate of IBNR and its volatility for the data included in this paper. However there are still some items, described in the upcoming sections that are not reviewed that need to be assessed so that we get the complete picture of the annuity loss.

The final step is to test the method itself by comparing it to existing claims reserving methods, and measure volatility, results, etc.

Another approach of interest is to compute a prediction interval on the predicted claims reserve, in particular finding the high quantiles of the predicted distribution by using either asymptotic approximations of standard errors, or through resampling. See Björkwall et al (2009) and references therein.

6.5.1 The Post Retirement Annuity

In Section 3.3 we discussed that the compensation for loss of income is usually divided into two annuities, the pre retirement and the post retirement. The pre retirement annuity is used as a basis for the post retirement annuity and the estimation of this should be done similarly to how claims handlers calculate this.

In addition, the post retirement annuity should be appended to the pre retirement annuity so that we get the estimated cost for the total annuity. Subsequently we can follow the steps in Section 6.5 to get the overall loss reserve.

6.5.2 Children

Throughout this thesis, on several occasions, we have discussed the difficulties in modeling personal injury claim, especially in cases where the injured was a child when the accident occurred. The fact is that children annuities are the most volatile and expensive claims and there is vital need for the insurance company to model these types of claims. Considering the information presented in this paper, we can state that there is a need to analyze Children separately and there are numerous alternatives we can use to come up with a model and apply it to reserving.

The first suggestion is to manually collect data and include all variables needed (discussed in Section 5.2) for a larger dataset, then fit a GLM and apply this to open claims. Another more pragmatic approach would be to use a severity and frequency approach where severity would be taken as an average of close claims for children.

6.5.3 Remaining Issues

Finally, for future development it is recommended to include important information that is not included in this paper such as: income disability percentages, worker compensation, reopening of claims and social insurance compensation (mentioned in Section 5.3). We believe that we will improve the fit considerably if all these factors are taking into account.

7. References

Brockman, M.J. and Wright,T.S (1992). Statistical Motor Rating : Making efficient use of your data. Journal of the Institute of Actuaries, vol.119, 457-458

Dahl P. Introduction to Reserving, Matematisk Statisktik , Stockholms Universitet, correct edition.

Olsson, U. (2002) : Generalized Linear Models, An Applied Approach, Studentlitteratur Ab.

McCullagh,P, Nelder, J.A. (1989). Generalized Linear Models. 2nd edition, Chapman and Hall, Boca Raton.

Verrall, R.J. (2000). An Investigation into Stochastic Claims Reserving Models and the Chain-Ladder Technique. Insurance: Mathematics and Economics, vol.26, 91-99

Ohlsson, E. Johansson, B.(2004) Prissättning inom sakförsäkring med Generaliserade linjära modeller. Kompendium, Version 3.3 Matematisk Statisktik, Stockholms Universitet.

Mack, T. and Venter, G. (2000). A comparison of stochastic models that reproduce the Chain Ladder reserve estimates, Insurance Mathematics and Economics, 26, pages 101-107.

Taylor, G. (2000). Loss reserving - an actuarial perspective. Buston: Kluwer Academic Press.

Björkwall, S.,Hössjer, O. and Ohlsson, E. (2009). Nonparametric and parametric bootstrap techniques for arbitrary age-to-age development factor methods in claims reserving. Scandinavian Actuarial Journal 2009(4), 306-331.

Trafikförsäkringsförening: www.tff.se

TrafikSkadenämnden: www.trafikskadenamnden.se

Försäkringskassan: www.forsakringskassan.se

Wikepedia: www.wikepedia.org

A. Appendix

A.1 THE ML-ESTIMATION OF B AND Ø IN GLM

To find the ML - estimate of β it is essential that we are familiar with the likelihood function defined as the probability of a sample dataset $y_1,\ldots,y_n$ given the probability densities $f_{Y_i}(y_i;\theta_i,\phi)$ and viewed as functions of the regression parameters $\beta_o,\beta_1,\ldots,\beta_k$ and the dispersion parameter ϕ.

We begin by writing the mathematical expression of the log likelihood function

$$\ell(\beta_o,\ldots,\beta_k,\phi) = Log\left(\prod_{i=1}^{n} f_{Y_i}(y_i;\theta_i,\phi)\right)$$

$$= Log\left(\prod_{i=1}^{n} \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(y_i,\phi,\omega_i)\right\}\right)$$

$$= \frac{1}{\phi}\sum_{i=1}^{n}\omega_i(y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{n} c(y_i,\phi,\omega_i) \qquad \text{(A.1)}$$

To continue it is convenient to use the score function known as the partial derivate of the log likelihood with respect to some regression parameter β_j.

$$\frac{\partial}{\partial\beta_j}\ell(\beta_o,\ldots,\beta_k,\phi)$$

With help of the chain rule the score function can now be written as

$$\frac{\partial}{\partial\beta_j}\ell(\beta_o,\ldots,\beta_k,\phi) = \sum_{i=1}^{n}\frac{\partial L}{\partial\theta_i}\frac{\partial\theta_i}{\partial\beta_j} = \frac{1}{\phi}\sum_{i=1}^{n}\omega_i(y_i - b'(\theta_i))\frac{\partial\theta_i}{\partial\beta_j} \qquad \text{(A.2)}$$

We know from section 4.1 that $\mu_i = b'(\theta_i)$ and that $g(\mu_i) = \eta_i = \sum_j x_{i,j}\beta_j$ (cf. (4.7)). We use this information and write the score function as

$$\frac{\partial}{\partial\beta_j}\ell(\beta_o,\ldots,\beta_k,\phi) = \frac{1}{\phi}\sum_{i=1}^{n}(\omega_i(y_i - b'(\theta_i)))\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_{i_i}}\frac{\partial\eta_i}{\partial\beta_j} \qquad \text{(A.3)}$$

We can in addition use the following identities from Section 4.3.1

$$\mu_i = b'(\theta_i)\,,\ \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i)\,,\ \eta_i = g(\mu_i) \text{ and } V(\mu_i) = b''(\theta_i) \tag{A.4}$$

and by substituting this in

$$\frac{\partial \theta_i}{\partial \mu_i} = \left[\frac{\partial \mu_i}{\partial \theta_i}\right]^{-1} = [b''(\theta_i)]^{-1} = \frac{1}{V(\mu_i)} \tag{A.5}$$

Further

$$\frac{\partial \mu_i}{\partial \eta_i} = \left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-1} = [g'(\mu_i)]^{-1} = \left[\frac{1}{g'(\mu_i)}\right]. \tag{A.6}$$

Finally we use $\eta_i = \sum_j x_{i,j}\beta_j$ to get the derivate

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{i,j} \tag{A.7}$$

By substituting (A.4)- (4.7) into (A.8) we get

$$\frac{\partial}{\partial \beta_j}\ell(\beta_o,\ldots,\beta_k,\phi) = \frac{1}{\phi}\sum_{i=1}^{n}\left(\frac{\omega_i(y_i-\mu_i)}{V(\mu_i)g'(\mu_i)}x_{i,j}\right) \tag{A.8}$$

The maximum likelihood estimate of $\beta_o,\ldots,\beta_k$ is found by setting the $k+1$ score functions to zero, i.e.

$$\sum_{i=1}^{n}\left(\frac{\omega_i(y_i-\mu_i)}{V(\mu_i)g'(\mu_i)}x_{i,j}\right) = 0 \qquad j = 0,\ldots,k+1. \tag{A.9}$$

Finally, the dispersion parameter ϕ is estimated by using for example

$$\hat{\phi} = \frac{\chi^2 \phi}{n-k-1} \tag{A.10}$$

where

$$\chi^2 = \sum_{i \neq 1}^{n} \frac{w_i (y_i - \hat{\mu}_i)^2}{\phi V(\hat{\mu}_i)} \tag{A.11}$$

Is Pearson's dispersion estimate, $\hat{\mu}_i$ is the estimated μ_i, calculated from the ML-estimates and $V(\mu_i)$ is the estimated variance function.

A2. Salary Group Classification

Code	Profession/Occupation	Occupation Group
01	Building worker	4 - Industrial/Construction/Police
02	Farmer	4 - Industrial/Construction/Police
03	Teacher/Day-care worker	2 - Public employee
04	Nursing staff	2 - Public employee
05	Factory worker	4 - Industrial/Construction/Police
06	Transport worker	4 - Industrial/Construction/Police
07	Shop assistent	3 - Shop assistant/Other manual workers
08	Office staff/Salaried employee	5 - White-collar worker
09	Other manual labour work	3 - Shop assistant/Other manual workers
10	Other non manual labour work	3 - Shop assistant/Other manual workers
11	Police/Service Man/Customs workers	4 - Industrial/Construction/Police
12	Driver	3 - Shop assistant/Other manual workers
13	unemployed/Student/Non labour active	1 - Unemployed/Student/Non labour active

A3. Model Selection including full model without gender and child group

Tables 17: Analysis of Parameter Estimates of Full Model without Gender and Child Group

Parameter		DF	Estimate	Error	Square	Pr>Chi
Intercept		1	10.0686	0.1855	2945.33	<.0001
Age_group	Adult	1	-0.2785	0.0825	4.37	0.0365
Age_group	Mature Adult	0	0.0000	0.0000	11.39	0.0007
Age_group	Young Adult	1	-0.5041	0.1200	.	.
Salary_group	low	1	-0.6256	0.1123	17.65	<.0001
Salary_group	Medium	0	0.0000	0.0000	31.04	<.0001
Salary_group	High	1	-0.4812	0.1386	.	.
MedDisPer_group	Low	1	-0.0919	0.1567	12.06	0.0005
MedDisPer_group	Medium Low	1	-0.1709	0.1999	0.34	0.5574
MedDisPer_group	Medium High	0	0	0	0.73	0.3927
MedDisPer_group	High	0	0.0000	0.0000	.	.
occupation_group	Unemployed/Student/Non Labour active	1	-0.2487	0.1142	4.74	0.0295
occupation_group	Public employee	1	-0.2996	0.0965	0	0.0019
occupation_group	Shop assistant/Other manual workers	1	-0.1729	0.0832	4.32	0.0376
occupation_group	Industrial/Construction/Police	1	-0.0828	0.0754	1.21	0.2717
occupation_group	White-collar workers	0	0.0000	0.0000	.	.
Scale		1	1.9894	0.0912		

A4. CHILDREN GLM OUTPUT

A GLM analysis was done for the Children data where we included all variables mention in Section 5.2 with the exception of Gender. We assumed that the dependent variable was Gamma distributed with a log link function applied to 31 observations.

The Proc Genmod analysis gave the following Output.

Table 22: Class Level Information

Class	Levels	Values
Salary_group	2	7 - Low 8 - Medium
MedDisPer_group	4	10 ╦ Low, 11 - Medium Low, 12 - Medium Hig, 13 - High
occupation_group	3	14 - Unemployed/Student/Non labour active, 16 - Shop assistant/Other manual workers, 17 - Industrial/Construction/Police

Table 23: Criteria for Assessing Goodness of Fit Group Children

Criterion	DF	Value	Value/DF
Deviance	24	5.5281	0.2303
Scaled Deviance	24	31.8929	1.3289
Pearson Chi-Square	24	4.4765	0.1865
Scaled Pearson X2	24	25.8259	1.0761
Log Likelihood		-300.3379	

Table 24: Analysis of Parameter Estimates of Full Model without Gender

Parameter		DF	Estimate	Error	Square	Pr>Chi
Intercept		1	9.8112	0.3268	901.31	<.0001
Salary_group	low	1	-0.7702	0.3589	4.60	0.0319
Salary_group	Medium	0	0.0000	0.0000	.	.
MedDisPer_group	Low	1	-0.1302	0.2644	0.24	0.6225
MedDisPer_group	Medium Low	1	-0.0693	0.2418	0.08	0.7744
MedDisPer_group	Medium High	1	0.0776	0.3016	0.07	0.7969
MedDisPer_group	High	0	0.0000	0.0000	.	.
occupation_group	Unemployed/Student/Non Labour active	1	0.2411	0.4837	0.25	0.6182
occupation_group	Shop assistant/Other manual workers	1	-0.7813	0.4745	2.71	0.0997
occupation_group	Industrial/Construction/Police	0	0.0000	0.0000	.	.
Scale		1	5.7692	1.4250		

Table 25: LR Statistics for Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
Salary_group	1	4.76	0.0292
MedDisPer_group	3	0.74	0.8637
occupation_group	2	9.29	0.0096