Assessment of Risks

in Automobile Insurance

Esther D. Gillissen
June 2001

Assessment of Risks

in Automobile Insurance

by: Gillissen, Esther D.

21-06-2001

At the department of Mathematics
Rijksuniversiteit Groningen,
Groningen, The Netherlands,

June, 2001.

Supervisors at the university:
Prof. dr. H. G. Dehling
Prof. dr . W. Schaafsma

Supervisors at the insurance company:
Mr. D. van Dyck
Mr. K. Verdoodt

A thesis submitted in fulfillment of the requirements for
the degree Master of Science
at the Rijksuniversiteit Groningen.

Preface

This thesis is submitted in fulfillment of the requirements for the degree Master of Science with a specialization of Statistics at the Rijksuniversiteit Groningen. The contents of this thesis is a product of research, during a six month period at the non-life department of an insurance company in Brussels, Belgium. The name of this insurance company cannot be mentioned because the results of this thesis are real-life business cases without any manipulation.

During my graduation in mathematics, at the same university, I was assigned as a student assistant to statistical courses. The resulting experience with statistics has opened my eyes: I started to appreciate its applicability. That is why I chose to graduate in statistics after my graduation in mathematics.

During the six month period of research I enjoyed my work and learned many things. I learned not only how to perform my research, but also how to work in a business environment and I became familiar with the real practice in car insurance. I want to thank all my colleagues for their support and their help, especially Mr. D. van Dyck, who helped me not only with my research, but also supported me in building interpersonal skills, and Mr. K. Verdoodt for helping me during my research. They provided support during a very busy period. Further, I want to mention Ms. E. de Cnodder and Mr. F. Achraf with whom I could discuss the statistical part of my research,.

Other people I would like to thank are Henrickus and his parents for their support and help even in the more difficult periods during my research. I would like to thank Mr. S. Keinälä and Mrs. L. J. Gustafson for proofreading this study.

Last but not least I would like to thank my supervisors during my research, Prof. Dr. H. G. Dehling and Prof. Dr. W. Schaafsma. Prof. Dr. H. G. Dehling was not only helpful during this period, but he also aided me during the studies with some personal problems.

I hope that the methods and results to be presented are of interest both to statisticians and actuaries.

Brussels, April 2001

Esther D. Gillissen

Abstract

The competition between insurance companies centers around the quantification of the premium to be asked of the individual client. The insurance company will avoid the acceptance of clients where the risk (expected loss) exceeds the (expected) premium.

Here the risk is far most uncertain. It consists of two parts: (1) the probability that a claim is made, (2) the distribution of the claim (or rather its expectation), under the condition that a claim is made (and paid, of course). To be precise, we concentrate the attention on a period of 3 years.

Logistic regression is used (in Chapter 3) to specify the probability of causing any damage leading to (the payment of) a claim, during this period, and as a function of explanatory variables like 'bonus malus' grade, etc. The theory of Generalized linear Models is used to study the conditional distribution of the claims (to be) paid as a function of the explanatory variables considered also before. The risk (expected loss) of interest is, of course, the product of the probability of causing a damage to be paid by the company and the conditional expectation of the damage caused, given its strict positivity.

An interesting phenomenon is that the damage claimed and paid depends on the characteristics which affect the probability of claiming a damage: damages caused by the less risky clients are, on the average, higher than those caused by the risky ones. This reflects the policy of clients not to report a damage if this might have their 'bonun malus' discount.

Samenvatting

De concurrentiestrijd tussen verzekeringsmaatschappijen draait om het bepalen van de te vragen premie aan de individuele klant. De verzekeringsmaatschappij wil dan ook de acceptatie van klanten, met een hoger risico (verwacht verlies) dan de (verwachte) premie, vermijden.

Hierbij is het risico heel erg onzeker. Het bestaat uit twee delen: (1) de kans dat een schade geclaimd wordt, (2) de verdeling van de claim (of liever de verwachtingswaarde van de claim) onder het gegeven dat een schade is geclaimd (en betaald natuurlijk). Om precies te zijn, concentreren we ons op een periode van drie jaar.

Logistische regressie wordt gebruikt (in Hoofdstuk 3) om de kans op het veroorzaken van een schade, welke leidt tot (de betaling van) een claim, te bepalen, en als functie van verklarende variabelen zoals bonusmalus graad, enz. De theorie van Generalized Linear Models is gebruikt om de conditionele verdeling van de betaalde (of nog te betalen) claims als functie van dezelfde verklarende variabelen, te bestuderen. Het risico, welke van belang is, is het product van de kans op het veroorzaken van schade en de conditionele verwachting van de veroorzaakte schade, gegeven dat deze strikt positief is.

Een interessant fenomeen is dat de geclaimde betaalde schade afhangt van de karakteristieken die de kans op het claimen van een schade beïnvloeden: schades veroorzaakt door klanten met een lage kans op schade zijn gemiddeld genomen, hoger dan de schades die veroorzaakt worden door klanten met een grote kans op schade. Dit geeft de manier van claimen van schades weer, van klanten die een lage bonusmalus graad hebben, om schades niet te claimen als het schadebedrag niet hoog genoeg is.

Table of Contents

Appendix IV

Introduction

Basic Facts about Risk

Like any business company, An insurance company wants to make a profit. This is the case, during a certain period, if the premiums paid by the customers exceed the amount of money (to be) paid for the damages they have caused during the period, plus the additional costs (administration, acquisition, etc.) The profit during one period may differ considerably from that during another period. These fluctuations have statistical and systematic components and are difficult to predict in detail though the insurance company is mainly interested in the average over the entire population. To improve the average profitness of this population, the company will try to avoid risky clients. Note that insurance companies exist due to willingness of people to pay premiums which exceed the expected amount of damage to be claimed. The reason is not only that they want to avoid to get hove but also to avoid not to be able to account for a damage if they themselves were the cause. The word 'moral expectation' is of interest in this respect.

1.1 Risk Aversion of People

People want to avoid catastrophic losses, such as that of their house burning down, excessive costs due to disease, a car accident possibly with medical injury, etc. Note that the word 'risk' is used here in the sense of a catastrophic loss of money. In mathematical statistics the word 'risk' is defined as loss to be expected. This is a theoretical concept used in the mathematical analysis.

People, who insure their risks, are willing to pay a higher premium than the net premium. The reason for this is that most people are risk averse. This means that these people want to protect themselves against losses, even if the risk (expected loss) is considerably less than the premium. The inconsistency of human behavior is manifest if a person buys insurance for his car, health care, etc., and, at another moment, goes to the casino.

Everyone has his own 'utility function'. A standard principle in the theory of economics behavior (von Neymann-Morgenstein) and in the theory of statistical decision functions (Wald) is that expected utilities are maximized.

This is only a theoretical principle because the expectation depends on the largely unknown probabilistic structure and on the specifications of a utility function. A theoretical elaboration is as follows. Let W denote the wealth (in BEF) of a person at the end of a insurance period, say 3 years. Let P denote the premium to be paid and let X denote the loss to be incurred during the period. Both W, P and X are (partially) at the beginning of the period. Let $u: \mathbb{R}^+ \rightarrow \mathbb{R}^+$ denote a person's utility function and suppose that it is known at the beginning of the period. The expected utility in the case of not buying insurance $Eu(W - X)$ has to be compared with the expected utility $Eu(W - P)$ in the case of buying. If

$$Eu(W - X) > Eu(W - P)$$

then the 'rationality principle' suggest not to buy insurance. If

$$Eu(W - X) < Eu(W - P)$$

then it is advantageous 'on the average' to buy security by paying the premium.

For the insurance company, the same can be done for accepting the client. An insurance company accepts a client if

$$E[U(W+P-X)] \geq U(W)$$

yields. Where $U: \mathbb{R}^+ \rightarrow \mathbb{R}^+$ stands for the utility function of the insurance company.

Though interesting, the discussion above is largely theoretical. An empirical basis has to be found for the specification of the (person's) utility function and (joint) distribution of (W, X, P). Most of the uncertainty is with respect to X (the loss to be incurred to damage, possibly 0).

Therefore, the insurance company needs to calculate their premiums in such a way that the insurance company does not have a large probability for (great) loss and that the premiums are not too high. An insurance company needs to take many different variables in consideration in order to calculate the premium. A person with a high risk, a large probability of claiming damage, has to pay a higher premium than a person with a small risk of damage. But how can you calculate this by taking into consideration the following three risks of an insurance company:

- product risk,
- cost-effectiveness risk and
- risk of losing a client.

To take these risks into consideration cards based on credit scoring are in use. The purpose of these score cards is assist in calculating how much a customer has to pay. In car insurance such score cards are less common than in mortgage and other areas of allowing credits. That is why the next section is included.

1.2 History of Credit Scoring

Credit scoring is a process where some information about a credit applicant or a credit account is converted into numbers that are combined to form a score. This

score is a measure of creditworthiness[1]. As the term 'creditworthiness' might be considered too personal, one prefers the less specific, somewhat doubtful and opposite term 'risk'. A person's credit application is rejected if the associated risk is too high.

The history of credit scoring is short and many good ideas from the past may have been lost. There is no doubt that many credit managers attempted at various times in the past to reduce their procedure to some sort of numerical form. But very little of a practical nature was reported before the Second World War, possibly because some secrecy was profitable for the company.

When the war ended, a number of events took place that enhanced the development of credit scoring. Computers became available for commercial purposes, the new field of Operations Research encouraged the quantitative examination of all sorts of business situations, and more and more people became adept to modern statistical methods. In addition, the end of the war brought about enormous changes in the economies of almost all of the countries of the western world, bringing all sorts of new problems to business management. The field of credit was no exception.

The growth of the use of credit scoring has been a component of the change in American business brought about by an increased awareness of the value of scientific analysis of problems of all sorts, not only those associated with credit. So management has become increasingly aware that technical methods must be used. Management has also been aware that competition is a broad term. Enterprises compete not only with pricing, variety and quality of their merchandise, but also in the manner in which they operate their business.

Much of the initial effort in the process of introducing credit scoring was directed towards finance companies, allowing loans and mortgages. This was because the

[1] Creditworthiness is a characteristic of an individual that makes him or her a suitable candidate for the extension of credit while someone who is not creditworthy is, conversely, unsuitable for credit.

problems of management and control were particularly acute in that area, at least in the view of the developers of the scoring systems. The finance companies had well-entrenched operations and recognized no pressing need for change. However, these companies slowly began to consider the ideas of scoring and to adapt them, at least as a component of the credit decision process.

Figure 1-1 Example of a Credit Score Card

Years on job	Less than 6 months -9	6 months to 1 yr 6 months 0	1 yr 7 months to 6 yrs 8 months 6	6 yrs 9 months to 10 yrs 5 months 13	10 yrs 6 months or more 25	Blank 0	
Own or rent	Own or buying 15	Rent -5	All other 2	Blank 0			
Banking	Checking account 5	Savings account 0	Checking and savings 14	None -17	Blank 0		
Major credit card	Yes 10	No -6	Blank 0				
Occupation	Retired 21	Professional 16	Clerical 7	Sales -2	Service -8	All other 7	Blank 0
Age of applicant	18 to 25 -3	26 to 31 -5	32 to 34 0	35 to 51 4	52 to 61 12	62 and over 18	Blank 0
Worst credit reference	Major derogatory -15	Minor derogatory -4	No record -2	One satisfactory 9	Two or more satisfactory 18	No investigation 0	

Source: Lewis (1992)

Nowadays, a scorecard is a table listing the characteristics that provide predictive information in the scoring system, the attribute of each characteristic, and the score points associated with each attribute.

For making a decision to accept or reject a client, at first, to find a sum which is called 'the score' all the score points assigned to for the appropriate attribute of each characteristic are added. Secondly the score needs to be compared with a cutoff value[2]. If the score lies above the cutoff value the client will be accepted and if the score lies under the cutoff value the client will be rejected.

For making a score card in the car insurance business, one needs information about the actual distribution of the probability of damage and the amount of claimed damage. Otherwise, one is not able to make an assessment of the expected amount of claimed / paid damage. This assessment is necessary to make a score card, because if a person has a high probability of damage and a high expected value of claimed damage, the insurance company does not want that person in its portfolio.

1.3 Theoretical Distribution of the Amount of Damage

A theoretical derivation of a probability distribution for the amount of damage does not exist, and comparative studies concerning this distribution are scarce. One could use a Poisson distribution, but when the number of accidents is large, it will be very complicated to use it in practice. Here are some remarks regarding this problem.

1. The distributions, which have the property that the sum of n random variables will follow the same probability distribution with modified parameters depending on n, will be recommended for reason of simplicity

[2] The score below which applications are either automatically rejected or are recommended for rejection. This score has been preset by analyzing the data.

Examples of such distributions are the gamma and the inverse Gaussian distributions.

2. Limiting distributions for small and for large number of accidents can be considered. These distributions take the form of a Bernoulli distribution or a standard normal distribution.
3. One can propose a model for deriving the distribution without making any assumptions about the type of the distribution. This problem could be formulated and solved as a regular Markov process.

A distribution which is often used in theoretical work is the exponential distribution, for reasons of simplicity. The sum of exponentially distributed random variables is gamma distributed, therefore it is very useful in deriving the distribution of the total amount of damage. Only this distribution is not skew enough: it underestimates the probability of large values of the amount of damage. Instead of the exponential distribution, the gamma distribution or the inverse Gaussian distribution could be used. For the large amount of damage the Pareto distribution can be used. This distribution has a long right tail and is therefore theoretically an interesting distribution. Also the Box-Cox normal distribution, Weibull, the Burr and the log-t distribution are distributions which can be used for the estimation of the amount of damage.

One sees that a real theoretical base for the distribution of the total amount of damage does not exists. For the most part it depends on the data one uses. So one just need to look at the results of the models to determine which distribution suits best.

1.4 Overview of the Report

The scope of this research is to create a useful acceptance policy for automobile insurance. This study has been complied by only using the information from the portfolio of this insurance company, thus one has a *select* sample of Belgian people who need automobile insurance. To draw up a useful acceptance policy

one needs to find out which variables relating to the client give us information indicating the risk of the client.

The insurance company thinks that it is important to take into consideration the area in which people live when accepting a new client. A consulting firm and another insurance company created a score based solely on this information. Before the insurance company will make a score of this information themselves, they would like to know if these two scores are indeed of any interest for indicating the risk of a policy. This will be examined separately in Chapter 2. In the subsequent chapters these scores will also be taken into consideration but more personal client information shall also be used.

Chapter 3 will try to find the variables which tell us something about the risk of a client by using multiple logistic regression. Chapter 4 will deal with looking at the premium of a client, where normally a high risk would pay a higher premium than a good risk. This premium is calculated by using the expected amount of damage and the probability of damage. Since these models which are calculated in Chapter 3 and Chapter 4 are not easy to use for determining the acceptance of a policy score card to show how a client is accepted will be drawn up in Chapter 5. The last chapter is a short review of this report with some remarks on this research.

The Introduction / Data / Method / Results / Discussion mode has been chosen to present our investigation. In the method section a short description of the used theory is given, for additional information please refer to the appendices.

Chapter 2

Using Region and Geo-Model score to Predict Causing Damage

Zip codes divide Belgium into 1150 neighborhoods with 3500 households in each. In contrast, the NIS-code, developed by the National Institute of Statistics, divides Belgium into 20,000 neighborhoods with 210 households in each. The region score, based on the zip codes of Belgium, has been developed by an insurance company. The geo-model score developed by a consulting firm for a credit company is based on this NIS-code. For both scores higher scores involves higher risks for the financial companies. The question one wants to examine is: "Are these scores of interest in risk prediction and which score is a best?"

2.1 Introduction

Financial service companies have various options to extend or strengthen their portfolio. One of these options is aiming at acquisition via a direct-mail campaign. Other options are to offer premium reductions to people from low-risk areas. What is a low-risk area? We consider two attempts to arrive at a quantification: the region score composed by an insurance company and the geo-model score made up by a consulting firm on behalf of a credit company. The region score is based on the zip-code of Belgium. This partition divides Belgium

into 1150 neighborhoods of 3500 households each, and has 5 categories: class 1 comprising the zip codes with the smallest risk, to class 5 for zip codes which are worst. The division of zip codes over these classes was based on professional knowledge rather than a computational analysis. In contrast, the geo-model score is based on a statistical analysis of data of a credit company. The consulting firm came up with a specific geo-model score which runs from 0.0 to 5.0 using the NIS-codes, which divide Belgium into 20,000 neighborhoods of 210 households each. This geo-model score tells us something about the risk a credit company incurs if it allows credit to people in a neighborhood with the score given. If the average risk is high the geo-model score is large, and if the risk is low the score is small.

The problem is to investigate whether these 'predictor variables' are of commercial interest. This depends on the financial product to be offered (automobile insurance, fire insurance, liability insurance and accident insurance). This research is about automobile insurance.

The major problem is the assessment of costs and benefits (for clients utilities are different and that is why financial products can find a market).

2.2 Data

For automobile insurance, data from the past three years is available for clients insured. Administrative characteristics, such as client number, address, policy number, and a dossier number if damage has been claimed, are transformed into the explanatory (or predictor) variables, par example:

x_1 = region score $\quad (1, 2, 3, 4, 5)$
x_2 = geo - model score $\quad (0, 0.1, 0.2, \ldots, 4.8, 4.9, 5)$
x_3 = geo band $\quad (0, 1, 2, \ldots, 7, 8, 9)$

The geo band is an aggregation of the geo-model score, according to a definition given by the consulting firm. When the geo-model score is 0, the geo band is also 0. The geo band continuous: 1 if the geo-model score is in $\langle 0,0.2\rangle$, 2 if the geo-model score is in $[0.2,0.3\rangle$, 3 if the geo-model score is in $[0.3,0.5\rangle$, 4 if the geo-model score is in $[0.5,0.7\rangle$, 5 if the geo-model score is in $[0.7,1.5\rangle$, 6 if the geo-model score is in $[1.5,2.2\rangle$, 7 if the geo-model score is in $[2.2,3\rangle$, 8 if the geo-model score is in $[3,5\rangle$, 9 if the geo-model score is 5. The geo band will be used, instead of the geo-model score, since it is more convenient in practice.

Because these scores are very important for the insurance company, more information about these scores is necessary for the analysis. Therefore, in the following part of this section, some information is given about these scores.

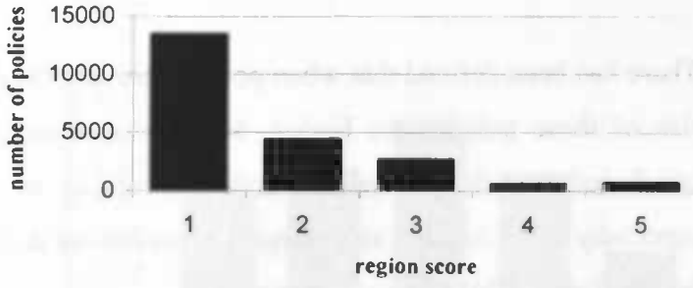


Figure 2-1 Frequency Distribution of Region Score in a Population of Car Policies

According to Figure 2-1, one must keep in mind that the policies are not equally distributed in the score bands of the region score. In band 1 there are more than 60% of the total number of policies of the car portfolio, while bands 4 and 5 together have only 6% of the portfolio. According to the region score, the market of this insurance company is that part of Belgium with the lowest risk.

Looking at Figure 2-2, it is obvious that the policies are not distributed equally in the bands of the geo band. So for this score there has to be kept in mind that bands 7, 8 and 9 have very few policies in relation to the other bands (together only 4% of the portfolio). According to the geo band, the market of the insurance company is located in that part of Belgium with low risk and normal risk.

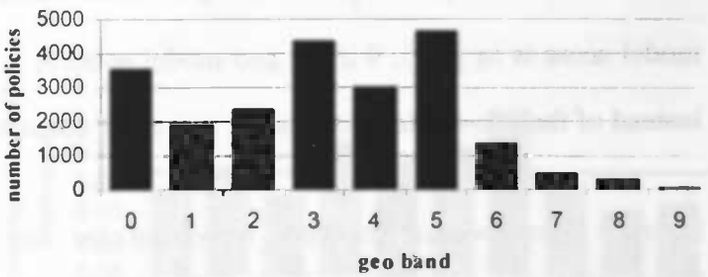
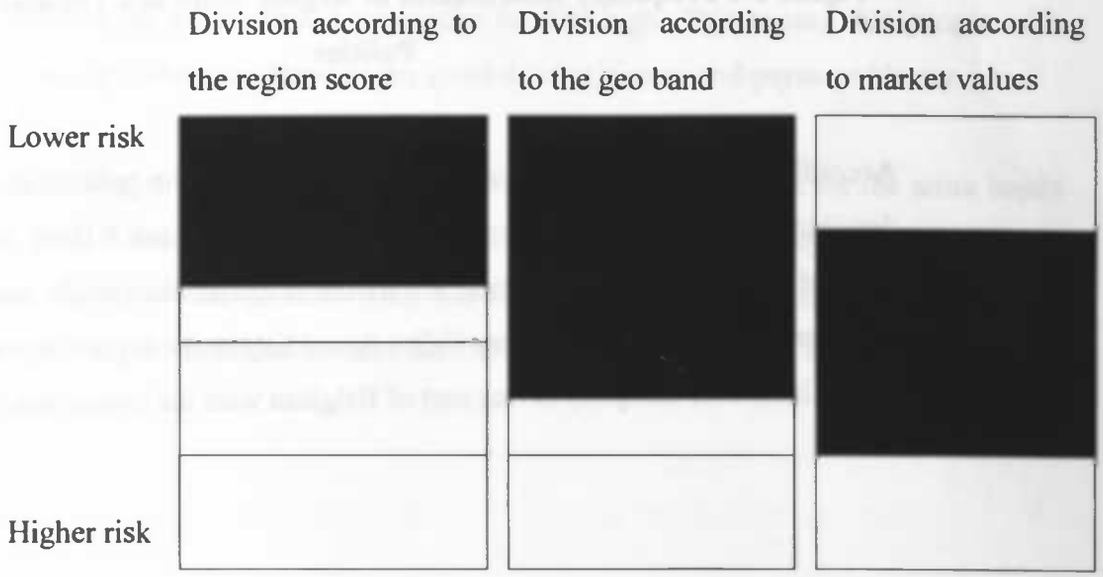


Figure 2-2 Frequency Distribution of Geo Band in a Population of Car Policies

There has been defined that when policies are in a band with a higher number, the risk of these policies are higher. So for this insurance company, there can be concluded that the acceptance policy according to these two scores is good, especially when looking at the region score, or the definitions of the risks must be very different. This can be illustrated as:



The division of the market value of risks, is the division the insurance company uses when they compare their portfolio with the portfolio of other financial companies. The definition of low risks and high risks which has been used for this division is not known. Probably it is an overall definition for all the financial companies given by the National Institute for Statistics. By comparing the conclusions of the scores over the division of the risks in the portfolio with this definition, there can be concluded that the definition of risk according to the geo-model score comes closer to the overall definition than the definition according to the region score. This is not unusual because the geo-model score covers almost the whole market. The region score only covers the best part of the market and gives us probably a too nice of a picture of the portfolio. This can mean two things; or the divisions of the zip codes over the classes has not been done correctly or the definition of the risks according to the region score is very different from the overall definition.

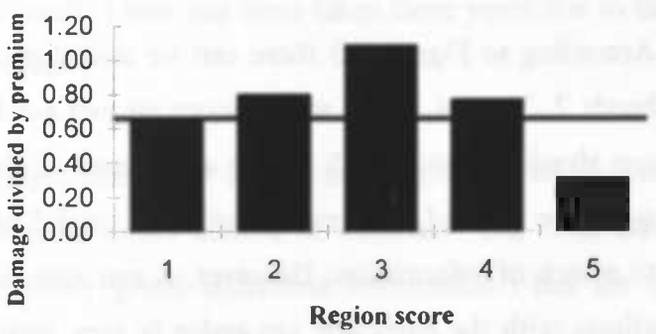


Figure 2-3 Economical Value of the Region Score

Figures 2-3 and 2-4 represents the 'total amount of damage paid divided by the total premium received (=D/P)' for each score band. The straight line in this figures represent the market value of D/P (=0.66). The value of D/P means that 66% of the premium is used for the amount of damage and the other 34% is for commission, salaries, administration costs, etc. When the D/P for the score bands are larger than the straight line, the insurance company does not have enough

money to pay the amount of damage and the commission, salaries, administration costs, etc. In order to arrive at this value there has not been taken into account the results of investments of the insurance company over a long time, since there will only be looking back over the past three years. To make sure that the company has enough money, they need to increase their premiums or make a better acceptance policy.

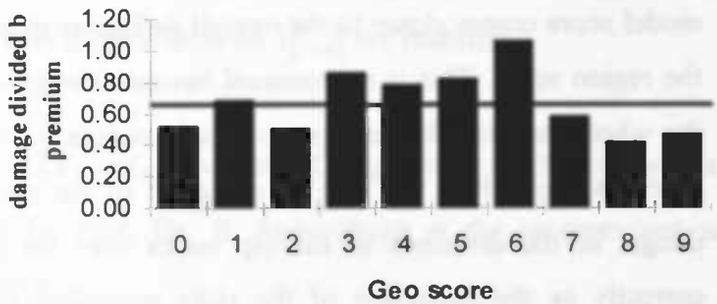


Figure 2-4 Economical Value of the Geo Band

According to Figure 2-3 there can be said that the policies which are in score bands 2, 3 and 4 of the region score are not good for the insurance company. If one should use this result for the acceptance of clients, almost 40% of the clients would be rejected. The reason that score band 5 has such small D/P could be due to a lack of information. However, it can also be that the acceptance policy of clients with the particular zip codes is very high. This means that people with those zip codes must have a very good driving history and payment history.

According to Figure 2-4, the same interpretation can be given for the score bands 1, 3, 4, 5 and 6 of the geo band. If all these clients would be rejected, only 40% of the clients would be left. What has been concluded about score band 5 of the region score, can also be said for the score bands 7, 8 and 9 of the geo-model score.

2.3 Method

We concentrate the attention on the distribution of

X_1 = region score

X_2 = geo - model score

X_3 = geo band

Y = whether ($Y = 1$) or not ($Y = 0$) a claim has been paid

in the population of car policies.

The definition of Y is based over the past three years. This means that if a claim has been paid to a client, sometime during the past three years, this client has a high risk. A claim has been paid if a client has been at fault in an accident and this amount could not be recovered from another person or if the client did not want to pay the amount of damage. The reason for choosing this definition is that, for approving a client, the insurance company looks at the damage history of a client over the past five years. If a client has been at fault, the probability of approving that client is very small. There has been taken three years due to the data of five years was not complete.

We note that the population is not exactly equal to that of all car policies and that, hence, some systematic errors cannot be excluded. Our 'risk analysis' (risk = 'expected loss, possibly given additional information') has the conditional or posterior probabilities

$$P(Y > 0 | X_h = x_h) \qquad (h = 1,2,3)$$

in its focus. The posterior expectations

$$E(Y | X_h = x_h, Y > 0)$$

are of interest as well. Note that

$$E(Y | X_h = x_h) = P(Y > 0 | X_h = x_h) E(Y | X_h = x_h, Y > 0)$$

because $E(Y | X_h = x_h, Y = 0) = 0$.

Here are the issues to be settled:

1) does X_h be of predictive interest in the sense that the null hypothesis $H_0 : P(Y > 0 | X_h = x_h) = P(Y > 0)$ (independently of x_h) is rejected.

2) do differences exist between the explanatory variables of their 'predictive values' are considered?

Note that we concentrate the attention on whether Y differs from 0. A more detailed analysis of the entire conditional distribution of the ('mixed' = 'continuous with discrete component at 0') response variable Y may also be of interest, especially if the effects on $1_{\{Y>0\}}$ are manifest.

To study $P(Y > 0 | X_h = x_h)$ a variety of suggestions can be followed, e.g.

1) (suggestion by Prof. Dr. B. Spanenburg at the presentation) use a neural-network approach

2) (suggestion by Prof. Dr. W. Schaafsma after the presentation) use Bayes's theorem where

$$P(Y > 0 | X_h = x_h) = \frac{P(Y > 0, X_h = x_h)}{P(Y = 0, X_h = x_h) + P(Y > 0, X_h = x_h)}$$
$$= \frac{P(Y > 0) P(X_h = x_h | Y > 0)}{P(Y = 0) P(X_h = x_h | Y = 0) + P(Y > 0) P(X_h = x_h | Y > 0)}$$

and the 'prior probabilities' $P(Y = 0)$ and $P(Y > 0)$ follow from the marginal distribution of Y while the class-specific distributions $\mathcal{L}(X_h | Y = 0)$ and $\mathcal{L}(X_h | Y > 0)$ require a separate investigation.

3) (suggested by K. Verdoodt and others) use (linear) logistic regression

4) (suggestion by Prof. Dr. W. Schaafsma) compose a function f, possibly with the requirement of non-decreasingness, such that the relationship between the logit and $f(x_h)$ is (as) linear (as possible); references is made to isotonic regression.

We have decided to elebrate on approach (3) because of its simplicity. However, we will start out with the method, that the consulting company who came up with the geo-model score, prescribes.

2.3.1 Method 1

For this method, one needs to define the proportion of low risks and high risks. These risks will be defined in the following way.

Low risks are people who have had no damage, so $Y_I = 0$. High risks are people with $Y_I = 1$.

In this method, the consulting company uses the Weight of Evidence (WoE). The Weight of Evidence must be calculated for each attribute of x_1 and x_3.

$$p_g(i) = \frac{g(i)}{G} \quad \text{the proportion of goods with the attribute i}$$

$$p_b(i) = \frac{b(i)}{B} \quad \text{the proportion of bads with the attribute i} \qquad \textbf{Eq 2-1}$$

$$WoE_i = \ln\left(\frac{p_g(i)}{p_b(i)}\right)$$

When the calculated values are plotted, the graph should look like Figure 2-5. When the bar is above the *x*-axis, the proportion of low risks with regard to the proportion of high risks, of score band *i*, is larger than the proportion of low risks with regard to the proportion of high risks in the whole portfolio and when the bar lies under the *x*-axis the proportion of high risks is larger with regard to the proportion of low risks, of score band *i*, than the proportion of high risks with regard to the proportion of low risks of the whole portfolio. The expectation is that the bars must be above the *x*-axis for the low scores and under the *x*-axis for the high scores. As high scores have a high risk according to the definition of the scores, people who are defined as high risks must be in the high scores.

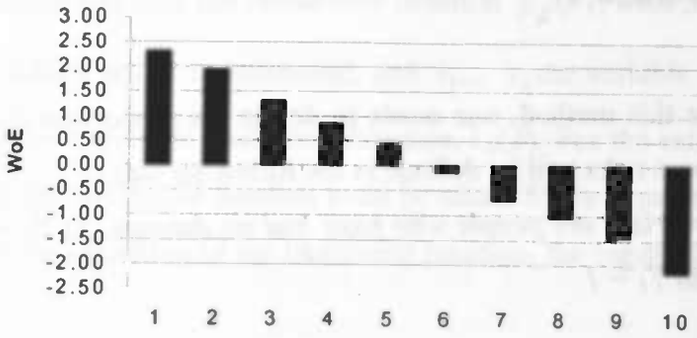


Figure 2-5 The Expected WoE

To give an extra proof of these results, another method can be used, the Information Value (IV).

$$IV = \sum_{i=1}^{n} \left(p_g(i) - p_b(i)\right) \ln\left(\frac{p_g(i)}{p_b(i)}\right) \qquad \text{with } n \text{ the total attributes} \qquad \textbf{Eq 2-2}$$

This value gives an idea of the difference between the proportions of high risks and the proportions of low risks. If this value is large it is easier to make two distributions, one for the high risks and one for the low risks. To look how much these two distributions differ, the divergence can be used.

$$div = \frac{\left(\mu_g - \mu_b\right)^2}{\frac{1}{2}\left(\sigma_g^2 + \sigma_b^2\right)} \qquad \textbf{Eq 2-3}$$

with μ_g = average score for the low risks

μ_b = average score for the high risks

σ_g^2 = variance of the score for the low risks

σ_b^2 = variance of the score for the high risks

This value must be large, because the difference of the two distributions is then very large and there is a good explanation by the classes for the high risks and the low risks.

Since it is not sure what the assumptions of this method are and what the meaning of this technique is, methods for nonparametric data, as the logistic regression and the Goodness of fit will also be used.

2.3.2 Method 2

One is interested in the probability of causing damage given a specific geo-model score or region score ($P(Y=1|X_h=x_h)$). To calculate that probability, logistic regression will be used. This means that the linear model

$$\text{logit}(Y)=\ln\left\{\frac{P(Y=1))}{1-P(Y=1)}\right\}=\beta_0+\beta_1 * x \qquad \textbf{Eq 2-4}$$

is used as the basis of the discussion. Note that the linearity is a doubtful assumption. It is very convenient, however.

Other link functions could also be used like probit or complementary log-log, but the logit is the canonical parameter of the binomial distribution and it gives a sufficient statistic for the parameter β. When this transformation is used, the following equation will specify the probability of having a high risk:

$$P(Y=1)=\frac{e^{\beta_0+\beta_1 * x}}{1+e^{\beta_0+\beta_1 * x}} \qquad \textbf{Eq 2-5}$$

The parameters β_0 and β_1 can conveniently be estimated by using the maximum likelihood method and regarding the score x_h as predetermined. The maximum

likelihood estimation uses the probability function $p_{\bar{\beta}}(\bar{s})$, with β the vector of the parameters, which has to be estimated, and $s_1,\ldots s_n$ the variable. This probability function will be called the likelihood function, $L_s(\beta)$. For the estimation of β, the maximum of the likelihood function must be taken. Since in most cases it is easier to maximize the logarithm of the likelihood function, the log-likelihood, $l_s(\beta)$, will be used.

$$l_{\bar{s}}(\bar{\beta}) = \ln L_{\bar{s}}(\bar{\beta}) = \sum_{i=1}^{n} (s_i \ln \theta_i + (1 - s_i)\ln(1 - \theta_i)) \qquad \textbf{Eq 2-6}$$

with $\theta_i = \dfrac{e^{\beta_0 + \beta_1 * x(i)}}{1 + e^{\beta_0 + \beta_1 * x(i)}}$.

To maximize the log-likelihood, Equation 2-6, the derivatives to β will be taken, and Equations 2-7 has to be solved.

$$\frac{d}{d\beta_o} l_{\bar{s}}(\bar{\beta}) = 0$$
$$\frac{d}{d\beta_1} l_{\bar{s}}(\bar{\beta}) = 0 \qquad \textbf{Eq 2-7}$$

Solving these equations provides the maximum likelihood estimates of β_0 and β_1.

To look if there is indeed an increase or a decrease of the values for the variables, the test whether the parameter β_1 is equal to 0 has to be performed. The likelihood ratio statistic will be used for this purpose:

$$G = -2 \ln \left[\frac{l_{\bar{s}}(\beta_0)}{l_{\bar{s}}(\beta_0, \beta_1)} \right] \qquad \textbf{Eq 2-8}$$

This statistic has a Chi-square distribution with 1 degree of freedom under the null hypothesis. So the Chi-square table can be used to look whether the hypothesis that $\beta_1 = 0$, can be rejected. In fact the standard normal distribution can be used as well.

If the hypothesis $\beta_1 = 0$ is rejected at some significance level, say $\alpha = 0.05$, then the expected values of the probability will be calculated and compared with the observed values. The expected values will be calculated, just by putting the estimated parameters in Equation 2-5, and fill in the values of the independent variable.

These expected values will be compared by using the goodness of fit procedure. First, the Chi-square must be calculated.

$$\chi^2 = \sum_{1}^{n} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \qquad \textbf{Eq 2-9}$$

For testing whether the expected model gives a good result, the value of Equation 2-8 needs to be compared with the critical value of the Chi-square distribution. For this comparison the degrees of freedom must be known and also the α, the level of significance, must be chosen. The degrees of freedom can be computed by taking the total number of classes and subtract 1 and the total number of parameters which must be estimated.

2.4 Results

This section is also divided into two paragraphs. The first paragraph gives the results of the method which is prescribed by the consulting firm, and the second paragraph gives the results of the logistic regression and the goodness of fit.

2.4.1 Results of Method 1

To give an impression of how the WoE has been built up and how the proportions of low risks and high risks are distributed over the score bands, graphs for both scores of the proportions of the good and the high risks will be shown in Figure 2-6 and Figure 2-7.

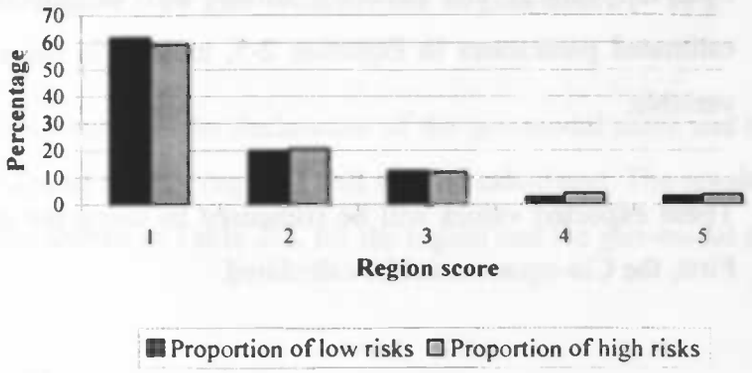


Figure 2-6 Proportion of Low Risks and High risks for the Region Score

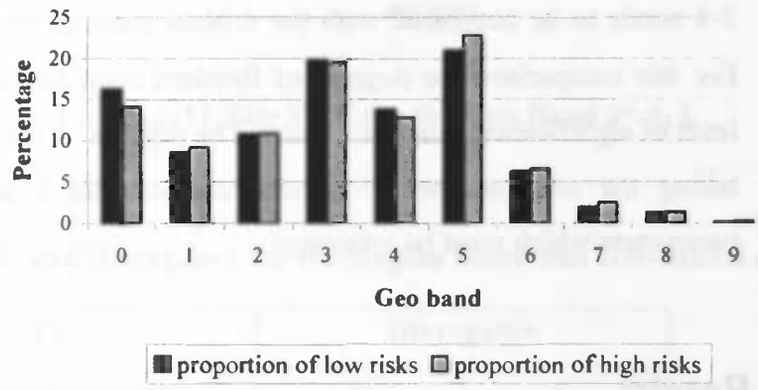


Figure 2-7 Proportion of Low Risks and High Risks for the Geo Band

For the calculation of the WoE, the natural logarithm of the proportions of the low risks and the high risks is necessary. To give an idea of the difference of these values the graphs of these values will be given.

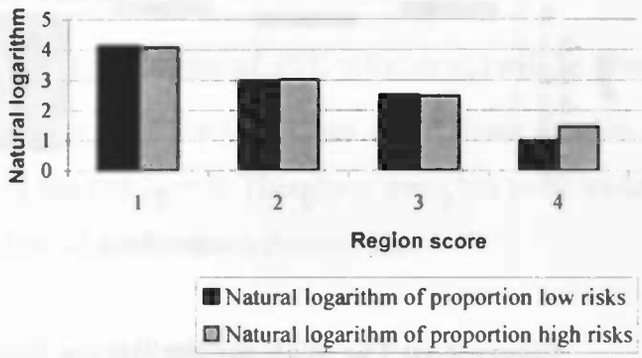


Figure 2-8 Natural Logarithm of the Proportions for the Region Score

Before the WoE will be calculated, the graph of the geo band, with the same meaning as Figure 2-8, will be given also.

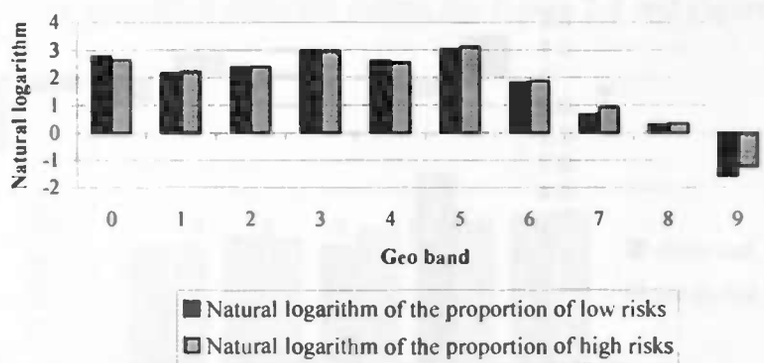


Figure 2-9 Natural Logarithm of the Proportions for the Geo Band

After having calculated these values, the only thing there has to be done is subtracting the natural logarithm of the high risks from the natural logarithm of the low risks which gives us the WoE. The graphs of the WoE will also be shown so that these results can be compared with Figure 2-5. The calculated values for the WoE will not be given.

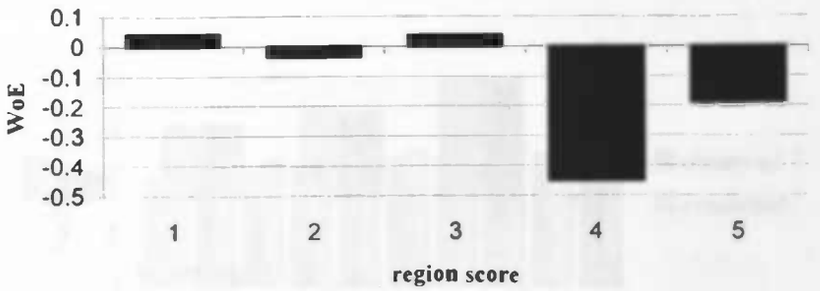


Figure 2-10 The WoE for the Region Score (= x_1).

To give an exact value of the declaration of the geo-model score and region score, the IV (eq 2-2) and the div (eq 2-3) will also be calculated. The results for the car policies will be shown in Table 2-1, for the region and the geo-model score.

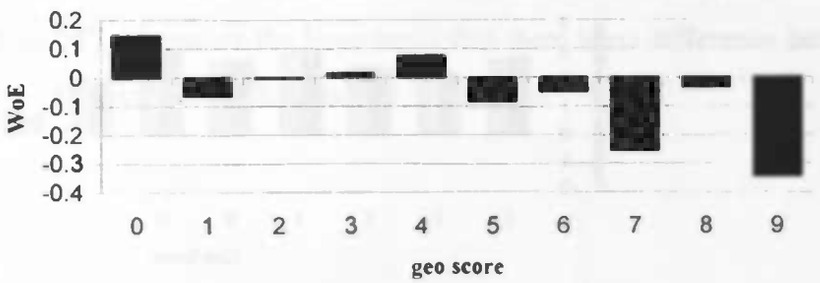


Figure 2-11 The WoE for the Geo Band (= x_3).

Table 2-1 IV and Divergence for the Region Score and Geo-model score

IV		Divergence	
x_1	x_3	x_1	x_3
0.009714	0.007401	0.47 %	0.35%

2.4.2 Results of Method 2

The results of the estimation of $P(Y_1 = 1|x_1 \text{ or } x_3)$ will be given in this section. The logistic regression gives for both cases a significant *P*-value, for x_1 0.0008 and for x_3 0.0038, for the test $\beta_1 = 0$. Therefore, there has to be looked at the Goodness of fit. The models which has been derived are:

$$P(Y_1 = 1|x_1) = \frac{e^{-2.0644+0.0657x}}{1+e^{-2.0644+0.0657x}} \qquad \text{model 1}$$

$$P(Y_1 = 1|x_3) = \frac{e^{-2.0454+0.0289x}}{1+e^{-2.0454+0.0289x}} \qquad \text{model 2}$$

For the calculation, the value of the region or geo-model score has to be put in model 1 respectively model 2. The graph of the models will be given, for an impression of the calculated expected values, see Figure 2-8 and Figure 2-9.

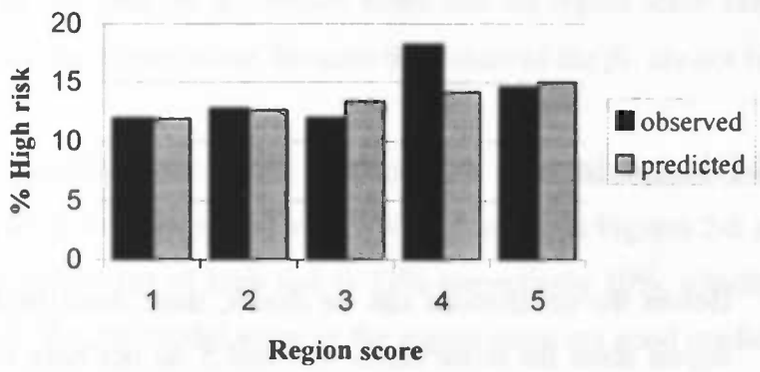


Figure 2-12 The Observed and Predicted Values of the Percentage High Risk for Model 1

By using Equation 2-9 the results of the Chi-square are:

$\chi^2_{\text{region},y_1} = 1.314$

$\chi^2_{\text{geo},y_1} = 1.827$

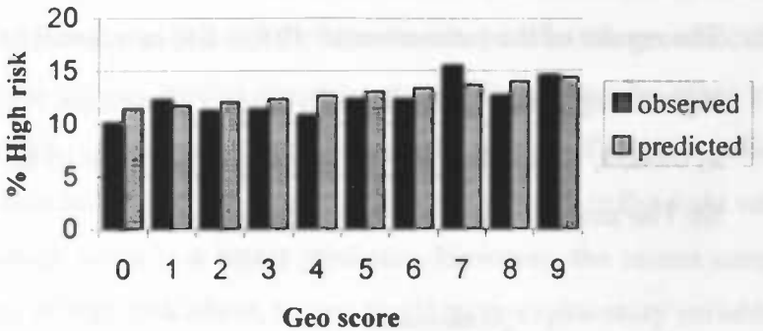


Figure 2-13 The Observed and Predicted Values of the Percentage High Risk for Model 2

The critical values, with $\alpha = 0.05$, $df_{region} = 5 - 1 - 2 = 2$ and $df_{geo} = 10 - 1 - 2 = 7$, because there are res. 5 and 10 classes and two parameters to be estimated, are res. 5.994 and 14.067. Therefore the hypothesis that there is no difference between the expected and the observed values cannot be rejected.

2.5 Discussion

Before the conclusions can be drawn, there must be kept in mind that for the region score the score bands 3, 4 and 5 do not have many policies and that are bands 7, 8 and 9 for the geo-model score.

The first step is to look at the graphs of the WoE. Figure 2-6 shows that the value of the WoE is very small for the bands 1, 2 and 3 of the region score. So there is not much difference between the proportion of high risks and low risks. In band 4 and 5 the proportion of high risks is indeed larger than the proportion of low risks, but it can also be due to a lack of data. This appears also for bands 7 and 9 of the geo-model score, see Figure 2-7. Band 0 gives a good result, as well as bands 4, 5

and 6. So for this method there might be concluded that the geo-model score is a better predictor than the region score.
However, looking at Table 2-1, the opposite could be said. This can be explained, since the bands with view policies are also taken into the calculation in these values. There has also been noticed that for the region score bands 4 and 5 gives a good result. The values of the IV and the Div are not very large, so there cannot be said whether the region score or the geo-model score predicts the probability of damage well.

Looking at the results of the second method, the logistic regression and the goodness of fit, there can be said more. In both cases, the test whether the hypothesis that $\beta_1 = 0$ is true, must be rejected, so there is a decrease or an increase among the classes. Indeed when the Goodness of fit is calculated, the model, which has been estimated by the logistic regression, can be accepted. This model does explain the damage in the way one would expect, because it is increasing when the risk of the geo and region score is getting higher. However, there must be said that the geo-model score and the region score cannot explain the damages of the clients alone, because the values of the β_1 are not large.

The models which are calculated, are adequate models, because the calculated Chi-squares are below the critical value. When looking at Figures 2-8 and 2-9, one sees that the percentage of high risk is 12% respectively 10%, which is too high for saying that the geo-model score or the region score are good predictors for the probability of high risk without other predictors. But they might help at predicting the probability for high risk if a model with more explanatory variables can be found.

So these results are in contradiction with the results of method 1. But there has already been said in the beginning of this chapter that one does not expect good results of method 1 since the interpretation of these results is not clear.

One thing which also has been mentioned is that the value of β, is not as large as expected. One explanation for this is of course that there is not enough data to give a good explanation. But another reason can be that the geo-model score and the region score cannot predict the risks alone. There must be other variables to explain the risks. The overall conclusion of the results of the car policies is that there can be concluded that the scores predict the damage in the right way, but one cannot say which score is a better predictor. However, the scores cannot predict the probability of high risk alone, so one needs more explanatory variables.

In Chapter 3, there will be looked whether there are other variables to explain this and to look which score is better since due to these results, a very hard evidence that the region score is better than the geo-model score or vice versa cannot be provided.

Chapter 3

Modeling the Probability of Causing Damage to be Paid

Luckily, the probability of causing damage and, more generally, the expectation of the damage caused does not depend only on the place where a client live. Variables like age and sex may also be of interest. An insurance company has more information about a client than the ones mentioned. Chapter 2 establishes that the geo score and the region score are not completely useless though their predictive value is only very small. Therefore, one wants to investigate whether other explanatory variables are of more interest. We shall see that, especially, the 'bonus-malus' grade is of interest.

3.1 Introduction

The region score and the geo score are constant in neighborhoods. Additional personal variables, like age, sex and 'bonus malus' will, hopefully, lead to a considerable improvement of the prediction of damage. An insurance company can even get more information by asking the client, or obtaining the history of the client by asking other insurance companies. But which information would be most relevant for the explanation of the risk? An easy but efficient approach consists in computing correlation coefficients The relationship of the variables with the definition of high risk will be tested by two non-parametric tests: Spearman's rank

correlation procedure for the correlation and the χ^2 test of independency in a 2*m table. The correlations between the variables will be tested only by Spearman's rank correlation. After rejecting the null hypothesis we shall make a liner-logistic model for the prediction of causing damage. This model provides an assessment of the probability of causing damage for any client with specific attributes of the explanatory variables.

3.2 Data

The same definition for high risk will be used, so the variable to be explained is:

Y_1 = whether ($Y_1 = 1$) or not ($Y_1 = 0$) a claim has been paid

For reminding what the meaning of this definition is, the explanation from Chapter 2 will be recalled: a claim has been paid if a client has caused damage by being at fault in an accident and the amount could not be recouped from another person or the client did not want to pay the amount of damage.

The geo score and the region score will also be used:

X_1 = region score (1, 2, 3, 4, 5)

X_3 = geo band (0, 1, 2, …, 7, 8, 9)

The geo band will be called *GEO*, and the region score *REG*. As there are various variables, it is easier to name them than to call them X_i.

Further, there are all kind of variables that can be used as predictor variables. First, the variables which tell something about the client will be given and then the variables which tell something about the car. The variables with their categories and the explanation can be found in Appendix I.

SEX

This variable is about the sex of the client and has two classes, male and female.

AGE

The AGE is a variable that has been classified into nine classes. The first class is for the young people up to 23, because these people are expected to have more damages than older people. The next class is until 30, and the classes after that increase by 10, until class 7. Class 8 is for people who are 80 years or older and class 9 is for business cars, these insurance policies do not have to give the year of birth as there are driving more than one peoples in the car. The reason why the classes has been created in such way is the fact that there is not much difference of the paid claims between the ages and also the insurance company thought that this was a good subdivision. Although there is not much difference, one thought that this variable could be important as it is used for the acceptance policy and for the calculation of the premium at the insurance company.

RAP

A client who does not pay the premium in time gets a dunning letter. After getting three dunning letters, the client gets a registered dunning letter. If the client does not pay that in time, the insurance policy can be suspended. Therefore, this variables tells us something about the payment history of the client. And if a client has his or her policy suspended by not paying the premiums, it is hard to get another insurance policy at another insurance company. This variable has two classes, class one for the people who have never had a registered dunning letter, in the past three years, and the second class which shows the people who have had at least one registered dunning letter in the past three years.

PAY

At this insurance company, people can pay monthly, biannually and annually. A client can only pay monthly if the premium is high enough and if they have never had a registered dunning letter. The people who have had a registered dunning letter must pay yearly so the insurance company knows that they got the premium for the whole year. If a client chooses to pay twice a year, the client has to pay 3% more. This variable might be of interest for declaring the probability of high risk

to see if the client is a good payer or not. However, this variable might be very strongly correlated with the variable RAP.

DUR

At the insurance business, there has been a research conducted about the duration of a policy. If a client goes to another insurance company after two years, this is a high risk for the insurance company, as the insurance company does not get the premiums, but would have to pay claims for that client. As one knows the yearly premiums do not cover a high amount of damages, so the insurance company loses money from such a client. Therefore, one has added this variable to the analysis to see how long a client has been insured at the same insurance company. This variable has been divided into 8 classes, the first seven classes are just for every year they are insured, and the eight class is for people who are insured for more than 6 years. The reason why one took six as the last group is because the research tells us that people who are longer than six years insured at one insurance company are low risks.

TOT

If a client also has another non-life insurance policy (as liability, fire or accident policies) at the same insurance company that client should have a overall risk that can be accepted. This variable tells something about the total policies a client has at this insurance company. The first class is for people with only one policy, and the last class for people with six policies or more.

NEW

One thinks that when a client is a new client and not a take-over from another company that the risks are higher. Therefore, the insurance company wants more take-overs than new clients.

PRI

Clients can use their car for business or privately. The risk of business and private use are not known, but one thought it might be a good variable.

BM

The Belgium 'bonus-malus' system has 23 classes. These classes tells something about how many accidents a person has been in fault and claimed.

If a person never had a automobile insurance policy and uses a car for private use, he or she will start in class 12 with 'bonus-malus' grade 11. After driving one year without damage, the grade will decrease with one, to 'bonus-malus' grade 10. This means that after eleven years of damage-free driving a person has 'bonus-malus' grade 0, the lowest grade a person can achieve. However, if a client has a caused damage and claimed, the grade will increase with five grades until the maximum of 'bonus-malus' grade 22.

This system is mandatory at every automobile insurance company, and if a person changes companies, the 'bonus-malus' grade stays the same. This system must be used for the calculation of a premium and well according to Table 3-1.

Table 3-1 Bonus-males system used for premiums

Grade	Premium level (according to base level 100)	Grade	Premium level (according to base level 100)
22	200	10	81
21	160	9	77
20	140	8	73
19	130	7	69
18	123	6	66
17	117	5	63
16	111	4	60
15	105	3	57
14	**100**	2	54
13	95	1	54
12	90	0	54
11	85		

The initial grade is 14 for professional use and 11 for private use of the car.

This variable has been divided into eight classes; in class 5 are people with 'bonus malus' equal to 11 and in class 7 is 'bonus malus' 14.

SPO

A sports car has of course a worse risk than other cars, so this variable needs to be included in the analysis.

KW

The power of the cars engine (measured in kilowatts) is an important factor in the calculation of the insurance premium. There is even a legal minimum premium depending on the power which changes every year. As the further calculation depends on the insurance company, one cannot give more details about the premium calculation, so this variable needs to be taken into account in the analysis. This variable has six classes defined by another insurance company. As one does not have enough data to make new classes, these classes will be used.

BRA

Since one does not have enough knowledge of the cars to classify the cars as middle class, luxury, small, large etc. One will look at the brand of cars. This variable has 27 classes, which have been grouped by looking at the largest brands, and by comparing the others. The classes are ordered according to the percentage damage of the portfolio used for the analysis of the risk, class 1 has the lowest percentage and class 27 the highest percentage, however there is not much difference between these percentages.

3.3 Method

The question of this chapter can be divided into two questions. The first question is what kind of information can give the variables of risk prediction. The second

question is which variables could be used for the estimate of the probability of causing loss to the insurance company and what are their estimated parameters so that one can make a model. The first question will be called the information of the variables, and the second question will be called the selection and estimation of the predictor variables.

3.3.1 Information of the Variables

To know something more about the relationship of the variables with the definition of high risk, some tests will be performed. First, one tries to find out the correlation to know what the variable predicts. Secondly, there will be considered whether a dependency exists between the variables.

The correlation can be tested by two methods, Kendall's tau and Spearman's rho. Kendall's tau compares all possible pairs in the observations. If most pairs are concordant, both values of the pair are larger (or smaller) than other pairs, the relationship is positive. If most pairs are discordant, one value of the pair is higher than other pairs and if the other value is smaller, there is a negative relation. There is no relation if most pairs are tied, one of the values of the pair is equal to other pairs, or if there is no difference between the total concordant pairs and the total discordant pairs.

Spearman's rho is a non-parametric correlation statistic, which gives the correlation of two ordinal variables or of two variables of which there cannot be assume that they are normally distributed. This method can be used if the standard correlation procedure, Pearson, can not be used.

As the performance of the Kendall's tau and Spearman's rho are very similar, and for large n the Spearman measure may be a bit better, the most known nonparametric correlation procedure the Spearman's rank correlation has been chosen for.

Spearman's Rank Correlation Procedure

Spearman's rank correlation is a non-parametric method to test whether there is a relation between two ordinal variables or two variables which do not have a normal distribution by using ranks.

For this procedure ranks need to be assigned to the observations. This means that for both variables separately the smallest value gets rank 1 and the largest value gets rank N, with N being the total number of observations. When values are equal, tied ranks must be calculated by taking the average of the ranks they should get and give that value to all observations with the same value.

To test if the variables have a correlation is to test if the hypothesis $\rho = 0$ against the alternative $\rho \neq 0$ cannot be accepted. Therefore, the correlation coefficient r_s needs to be estimated.

$$r_s = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N^3 - N} \qquad \textbf{Eq 3-1}$$

with d_i = difference between the ranks of the variables of observation i.

As one is sure that there are tied ranks, Equation 3-1 cannot be used, instead Equation 3-2 must be used.

$$(r_s)_c = \frac{(N^3 - N)/6 - \sum d_i^2 - \sum t_X - \sum t_Y}{\sqrt{\left[(N^3 - N)/6 - 2\sum t_X\right]\left[(N^3 - N)/6 - 2\sum t_Y\right]}} \qquad \textbf{Eq 3-2}$$

with $\sum t_X = \frac{\sum (t_i^2 - t_i)}{12}$ where t_i is the number of tied values of variable X in a group of ties, and $\sum t_Y = \frac{\sum (t_i^2 - t_i)}{12}$ where t_i is again the number of tied values in a group of ties only now for variable Y.

Instead of Equation 3-2, an equation which is related to the correlation procedure for normal distributed variables should be used.

$$r_s = \frac{\sum_{i=1}^{n}(\text{rank of } X_i)(\text{rank of } Y_i) - \frac{n(n+1)^2}{4}}{\sqrt{\left(\sum_{i=1}^{n}(\text{rank of } X_i) - \frac{n(n+1)^2}{4}\right)\left(\sum_{i=1}^{n}(\text{rank of } Y_i) - \frac{n(n+1)^2}{4}\right)}} \qquad \textbf{Eq 3-3}$$

To test whether the estimated r_s is equal to zero, one can consult a special table for the Spearman's rank correlation criteria and use the total number of observations to look up the critical value for the estimated correlation coefficient. If the estimated correlation coefficient is larger than the critical value, one can say that the hypothesis, that ρ=0, cannot be accepted. Thus, there might be a correlation between the two variables.

Test for Independence

The data which will be used for this test can be put in a 2*k contingency table, see Table 3-1. Where n_{ij} means the observation of cell (i, j), and when there is a point instead of a number, it then means the sum over the row or column has been taken.

Table 3-2 An Example of 2*k Cross Table

	1	2	…	…	k	total
1	n_{11}	n_{12}			n_{1k}	$n_{1.}$
2	n_{21}	n_{22}			n_{2k}	$n_{2.}$
total	$n_{.1}$	$n_{.2}$			$n_{.k}$	$n_{..}$

One wants to test if there is independence between the row and the column variable. Independence means that the rows of probability are proportional and that the columns of probability are proportional. This can be written in hypothesis form as follows:

$$H_0 : p_{ij} = p_{i.} p_{.j}$$
$$H_1 : p_{ij} \neq p_{i.} p_{.j}$$

where p_{ij} stands for the probability of cell (i, j).

The expected numbers in a cell needs to be calculated; this can be done by using $e_{ij} = np_{ij}$. Under the assumption that the variables are independent, there yields that $e_{ij} = np_{i.} p_{.j}$. For the estimation of the probabilities $p_{i.}$ and $p_{j.}$, Maximum Likelihood can be used and have $\hat{e}_{ij} = \frac{n_{i.} n_{.j}}{n}$.

Now the independence can be tested by using the Likelihood ratio Λ.

$$\Lambda = \frac{\prod\prod (\hat{e}_{ij})^{n_{ij}}}{\prod\prod (n_{ij})^{n_{ij}}} \qquad \textbf{Eq 3-4}$$

If the natural logarithm of the Likelihood ratio is taken and multiplied by –2, then the likelihood ratio has a Chi-square distribution with the difference between the two estimated parameters under the hypothesis and under the alternative hypothesis as degrees of freedom. The test statistic becomes:

$$-2 \ln \Lambda = -2 \sum\sum n_{ij} \left(\ln \hat{e}_{ij} - \ln n_{ij} \right) \qquad \textbf{Eq 3-5}$$

which is the G-test statistic.

If one now wants to know whether the variables are independent or not, the critical value table of the Chi-square can be used. For that purpose the number of degrees of freedom must be determined, which is $(2k - 1) - k = k - 1$. If the null-hypothesis will be rejected, one might say that the variables are dependent.

3.3.2 Risk Prediction with all Variables

The risk prediction, using all the variables, will be done by multiple logistic regression. The selection of the variables to get the best model will be done by stepwise selection, forward selection and backward elimination.

Multiple logistic regression is used for the same reason as one uses logistic regression, described in Section 2.3.2. Due to the fact that there are now more than one explanatory variable multiple logistic regression has to be used.

The same transformation as given in Equation 2-4 needs to be executed, but the result is different as there are more independent variables, see Equation 3-6.

$$\text{logit}(Y) = \ln\left\{\frac{P(Y=1))}{1-P(Y=1)}\right\} = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \ldots + \beta_k * x_k \qquad \textbf{Eq 3-6}$$

The only thing that is different is that there are more explanatory variables, and of course more parameters to be estimated. The probability function becomes

$$P(Y=1) = \frac{e^{\beta_0+\beta_1*x_1+\beta_2*x_2+\ldots+\beta_k*x_k}}{1+e^{\beta_0+\beta_1*x_1+\beta_2*x_2+\ldots+\beta_k*x_k}} \qquad \textbf{Eq 3-7}$$

The estimation of the parameters goes exactly the same, only one needs to take more derivatives to every parameter. Therefore, there are k+1 equations to solve. After the estimation of the parameters one also needs to test whether one of the β's is equal to zero, by implementing the likelihood ratio test.

$$G = -2\ln\left[\frac{l_x(\beta_0)}{l_x(\beta_0,\beta_1,\ldots,\beta_k)}\right] \qquad \textbf{Eq 3-8}$$

In addition, this test statistic has, under assumption of the null hypotheses:

$H_0 : \beta_1 = 0$ and $\beta_2 = 0$ and $\beta_3 = 0$ and ... and $\beta_k = 0$

a Chi-square distribution, with k degrees of freedom.

In Section 2.3.2 the model has been tested by using Goodness of fit, which cannot be used for this model because there are a lot of classes, so the probability is high that one of the classes is smaller than five. As the Goodness of fit procedure, using Chi-square, cannot guarantee that the results will be good when classes are smaller than five. Therefore, a different method will be used.

The statistic program which is used, gives us a nice result to check whether the model is a model which can be accepted, namely the concordant and the discordant.

Concordant and Discordant

The statistic program has estimated a model for the probability that $Y = 1$. To look if this model is good, pairs are formed with one observation where $Y = 1$, say X_1 and with one observation where $Y = 0$, say X_0. So there will be $n_1 * n_0$ pairs.

Each observation has an estimated probability of having $Y = 1$. For each pair these probabilities will be compared in the following way:

If $P(X_1 = 1) < P(X_0 = 1)$ this pair will be called discordant.

If $P(X_1 = 1) > P(X_0 = 1)$ this pair will be called concordant.

These probabilities can only be equal if the observations have the same attributes and so the same chance, or there is not much difference among the attributes. If the probabilities are equal there are tied pairs.

To look whether the model gives a good representation of the dependent variable, the percentages concordant, discordant and tied pairs are calculated. The model gives a good representation if the percentage concordant is large and the percentage discordant is small. If the percentage tied pairs is large then there are a lot of observations with the same attribute, or it can indicate that there is not much difference among the observations.

Selection of Variables

The variables can be selected in two ways: by just adding the dependent variables which have been found by the independency test, or by putting all variables in. In both cases a selection method has to be used. If only the dependent variables will be used, it is possible that one does not get the best prediction, because independent variables could in combination with other variables predict the damage also. So both ways will be tried and there will be examined which variables give the best model.

There are three different selection methods one can use; stepwise selection, forward selection or backward elimination. The base of these methods is to find the best model with the independent variables that explains the dependent variable the best.

Forward Selection

The first step of this method is to add an intercept and calculate the residual Chi-square. The next step is to calculate for all the independent variables the Chi-square, then the variable with the smallest Chi-square will be added to the model, and the residual Chi-square will be calculated again. After this, it continuous until there is no variable which has a significant p-value of the Chi-square, or until the residual Chi-square becomes larger. In the end, the variables that entered the model are given with their parameter estimation, standard error, Chi-square and the probability of the Chi-square. Also the concordant, discordant and tied percentages are calculated.

Backward Elimination

The first steps of the backward elimination does not add only one intercept, but starts with all the variables, and throw out the variable with the largest p-value until there are no more variables in the model which are not significant. The output of this method is the same as the output of the forward selection, only now the variables which stayed in the model are given. Normally, these variables must be the same as the variables found in the forward selection. When the results are

not the same, it is better to use the backward elimination because this method can detect relationships between two variables easier as all the variables are already in the model.

Stepwise Selection

Stepwise selection uses both methods. The first steps are similar to forward selection, however, when a variable has been added to the model and another variable becomes insignificant, backward elimination is used to drop that insignificant variable again and another variable will be entered by using the forward selection method. This continuous until nothing can change the model any more and one has the explanatory variables, with their parameters, for the dependent variable.

3.4 Results

Before there will be started with the high risk prediction, one first wants to know more about the variables which will be used. If the variables are indeed related to the definition of high risk and are not used in the model as declaring variables, it might be useful to keep them in mind for the acceptance of a client.

3.4.1 Information of the Variables

In this result, only the variables which have significant P-values will be given. When the test of Spearman's rank correlation gives a significant result, then there might be a correlation. By looking at the value of the r_s there can be seen whether the correlation is positive or negative. For the test of independence, there can be said that if the test is significant then the definition of high risk and the explanatory variable might be dependent.

Spearman's Rank Correlation

The P-values and the r_s will be presented, in order to get an impression of which way the variables are correlated with the definition of high risk.

Variable	*Spearman's r_s*	*P-value*
BM	0.2891	0.0001
NEW	-0.0306	0.0001
PAY	0.0299	0.0001
BRA	0.0248	0.0002
KW	0.0217	0.0013
VER	0.0199	0.0032
REG	0.0193	0.0042
GEO	0.0191	0.0047
AGE	-0.0152	0.0243
RAP	-0.0141	0.0362

To have an idea of the relationship amongst the variables, the Spearman's rho has also been calculated and tested if there might be any correlation, see Tables 3-3 and 3-4 for the results.

Test for Independence

The variables for which the test for independence with the dependent variable y_1 was significant are BM, VER, PAY, RAP, NEW, AGE and KW. Also the variables REG and GEO are dependent.

Table 3-3 Correlation Among the Explanators

r_s p-value	SEX	RAP	SPO	NEW	PRI	PAY	KW
SEX		-0.018 0.009	0.019 0.005	-0.166 0.0001	0.049 0.0001	-0.070 0.0001	0.153 0.0001
RAP	-0.018 0.009		0.029 0.0001	0.043 0.0001	-0.018 0.006	-0.083 0.0001	-0.015 0.022
SPO	0.019 0.005	0.029 0.0001		0.033 0.0001	-0.025 0.0001	0.021 0.002	0.170 0.0001
NEW	-0.166 0.0001	0.043 0.0001	0.033 0.0001		0.006 0.370	0.038 0.0001	-0.069 0.0001
PRI	0.049 0.0001	-0.018 0.006	-0.025 0.0001	0.006 0.370		0.006 0.347	-0.114 0.0001
PAY	-0.070 0.0001	-0.083 0.0001	0.021 0.002	0.038 0.0001	0.006 0.347		-0.003 0.677
KW	0.153 0.0001	-0.015 0.022	0.170 0.0001	-0.069 0.0001	-0.144 0.0001	-0.003 0.677	
AGE	0.094 0.0001	-0.062 0.0001	-0.028 0.0001	-0.208 0.0001	-0.066 0.0001	-0.244 0.0001	-0.017 0.012
BRA	0.001 0.896	-0.014 0.044	-0.003 0.650	0.007 0.273	-0.008 0.217	-0.004 0.570	-0.073 0.0001
GEO	-0.004 0.567	0.054 0.0001	0.008 0.220	0.012 0.084	-0.004 0.512	0.067 0.0001	-0.014 0.645
REG	-0.004 0.540	0.024 0.0001	0.010 0.127	-0.006 0.535	-0.023 0.0001	0.107 0.0001	-0.0001 0.969
DUR	0.138 0.0001	-0.151 0.0001	-0.034 0.0001	-0.092 0.0001	0.004 0.599	-0.156 0.0001	0.038 0.0001
BM	-0.103 0.0001	0.128 0.0001	0.021 0.002	0.209 0.0001	-0.013 0.051	0.145 0.0001	-0.050 0.0001
TOT	0.008 0.231	-0.088 0.0001	0.002 0.735	-0.007 0.281	0.008 0.214	0.069 0.0001	0.003 0.659

Table 3-4 Continuation of Table 3-3

r_s p-value	AGE	BRA	GEO	REG	DUR	BM	TOT
SEX	0.094 0.0001	0.001 0.896	-0.004 0.567	-0.004 0.540	0.138 0.0001	-0.103 0.0001	0.008 0.231
RAP	-0.062 0.0001	-0.014 0.044	0.054 0.0001	0.024 0.0001	-0.151 0.0001	0.128 0.0001	-0.088 0.0001
SPO	-0.028 0.0001	-0.003 0.650	0.008 0.220	0.010 0.127	-0.034 0.0001	0.021 0.002	0.002 0.735
NEW	-0.208 0.0001	0.007 0.273	0.012 0.084	-0.006 0.535	-0.092 0.0001	0.209 0.0001	-0.007 0.281
PRI	-0.066 0.0001	-0.008 0.217	-0.004 0.512	-0.023 0.001	0.004 0.599	-0.013 0.051	0.008 0.214
PAY	-0.244 0.0001	-0.004 0.570	0.067 0.0001	0.107 0.0001	-0.156 0.0001	0.145 0.0001	0.069 0.0001
KW	-0.017 0.012	-0.073 0.0001	-0.014 0.645	-0.001 0.969	0.038 0.0001	-0.050 0.0001	0.003 0.659
AGE		-0.011 0.108	-0.005 0.419	0.027 0.0001	0.311 0.0001	-0.283 0.0001	-0.054 0.0001
BRA	-0.011 0.108		-0.003 0.661	-0.013 0.058	-0.006 0.386	0.016 0.020	-0.005 0.421
GEO	-0.005 0.419	-0.003 0.661		0.216 0.0001	-0.065 0.0001	0.092 0.0001	-0.072 0.0001
REG	0.027 0.001	-0.013 0.058	0.216 0.0001		-0.050 0.0001	0.058 0.0001	-0.072 0.0001
DUR	0.311 0.0001	-0.006 0.386	-0.065 0.001	-0.050 0.0001		-0.337 0.0001	0.091 0.0001
BM	-0.283 0.0001	0.016 0.020	0.092 0.0001	0.058 0.0001	-0.337 0.0001		-0.090 0.0001
TOT	-0.054 0.0001	-0.005 0.421	-0.071 0.0001	-0.072 0.0001	0.091 0.0001	-0.090 0.0001	

3.4.2 Risk Prediction with all Variables

First, all the described variables will be added to the model. After that the concordance, discordance and the tied percentages are given, to compare both models.

Definition: High Risk When Damage Has Been Claimed

The stepwise selection procedure gives the following results:

Variable	*number in*	*parameter Estimate*	*standard error*	*Wald Chi-square*	*P > Chi-square*
Intercept		-3.7630	0.1356	770.01293	0.0001
BM	1	0.5139	0.0150	1174.4321	0.0001
DUR	2	0.1515	0.0107	200.2482	0.0001
RAP	3	-0.6021	0.1280	22.1163	0.0001
KW	4	0.0659	0.0154	18.4233	0.0001
BRA	5	0.0133	0.0036	13.8178	0.0002
PRI	6	-0.3021	0.0849	12.6509	0.0004
NEW	7	-0.1025	0.0454	5.1015	0.0239

Concordant = 70.3%

Discordant = 28.6%

Tied = 1.1 %

The results of the two other procedures give exactly the same results as above.

If only the dependent variables are used, the variables which give the probability on high risk are: BM, VER, KW, AGE and RAP with the following results:

Variable	*number in*	*parameter Estimate*	*standard error*	*Wald Chi-square*	*P > Chi-square*
Intercept		-4.0323	0.1068	1424.3436	0.0001
BM	1	0.5079	0.0147	1186.9417	0.0001
DUR	2	0.1457	0.0110	175.4207	0.0001

RAP	3	-0.6001	0.1277	22.0969	0.0001
KW	4	0.0697	0.0151	21.2277	0.0001
AGE	5	0.0337	0.0149	5.1501	0.0232

Concordant = 70.1%

Discordant = 28.5%

Tied = 1.4 %

One sees that the variable AGE is suddenly in the model. As one sees in Tables 3-3 and 3-4 AGE is correlated with PRI and NEW, so a model without PRI and NEW and with all the other variables is tried, the results are:

Variable	*number in*	*parameter Estimate*	*standard error*	*Wald Chi-square*	*P > Chi-square*
Intercept		-4.2261	0.1189	1263.0682	0.0001
BM	1	0.5076	0.0148	1183.5205	0.0001
DUR	2	0.1457	0.0110	175.2663	0.0001
KW	3	0.0747	0.0152	24.1608	0.0001
RAP	4	-0.5918	0.1277	21.4819	0.0001
BRA	5	0.0136	0.0036	14.3144	0.0002
AGE	6	0.0345	0.0149	5.3868	0.0203

Concordant = 70.5%

Discordant = 28.5%

Tied = 1.0 %

There are less variables in the model, yet it predicts better.

3.5 Discussion

Not all the variables which were introduced have a relationship with the definition of high risk. The relationships of the variables with the definition will only be described if there is a relationship.

The variables BM, REG, GEO, PAY, DUR, NEW and KW have all the same relationship with the definition of high risk. This relationship means that the definition might be dependent on these variables and they have a positive correlation. This means, for example, that the higher the 'bonus malus', the larger the probability of high risk. Looking at the payment history there can be said that a person who pays monthly, has a smaller probability of high risk than a person who pays annually. The same as in Chapter 2 can be mentioned for the two scores as they have a positive relation with the definition of high risk.

The variables RAP and AGE are dependent and have a negative correlation with the definition of high risk. For example, the older the person, the smaller the probability of high risk, and the probability of high risk is lower for a new case than for a take-over. This is in contradiction with what has been expected.

One variable, BRA, has a correlation with the definition of high risk, but it is not dependent. The reason for this can be the way in which the classes of this variable are defined.

Although some variables have almost the same relationship with the definition, they do not have all the same correlation with each other. There are even several variables like REG with NEW, GEO with NEW and KW and PAY with KW who do not have any correlation.

In several tests one has seen that the region score and the geo score have a relation with the definition of high risk. But if one looks at the result of the multiple logistic regression, one sees that they are both NOT in the model. So the scores do not predict strongly enough to be added to the model with personal variables for the probability of high risk.

If one adds the results of the multiple logistic regression into the model, the model becomes:

$$P(Y_1 = 1) = \frac{e^{-3.7630+0.5139*BM+0.1515*DUR-0.6021*RAP+0.0659*KW+0.0133*BRA-0.03021*PRI-0.1025*NEW}}{1+e^{-3.7630+0.5139*BM+0.1515*DUR-0.6021*RAP+0.0659*KW+0.0133*BRA-0.03021*PRI-0.1025*NEW}}$$

This model predicts for 70.3% in the right way.

If only the dependent variables will be used, one dependent variable has been added to the model extra. Thus the independent variables BRA, NEW and PRI, have been chosen above the dependent variable AGE. It can be that the variables PRI and NEW has been chosen above AGE, as these variables have a correlation. However, the variable BRA does not have a correlation with the variable AGE, thus this is only additional information. A new model without PRI and NEW has been constructed:

$$P(Y_1 = 1) = \frac{e^{-4.2261+0.5076*BM+0.1457*DUR+0.0747*KW-0.5918*RAP+0.0136*BRA+0.0345*AGE}}{1+e^{-4.2261+0.5076*BM+0.1457*DUR+0.0747*KW-0.5918*RAP+0.0136*BRA+0.0345*AGE}}$$

This model predicts for 70.5% in the right way and has less tied and discordant pairs. Therefore, this model will be advised to be used.

After this analysis there can indeed a conclusion about the scores been drawn: none of the scores predict well enough to predict the probability of high risk. But it can be possible that the scores do predict the amount of damage which the insurance company must be expected to pay for a policy. This research will be done in the next chapter.

Chapter 4

Premium: Another Way to Determine High risk

An insurance company needs to make rules for accepting or rejecting a new client. This cannot be done with only the probability of causing damage is one aspect, the amount of damage is of considerable interest. Moreover, the premium can even be calculated by using the predicted amount of damage and the probability of having damage. As the premium would be a better method for an acceptance policy than the amount of damage, the premium needs also to be calculated. Therefore, the question of this chapter is:"What are the independent variables in a model for a premium?"

4.1 Introduction

Whether an insurance company rejects or accepts a client cannot only be based on the prediction of high risk. If a client has had an accident, it does not necessarily mean that the damage of this accident was large, and if it was large it could have been that the premium of the client was higher than the premium of other clients, so that in the end it might have been a small loss. To tell something more about the new client, a company needs to calculate the expected amount of damage conditionally declaring variables. A premium is calculated by using the probability of damage and the predicted amount of damage, like:

$$E(S|\text{covariables}) = P(Y = 1|\text{covariables})E(S|Y = 1,\text{covariables})$$

Where S stands for the amount of damage.

Theoretically it would be nice to do it like this for a premium, but in practice there are minimum premiums defined by law. This premium can be used in the decision of accepting or rejecting a client.

4.2 Data

For the premium, a model for the expected amount of damage and the average probability of damage over the past three years needs to be estimated. The dependent variables will be:

Y_2 = average amount of damage (continuous)

$$Y_{98} = \begin{cases} 1 & \text{if an amount of damage has been paid in 1998} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_{99} = \begin{cases} 1 & \text{if an amount of damage has been paid in 1999} \\ 0 & \text{if not} \end{cases}$$

$$Y_{00} = \begin{cases} 1 & \text{if an amount of damage has been paid in 2000} \\ 0 & \text{if not} \end{cases}$$

Y_3 = average of having an amount of damage paid over the past three years

$$P(Y_3 = 1) = \frac{P(Y_{98} = 1) + P(Y_{99} = 1) + P(Y_{00} = 1)}{3}$$

To estimate the amount of damage per damage, one needs to use as a dependent variable the amount of damage per damage. Therefore the amount of damages, per policy, has been added together and divided by the total damages, to arrive at the average amount of damage and dependent variable Y_2. After calculating the average amount of damage, the amount has been truncated at the re-insurance value of 12,000,000 BEF, as the amount above that re-insurance value is not paid by the insurance company.

For the calculation of the predicted amount of damage, only the policies which had a positive amount of damage paid over the past three years can be used.

For the four models which need to be estimated, the same variables as in Chapter 3 will be needed. The variable names and a short prescription of them will only be given.

GEO = geo band

SEX = sex of the client

AGE = age of the client

RAP = registered duns

PAY = frequency of the payment

DUR = duration of an insurance policy

TOT = total number of policies at this insurance company

PRI = private or professional use of the car

NEW = new case or a take-over from another insurance company

BM = bonus males

KW = Power of the engine of the car measured in kilowatts

SPO = sports car or normal car

BRA = brand of car

Further, two more variables are needed: the real premium for the comparison of the expected premium with the real premium and the definition of high risk, Y_1, to look if the premium can be used for the acceptance of a client.

4.3 Method

Before being able to describe the method for calculating the amount of damage, the relationship between the dependent and the independent variables needs to be known. This relationship will be tested by the Kruskal-Wallis test.

4.3.1 Kruskal-Wallis Test

As the explanatory variables are discrete, their levels define sub-samples, e.g. the region score has 5 classes defining thus 5 samples. In order to test whether an

explanatory variable affects the response variable, the homogeneity of these samples must be tested.

The Kruskal-Wallis test is a non-parametric test for the k-sample problem, and as such, is an alternative for the one-factor-ANOVA F-test. The Kruskal-Wallis test can be used since the assumption of normality underlying the F-test cannot be made for our data. Ranks are assigned to the original observations, from rank 1 for the smallest to rank N, the total number of observations, for the largest value. When values are equal, the ranks are tied.

The test statistic to be calculated is:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \qquad \textbf{Eq 4-1}$$

where n_i is the number of observations in class i, N is the total number of observations in all k classes and R_i is the sum of the ranks of the n_i observations in class i.

When there are tied ranks the test statistic H must be a little higher, therefore a correction factor C needs to be calculated.

$$C = 1 - \frac{\sum_{i=1}^{m} (t_i^3 - t_i)}{N^3 - N} \qquad \textbf{Eq 4-2}$$

with t_i the number of ties in class i, and m the number of groups of tied ranks.

The corrected value of H will be $H_C = \frac{H}{C}$.

To know whether the results are significant, the result of the Chi-square with k-1 degrees of freedom can be used. (It can only be used when the sample sizes are large enough, but in this study we do not worry about that.)

The hypothesis of this test is whether the distributions of the categories of the variables are equal: $H_0 : F_1 = F_2 = \ldots = F_k$

with k the total classes of the variable and F_i the distribution function of class i.

If the test rejects H_0, then there can only be said that one of the means is different from the others, but there cannot be said which mean. If there is a difference between the classes, then there might be an increasing or decreasing correlation with the definition of high risk.

4.3.2 Calculation of the Premium

The estimation of the premium can be calculated, by using the probability of damage and the expected amount of damage, and by the following fundamental formula of non-life insurance mathematics.

$$ES = EX * EN \qquad \textbf{Eq 4-3}$$

Here X stands for the damage, S stands the amount of damage that must be paid and N stands for the frequency of damage. Which in our terminology will be rewritten as:

$$E(S|\text{covariables}) = P(Y = 1|\text{covariables})E(S|Y = 1,\text{covariables})$$

So that we indicate the dependency of the covariables and that we only use the policies which had damage.

There are several ways to calculate the premium ($=\pi(S|\text{covariables})$), these are called the premium principles:

a) Net Premium $\pi(S|\text{covariables}) = E(S|\text{covariables})$

which is also known as the equivalence principle. This is the premium which is good for the risk neutral insurer[1].

b) Expectation Principle $\pi(S|\text{covariables}) = (1+\alpha)E(S|\text{covariables})$

Here $\alpha E(S)$ is a safety advance, with $\alpha > 0$ the safety advance parameter

c) Variance Principle

$\pi(S|\text{covariables}) = E(S|\text{covariables}) + \alpha Var(S|\text{covariables})$

The safety advance in this principle is proportional to $\alpha Var(S|\text{covariables})$, with $\alpha > 0$.

d) Standard Deviation Principle

$\pi(S|\text{covariables}) = E(S|\text{covariables}) + \alpha\sigma(S|\text{covariables})$

Now, the safety advance is proportional to $\alpha\sigma(S|\text{covariables})$, with again $\alpha > 0$.

e) Exponential Principle $\pi(S|\text{covariables}) = \frac{1}{\alpha}\log(m_S(\alpha))$

The parameter $\alpha > 0$, is the risk aversion and $m_S(\alpha)$ is the moment generating function.

There are more premium principles, but these are the most common and most important principles.

The following five properties can be used to look which premium principles can be used by the insurance company according to the business rules of that insurance company.

1) Non-Negative Safety Advance $\pi(S|\text{covariables}) \geq E(S|\text{covariables})$

A premium without safety advance will lead to ruin, this means that the capital of the insurance company becomes negative.

2) No Swindle $\pi(S|\text{covariables}) \leq \max(S|\text{covariables}) = \min\{P|F_S(P)=1\}$

[1] Risk neutral insurer = for every risk X a premium of $E(X)$ is enough, so the benefit function is linear.

3) Consistency $\pi(S+c|\text{covariables}) = \pi(S|\text{covariables}) + c \qquad \forall c$

If the amount of damage that must be paid increases with a fixed amount c, the same has to be done with the premium.

4) Additive $\pi(S+T|\text{covariables}) = \pi(S|\text{covariables}) + \pi(T|\text{covariables})$

Independent risk does not change the total premium.

5) Iterative $\pi(S|\text{covariables}) = \pi(\pi(S|T, \text{covariables})) \qquad \forall S, T$

The premium for S can be calculated in two steps.

In the Table 4-1 a summary of the premium principles and their properties will be given. If there is a '+', the premium has that property and if there is a '-', the premium does not have that property.

Table 4-1 The Premium Principles with Their Properties.

Principles / Properties	**Net Premium**	**Expectation Principle**	**Variance Principle**	**S.D. Principle**	**Exponential Principle**
Non-Negative Safety Advance	+	+	+	+	+
No Swindle	+	-	-	-	+
Consistency	+	-	+	+	+
Additive	+	+	+	-	+
Iterative	+	-	-	-	+

Only two premium principles have all the properties, the net premium and the exponential principle. Although the expectation principle does not have all the properties, this premium principle will be used. As the market value of D/P (=damage divided by premium) needs to be worked with, one needs to multiply the predicted amount of damage paid by P/D, so the α = P/D-1.

4.3.3 Calculation of the Predicted Amount of Damage

The expectation of the damage cannot be calculated by using logistic regression, because one does not have a dichotomous variable as depended variable, but a continuous variable Y_2, the average amount of damage claimed by an individual over the past three years. To achieve this, generalized linear models (GLM) have been used, which are implemented in the SAS-procedure GENMOD.

The class of generalized linear models is an extension of traditional linear models. For traditional linear models, the assumption that the data is normally distributed must be made. In our case it is an exponential family with a nonlinear link function because the generalized linear models allows the mean of a population to depend on a linear predictor through a nonlinear link function and allows the response probability to be any member of an exponential family of distributions. A generalized linear model will be used.

When GLM is used one gets a linear component $\eta_i = \mathbf{x}_i'\beta$. A monotonic differential link function g is needed to describe the relationship between the expected value of Y_i, the dependent variable for observation i, and the linear predictor,

$$g(Ey_i) = g(\mu) = \eta_i = \mathbf{x}_i'\beta .$$

The variable Y_i is independent for all i and has a probability distribution of an exponential family. The variable in this case will be the average amount of damage that has been claimed over the past three years. This variable has theoretically a Gamma distribution, as seen in the introduction section 1.3. The logarithm as link function is used because the mean must be a positive value without a boundary.

To estimate the parameters of vector β, the procedure uses Maximum Likelihood estimation. The parameters are estimated numerically through an iterative process.

An important aspect of the GLM is the selection of the explanatory variables in the model. Because PROC GENMOD does not have the possibility to use backward elimination or forward selection, this has to be done by oneself.

The first step is to look which variable is the most explanatory variable for the dependent variable. This will be done by just putting in one variable and letting the procedure calculate the Chi-square and the deviance. The deviance, defined to be twice the difference between the maximum attainable log likelihood and the log likelihood of the model under consideration, is often used as a measure of goodness of fit. First the P-value of the chi-square will be looked at and this value must be smaller than 0.05 to be significant. When variables have the same significant P-value, one should look at the deviance. If the deviance is small according to its decrease of freedom then it is a possible indication of a good model. The variable with the most significant P-value and with the smallest deviance will be added to the model first. These steps will be repeated until none of the variables are significant anymore, and then one might have the best predictable model.

Because Y has a gamma(ν, α) distribution with the logarithm as link function, the probability function of the dependent variable will be Equation 4-4.

$$f(y)=\frac{\alpha^{\nu}}{\Gamma(\nu)}y^{\nu-1}e^{-\alpha y} \qquad \textbf{Eq 4-4}$$

$$\text{with } \alpha=\frac{\nu}{\mu},\ E(Y)=\frac{\nu}{\alpha} \text{ and } Var(Y)=\frac{\nu}{\alpha^2}$$

Here the scale parameter ν is estimated by maximum likelihood and is also given in the output.

After using this distribution for the estimation of the parameters by maximum likelihood, the expected amount of claimed damage can be written as Equation 4-5.

$$EY = \mu = e^{x'\beta} \qquad \text{Eq 4-5}$$

For the model of the probability of damage, one should do the same as in Chapter 3. Now it will need to be done three times, separately every year. For the average probability which is necessary for the calculation of the premium, the average of these probabilities will be taken.

Now one has the predicted amount of damage and the expected damage frequency, so the premium can be calculated. Only one thing must not be forgotten. It has already been noted, that only 66% of the premium people pay is for the payment of the amount of damages. It has been stated that one should use the expectation principle, with $\alpha = P/D - 1$, that will be in this case:

$$\alpha = 3/2 - 1 = ½.$$

This premium will be compared to the real yearly premium by just taking the difference. Further, one needs to know whether the premium is indeed a good measure for a high risk. The premium of a high risk has to be compared with the premium of a low risk. The premium of a high risk must be higher than one for a low risk.

4.4 Results

This section will be divided into three parts. First, the results of the expected amount of damage and the relation between the dependent and the independent variables will be given. Secondly, the probability of damage will be given and in the last part the premium will be calculated and looked at if this premium can be a measure for the risk of a client. Also in this part will be looked at how the insurance company has set their premiums, if they get enough premiums to pay all the expected amount of damages.

4.4.1 Predicted Amount of Damage

At first the results of the Kruskal-Wallis test will be given. There are only three variables which might be said that their classes are not homogeneous. This means that there is a difference between the classes of the claimed amount of damage. The variables that are not homogeneous are AGE, BM and DUR. The P-value, the Chi-square, the degrees of freedom, the total observations per class and the mean of the ranks per class will be given in Table 4-2.

Table 4-2 Results of the Kruskal-Wallis Test

Variable	Chi-square	d.f.	p-value	Number of observations	Mean of the ranks
AGE	20.307	8	0.009	2722	
1				28	1654.55
2				189	1546.12
3				743	1372.88
4				755	1343.29
5				494	1338.13
6				267	1552.20
7				167	1292.73
8				38	1138.29
9				41	1268.59
DUR	19.405	7	0.007	2722	
0				25	1786.68
1				152	1536.53
2				295	1578.91
3				249	1363.25
4				169	1389.26
5				138	1400.97
6				156	1266.39
7				1538	1336.72

BM	17.744	7	0.013	2722	
1				861	1428.75
2				925	1307.66
3				428	1304.49
4				190	1409.06
5				84	1456.49
6				108	1268.17
7				51	1393.68
8				75	1464.43

For the prediction of the amount of damage, the procedure GENMOD will be used, with the selection as in Section 4.3.3. The first variable which needs to be added to the model is BM, the second is SPO, then DUR, SEX, KW, PAY, BRA, GEO and last the variable, RAP. The estimated parameters, the degrees of freedom, their standard error, the Chi-square value and the P-value of the estimation are given in Table 4-3.

Table 4-3 Analysis of Parameter Estimates

Variable	Degrees of freedom	Parameter estimation	Standard error	Chi-square	P-value
Intercept	1	11.9046	0.1301	8368.7117	0.0001
BM	1	-0.0968	0.0129	56.3536	0.0001
SPO	1	1.3777	0.2069	44.3437	0.0001
VER	1	-0.0515	0.0107	23.2816	0.0001
SEX	1	0.2340	0.0518	20.4461	0.0001
KW	1	-0.0736	0.0173	18.1313	0.0001
PAY	1	0.0717	0.0259	7.6323	0.0057
BRA	1	0.0131	0.0039	11.2055	0.0008
GEO	1	0.0305	0.0119	6.5618	0.0104
RAP	1	0.2838	0.1369	4.3001	0.0381

For all the variables, there is a significant result of the estimation of the parameters. The estimated scale value ν is equal to 0.6868 with a standard deviation of 0.0158.

The model which has been found has been used to calculate for every policy the predicted amount of damage. As one would like to know if this could be a good method for accepting or rejecting a client, one needs to make two plots. Figure 4-1 shows the distribution of the predicted amount of damage for the high risks, thus for the people to whom an amount of money has been paid over the past three years. The second plot, Figure 4-2, also shows the distribution of the predicted amount of money, only then for the low risks, so for the people who never received any amount of money over the past three years.

To give an impression of the statistical values of these distributions a short summary will be given in Table 4-4.

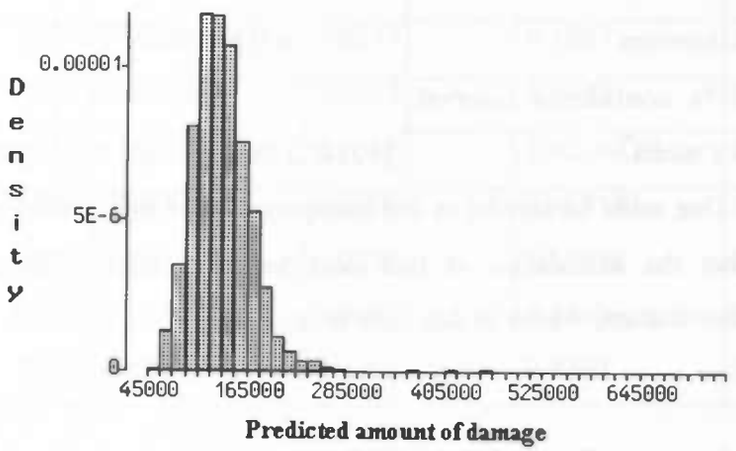


Figure 4-1 Predicted Amount of Damage for High Risks

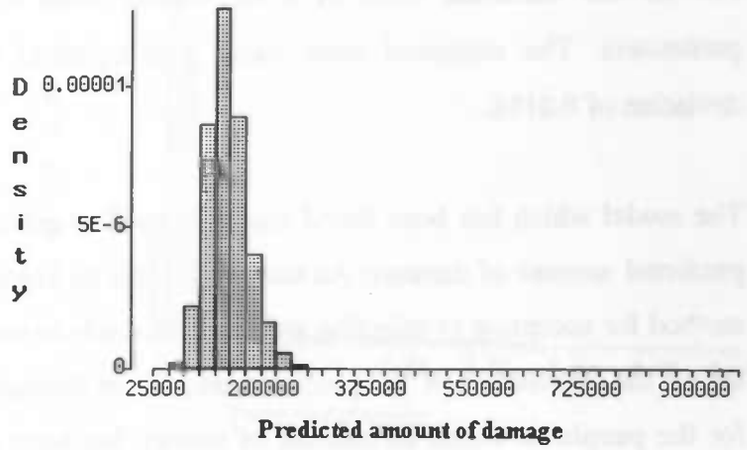


Figure 4-2 Predicted Amount of Damage for Low Risks

Table 4-4 Statistics for the Expected Amount of Damage

Statistics	High Risk (Fig 4-1)	Low Risk (Fig 4-2)
Mean	137810.85	149702.62
Standard deviation	52148.6587	53933.4122
Maximum	747059.42	967141.41
Minimum	47586.29	42616.64
95% confidence interval for mean*	[135852.00 ; 139769.70]	[148939.93 ; 150456.31]

* One must be careful in the interpretation of this confidence interval, as it can be that the calculation of this confidence interval is based on normality of the distribution, which in this case is not true.

4.4.2 Average Probability of Damage

The model for 1998 gives a declaration of 62.1% with the variables BM, PRI, KW, NEW and PAY. The model for 1999 has a declaration of 64.1% with the variables GEO, PAY, BM, KW, BRA, AGE and TOT. The model for 2000 has the largest declaration of 65.3%. The variables which need to be in this model are BM, KW, NEW and REG.

The parameter estimates of the three models will be given in Table 4-5 with the standard error, Chi-square and the P-value.

Table 4-5 Parameters Estimates for 1998, 1999 and 2000

Variables		Parameter Estimates	Standard Error	Chi-square	P-value
1998	intercept	-3.3073	0.1696	380.0984	0.0001
	PAY	0.0747	0.0370	4.0777	0.0435
	BM	0.3235	0.0200	260.8915	0.0001
	KW	0.0487	0.0240	4.1306	0.0421
	PRI	-0.4116	0.1210	11.5794	0.0007
	NEW	-0.1455	0.0715	4.1410	0.0419
1999	intercept	-4.1943	0.2026	428.7942	0.0001
	GEO	0.0349	0.0157	4.9283	0.0264
	PAY	0.0923	0.0364	6.4525	0.0113
	BM	0.3246	0.0189	295.6847	0.0001
	KW	0.0607	0.0227	7.1307	0.0076
	BRA	0.0189	0.0054	12.4469	0.0004
	AGE	0.0609	0.0237	6.5807	0.0103
	TOT	-0.0652	0.0268	5.8973	0.0152
2000	intercept	-3.9597	0.1065	1382.0827	0.0001
	BM	0.3439	0.0109	326.8193	0.0001
	KW	0.0840	0.0235	12.7856	0.0003
	NEW	0.1420	0.0697	4.1548	0.0415
	REG	0.0961	0.0305	9.9362	0.0016

When the average of these models is taken one gets the probability of damage. For the distribution of the probability of damage two graphs are again given in Figure 4-3 and Figure 4-4.

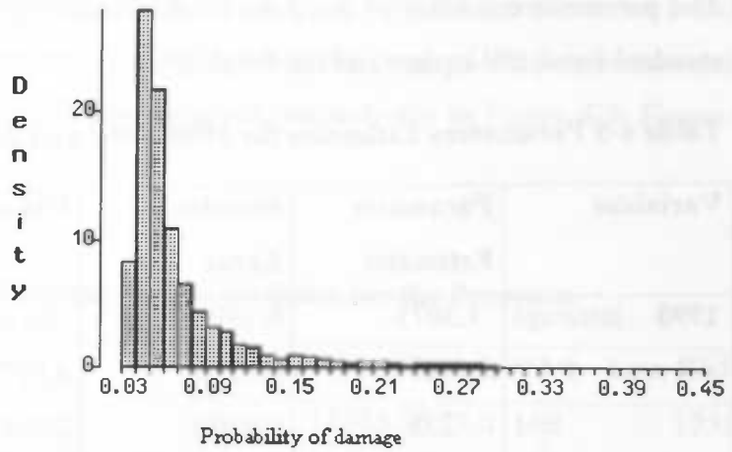


Figure 4-3 Distribution of the Probability for High Risks

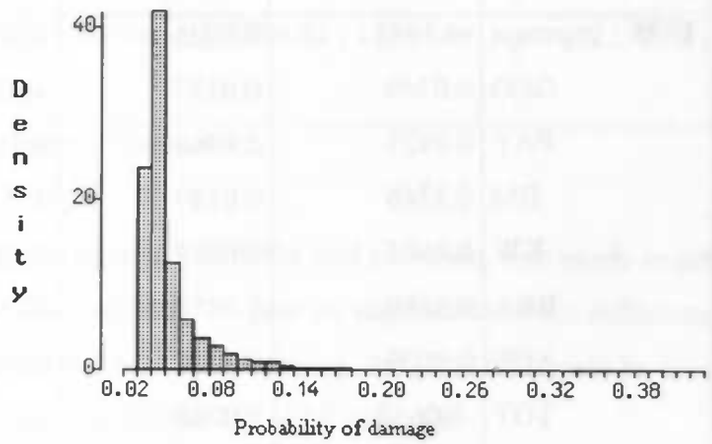


Figure 4-4 Distribution of the Probability for Low Risks

Also for a better interpretation the most important statistics will be given in Table 4-6.

Table 4-6 Statistics for the Probability of Damage

Statistics	High Risk (Fig 4-3)	Low Risk (Fig 4-4)
Mean	0.0745	0.0562
Standard deviation	0.0522	0.0341
Maximum	0.4494	0.4255

Minimum	0.0302	0.0282
95% confidence interval for mean**	[0.0725 ; 0.0765]	[0.0557 ; 0.0567]

** See commentary at Table 4-4.

4.4.3 Premium

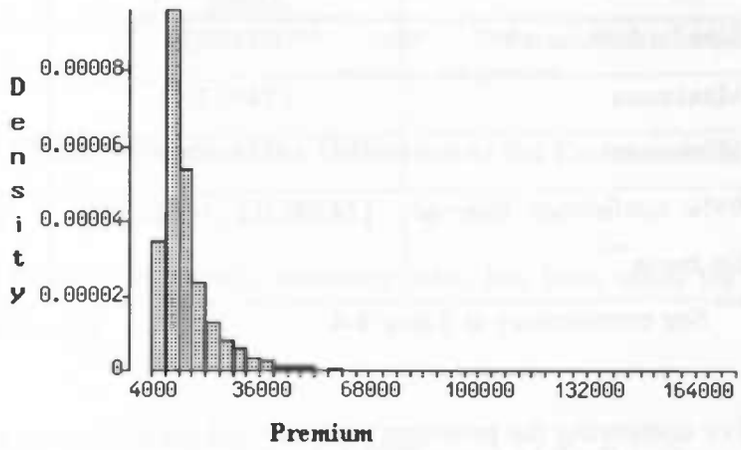


Figure 4-5 Distribution of the Premium for High Risks

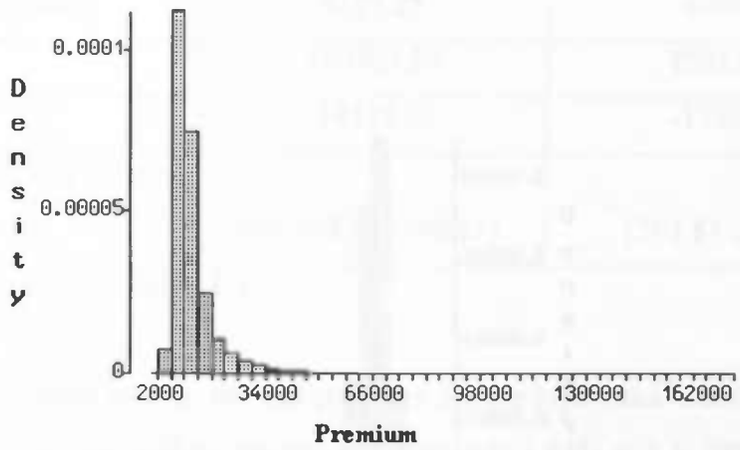


Figure 4-6 Distribution of the Premium for Low Risks

After multiplying the results of 4.4.2 with the results of 4.4.1 and with 3/2 one gets the premium. For the premium the graph of the distribution and a summary of the statistics will also be given, respectively in Figure 4-5, Figure 4-6 and Table 4-7.

Table 4-7 Summary of the Statistics for the Premium

Statistics	High Risk (Fig 4-5)	Low Risk (Fig 4-6)
Mean	14585.06	12364.95
Standard deviation	10090.70	8087.46
Maximum	174953.52	173732.68
Minimum	4456.04	4163.00
95% confidence interval for mean***	[14206.02 ; 14964.09]	[12250.58 ; 12479.31]

*** See commentary at Table 4-4.

For comparing the premium with the real premium, one needs to subtract the real premium from the premium. To give an impression of the difference between the real premium and the expected premium, the plots will again be divided into high risks and low risks, see Figure 4-7 and Figure 4-8.

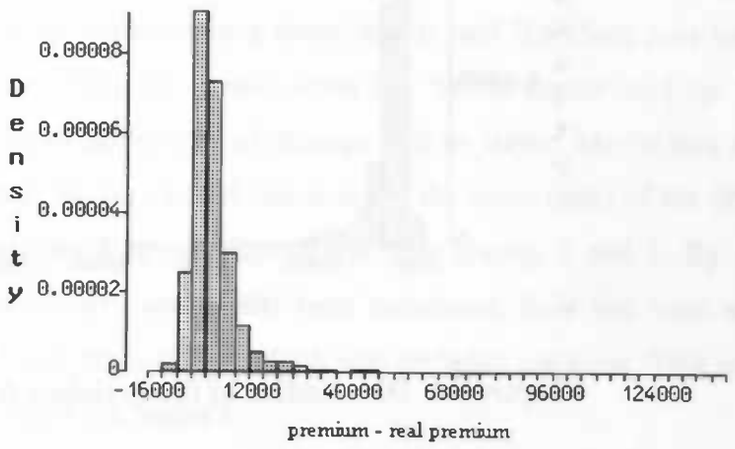


Figure 4-7 Distribution of the Difference of the Premiums for High Risks

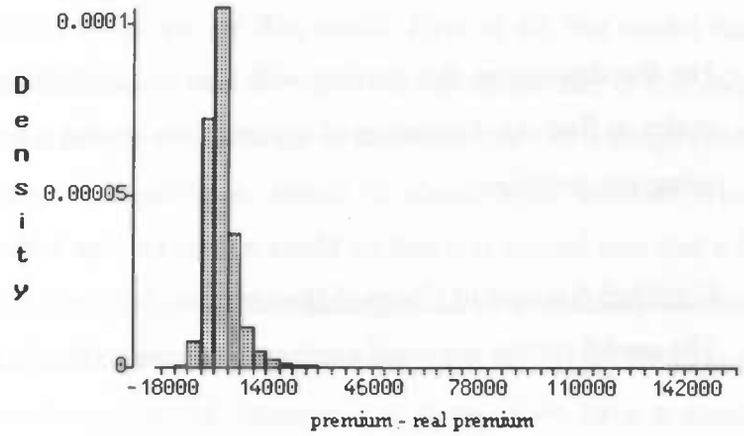


Figure 4-8 Distribution of the Difference of the Premiums for Low Risks

For more statistics, a statistic summary table has been made for the difference again in Table 4-8.

Table 4-8 Summary of Statistics for Difference of the Premiums

Statistics	High Risk (Fig 4-7)	Low Risk (Fig 4-8)
Mean	1932.97	689.00
Standard deviation	8177.25	6797.62
Maximum	140765.52	156120.69
Minimum	-14519.21	-17552.09
95% confidence interval for mean[+]	[1625.81 ; 2240.13]	[592.87 ; 785.12]

[+] See commentary at Table 4-4.

According to these results, the real premium for the high risks is between 11.44% and 14.97 % too low, and for the low risks between 4.84% and 6.29% too low.

4.5 Discussion

For the discussion this section will also be divided into three parts, so that it is easier to find the discussion of a part of the research to determine the high risk by using the premium.

Expected Amount of Claimed Damage

The model for the expected amount of claimed damage is:

$$EY = e^{11.9046+1.3777*SPO+0.2838*RAP+0.234*SEX-0.0968*BM-0.0736*KW-0.0515*VER+0.0717*PAY+0.0305*GEO+0.0131*BRA}$$

Eq 4-6

The strangest part of this is that the variable AGE is not in the model although the classes have different means. The cause for this can be that the Kruskal-Wallis test works with ranks instead of with the real values.

When looking just at the declaring variables, one sees that two variables tells something about the payment of the client, and also the geo-model score tells something about the payment of the client because it is made for a credit company. Probably the payment behavior of a client is very important for the amount of damage, as one can see that the expected amount of damage is higher if a client had a registered dunning letter, and as well if a client pays less frequently a year. One can also see is that when the 'bonus malus' and the kilowatt are smaller, the expected amount of damage will be larger. Maybe this demonstrates the strange part of this case. If one looks at the mean ranks of the BM, there can be seen that class 8 is the highest and then classes 5 and 1. By the way the predicted amount of damage has been calculated, their has been assumed that there was a linear relationship, which was probably not good. This comment will be looked at again in Chapter 6.

The odd thing that one sees in this case, is that the expected amount of money for the low risk is higher than the expected amount of money for the high risk. Two explanations can be given for this result. First of all, the model has been made only with the information of the high risk. Because this is the only part of the portfolio for the last three years which had an amount of damage paid. This part is about 12.5% of the portfolio, which is about 2,722 policies which could be problematic. But another reason could be that it is indeed true that a low risk does have in higher expected amount of damage, but its probability of damage must lie much lower than the probability of damage for a high risk. This means that a low risk, does not have a lot of damages, but if one does have a damage, then the amount of damage is very high. There are indeed a lot of people who have had only one damage claimed in the past ten years and the amount of that damage was indeed very high. So, it might be wrong or it might be good. In the next discussion both explanations will be examined to draw some conclusions.

It is known that the maximum of the real amount of damage is 12,000,000 BEF, and that the average of the expected amount of damage of high risk, must be almost equal to the average of the real amount of money. The averages must be almost equal because the model of the expected amount of damage is based on the real amount of damage. Therefore, one can say that the expected amount of damage has a smaller standard deviation than the real amount of money.

Looking at Figures 4-1 and 4-2, there can be seen that there is a lot of difference between the expected amount of damages. The minima for respectively high risks and low risks are 47,586.29 BEF and 426,192.64 BEF while the maxima for respectively high risks and low risks are 747,059.42 BEF and 967,141.41 BEF. For high risk that is a difference of almost 700,000 BEF and for the low risks that is even more, about 924,500 BEF. Although the standard deviation must be smaller than the standard deviation of the real amount of damage, there is still a big difference.

For accepting a policy it is hard to use this result on its own, because one does not know if the predicted amount of damage for the low risks must indeed be larger

than the predicted amount of damage for the high risks. Therefore one needs to take a look at the average probability of damage over the past three years.

Probability of Damage

The first thing that can be seen, is that the models for each year are very different. First the models will be shown and then something will be explained about it.

$$P(y_{98}=1)=\frac{e^{-3.3073+0.0747*PAY+0.3235*BM+0.0487*KW-0.4116*PRI-0.1455*NEW}}{1+e^{-3.3073+0.0747*PAY+0.3235*BM+0.0487*KW-0.4116*PRI-0.1455*NEW}}$$

$$P(y_{99}=1)=\frac{e^{-4.1943+0.0349*GEO+0.0923*PAY+0.3246*BM+0.0607*KW+0.0189*BRA+0.0609*AGE-0.0652*TOT}}{1+e^{-4.1943+0.0349*GEO+0.0923*PAY+0.3246*BM+0.0607*KW+0.0189*BRA+0.0609*AGE-0.0652*TOT}}$$

$$P(y_{00}=1)=\frac{e^{-3.9597+0.3439*BM+0.0840*KW+0.1420*NEW+0.0961*REG}}{1+e^{-3.9597+0.3439*BM+0.0840*KW+0.1420*NEW+0.0961*REG}}$$

There are only two variables which are in all the models, namely BM and KW. That is not odd, because these variables are very important for the damage frequency. According to the law, the bonus males and the kilowatt must be taken into the premium calculation. It is good that these variables are also in these three models. Hopefully this tells something about the correctness of the three models. Then there are also two variables which are in two of the three models, namely PAY and NEW. Luckily those variables are used in the same way for the two models. The way in which the variable PAY is used tells us that the less frequent a person pays the higher the probability is of damage, and this yields also for a new case. It can be that the new cases are mostly for young people and that they mostly start at a bonus males of 11. So there might be a correlation with the variable BM and AGE. One sees that if NEW is not in the model, the variable AGE is in the model. Furthermore there are in two models the region score or the geo-model score and some other declaring variables. There can also be seen that the declaring percentage of the model of 1998 is the worst and the declaring percentage of model 2000 is the best. Whether the models has a good declaration does not depend on how much declaring variables are in the model.

When this information will be used for looking at the average probability of damage. One sees that the average probability of high risks lies about 1.8% higher than the average probability of the low risks. So the conclusion that can be drawn for the expected amount of damage can indeed be correct, the expected amount of damage for a low risk is higher than for a high risk.

Premium

Indeed when one looks at the expected premium, there can be seen that the premium for a high risk is higher than the premium for a low risk. The premium could be used to determine a high risk. One can say, the higher the premium, the worse the risk. For the acceptance of a client the insurance company must declare a maximum premium. If a new client exceeds this premium than the client has a high probability of being rejected. This result can also be used for calculating the premium a person need to pay for their insurance policy.

To give an impression of this expected premium, it has been compared to the real average year premium. Looking at Figures 4-7 and 4-8, one can see that the insurance company asks a too small premium to cover the amount of damages. The most important thing that one can see is the insurance solidarity: low risks pay for the high risks. This solidarity is good, but in this case a bit too much.

So the premiums for automobile insurances must be increased. The premiums for a high risk must go up much more than the premiums for a low risk. And if one does not accept the fact that the expected amount of damage of a low risk is higher than for a high risk, the premium for a low risk does not have to increase very much. Since the expected premium will then be much lower.

Since it is not easy to use models like this for the acceptance of a client, one could make it easier by making a score card. This will be done in Chapter 5.

Chapter 5

The Scorecard for Accepting or Rejecting a Client

An insurance company needs to make a decision about whether they should accept a client or reject a client. Normally they look at a few variables whether an applicant looks good or not, for example, the claimed damages of the client, the years of no damage, the way in which the client paid the premiums or the' bonus malus' grade of the client. However, it would be much easier and better if they had a card to give a better idea of what the risk of the client would be according to the information the insurance company could get from the client and whether it results in a low risk or a high risk. It is not necessary that this card would give a direct decision, but it would give a better indication for the acceptance of insurance policies. The focus of this chapter will be how one can make a card which can help to accept or reject a client.

5.1 Introduction

In Chapter 3 the probability that a client has a high risk has been estimated. In Chapter 4 the premium was estimated and one has seen that it gives a good indication of a high risk. These two results can be used to accept or reject a client. It is even a better method, but it is not an easy method to use. Therefore, one needs to find an easier way for the acceptant of the insurance company. A scorecard needs to be made. A scorecard is a table that shows for each attribute the number of points that are to be awarded to a client showing that attribute. This means that for every explanatory variable there are classes, each class having a

specific number of points. These number of points, for the different attributes of which the client has, needs to be totaled. The sum, which will be called the score, must be smaller than a preset value, which is in statistical terms the critical value and in scoring terms the cutoff value, to automatically reject the client or to recommend to reject the client. Of course, when the sum lies around the cutoff value one will still need to look what one thinks is the best thing to do, accept the client or reject the client, because the card cannot be the only part used in the decision to reject a client.

5.2 Data

For making a scorecard, almost the same dependent variable as in Chapter 3 will be usde. The variable to be declared is:

$$Y_4 = \begin{cases} 1 & \text{if a claim has been paid per year} \\ 0 & \text{if not} \end{cases}$$

The difference of this variable with Y_1 is that it will be used for every year, so a client might have a high risk in 1998, and a low risk in 1999 and 2000. We do this for the reason that the frequency of damage must also be taken into account and this will give a practical scorecard and not a theoretical scorecard. A theoretical score card means that it can be assumed that if a client has had a claim been paid, that client will again have a claim paid over x years. While a practical score card assumes that there are clients who never will have had a claim paid and other clients who frequently have a claim paid.

Next to these variable there are, of course, the independent variables, which are the same as in the previous chapters. Only now the classes will really be used and the whole variable will not be seen as one.

When one has the estimated scores, a cutoff value needs to be found, and one can try to determine this value by using the definition of low and high risks over the three years together, Y_1 because that is the accepting or rejecting base.

5.3 Method

This section will be divided in two subsections. The first section will tell how one makes a scorecard, and the second section explains the theory of finding the best cutoff value.

5.3.1 Making a Scorecard

The first part of this chapter will be devoted to making a scorecard. This can be done in various ways. However, logistic regression has been chosen as there is again a dichotomous dependent variable. Only now one does not want to use the whole variable as one, but one wants to use the classes. This means that dummy variables have to be created. The logistic regression function used in Chapter 3 looked like Equation 5-1.

$$\text{logit}(Y) = \ln\left\{\frac{P(Y=1))}{1-P(Y=1)}\right\} = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \ldots + \beta_k * x_k \quad \textbf{Eq 5-1}$$

Equation 5-1 is rewritten with dummy variables one gets Equation 5-2.

$$\text{logit}(Y) = \ln\left\{\frac{P(Y=1))}{1-P(Y=1)}\right\} = \beta_0 + \beta_{11} * x_{11} + \beta_{12} * x_{12} + \ldots + \beta_{jk} * x_{jk} \quad \textbf{Eq 5-2}$$

In Equation 5-2 every x_k is written as x_{jk}, where the j stands for the variable number and the k for the number of the class of the variable. These variables are, just like the dependent variable, dichotomous variables. A client either has that attribute or does not have that attribute.

One now has more parameter estimates to make as their must be a parameter estimate for every dummy variable. If there are five variables which each four classes, there have to be estimated twenty parameters β.

Dummy variables could be created by using the procedure for logistic regression, but it is much easier to use PROC GENMOD for this purpose. As one can enter in the variables as classes, these classes automatically become dichotomous variables.

One is allowed to use PROC GENMOD, as the binomial distribution is part of the exponential family. So the distribution function to be used is binomial and the link function is the logit, as one then again uses the logistic regression, which should be used as the dependent variable is dichotomous.

To find the best model for the predicting the dependent variable, the same method as prescribed in Chapter 4 is used and the parameters are also estimated in the same way. The only difference is that in the last class the variable will be set at zero, as one class is always dependent on all of the other classes. The PROC GENMOD always uses for that class the last class and that class will be used to calculate the other parameter estimates.

The parameters will be estimated for each class of variables. These values, β_{ij}, will be used for the number of points for the scorecard. One wants a scorecard where the score must lie between 0 and 1000. Therefore the parameter estimates needs to be transformed to get useful number of points. The steps for making the transformation are as follows:

1. First, multiply the parameter estimates, β_{ij}'s, by 1000 and round the values.
2. Then looking at each variable, find which class has the smallest value. These values from all the variables must be totaled and this sum divided by the total number of variables in the score card. This absolute value must then be added to all the values.
3. Now, for these values, the result of step 2, one needs to look which class has the largest values. These values of all the variables must also be totaled. Then divide 999 by the sum of the maximum values to arrive at a multiplication factor.

4. Finally: multiply all the values, from step 2, by the multiplication factor, and round the values to get the number of points for a scorecard which runs from 0 to 999.

Here is a short example:

A variable X has three classes, 1 , 2 and 3. The parameter estimates are respectively 0.2354, 0.0505 and 0. Also there is a variable Y which has four classes 1, 2, 3 and 4, with parameter estimates respectively –0.1302, 0.2001, 0.0065 and 0.

1. X 1 0.2354*1000 = 235.4 => 235
   X 2 0.0505*1000 = 50.5 => 51
   X 3 0*1000 = 0 => 0
   Y 1 -0.1302*1000 = -130.2 => -130
   Y 2 0.2001*1000 = 200.1 => 200
   Y 3 0.0065*1000 = 6.5 => 7
   Y 4 0*1000 = 0 => 0

2. MIN(X1, X2, X3) + MIN(Y1, Y2, Y3, Y4)= 0 +(-130) = -130
   -130/2 = -65
   X 1 235 + 65 = 300
   X 2 51 + 65 = 116
   X 3 0 + 65 = 65
   Y 1 -130 + 65 = -65
   Y 2 200 + 65 = 265
   Y 3 7 + 65 = 72
   Y 4 0 + 65 = 65

3. MAX(X1, X2, X3) + MAX(Y1, Y2, Y3, Y4) = 300 + 265 = 565
   999/565 = 1.76814

4.	X 1	300 * 1.76814 = 530.442 =>	530
	X 2	116 * 1.76814 = 205.104 =>	205
	X 3	65 * 1.76814 = 114.929 =>	115
	Y 1	-65 * 1.76814 = -114.929 =>	-115
	Y 2	265 * 1.76814 = 468.557 =>	469
	Y 3	72 * 1.76814 = 127.306 =>	127
	Y 4	65 * 1.76814 = 114.929 =>	115

Then one has a scorecard. One needs to check whether the card indeed runs from 0 to 999. This can be done by recalling the first part of Step 2, if this sum is 0, the minimum is alright. For checking the maximum of the card, just recall the first part of Step 3, if this sum is 999, the score card is ready. Now each client can be given a score but a cutoff score to determine whether to accept or reject a client still needs to be determined.

5.3.2 Determine a Cutoff Value

Determining a cutoff value can also be done in very different ways. When it is not important what the cost is for wrongly accepting or rejecting a client, one can use a decision theory by just looking at the distribution and deciding where the cutoff value should be. In this study's it is important what the costs are, so the cutoff value was determined by using a decision theory with examining the costs of making a wrong decision.

A decision theory can be divided into several parts. This study will just look at one part of the decision theory, the hypothesis testing problem. This means that one has an action space A, the set of allowable decisions, with only two possibilities, to accept or to not accept the hypothesis. So the action space is the set $A = \{a_0, a_1\}$, where a_0 stands for accepting the hypothesis, in this case accepting a client, and a_1 stands for not accepting the hypothesis or rejecting a client.

In the decision theory a loss function must also be specified. This means that if θ is the true state of nature, which means it has a low risk or a high risk, and decision a is taken that $L(\theta,a)$ gives the loss incurred in that decision. If $L(\theta,a)=0$ then the decision is correct. So if one knows that a client has a low risk and is accepted, then the decision is correct.

The loss function must specify the costs for making incorrect decisions. There are two incorrect decisions:

1) if $\theta \in \Theta_0$ and decision a_1 is made this means that the client is a low risk and the client will be rejected

2) if $\theta \in \Theta_0^c$ and decision a_0 is made this means that the client is a high risk and the client will be accepted.

If one of these two decisions is made the loss function must specify this. For the specification of the loss function a generalized 0-1 loss is used.

$$L(\theta,a_0)=\begin{cases}0 & \theta \in \Theta_0 \\ c_{II} & \theta \in \Theta_0^c\end{cases}$$
$$L(\theta,a_1)=\begin{cases}c_I & \theta \in \Theta_0 \\ 0 & \theta \in \Theta_0^c\end{cases} \qquad \textbf{Eq 5-3}$$

With c_I the loss for a type I error, falsely not accepting H_0, thus a client with a low risk will be rejected and c_{II} the loss for a type II error, falsely accepting H_0, thus a client with a high risk will be accepted.

To specify which action to take, there is a decision rule $\delta(x)$. The decision rule specifies what action $\mathbf{a} \in A$ will be taken, if $X = x$ is observed, where X stands for the observed score of a new client, which lies between 0 and 1000.

$\delta(x)=a_0$ $\quad \forall x$ that are in the acceptance region of the test

$\delta(x)=a_1$ $\quad \forall x$ that are in the rejection region of the test

Here the acceptance region means that the score of a client, X, must be equal to or larger than the critical value c. And the rejection region is where the score of the client is smaller than that of the critical value.

The quality of a decision rule is quantified in the risk function of the decision rule.

$$R(\theta,\delta)=E_{\theta}L(\theta,\delta(X)) \qquad \textbf{Eq 5-4}$$

This means that at a given θ, the risk function is the average loss that will be incurred if the decision rule $\delta(x)$ is used. For this risk function one needs a power function, $\beta(\theta)$, based on the decision rule $\delta(x)$.

$$\beta(\theta)=P_{\theta}(\delta(X)=a_1) \qquad \textbf{Eq 5-5}$$

When one puts this result and the specified loss function in the risk function, one gets the specified risk function, as in Equation 5-6, which is also called the Bayes Risk.

$$\begin{aligned} R(\theta,\delta(X)) &= 0 * P_{\theta}(\delta(X)=a_0)+c_I * P_{\theta}(\delta(X)=a_1)=c_I\beta(\theta) && \text{if } \theta\in\Theta_0 \\ R(\theta,\delta(X)) &= c_{II} * P_{\theta}(\delta(X)=a_0)+o * P_{\theta}(\delta(X)=a_1)=c_{II}(1-\beta(\theta)) && \text{if } \theta\in\Theta_0^c \end{aligned}$$

Eq 5-6

To get the minimum loss the results of the risk function have to be minimized, which is not that easy, as minimizing the loss of an incorrectly accepted client, increases the loss of incorrectly rejecting a client.

As the probability function from the company's portfolio is known, this study has an unique solution for the best cutoff value. But this does not mean that this is also an unique solution for the whole of Belgium, as one is working with a select sample of the Belgium people. In this case, with this portfolio one has the distribution functions, but it can be that when it is used for the new real clients it is

not good. However, one now has an overall probability of high risk of 12.5% and this study does not think that the probability of having a high risk for all the people in Belgium lies much lower, so it might give a good result.

5.4 Results

The variables for predicting the dependent variable Y_4 are BM, VER, KW, REG, AGE and PAY. After having estimated the parameters and made the calculation as presented in the Method section, the scorecard is created, see Table 5-1.

Table 5-1 Scores for the Separate Classes

Declaring Variables	Classes	Scores for the Classes
'Bonus Malus'	1, …. , 8	665, 449, 476, 419, 437, 186, 63, 39
Insurance Duration	0, …. , 7	-50, 42, -2, 15, 33, 48, 39, 39
Kilowatt	1, …. , 6	60, 54, 54, 21, 36, 39
Region Score	1, …. , 5	72, 60, 72, -11, 39
Age	1, …., 9	105, 114, 69, 66, 69, 84, 45, -17, 39
Way of Payment	1, 2, 3	33, 18, 39

To give an idea why these scores are as prescribed the scores in Table 5-1, the plots of the distribution of these variables over the variable Y_4 will be given.

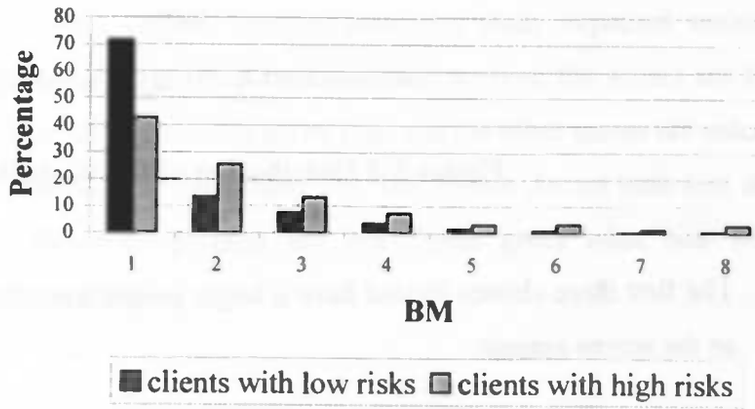


Figure 5-1 Distribution of the Variable BM

We can see in Figure 5-1 the reason why the scores decrease.

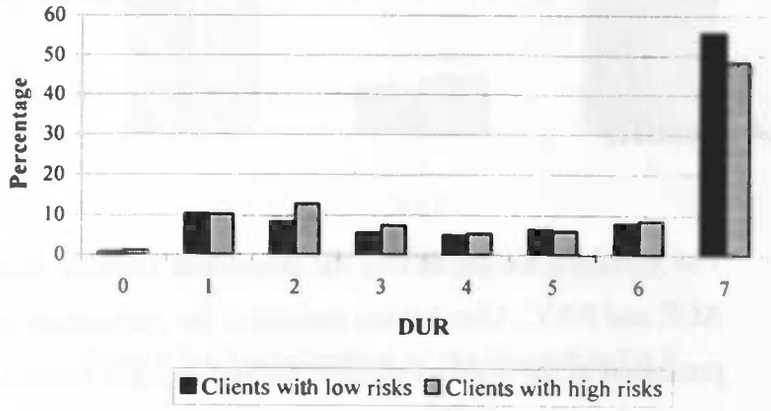


Figure 5-2 Distribution of the Variable DUR

This figure also gives a little indication of the values of the scores for the variable DUR.

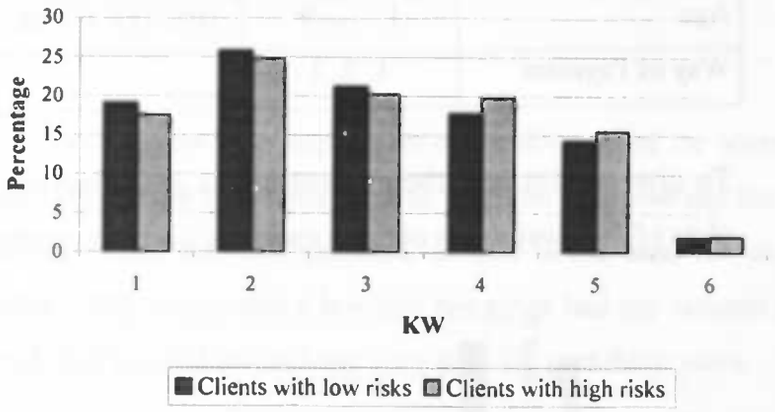


Figure 5-3 Distribution of the Variable KW

The first three classes indeed have a larger proportion of low risks than high risks as the scores assume.

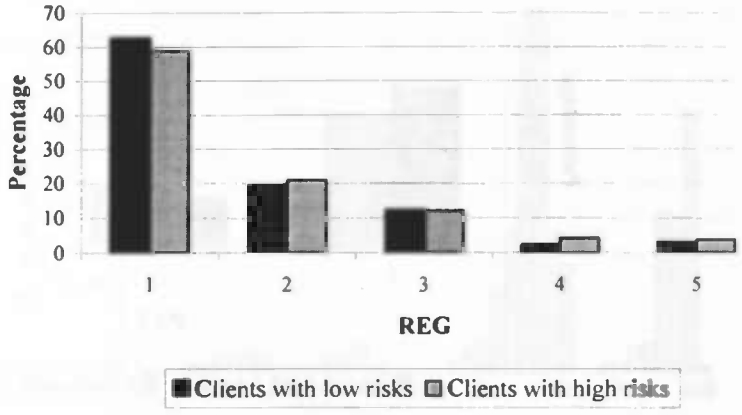


Figure 5-4 Distribution of the Variable REG

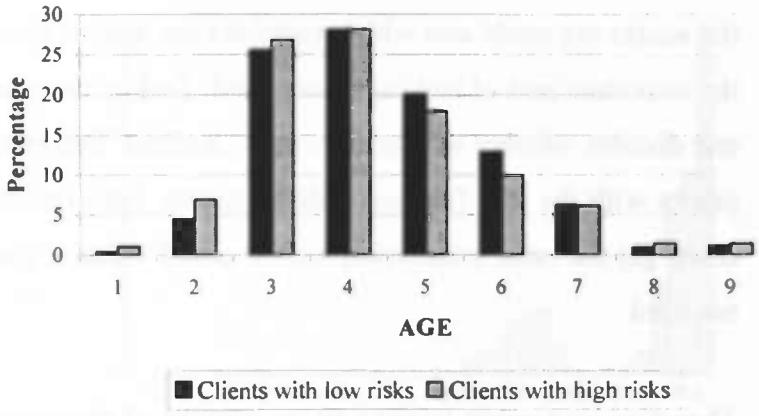


Figure 5-5 Distribution of the Variable AGE

The scores of this variable are not what this study expected according to this figure. However, as described in the method section, the scores are based on the fact that the last class has been set to zero and the other scores are calculated given that class. This can be the reason why the figures do not look like the values of the scores. However, perhaps the last figure gives what one would expect according to the scores.

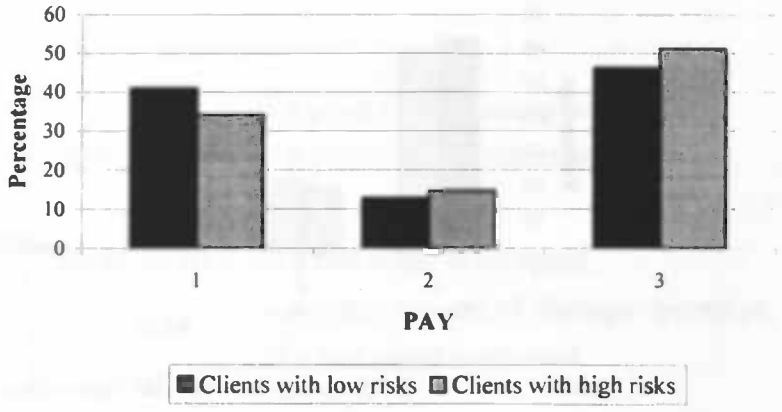


Figure 5-6 Distribution of the Variable PAY

Now that the distribution of the variables is known to get an impression of how the scores are made and which variables are indeed most important one can go to the important part of making a scorecard: finding the cutoff value, the value where one decides whether to accept or reject a client. Before this results of the decision theory with the loss function will be shown. First one needs to look whether one could get the same conclusion for the cutoff value if the generalized 0-1 loss was not used.

Therefore, one needs to have an impression of the division of the scores, and the graphs for the low and the high risks need to be given. The low and the high risks are the same as in the previous chapters, as this is the base for accepting or rejecting a client. This means that a low risk has never had any amount of damage paid and a high risk has had one at least once over the past three years.

Looking at Figures 5-7 and 5-8, one can try to find the cutoff value, but the best way to find the cutoff value is to use the loss function.

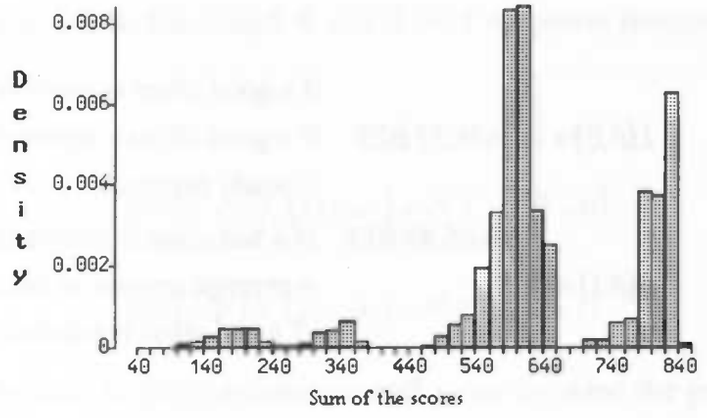


Figure 5-7 Distribution of the Sum of the Scores for High Risks

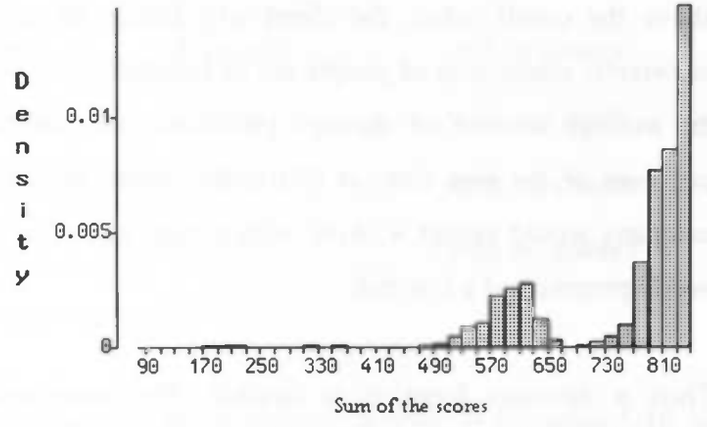


Figure 5-8 Distribution of the Sum of the Scores for Low Risks

First notations used in the method will be specified.

The hypothesis to be tested is:

$H_0 : \theta \in 0$ the client is a low risk

$H_1 : \theta \in 1$ the client is a high risk

This means that one accepts the H_0 if the score of the client, X, lies above c, the cutoff value. The hypothesis of this study is different for every cutoff value. The cutoff value needs to be found, where the loss is kept at its minimum.

The loss function is defined as in Equation 5-7.

$$L(\theta,0)=\begin{cases}0 & \text{if a good client is accepted}\\ 11676{,}52\ \text{BEF} & \text{if a good client is rejected,}\\ & \text{= yearly premium}\end{cases}$$

$$L(\theta,1)=\begin{cases}131686{,}40\ \text{BEF} & \text{if a bad client is accepted,}\\ & \text{= average amount of damage - premium}\\ 0 & \text{if a bad client is rejected}\end{cases}$$

Eq 5-7

The interpretation of the loss function is that if one knows that a client is a low risk but the score of the client lies below the cutoff value, a client is incorrectly rejected. Vice versa, if one knows that a client is a high risk but the score lies above the cutoff value, the client will falsely be accepted. The costs for this incorrectly classifying of people are in Equation 5-7. The c_I is calculated by using the average amount of damage per claim and subtracting the average yearly premium of the high risk, as this is the money one would lose if the insurance company would accept a client with a high risk. The c_{II} is equal to the average yearly premium of a low risk.

Then a decision function is needed. The acceptance region can be either 'accepting' or 'rejecting', so A = {accept, reject}. Here it can be said that if one accepts a client $a_0 = 0$ and if one rejects a client $a_1 = 1$. Thus the decision function will be the following:

$\delta_c(x)=0$ $\forall x$ that are in the acceptance region of the test, thus $X \geq c$.

$\delta_c(x)=1$ $\forall x$ that are in the rejection region of the test, thus $X < c$.

This decision function depends on the cutoff value, as does the acceptance and rejection regions. (depends on the cutoff value)

The only thing still needed to be specified is the power function $\beta(\theta)$. If $\theta = 0$ the power function will be Equation 5-8, and if $\theta = 1$ the power function will be equal to Equation 5-9.

$$\beta(\theta) = P_0(\delta_c(x) = a_1) = P(X < c|\theta = 0) \qquad \textbf{Eq 5-8}$$

$$\beta(\theta) = P_1(\delta_c(x) = a_1) = P(X < c|\theta = 1) \qquad \textbf{Eq 5-9}$$

One needs to look at every available cutoff value for what the probability is of being a high risk given whether the client has a low risk or a high risk. These probabilities will be calculated using the fact that the portfolio an be divided, with a total of n clients, into four parts:

Part I:	$\theta = 0$	$\delta_c(x) = 0$	with n_{I} clients
Part II:	$\theta = 0$	$\delta_c(x) = 1$	with n_{II} clients
Part III:	$\theta = 1$	$\delta_c(x) = 0$	with n_{III} clients
Part IV:	$\theta = 1$	$\delta_c(x) = 1$	with n_{IV} clients

with $n = n_{\text{I}} + n_{\text{II}} + n_{\text{III}} + n_{\text{IV}}$.

As the probability of X given Y can be written as Equation 5-10, one can use this to calculate the power function.

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} \qquad \textbf{Eq 5-10}$$

This gives the following results for the power function:

$$P(X < c|\theta = 0) = \frac{n_{II}}{n_I + n_{II}}$$

$$P(X < c|\theta = 1) = \frac{n_{IV}}{n_{III} + n_{IV}}$$

Thus, each cutoff value has its own probability of having a high risk given that the client is a low risk or a high risk since the number of observations in the four parts changes when the cutoff value changes.

Now one is able to determine the risk function. By using Equation 5-6 and filling in everything known, the result is Equation 5-11 with $\beta(\theta)$ as in Equations 5-8 and 5-9.

$$R(\theta, \delta_c(X)) = 11676.52 * \beta(\theta) \quad \text{if } \theta = 0$$
$$R(\theta, \delta_c(X)) = 131686.40 * (1 - \beta(\theta)) \quad \text{if } \theta = 1 \qquad \textbf{Eq 5-11}$$

To find the best cut off value, the minimum of these risk function must be found. Therefore, a graph, Figure 5-8, will be used which tells something about the loss per cutoff value.

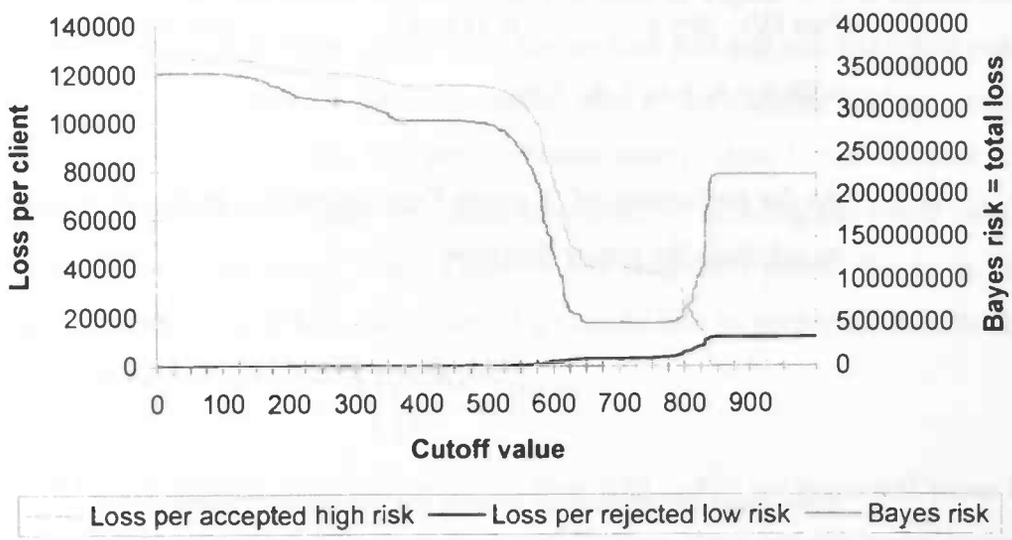


Figure 5-9 Loss of Accepting or Rejecting a Client Incorrectly, per Cutoff Value

And another graph, Figure 5-9, gives an idea of the distribution of high risks in the portfolio, but also of the percentage of people who are incorrectly classified and the percentage people who are rejected.

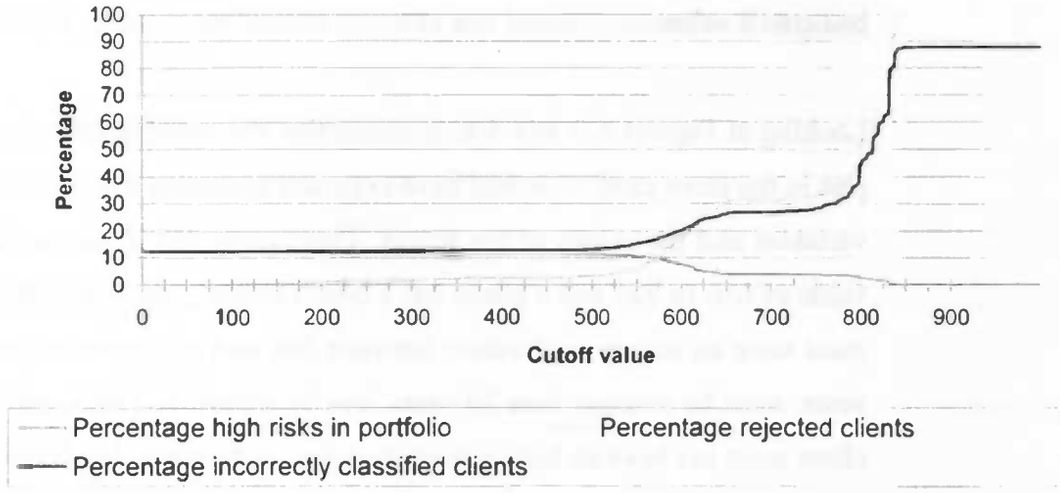


Figure 5-10 Several Percentages Under Influence of the Cutoff Value

5.5 Discussion

There are six variables which are important for accepting or rejecting a client. These variables are BM, DUR, KW, REG, AGE and PAY. When looking at the figures in the distribution of the variables, one sees that the biggest difference between the good and the high risks per class is in the BM. Only in Class 1 is the percentage of low risks larger than the percentage of the high risks. This can also be seen in the scores. If a client has 'bonus malus' grade 0, 1 or 2, the score is very high. For other variables, there is of course a difference, otherwise the variables would not be in the model, yet the difference is not that large as the difference for the variable BM. One must keep this in mind in the further conclusions of the scorecard and the cutoff value.

To getting an idea of where the cutoff value could be put, one should first look at the distribution of the scores. According to Figure 5-1, it would of course be best

to put the cutoff value at 840, but then one would not have many clients in the portfolio anymore. So the second attempt would be between 650 and 720. If one then looks at Figure 5-8, one sees that the best cutoff value would lie below 480. But if one takes that as cutoff value, a lot of high risks would be accepted. A second look is needed. Also in this case the range of 650 to 720 would not be a bad cutoff value.

Looking at Figures 5-7 and 5-8, it seems that the variable BM plays a very large part in the score card, as would have expected according to the distributions of the variables and the values of the scores. This means that if one would use a cutoff value of 650 to 720 and a client has a bonus-malus grade larger than 2, the client must have an insurance duration between one and two years or longer than four years, must be younger than 70 years, live in regions 1, 2 or 3, and the car of that client must not have an high-powered engine. If the client has bonus-malus 0, 1 or 2, hardly anything can go wrong with accepting this client.

To know what the best cutoff value is one can look at Figure 5-9. It shows that there are not many high risks with a score higher than 840 and not many low risks with a score lower than 520. This is probably due to the variable BM as a large part of the clients with a low risk are in 'bonus malus' class 1, and the loss for accepting a high risk are indeed very high. However, for deciding the cutoff value, one must look at the Bayes Risk. There is one minimum point, that is in the grange between 650 and 850 and is almost a straight line so perhaps it is better to consider the region instead of a single point.

Look at Figure 5-10 one sees that the range from 650 to 850 for the cutoff value is a bit too large, because if 850 is used as cutoff value, there are not many clients left. The best range is perhaps from 650 to 750. Here the loss is about 50 million BEF, if one would use the card on the clients which are in the portfolio. One should then reject about 32% of the clients. However, the good part is that one would have only 4% of high risks in the portfolio, which is a decrease of more than 8%.

It is better to test this scorecard on the data from next year to get a better impression of the way the scorecard works, as the scorecard has now been made and tested on the same data as it has been built upon. The only thing that is different between the test data and the data of building the score card is the definition of a high risk and a low risk. This will be commented upon in Chapter 6 where overall conclusions, remarks and lessons learned will be given.

Chapter 6

Conclusions, Remarks and Lessons learned

Hopefully, now the most important part of the research is finished; what could be done In this chapter the most important conclusions of the research will be given and the way in which the insurance company already wants to use the results of this research. In the second part some remarks on this research will be made and the last part will tell about the lessons learned during this research.

6.1 Conclusions and Possible Use of the Results

Before giving the possibilities of implementing the results according to the insurance company, the most important conclusions of this research will be summarized chapter by chapter.

6.1.1 Conclusions

The first part of the research was to find out if the geo-model score and the region score have a relationship with the definition of high risk and low risk and if they predict the probability of having a high risk correctly. From Chapter 2, one important conclusion can be given: *The geo-model score and the region score do have a positive relationship with the definition of the risk of a client, but this relationship is not strong enough to predict the risk alone.* However, these scores

were not made to predict the relationship alone, but for use with other predicting variables.

Yet, according to the results of Chapter 3, *the region score and the geo-model score do not even give a strong enough prediction to be in the logistic model with personal variables for the probability of having a high risk*. The variables which should be in the model according to the selection procedures are:

BM = the 'bonus malus' grade of a client

DUR = insurance duration of the automobile policy

KW = power of the engine of the car measured in kilowatts

RAP = whether a client has had a registered dunning letter or not

BRA = brand of car

AGE = the age of the client

Only the variable BRA is not dependent, but this variable has a positive correlation with the dependent variable.

This model is not a result of just using the logistic regression with selection procedures. It is built upon three steps. First, logistic regression is used with selection procedures for all the variables. Then, the same is done again but only with variables which are said to be dependent with the definition of high risk and low risk. Then one noticed that the variable AGE was not in the first model but was in the second model. Therefore, the variables which were in the first model without a relationship with the dependent variable and with a relationship with the variable AGE were rejected. The model arrived at variables AGE and BRA in it, and predicts better than the other two models. Thus, it is better to use only the variables which have a relationship with the dependent variable.

Although the variables REG and GEO are not in the logistic model in Chapter 3, *the variable GEO is in the model of declaring the amount of damage* in Chapter 4. The variables PAY and RAP, which tell something about the payment history of the client, are also in this model, it is not strange that the variable GEO is in the model. The variable GEO was based on the data of a credit company which looks at the payment history of a client for the definition of low risks and high risks. The

variables BM, KW, DUR, BRA, SPO, and SEX are also in the model. The variable AGE, which has a relationship with the dependent variable, the average amount of damage paid per damage, is not in the model.

The oddest part of this model is that the mean of *the expected amount of damage is larger for the low risks than for the high risks*. However, when one looks at the average probability of damage, one sees that this average is lower for the low risks than for the high risks. This results in the fact that *the average premium of the low risks lies lower than the average premium of the high risks*, as one can calculate the premium by Equation 6-1.

$$\pi(S) = (1+\alpha) * ES = (1+\alpha) * EX * EN \qquad \textbf{Eq 6-1}$$

Here α = 0.5 and X stands for the damage, S stands the amount of damage that must be paid and N stands for the frequency of damage.

For the probability of damage per year, there are three different models. This can indicate that there is a lot of difference between having a claim paid and the variables per year. This may indicate that there must be a lot of more research conducted before these results can be taken into consideration for the use of accepting or rejecting a client.

If these results would be accepted and the premium calculated by using Equation 6-1, one would see that *the premium which the insurance company now asks are too low*. The insurance solidarity is too large, since the difference between the premium which the insurance company asks for low risks and high risks is much smaller than the difference between the predicted premium for the low risks and the high risks. As the insurance company needs to keep their premiums low, they have to find a better acceptance policy.

Maybe a better acceptance policy can be drawn up with the results and conclusions from Chapter 5; making a scorecard for accepting or rejecting a

client. The variables which are important for accepting or rejecting a client, according to the statistical procedure, are BM, DUR, KW, REG, AGE and PAY. *The most important factor is the variable BM*, as concluded from Figures 5-1, 5-7 and 5-8. This is not an unusual fact as the 'bonus malus' grade is a representation of the caused claimed damages over the years, and one uses as dependent variable whether a claim has been paid or not.

The scorecard must be used with a cutoff value, which lies in this study's case between 650 and 750. The reason why a range is made instead of defining one value as the cutoff value is that the minimum loss of that value is almost the same as the minimum loss of the values between the cutoff values of 650 and 750. The important role of the variable BM is probably the cause.

Before this card can be used, it is necessary that this card is tested on the data from next year or during the acceptance from this year and look if the acceptance policy which already is being used, gives the same results as the score card for accepting or rejecting a client.

The overall conclusion is that it is possible to make a good acceptance policy by using all three models from Chapters 3, 4 and 5. However, it cannot be used this year as the models must be tested on other automobile insurance data than the data used for making the models. The most important variables for accepting or rejecting a client are BM, KW, PAY, DUR and AGE. Although the last variable is not in the model for declaring the expected amount of damage in Chapter 4, it does have a relationship with that dependent variable.

6.1.2 Possible Use for the Results

Although it is believed that the results of this research cannot directly be used for accepting or rejecting a client, the insurance company sees some important possibilities in the results of this research.

Possibility 1

The logistic models in Chapters 3 and 5 can be used for the *acceptance policy*. This does not mean that the insurance company will only use these models. The insurance company will use the relationship of the declaring variables of these models with the dependent variable of having a low risk or a high risk in the acceptance policy. The reason is that the insurance company finds the way in which the declaring variables predict the high risk, in a way as they would have expected, by using their own feelings and results.

Possibility 2

The logistic models of Chapter 3 and 5, and also the model of declaring the amount of damage in Chapter 4, can be used for *reorganizing the portfolio*. One way of reorganizing the portfolio is that an insurance company is allowed to terminate a policy after the client of that policy has had a claim paid.

A client which had a claim paid and the attributes of the client indicates, according to the logistic models in Chapters 3 and 5, that the probability of high risk will be large, and also according to the model in Chapter 4 the client has a large expected amount of damage. The probability of terminating the automobile policy by the insurance company will be very large.

Possibility 3

The insurance company wants to use *cross-selling* with other financial companies. According to the model for predicting the amount of claimed damage in Chapter 4, it is noted that the high risks of an insurance company do have a relation with the high risks in other financial companies.

This means that the financial companies can use each others client information for rejecting or accepting a client or even to use this result to get new clients at the insurance company. For example, if a person takes a loan at a credit company for buying a car, the insurance company can offer that person a automobile policy, since the insurance company knows that the client could be a low risk for the

insurance company because the client already obtained a loan, meaning that the client was seen as a low risk for the credit company.

6.2 Remarks on the Research

The first and most important remark is the fact that this research is based on a *select sample* of the Belgian population. This can have an influence on the results. However, there is nothing that one can do about it.

A second remark on the data is the fact that the *ratio of clients with high risk versus the ratio of clients with low risks* is too small to have solid results. This ratio is 1:7, so if there are eight clients, seven of those clients will be a low risk and only one client will be a high risk. For more reliable results it is better to have a larger ratio, then the results of the two groups can be better compared. However, for the insurance company this is a sign that the acceptance policy they already use is satisfactory.

There are also remarks on the research itself. The first remark is the fact that make a *test dataset* was not made. The reason why it was not made was that there was not enough data. However, if one could have been made the results could have been better. One could have tested the models on the test dataset and draw better conclusions. Maybe the results could already be used for the acceptance of clients this year.

The second remark on the research is that in Chapters 2, 3 and 4 all the variable were used as one. This would not have been a problem if one did not *assume that the classes of the variables are linear related.* This cannot be assumed if one looks at the score values of Table 5-1 and the figures of the distribution of these variables, as the relationships of the classes of these variables are far from linear.

Dummy variables could have been used as in Chapter 5, but then one would not use the variable as whole, but as different variables. This is not what this study wanted to do for the models of Chapters 2, 3 and 4.

However it could have been done in a different way such that the variable will still be used as a whole and not used as the separate classes by assuming a quadratic or cubic relationship. This means that the logit(Y) must be written as Equation 6-1.

$$\text{logit}(Y) = \beta_0 + \beta_{11}x_1 + \beta_{12}x_1^2 + \beta_{13}x_1^3 + \ldots + \beta_{i1}x_i + \beta_{i2}x_i^2 + \beta_{i3}x_i^3 \qquad \textbf{Eq 6-2}$$

with $i = 1,2,\ldots$

If Equation 6-2 is used, one is still able to use the variable as whole, and it is possible that class k has a smaller value than class k - 1 and class k + 1. Although the models do predict reasonably well according to the data used, implementing Equation 6-1 could have given better models.

6.3 Lessons Learned

The most important part of doing research is the *scope of the research*. The scope must be preset and there must be a certainty about getting the information one needs. The scope of this research was first to test whether the geo-model score and the region score have a relationship with the high risks and the low risks of the automobile policies and then extend this research to the other non-life insurance policies.

The last part should have been checking whether these variables could have an impact on cross-selling with a credit company. The scope has been altered as the client information from the credit company could not be compared with the information from the clients of automobile policies. So the cross-selling part of this research could not be done. Therefore, we have chosen for a depth analysis

the use of the region score and the geo-model score in the acceptance policy of the insurance company which has resulted in this research.

The second important part of doing research is the *database*. This database must be clean; there must be one business meaning for each data element and the business data must be understood well. As this was not the case for the data used, it was very hard to make a good database. At one point in the research some facts of the database had to be accepted before the research could continue. If a database contains a lot of mistakes and a lot of ambiguous data, one is unable to clean the database completely, as some mistakes cannot be recovered.

Another thing that has been learned is that first, the data must be taken into consideration before the *theory* can be used. Another difference with the theory is the distribution of the data. This distribution is not always a theoretical distribution as assumed in some textbooks. Also the amount of data elements must be a lot larger than suggested in some textbooks to obtain reliable results.

The last point to be made is that it is hard to determine a *definition for a high risk of a automobile policy*, as there are so many variables which have an important role for having a claim paid or not.

Of course there were many more lessons learned, like working in a business, the way a business is handled, and other personal parts of doing research in a business environment. Just the most important facts of the lessons learned during the six month period of research have been written down.

References

Casella, George and Roger L. Berger (1990)
Statistical Inference. Duxbury Press.

Cox, D. R. and E. J. Snell (1989)
Analysis of Binary data 2nd edition. Chapman and Hall.

Dehling, H. G. and J. N. Kalma (1995)
Kansrekening. Epsilon Uitgaven.

Dehling, H. G. (1996)
Inleiding Statistiek. Rijksuniversiteit Groningen.

Finger, Robert J., Matthew Rodermund and others (1996)
Foundations of Casualty Actuarial Science. United Book Press, Virginia.

Frees, Edward W. (1996)
Data analysis using regression models. Prentice-Hall, Inc.

Ferguson, Thomas S. (1967)
Mathematical Statistics, a decision theoretic approach. Academic Press, Inc.

Ferguson, Thomas S. (1996)
A course in large sample theory. Chapman & Hall.

Goovearts, M. J. and R. Kaas (1998)

Inleiding Risico theorie. Instituut voor actuariaat en economtrie, universiteit van Amsterdam.

Greene, William H. (1997)

Econometric Analysis 3rd edition. Prentice-Hall, Inc.

Hettmans, Thomas P. (1984)

Inference based on ranks. John Wiley & Sons.

Hosmer, David W. and Stanley Lemeshow (1989)

Applied logistic regression. John Wiley & Sons.

Knypstra, S. (1999)

Syllabus by het college statistiek 2C. Rijksunversiteit Groningen.

Lewis, Edward M. (1992)

An introduction to credit scoring. The Athena Press.

Lindgren, Bernard W. (1993)

Statistical theory 4th edition. Chapman & Hall, Inc.

McCullagh, P and J. A. Nelder (1989)

Generalized linear Models 2nd edition. Chapman and Hall.

Menard, Scott (1995)

Applied Logistic Regression Analysis. Thousand Oaks Sage.

SAS, SAS/STAT® Software (1996)

Changes and enhancements through release 6.11. SAS Institute Inc.

Zar. Jerrold H. (1996)

Biostatistical Analysis 3rd edition. Prentice-Hall, Inc.

Appendix I

The Classes of the Explanatory Variables

In the first appendix of five appendices the explanation of the classes of all the explanatory variables will be given. The first two variables GEO and REG cannot be described precisely as they are based on the NIS-codes and the ZIP-codes of Belgium and we are not allowed to give the division of these codes over the classes. However, the variable GEO can be explained a little. As the consulting firm gave rates to every NIS-code from 0.0 until 5.0, the last one called bad rates, and these bad rates have been used for making the classes of the variable GEO, this division can be given.

GEO variable for the geo score based on the data of a credit company.

Class:		
	0	if bad rate = 0
	1	if 0 < bad rate < 0.20
	2	if 0.20 <= bad rate < 0.30
	3	if 0.30 <= bad rate < 0.50
	4	if 0.50 <= bad rate < 0.70
	5	if 0.70 <= bad rate < 1.50
	6	if 1.50 <= bad rate < 2.20
	7	if 2.20 <= bad rate < 3.00
	8	if 3.00 >= bad rate < 5.00
	9	if bad rate = 5.00

The other variables will be explained totally.

GES	variable for the sex of the client	
Class:	0	if a client is a female
	1	if a client is a male
AGE	variable for the age of the client	
Class:	1	if 18 <= age of the client <= 22
	2	if 23 <= age of the client <=29
	3	if 30 <= age of the client <= 39
	4	if 40 <= age of the client <= 49
	5	if 50 <= age of the client <= 59
	6	if 60 <= age of the client <= 69
	7	if 70 <= age of the client <= 79
	8	if age of the client >= 80
	9	if the client is a firm without an age
RAP	variable for the payment history of the client	
Class:	0	the client has never received a registered dunning letter
	1	the client has received at least once a registered dunning letter
PAY	variable for the frequency of paying the premiums per year	
	1	the client pays monthly, with an automatic order
	2	the client pays twice a year with a raise of 3%
	3	the client pays yearly
DUR	variable for the insurance duration of one policy of a client	
Class:	0	less than 1 year insured
	1	between 1 and 2 years insured
	2	between 2 and 3 years insured
	3	between 3 and 4 years insured
	4	between 4 and 5 years insured
	5	between 5 and 6 years insured

	6	between 6 and 7 years insured
	7	more than 7 years insured

TOT variable for the total of non-life policies at the insurance company

Class:	1	1 policy at the insurance company
	2	2 policies at the insurance company
	3	3 policies at the insurance company
	4	4 policies at the insurance company
	5	5 policies at the insurance company
	6	6 or more policies at the insurance company

PRI variable for the use of the car

Class:	0	professional use
	1	private use

NEW this variable tells whether a client has had another automobile insurance policy or not

Class:	0	take over, thus the client has had another automobile policy
	1	new case, thus the client has never had another automobile policy

BM this variable gives the 'bonus malus' grade of a client

Class:	1	if the client has a 'bonus malus' grade = 0, 1, 2
	2	if the client has a 'bonus malus' grade = 3, 4, 5
	3	if the client has a 'bonus malus' grade = 6 ,7 ,8
	4	if the client has a 'bonus malus' grade = 9, 10
	5	if the client has a 'bonus malus' grade = 11
	6	if the client has a 'bonus malus' grade = 12, 13
	7	if the client has a 'bonus malus' grade = 14
	8	if the client has a 'bonus malus' grade >= 15

KW the variable which tells something about the power of the engine measured in kilowatts

1	if 0 <= kilowatt <= 40
2	if 41 <= kilowatt <= 51
3	if 52 <= kilowatt <= 62
4	if 63 <= kilowatt <= 73
5	if 74 <= kilowatt <= 103
6	if kilowatt >= 104

SPO this variable gives the indication whether the car is a sports car or not

Class:

0	no sports car
1	sports car

BRA this variable divides the brand of car into classes

Class:

1	if the brand of car is Suzuki
2	if the brand of car is BMW
3	if the brand of car is Volvo
4	if the brand of car is Toyota
5	if the brand of car is Mitsubishi
6	if the brand of car is Daihatsu
7	if the brand of car is Fiat
8	if the brand of car is Honda
9	if the brand of car is Ford
10	if the brand of car is Lada
11	if the brand of car is Opel
12	if the brand of car is Citroën
13	if the brand of car is Renault
14	if the brand of car is Peugeot, Simca or Talbot
15	if the brand of car is Chevrolet, Chrysler, Daewoo, Kia, Lancia, Landrover, Rover, Austin, Sunbeam, Saab, Jaguar, Lexus, Vauxhall, Xedos, MG, Morgan, Porsche, Triumph, Buick, Pontiac, Cadillac, Oldsmobile,

	Plymouth, Asia, Dodge, Isuzu, Jeep, Minerva, Range rover, Ssangyoung, Aixa, Apal, Burster, Daf, Iveco, Ligier, Malvern, Mega, Packard, Proton, Saviem, Smart, Trabant, Vanclee, Vanhool, Wartburg, Yugo or Zastava
16	if the brand of car is Subaru
17	if the brand of car is Mazda
18	if the brand of car is Hyundai
19	if the brand of car is Volkswagen
20	if the brand of car is Mercedes
21	if the brand of car is Audi
22	if the brand of car is Nissan
23	if the brand of car is Seat
24	if the brand of car is Skoda

This are all the explanatory variables which are used in this research.

Appendix II

The Theory Behind the Generalized Linear Models

Regression models are an important part of any data analyses, concerned with describing the relationship between a dependent variable and one or more explanatory variables. It is often the case that the outcome variable is discrete. Over the last decade the logistic regression has become the standard method of analysis in this situation. In the chapters where logistic regression and multiple logistic regression has been used, was a short explanation of this theory. Only the exact theory of the multiple logistic regression will be given in this appendix as the one explanatory case can easily be derived form the multiple explanatory case. There will be started with linear regression, and then translate it to logistic regression. The test for the null hypothesis, that $\beta = 0$, will also be explained.

As logistic regression is part of the generalized linear models the theory of the generalized linear models with the gamma as distribution function and link function log will also be given, instead of the binomial distribution function with link function logit.

II.1 Linear Regression

The difference between Generalized Linear Models (GLM) and linear regression (classical linear models) is in the choice of a parameter model and in the assumptions. Once this difference is accounted for, the method employed in an

analysis using logistic regression follow the same general principles as used in linear regression. The techniques used in linear regression will motivate our approach to logistic regression.

In linear regression analysis, there can be tested whether two variables are linearly related and the strength of the relationship can be calculated. This relationship can be described by equation II-1.

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_i x_i \qquad \textbf{Eq II-1}$$

The variable Y, the dependent or endogenous variable, has to be predicted by the independent variables x_i (the predictor or exogenous variables). In Equation II-1 there are j+1 population parameters to be estimated, β_0 until β_j. The population parameter β_0, the intercept, represents the value of Y when all the x_i 's are zero. The population parameters β_i, represents the partial slopes of the line that provides the best linear estimate of Y from x_i.

Equation II-1 is sometimes written in a form, such that there can be recognized that the prediction of Y by x_i can be imprecise.

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_i x_i + \varepsilon \qquad \textbf{Eq II-2}$$

Where ε is the error term, a random variable, which represents the error in predicting Y from the x_i's.

Equation II-1 and II-2 are being used to describe the relationship among the variables for all of the cases in the sample or the population. When one wants to describe only one case, Equation II-3 can be used. When one drops the error term then Equation II-3 will be the rewriting of Equation II-1 for a special case j.

$$Y_j = \beta_0 + \beta_1 x_{1j} + \ldots + \beta_i x_{ij} + \varepsilon_j \qquad \textbf{Eq II-3}$$

To use the Equations for prescribing the relationship among the variables, some assumptions must be made.

II.2 Assumptions of Regression Analyses

The assumptions of linear regression:

1 Measurement

All independent variables are interval, ratio or dichotomous. De dependent variable is continuous, unbounded and measured on an interval or ratio scale. All variables are measured without errors.

2 Specification

a. All relevant predictors of the dependent variable must been taken into the analysis

b. No irrelevant predictors of the dependent variable are allowed in the analysis

c. The form of the relation is linear

3 Expected value of the error

$$E(\varepsilon) = 0$$

4 Homoscedasticity

The variance of the error term is the same or constant, for all the values of the independent variables

5 Normality of Errors

The errors are normally distributed for all set of values of the independent variables

6 No Autocorrelation

There is no correlation between the error terms produced by different values of the independent variables, $E(\varepsilon_i, \varepsilon_j) = 0$

7 No Correlation Between the Error Terms and the Independent Variables

The error terms are uncorrelated with the independent variables $E(\varepsilon_j, x_j) = 0$

8 Absence of Perfect Multi Co-linearity

For multiple regression, none of the independent variable is a perfect linear combination of the other independent variables, for any *i*, $R_i^2 < 1$, where R_i^2 is the variance in the independent variable x_i that is explained by all other independent variables $x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k$. When there is only one independent variable, then multi co linearity is no issue.

II.3 From Linear Regression to Logistic Regression

When the dependent variable is a dichotomous variable, the model will be called a linear probability model. The mean of the variable is the proportion of the cases in Y=1 given the independent variables $\boldsymbol{x}$ and the assumption that the variables are linearly related. This prediction does have values from minus infinity until infinity. However, the values of a probability may not be below 0 and above 1, so one needs to adapt the formula for the description of a probability.

If the dependent variable is a dichotomous variable, assumption 1 does not hold, as this variable in not continuous. And assumption 5 does not hold either, as the errors have a binomial distribution. To solve this problem Generalized Linear Models (GLM) can be used instead of classical linear models, as linear regression. GLM allows to use a continuous link function such that one can make a transformation, and assumption 1 holds for that transformation. And for GLM it is enough that the error term has a distribution which is part of the exponential

family, which yields for the binomial distribution. The steps of the transformation are written below.

A first step for solving the problem, is to replace the probability that Y=1 by the Odds that Y=1, see Equation II-4.

$$Odds(Y=1)=\frac{P(Y=1)}{1-P(Y=1)} \qquad \textbf{Eq II-4}$$

Using Equation II-4, there is no fixed maximum for the depended variable, but still a minimum of zero. The next step must be to get rid if the fixed minimum. The natural logarithm of the Odds can be taken and will result in Equation II-5.

$$\ln(Odds(Y=1))=\text{logit}(Y)=\beta_0+\beta_1 x_1+\ldots+\beta_j x_j \qquad \textbf{Eq II-5}$$

The logit(Y) becomes negative and increasingly large in absolute value as the odds decrease from 1 to 0, and becomes increasingly large in the positive direction as the odds increase from 1 to infinity. So, if logit(Y) is used, one has no longer the problem that the estimated probability may exceed the maximum or the minimum values.

For the value of the probability, Equation II-7 can be used.

$$P(Y=1)=\frac{Odds(Y=1)}{1+Odds(Y=1)} \qquad \textbf{Eq II-6}$$

Because $\text{logit}(Y)=\ln(Odds(Y=1))=\beta_0+\beta_1 x_1+\ldots+\beta_j x_j$ yields $Odds(Y=1)=e^{\beta_0+\beta_1 x_1+\ldots+\beta_j x_j}$. If this will be put into Equation II-6 one gets for the probability that Y=1 given the dependent variables Equation II-7.

$$P(Y=1|\mathbf{x}) = \frac{Odds(Y=1)}{1+Odds(Y=1)} = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_j x_j}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_j x_j}} \qquad \textbf{Eq II-7}$$

The important property of this link function is the fact that it gives a linear sufficient statistic for β with the same dimension as the parameter space, independent of the sample space. In this way we will not loose any information by using the linear statistic. If one is interested in the parameter of a distribution, where the distribution is part of the exponential family, the canonical parameter will be used as link function.

The distribution is a binomial distribution, because the dependent variable Y given x can be written as $P(Y|\mathbf{x}) = \pi(\mathbf{x}) + \varepsilon$. Here $\pi(\mathbf{x}) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_j x_j}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_j x_j}}$ with

$\varepsilon = 1 - \pi(\mathbf{x})$ if $Y = 1$ with probability $\pi(\mathbf{x})$

$\varepsilon = -\pi(\mathbf{x})$ if Y = 0 with probability $1 - \pi(\mathbf{x})$

So ε has a distribution with mean zero and variance equal to $\pi(\mathbf{x})[1-\pi(\mathbf{x})]$ and thus the conditional distribution of Y is $Bin[1, \pi(\mathbf{x})]$.

A binomial distribution is part of the exponential family, as the conditional distribution of Y can be written as Equation II-8.

$$p(y) = p^y(1-p)^{1-y} = \left(\frac{p}{1-p}\right)^y (1-p) = e^{\log\left(\frac{p}{1-p}\right)y+\log(1-p)} = e^{\theta y+\varphi(\theta)} \qquad \textbf{Eq II-8}$$

which is the canonical parameterization, and according to the theory of generalized linear models is $\log\left(\frac{p}{1-p}\right)$ the canonical parameter.

Lets assume that $\psi(\theta)=\sum_{j=1}^{p} x_j \beta_j$ where β_j is the unknown parameter and x_j are the explanatory variables. Thus $\theta=\psi^{-1}\left(\sum_{j=1}^{p} x_j \beta_j\right)$ can be written for the ith observation as $\theta_i=\psi^{-1}\left(\sum_{j=1}^{p} x_{ij} \beta_j\right)$ where x_{ij} is the value of x_j for the ith observation. Adding this to the distribution function for observation i, the last part of Equation II-8 becomes Equation II-9.

$$p_{\theta_i}(y_i)=e^{\theta_i y_i+\varphi(\theta)} \qquad \textbf{Eq II-9}$$

The joint distribution can be written as Equation II-10, since one assumes that the observations are independent and knows that $\theta_i=\psi^{-1}\left(\sum_{j=1}^{p} x_{ij} \beta_j\right)$.

$$\prod_{i=1}^{n} p_{\theta_i}(y_i)=e^{\sum_{i=1}^{n}\psi^{-1}\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)y_i+\sum_{i=1}^{n}\varphi(\theta_i)} \qquad \textbf{Eq II-10}$$

To find a sufficient statistic for β_j one has to rewrite the first summation $\sum_{i=1}^{n}\psi^{-1}\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)y_i$. The way in which this summation gives a sufficient statistic for β_j is if $\psi^{-1}(\theta)=\theta$. Then the summation can be rewritten as Equation II-11, and according to Neymans factorization criterium, is $\sum_{i=1}^{n} x_{ij} y_i$ a sufficient statistic for β_j.

$$\sum_{i=1}^{n}\psi^{-1}\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)y_i=\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)y_i=\sum_{j=1}^{p}\beta_j\sum_{i=1}^{n} x_{ij} y_i \qquad \textbf{Eq II-11}$$

Although, the link function logit gives a sufficient statistic, one could use other link functions like probit or complementary log-log. According to McCullagh and Nelders there is no preconceived opinion as reason, in spite for the fact that the canonical links are often eminently sensible on scientific grounds, why one should use the canonical link function.

To calculate the probability given in Equation II-7, one needs to estimate the parameters β_j. Because a transformation has been submitted, the values of the probability will not be greater than 1 and less than 0, so the leased squares error method to estimate the parameters cannot be used. The desirable statistical properties which the estimated parameters get by using least squares method, will not be there if the same method is used for a model with dichotomous dependent variable, because the variables are not linearly related, but the logit of the variables are linearly related. Instead, the Maximum Likelihood techniques will be used.

II.4 Estimating the Parameters

The maximum likelihood method gives estimates for the unknown parameters, which are called the maximum likelihood estimates. These estimates are, according to Ferguson(1996), under fairly general conditions strongly consistent as the sample size tends to infinity. The maximum likelihood estimator is said to be asymptotic normal if the second derivative of the likelihood function to the unknown parameters exist and continuous is. And no sequence of estimates, which satisfy the asymptotic normality condition, can have a smaller variance asymptotically for any unknown parameter. Therefore this method is a good method to use to estimate the unknown parameters of the logit model.

For the estimation of the parameters first, the likelihood function is needed, which expresses the probability of the observed data as a function of the parameters. This likelihood function becomes as Equation II-12 for the case of logistic regression.

$$L(\vec{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \qquad \textbf{Eq II-12}$$

$$\pi(\mathbf{x}) = \frac{e^{\beta_0 + \beta_1 * x + \ldots + \beta_j x_j}}{1 + e^{\beta_0 + \beta_1 * x + \ldots + \beta_j x_j}} \qquad \textbf{Eq II-13}$$

The product of the probability functions of the x_i may be used as the observations are assumed to be independent.

The principle of maximum likelihood states that the estimate of β must be used, the value which maximizes the expression L(β). To maximize that expression it is easier to use the logarithm of that expression.

$$l_x(\vec{\beta}) = \ln(L(\vec{\beta})) = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\} \qquad \textbf{Eq II-14}$$

To find the maximum of L(β),Equation II-14 must be differentiated with respect to β, and the resulting expressions must be set equal to zero, Equation II-16. As one knows that the logit link function, is the canonical link function, one can rewrite Equation II-14 see Equation II-15.

$$l(\vec{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{p} y_i x_{ij} \beta_j - \sum_{i=1}^{n} m_i \log\left(1 + e^{\sum_{j=1}^{p} x_{ij} \beta_j}\right) \qquad \textbf{Eq II-15}$$

In such a way that the log-likelihood only depends on y, through the linear combinations $X^T y$. These p combinations are said to be sufficient for β, as seen in the explanation of the use of the logit as link function.

$$\frac{d}{d\beta_o} l_x(\bar{\beta}) = \frac{d}{d\beta_o}\left\{\sum_{i=1}^{n}\{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\}\right\} = 0$$

$$\frac{d}{d\beta_1} l_x(\bar{\beta}) = \frac{d}{d\beta_1}\left\{\sum_{i=1}^{n}\{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\}\right\} = 0 \qquad \textbf{Eq II-16}$$

$$\vdots$$

$$\frac{d}{d\beta_j} l_x(\bar{\beta}) = \frac{d}{d\beta_j}\left\{\sum_{i=1}^{n}\{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\}\right\} = 0$$

In linear regression the likelihood Equations are linear in the unknown parameters and thus are easily solved. For logistic regression the expressions are non-linear and require special methods for their solution. These methods are iterative in nature and have been programmed into available logistic regression software. This means, beginning with a tentative solution, revising it slightly to see if it can be improved and repeating the process until the change in the likelihood function from one step of the process to another step is negligible. The process of repeated estimation, testing and re-estimation is called iteration, and the process of obtaining a solution from repeated estimation is called iterative process.

After these estimation we need to test whether the found estimates of the unknown parameters are equal to zero, and if not whether the found model declares the probability of $Y = 1$ in a way we can accept.

II.5 Testing of Significance

Testing of significance, means that we need to compare observed values of the response variable to predicted values obtained from models with and without the variable(s) in question. This comparison is based on the log-likelihood function from Equation II-14.

When there are n observations one can fit models to them containing up to n parameters. There is a simple model, null model, which has only one parameter

for representing a common mean for all the observations, and there is another extreme model, the full model, which contains all parameters. In practice, the null model is too simple and the full model does not summarize the data, but repeats the data. This last model can be used for the difference among the real model and a summarized model with only p parameters, where of course $p < n$. This difference can be tested by using the deviance, see Equation II-17.

$$D = -2\ln\left[\frac{\text{likelihood of the current model}}{\text{likelihood of the full model}}\right] = \\ = -2\sum_{i=1}^{n}\left[y_i \ln\left[\frac{\hat{\pi}(x_i)}{y_i}\right] + (1-y_i)\ln\left(\frac{1-\hat{\pi}(x_i)}{1-y_i}\right)\right] \qquad \textbf{Eq II-17}$$

with $\hat{\pi}(x_i)$ is the estimates of $\pi(x_i)$ by using Equation II-13 and instead of β one uses the estimates of β.

The multiplication of –2ln will be used since this gives a quantity whose distribution is known, namely the chi-square distribution with 1 degree of freedom, and thus can be used for hypothesis testing purposes.

Equation II-17 is used for testing whether the estimated model is a good summary of the real observations. If one wants to test whether an estimates parameter is equal to zero Equation II-18 must be used, the G-statistic, which is almost the same as Equation II-17 and has the same chi-square distribution.

$$G = -2\ln\left[\frac{\text{likelihood of the model without the variable}}{\text{likelihood of the model with the variable}}\right] = \\ = -2\sum_{i=1}^{n}\left[y_i \ln\left[\frac{\tilde{\pi}(x_i)}{\hat{\pi}(x_i)}\right] + (1-y_i)\ln\left(\frac{1-\tilde{\pi}(x_i)}{1-\hat{\pi}(x_i)}\right)\right] \qquad \textbf{Eq II-18}$$

with $\tilde{\pi}(x_i)$ is the estimates of $\pi(x_i)$, without the variable for testing the hypothesis that the estimates of the unknown parameter is equal to zero and $\tilde{\pi}(x_i)$ is the estimates of $\pi(x_i)$with that variable.

The G-statistic and the deviance can be analyzed in the same way. Larger values of both statistics indicates worse prediction of the dependent variable. For looking whether these statistics give significant results, critical value table of the chi-square distribution with one degree of freedom and for a preset α, normally we use $\alpha = 0.05$, can be used.

II.6 Gamma Distribution with Log as Link Function

The gamma distribution of Y, Gam(μ, ν), can be written as Equation III-1.

$$f(y) = \frac{\alpha^{\nu}}{\Gamma(\nu)} y^{\nu-1} e^{-\alpha y} \qquad \textbf{Eq II-19}$$

$$\text{with } \alpha = \frac{\nu}{\mu},\ E(Y) = \frac{\nu}{\alpha} \text{ and } Var(Y) = \frac{\nu}{\alpha^2}$$

The value of ν determines the shape of the distribution. If $0 < \nu < 1$ the density has a pole at the origin and decreases monotonically as $y \rightarrow \infty$. The special case $\nu = 1$ corresponds to the exponential distribution. If $\nu > 1$ the density is zero at the origin and has a single mode at $y = \mu - \mu/\nu$.

As there are observations which tends from 0 to infinity (the re-insurance value is used, so practical the observations cannot be larger than 12.000.000, but there will be assumed that these values can go to infinity), the logarithm as link function has to be used. Where Equation II-7 becomes for the log link function Equation II-20.

$$EY = \mu = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j} \qquad \textbf{Eq II-20}$$

Although the logarithm is not the canonical link function of the gamma distribution, it can give good results, according to the remark of McCullagh and Nelders.

As the theory of Maximum likelihood and the test of significance can also be used for the gamma distribution with logarithm as link function, only the statistics for this distribution will be given.

The likelihood function can be written as the product of the distribution function of the observations, this results in Equation II-21 as the likelihood function for the gamma distribution.

$$L(\bar{\beta}) = \prod_{i=1}^{n} \frac{\alpha^{\nu}}{\Gamma(\nu)} y^{\nu-1} e^{-\alpha y} \qquad \textbf{Eq II-21}$$

$$\text{with } \alpha = \frac{\nu}{\mu} = \frac{\nu}{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j}}$$

The log-likelihood can be found by just taking the logarithm of the likelihood function, Equation II-21. As one assumes that the scale parameter is the same for all observations and after leaving behind the constant factors one gets Equation II-22 as the log-likelihood for the gamma distribution. For the estimation of the unknown parameters Equation II-22 needs to be differentiated with respect to the unknown parameters, as described in Equation II-16.

$$l(\bar{\beta}) = \sum_{i=1}^{n} \left(-\frac{y_i}{\mu_i} - \log \mu_i \right) = \sum_{i=1}^{n} \left(-\frac{y_i}{\mu_i} - \left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_j x_{ij} \right) \right) \qquad \textbf{Eq II-22}$$

The deviance, which is proportional to twice the difference between the log-likelihood achieved under the estimated model and the full model, is prescribed in Equation II-23.

$$D = -2 \ln \left[\frac{\text{log - likelihood of the current model}}{\text{log - likelihood of the full model}} \right] =$$

$$= -2 \sum_{i=1}^{n} \left[\ln \left[\frac{y_i}{\hat{\mu}_i} \right] + \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i} \right] \qquad \textbf{Eq II-23}$$

Equation II-23 yields only if the observations are all strictly positive.

The G-statistic can now easily be described by using Equations II-18 and II-23, with $\hat{\mu}(x_i)$ is the estimates of $\mu(x_i)$, without the variable for testing the hypothesis that the estimates of the unknown parameter is equal to zero and $\tilde{\mu}(x_i)$ is the estimates of $\mu(x_i)$ with that variable.

$$G = -2\ln\left[\frac{\text{log - likelihood of the model without the variable}}{\text{log - likelihood of the model with the variable}}\right] =$$

$$= -2\sum_{i=1}^{n}\left[\ln\left[\frac{\hat{\mu}_i}{\tilde{\mu}_i}\right] + \frac{(\hat{\mu}_i - \tilde{\mu}_i)}{\tilde{\mu}_i}\right]$$

Now one has everything to estimate the unknown parameters and testing the significance of the model.

Appendix III

Nonparametric Tests

For several data, the assumption that the data has a normal distribution may not be made. For the data in this study, in our research, is that also the case. For a research it is necessary to know what the relationship is of the dependent variable with the declaring variables. Luckily, there are nonparametric tests. These test do not need the assumption of normally distributed data. In chapters 3 and 4 several nonparametric tests has been used to find the relationship between the variables. These tests have been described for a bit in the sections method of these chapters. In this appendix, the whole theory of the tests will be described and the difference between Kendal's tau and Spearman's rho.

III.1 Brief History of the Nonparametric Theory

The history of nonparametric tests based on ranks is extending back roughly fifty-five years. A few of the major contributions will be mentioned to give an idea of the history.

The systematic development and assessment of the nonparametric methods began with the work of F. Wilcoxon in 1945 and H. B. Mann and D. R. Whitney in 1947. For the next ten years, nonparametric tests were studied using Pitman's asymptotic efficiency to asses their local power properties. J. L. Hodges and E.L. Lehmann discovered the result that rank tests suffer negligible efficiency los when

compared to the *t*-test at the normal model and maybe much more efficient at heavy-tailed models. At about this time, nonparametric tests began to gain some acceptance from data analyst.

In the 1960's, Hodges and Lehmann derived point estimates and confidence intervals for location parameters form rank test statistics. They also showed that the estimation methods get their efficiency properties from the parent test statistics. These estimates are also robust according to the new criteria proposed by J. W. Tukey, P. J. Huber and F. Hampel for assessing stability of estimates. During this period, J. Hajek developed a new and powerful approach to the asymptotic distribution theory needed for the construction of general rank score test statistics.

Aligned rank tests for analysis of designed experiments were introduced in the early 1960's by Hodges and Lehmann and developed by M. L. Puri and P. K. Sen. The rank tests and the corresponding estimates for simple regression models were proposed and studied by J. N. Adichie.

In the 1970's the previous work was consolidated and extended to rank-based tests and estimates in the linear model. Much of the asymptotic distribution theory needed for the linear model derives from basic results due to J. Jureckova. Therefore, it is possible to develop unified approaches based on ranks to the analysis of complex data sets. During the 1980's and later, there was seen the computer implementation and more widespread use of these efficient and robust statistical methods.

III.2 Kruskal-Wallis Test

If a set of data is collected according to a completely randomized design where k > 2, it is possible to test nonparametrically for difference among groups. The Kruskal-Wallis test can be used to test the null hypothesis that all samples come

form the same population. In this case it means that to test whether all classes come from the same population, thus have the same statistics as the whole population. This test is often called an analysis of variance by ranks. The nonparametric analysis is especially desirable when the k samples do not come from normal populations and it may be used when the k population variances are somewhat heterogeneous.

The sampling model consists of k samples $X_{11},\ldots,X_{n_1 1},\ldots X_{n_k k}$ from $F(x-\theta_1),\ldots.$, $F(x-\theta_k)$, respectively. One wishes to test H_0: $\theta_1 = \ldots. = \theta_k$ versus H_A: $\theta_1 ,\ldots. , \theta_k$ not all equal. This null hypothesis simply specifies that the locations are all equal without specifying the common location.

The data can be seen as a two-way array in which each column is a sample. The basic strategy is to rank the combined data set of size $N = \sum_{j=1}^{k} n_j$ and compare the column rank sums or averages. Let R_{ij} denote the rank of X_{ij} in the combined data and let

$$R_{\cdot j} = \sum_{i=1}^{n_j} R_{ij} \qquad \text{and} \qquad \overline{R}_{\cdot j} = \frac{R_{\cdot j}}{n_j}. \qquad \textbf{Eq III-1}$$

Here $R_{\cdot j}$ denotes the sum of the ranks for column j, and $\overline{R}_{\cdot j}$ denotes the average of the ranks of column j.

Under the null hypothesis, R_{ij} has the following distributional properties:

$$P\left(R_{ij} = s\right) = \frac{1}{N} \qquad s = 1,\ldots,N$$

$$P\left(R_{ij} = s, R_{lk} = t\right) = \begin{cases} \dfrac{1}{N(N-1)} & s \neq t \\ 0 & s = t \end{cases}$$

with the following statistics:

$$ER_{.j} = \frac{n_j(N+1)}{2} \qquad E\bar{R}_{.j} = \frac{(N+1)}{2}$$

$$VarR_{.j} = \frac{n_j(N-n_j)(N+1)}{12} \qquad Var\bar{R}_{.j} = \frac{(N-n_j)(N+1)}{12n_j}$$

$$Cov(R_{.i}, R_{.j}) = \frac{-n_i n_j (N+1)}{12} \qquad Cov(\bar{R}_{.i}, \bar{R}_{.j}) = \frac{-(N+1)}{12}$$

Now $\bar{R}_{.j} - \frac{N+1}{2}$ shows the difference of what has been expected under the null hypothesis. When these values are too large one wishes to reject the null hypothesis, that all samples come form the same population. This gives a test statistic of the form of equation III-2.

$$H = \sum_{j=1}^{k} c_{jN}^2 \left\{ \frac{\bar{R}_{.j} - \frac{(N+1)}{2}}{\sqrt{Var\bar{R}_{.j}}} \right\}^2 \qquad \textbf{Eq III-2}$$

The weighting constants $c_{1N},\ldots, c_{kN}$ are chosen so that the test statistic H is asymptotically chi-squared with k-1 degrees of freedom. A logic choice would be 1. Then H would be the sum of squares standardized rank averages. However the rank averages are correlated and this will require some adjustment. To chose the values of these weighted constants, asymptotic distribution theory must be used. Just the theorems which are necessary will be given, the proofs of these theorems can be found in Hetmans(1984).

Theorem 1

Suppose the k samples come from a common distribution. Suppose also $n_j \rightarrow \infty$, $j = 1,\ldots, k$ in such way that $n_j / N \rightarrow \lambda_j$, $0 < \lambda_j < 1$, where $N = \sum_{j=1}^{k} n_j$.

Suppose $c_{jN} \rightarrow n_j$ for $j = 1,\ldots, k$, define $\boldsymbol{T}' = (T_1,\ldots,T_k)$ where

$$T_j = c_{jN} \frac{1}{\sqrt{N}} \left(\bar{R}_{.j} - \frac{N+1}{2} \right).$$

Then $\boldsymbol{T}$ is asymptotically $MVN(\boldsymbol{0},\boldsymbol{B})$, where

$$b_{ij} = \begin{cases} c_i^2(1-\lambda_i)/(12\lambda_i) & \text{if } i = j \\ -c_i c_j /12 & \text{if } i \neq j \end{cases}$$

Theorem 2, form Normal Distribution Theory

Suppose ($Z_1,\ldots,Z_k$)' has a $MVN(\mathbf{0},\boldsymbol{A})$ distribution. Suppose that $\boldsymbol{A}$ is idempotent ($\boldsymbol{A} = \boldsymbol{A}^2$) with rank r. Then $\sum_{i=1}^{k} Z_i^2$ has a chi-square distribution with r degrees of freedom, denoted $\chi^2(r)$.

Using these theorems one can give the weighted values such that the test statistic H has an asymptotic chi-square distribution. By using $c_{jN}^2 = 1 - \frac{n_j}{N}$ the test statistic can be rewritten as equation III-3.

$$H = \sum_{j=1}^{k}\left(1-\frac{n_j}{N}\right)\left\{\frac{\bar{R}_{.j} - (N+1)/2}{\sqrt{(N-n_j)(N+1)/(12n_j)}}\right\}^2$$

$$= \frac{12}{N(N+1)}\sum_{j=1}^{k} n_j\left[\bar{R}_{.j} - (N+1)/2\right]^2 \qquad \textbf{Eq III-3}$$

$$= \frac{12}{N(N+1)}\sum_{j=1}^{k}\frac{R_{.j}^2}{n_j} - 3(N+1)$$

which has an asymptotic chi-square distribution with k-1 degrees of freedom.

H is the Kruskal-Wallis statistic and rejects H_0: $\theta_1 = \ldots. = \theta_k$ at approximate level α when $H \geq \chi_\alpha^2(k-1)$ where $\chi_\alpha^2(k-1)$ is the 1-α percentile of the chi-square distribution with k-1 degrees of freedom.

III.3 Rank Correlation

Relationship between two variables is not always one of dependence. The magnitude of one of the variables changes as the magnitude of the second variable

changes, but it is not reasonable to consider there to be an independent and a dependent variable. In such situations correlation, rather than regression, analysis are called for, and both variables are theoretically to be random variables.

A positive correlation implies that for an increase in the value of one of the variables, the other variables also increase in value; a negative correlation indicates that an increase in value of one of the variables is accompanied by a decrease in value of the other variable. If the correlation is zero then there is no linear relationship between the variables.

The correlation coefficient r, which is calculated from a sample is an estimation of a population parameter, namely the correlation coefficient in the population that was sampled. If the data is normally distributed one could use Pearson's correlation statistic, as the data is not normally distributed another correlation statistic is needed, which is based on the ranks of the observations.

The sampling model consists in a sample $(X_1, Y_1),\ldots, (X_n, Y_n)$ from $F(x,y)$ where $F(.,.)$ is absolutely continuous with absolutely continuous marginal cdfs $F_x(.)$ and $F_y(.)$. One thinks of the data arranged in two columns, which each observation has his own rank:

X_1	Rank of $X_1=R_{X1}$	Y_1	Rank of $Y_1=R_{Y1}$
...	...	...	...
X_n	Rank of $X_n=R_{Xn}$	Y_n	Rank of $Y_n=R_{Yn}$

There are two methods for assessing the degree of agreement between the two sets of rankings, namely Spearman's rho and Kendall's tau. Spearman's rho is simply the product-moment correlation coefficient computed on the ranks:

$$r_s = \frac{\sum_{i=1}^{n}\left(R_{Xi} - \frac{n+1}{2}\right)\left(R_{Yi} - \frac{n+1}{2}\right)}{\sqrt{\sum_{i=1}^{n}\left(R_{Xi} - \frac{n+1}{2}\right)^2 \sum_{i=1}^{n}\left(R_{Yi} - \frac{n+1}{2}\right)^2}} \qquad \textbf{Eq III-4}$$

If the sum of n ranks is equal to $n(n+1)/2$, equation III-3 may be written as equation III-5.

$$r_s = \frac{\sum_{i=1}^{n} R_{Xi} R_{Yi} - \frac{n(n+1)^2}{4}}{\sqrt{\left(\sum_{i=1}^{n} R_{Xi} - \frac{n(n+1)^2}{4}\right)\left(\sum_{i=1}^{n} R_{Yi} - \frac{n(n+1)^2}{4}\right)}} \qquad \textbf{Eq III-5}$$

However the best equation to use for the computation of the correlation is equation III-6, as $\sum_{i=1}^{n}(R_{Xi} - (n+1)/2)^2 = n(n^2-1)/12$ and $\sum_{i=1}^{n}[R_{Xi} - (n+1)/2] = 0$.

$$r_s = \frac{12}{n(n^2-1)}\sum_{i=1}^{n}[R_{Xi} - (n+1)/2]R_{Yi}$$
$$= 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_{Xi} - R_{Yi})^2 \qquad \textbf{Eq III-6}$$

Since r_s is a correlation coefficient it has the property that $-1 \leq r_s \leq 1$. It is easy to check that the extremes are attainable. If there is independency, so that the two rankings are independent, the joint distribution of the ranks is uniform on the $n!$ permutations. If there is no association, the expected value of the rank correlation is equal to zero.

Kendall's tau is the second measure, this measure looks whether pairs (X_i, Y_i) and (X_j, Y_j) are concordant or discordant. The pairs are concordant if $X_i > X_j$ and $Y_i > Y_j$ or if $X_i < X_j$ and $Y_i < Y_j$. This means that if sgn(x) =1, 0, -1 as $x > 0$, 0, < 0, then the pairs are concordant if sgn(X_j - X_i)sgn(Y_j - Y_i) = 1. The pairs are discordant if sgn(X_j - X_i)sgn(Y_j - Y_i) = -1. Let P and Q denote the number of concordant and discordant pairs, respectively, and let $S = P - Q$. Then the possible values of S range form $-n(n-1)/2$ to $n(n-1)/2$. Kendall suggested the coefficient:

$$\tau = \frac{S}{\max S} = \frac{2(P-Q)}{n(n-1)}$$
$$= 1 - \frac{4}{n(n-1)}Q \qquad \text{since } P + Q = n(n-1)/2 \qquad \textbf{Eq III-7}$$

If $s(x)$=1 if $x > 0$ and 0 otherwise, equations III-6 and III-7 can be rewritten to be compared with each other.

$$r_s = 1 - \frac{12}{n(n^2-1)} \sum\sum_{i<j} (R_{Xi} - R_{Xj}) s(R_{Yi} - R_{Yj})$$

$$\tau = 1 - \frac{4}{n(n-1)} \sum\sum_{i<j} s(R_{Yi} - R_{Yj})$$

since $X_1 < \cdots < X_n$, $\quad Q = \sum\sum_{i<j} s(Y_i - Y_j) = \sum\sum_{i<j} s(R_{Yi} - R_{Yj})$

For the proof of this see Hetmans(1984).

This shows that except in the extreme cases, r_s and τ will not generally be equal. Since Spearman's statistic gives larger weights to inversions of ranks that are farther part, thus this absolute value will be larger than the value of Kendall's statistic. However when X and Y are independent these two statistics are highly correlated, which tends to 1 if n increases. Hence, for testing independence they are asymptotically equivalent under the null hypothesis.

Appendix IV

Some Basic Facts of the Decision Theory

In Chapter 5 the elementary theory of the hypothesis testing problem with 0-1 generalized loss was described. There was mentioned that there are other forms of inference, these will be described in this appendix. Also the general theory of the Bayes risk, which is an important part of the decision theory, will be given and some other parts of decision theory will be explained. However for more theory and proofs there must be referred to Ferguson (1967) and Casella and Berger (1990).

IV.1 Other Decision Problems

Before the other forms of inference will be mentioned, all the elements necessary for a decision theoretic formulation will be described. The data are described by a random vector **X**, with sample space $\boldsymbol{X}$. The set of possible probability distributions for **X**, indexed by θ, which is the true but unknown state of nature about one wishes to make an inference, is called the model. The set of possible values for θ is called the parameter space Θ. So the model can be described as a set $\{f(\mathbf{x}|\theta): \theta \in \Theta\}$ with each $f(\mathbf{x}|\theta)$ describes the pdf or pmf on $\boldsymbol{X}$.

The action space A, which is the set of all allowable decisions regarding θ, is known. This action space determines whether the inference problem is a point

estimating problem, a hypothesis testing problem or an interval problem. There can be other problems but only these three problems will be looked at.

After making a decision one needs to know whether that decision was right or wrong. Therefore the loss function $L(\theta,a)$ can be used, this gives the loss incurred if θ is the true state of nature and action a is taken.

However for making a decision a decision rule $\delta(\mathbf{x})$ must be specified, this rule specifies for each $\mathbf{x} \in \boldsymbol{X}$ what action $a \in \boldsymbol{A}$ will be taken if $\boldsymbol{X} = \mathbf{x}$ is observed. With D is the set of all allowable decision rules. The quality of a decision rule is quantified in the risk function of the decision rule, $R(\theta,\delta) = E_\theta L(\theta,\delta(\mathbf{X}))$.

One other element of the decision theory can be the prior distribution $\pi(\theta)$, this can be used as an opinion summary or to summarize the information about a decision rule that is in the risk function. The Bayes risk of a decision rule with respect to a prior distribution is $B(\pi,\delta) = E_\pi R(\theta,\delta)$.

This are all the basic elements of the decision theory. These elements will be used to describe point estimation and interval estimation.

Point Estimation

The possible actions are the various possible values of θ, so $\boldsymbol{A} = \theta$. This specification is what characterizes a decision problem as a point estimation problem. The loss function reflects the fact that if an action a is close to θ, the decision is reasonable and little loss is incurred. The loss function generally increases as the distance between a and θ increases. If one assumes that θ is a real valued function, then there are two common loss functions:

$L(\theta,a) = |a-\theta|$ absolute error loss

$L(\theta,a) = (a-\theta)^2$ squared error loss

Squared error loss gives relatively more penalty for large discrepancies and absolute error loss gives relatively more penalty for small discrepancies. The risk function for squared error loss is equal to the mean squared error (MSE) as

$$R(\theta,\delta)=E_\theta L(\theta,\delta(\mathbf{X}))=E_\theta(\delta(\mathbf{X})-\theta)^2=MSE(\theta)$$
$$=Var_\theta\delta(\mathbf{X})+(E_\theta\delta(\mathbf{X})-\theta)^2=Var_\theta\delta(\mathbf{X})+(Bias_\theta\delta(\mathbf{X}))^2$$

Interval Estimation

The action space will consists of subsets of the parameter space. So one should talk about set estimation instead of interval estimation, since the optimal rule may not necessary be an interval. We will use C to denote elements of A, with the meaning of the action C being that the interval estimate $\theta \in C$ is made. A decision rule in this problem simply satisfies which set $C \in A$ will be used as an estimate of θ if $X = \mathrm{x}$ is observed.

The loss function usually has two quantities, a measure of the set estimates includes the true value θ and a measure of the size of the set. Since mostly an interval will be used the size of a set will be determined as Len(C) = length of C.

$$L(\theta,C)=bLen(C)-I_C(\theta)$$

$$I_C(\theta)=\begin{cases}1 & \theta\in C\\ 0 & \theta\notin C\end{cases}$$

The constant b can be small if there is more concern with correct estimates, or b could be large if there is more concern with interval length.

The risk function has also two components, the expected length of an interval and the coverage probability of the interval estimator.

$$R(\theta,C)=bE_\theta[Len(C(\mathbf{X}))]-E_\theta I_{C(\mathrm{X})}(\theta)$$
$$=bE_\theta[Len(C(\mathbf{X}))]-P_\theta(I_{C(\mathrm{X})}(\theta)=1)$$
$$=bE_\theta[Len(C(\mathbf{X}))]-P_\theta(\theta\in C(\mathbf{X}))$$

The use of this estimation problem is not as widespread as the use of point estimation or hypothesis testing, as the choice of b is very difficult to make.

IV.2 Admissibility and Completeness of Decision Rules

As the class D, of allowable decision rules is usually very large, it would be better if one could make a subclass of D, which contains only the good rules. This can be done by using the criterion admissibility. A decision rule is said to be admissible if there is no other decision rule that is better. Where a better decision rule means that for all θ yields that the risk function of that decision rule is smaller or equal to the risk function of the other decision rule and there must be some θ, wherefore yields that the risk function is strict smaller than the risk function of the other decision function, in symbols:

δ is better than δ' if for all $\theta \in \Theta$ yields $R(\theta,\delta) \leq R(\theta,\delta')$ and for some $\theta \in \Theta$ yields $R(\theta,\delta) < R(\theta,\delta')$.

The subclass of D are all the decision rules which are admissible. This does not mean that one has found the unique superior decision rule, but it can be easier to find that unique decision rule.

Admissibility is an optimum property, although in very week sense. In very real sense the word 'admissible' is a synonym for the word 'optimal'.

There is also another important property of the decision rules, completeness. The definitions belonging to this property will be given.

A class C of decision rules is said to be complete if, given any rule $\delta \in D$ not in C, there exists a rule $\delta_0 \in C$ that is better than δ. A class of C is minimum complete if C is complete and if no proper subclass of C is complete.

Completeness and admissibility of a class are related to each other, as if a minimal class exists, it consists exactly of the admissible rules.

IV.3 Bayes Rules

In decision theory one has to make a decision between the decision rules by using their risk functions. But that is not always that easy, as the risk function depend on θ. A way for comparing this decision rules is to use the Bayes risk. This means that the risk function is summarized by the average risk $B(\pi,\delta)=E_{\pi}R(\theta,\delta)$, and the decision rule with the smaller Bayes risk is preferred to a rule with a larger Bayes risk. The Bayes risk has to be minimized.

In using the Bayes principle, there will be assumed that the parameter is random with a known distribution. For a fixed distribution $\pi \in \Theta$ a decision rule is preferred if it has a smaller Bayes risk than another decision rule. This sets up a linear ordening on the space of discussion rules. A Bayes decision rule is one that is the best with respect to the ordening.

The Bayes rule is the decision rule that minimizes the Bayes risk among all possible decision rules, $B(\pi,\delta^{\pi})=\inf_{\delta\in D} B(\pi,\delta)$. So the Bayesian version of a decision theory is to find the rule δ^{π} that minimizes $B(\pi,\delta)$. The advantage of this rule is that summary is in terms of a single number and any two decision rules can be compared by using the Bayes risk.

IV.4 Minimax Rule

Another commonly used summary is the maximum value of the risk function. First the definition of this summary, the minimax decision rule, will be given.

A decision rule δ' is called a minimax decision rule if

$$\sup_{\theta\in\Theta} R(\theta,\delta') = \inf_{\delta\in D}\sup_{\theta\in\Theta} R(\theta,\delta).$$

A minimax decision rule has the smallest possible maximum risk, which means that the risk of any other decision rule is at least as big as the maximum for a minimax decision rule. For each decision rule, the minimax criterion looks at the worst value of θ that could be true, and guards against this worst case.

Many people find a decision rule that is both admissible and minimax very desirable, as the admissible selects the better decision rules and the minimax rule makes sure that the risk is not too large at any value of θ. There is also a relationship between the Bayes rule and the minimax rule. Lets assume that δ^{π} is a decision rule that is Bayes with respect to some prior π. If the risk function satisfies $R(\theta,\delta^{\pi}) \leq B(\pi,\delta^{\pi})$ for all $\theta \in \Theta$, then δ^{π} is a minimax rule.

Of course there is much more theory about the decision theory, but there has been chosen to only recall the most important facts. For more theory and proofs we refer to Ferguson (1967) and Casella and Berger (1990).