Mathematical Statistics
Stockholm University

Application of statistical methods in the determination of health loss distribution and health claims behaviour

Vasileios Keisoglou

Examensarbete 2005:8

Postal address:

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

Internet:

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

Mathematical Statistics
Stockholm University
Examensarbete **2005:8**,
http://www.math.su.se/matstat

Application of statistical methods in the determination of health loss distribution and health claims behaviour

Vasileios Keisoglou*

September 2005

Abstract

This paper describes a method of analyzing health loss data in order to determine the claim behavior and using it for forecasting and budgeting. For the purpose of this paper, health loss data are retrieved from the health products portfolio of a company in the Greek market.

The company is currently selling morbidity risk type products like health and personal accident coverage. The company has developed some approaches/methodologies to quantify the morbidity risk. The appropriateness of each approach depends on product features and availability of data.

As this company is still developing a methodology for morbidity risk measurement, further investigation of this subject is needed. This investigation requires the application of statistical methods.

Morbidity insurance products are products that cover the financial risk of sickness. Morbidity risk is the risk of variations in claim levels and timing due to fluctuations in policyholder morbidity.

The goal of this diploma work is not to cover the whole range of health insurance products but to study the claim behavior of a certain health insurance product from past experience and to apply the most appropriate methods that fit the available data capturing all the volatility and uncertainty.

*Postal address: Dept. of Mathematical Statistics, Stockholm University, SE–106 91 Stockholm, Sweden. E-mail: vkeisoglou@gmail.com. Supervisor: Anders Martin-Löf.

Preface

This is a thesis in mathematical statistics and is done at Stockholm University and the Company in Greece.

I would like to thank my supervisors for helping me with the theory of mixed models, literature recommendations, report writing and for being supportive.

I would also like to thank my supervisor, Anders Martin-Löf from the department of mathematical statistics at Stockholm University.

Further thanks go to my coordinator Mikael Andresson for granting me the permission to complete my thesis in Greece.

Finally I wound like to thank actuarial department of the company.

Contents

1. The Company's Background in Greece

1.1 The Experience from Greek insurance market

In the last two decades, the private insurance industry in Greece showed rapid growth especially in the health sector as a result of inadequate social security systems. In response to cover this demand, private health providers emerged and supplied the necessary services.

High demand increased the cost of private health treatment resulting in an increase overall cost of health insurance. Thus the need for measuring morbidity risk is a key condition for risk management by insurance companies.

1.2 General of Hospitalization product

This product is issued in order to ensure to the insured a hospitalization of high prescription. The cost of Room and Board in the private hospitals has been increased lately. As a consequence, the client, who has signed a contract with the Company and is insured with some of the available hospitalization products, must pay the surplus over the defined Room and Board within the Company's existing products that have been purchased. Thus, making up the difference between the insured cost and the real costs incurred. On the other hand, in most cases the client wishes to have the hospital treatment of his satisfaction, which is directly dependent upon the hospitalization class.

<u>Description of a Typical Hospitalization Product:</u>

1. The Company covers the risk of hospital treatment of the insured person and the member of his/her family eventually covered, due to illness or accident.
2. The Company agrees to pay fully or partly his recognized expenses realized during his hospital treatment that correspond to the

hospitalization class the insured has chosen. A typical of hospitalization classes are:

Class C (three bed room)

Class B (two bed room)

Class A (single bed room)

Class Luxury

Class Suite

3. The company covers X% of the expenses for the room and board in a hospitalization for the insured or any member of his family covered by the insurance, after deduction of the eventual policyholder's participation, according to the hospitalization class that is included in his contract.
4. The Company will pay double the amount of the expenses that correspond to the hospitalization class that is described in his contract, in case of the insured or any member of his family covered by this policy is under treatment in an intensive care unit in Greece or abroad, if that is considered necessary.
5. The Company covers the X% of the hospital fees for the insured or any member of his family covered by the insurance in Greece, after deduction of the eventual policyholder's participation, according to the hospitalization class that is inscribed in his contract. If the client wishes to have a treatment in an upper hospitalization class than the one he has chosen, he has to participate in the hospital fees for each upper hospitalization class, beyond the eventual policyholder's participation.
6. In case of surgery expenses in Greece or abroad the Company will pay, after deduction of any participation of the policyholder for the cost of hospitalization.
7. The rider product usually includes benefits for AIDS.

1.3 General about Claims

When an incident occurs which requires hospitalization, the customer must complete a claim form which was provided at the time of the signed health contract. The claim form document must then be submitted of the Company Claim department in order to be assessed for validity. The Company proceeds to establish an insurance provision for the claim. Claims payments, over the course of claim settlement, are then deducted from insurance provision until the final settlement of the claim. This procedure may take a few months.

1.4 Chosen cover

The Company Health portfolio has two general categories: Inpatient and Outpatient products. The first of the two products, Inpatient, compensates the insured for being hospitalized. Meanwhile, the Outpatient products compensate the insured for having medical examinations without the need for hospitalization. This paper assumes the first category and in particular the **Daily indemnity Insurance** of which a short description is being provided below.

Daily Indemnity Insurance contains the following components:

1. Hospitalization can denote all public and privately held hospital facilities.
2. *Recover due to sickness*: denotes all non-pre existing conditions, which present themselves during the coverage period, but not before 30 days after the contract start day.
3. *Recover due to accident*: Accident is defined as all bodily conditions that occur and are not a result of either a genetic or pre-existing condition.
4. *Dependent member* defines the insured and declared spouse and children. Children must be over 3 months old and under 20 years old, or in the case of university students, under 25 years old.

2. Analysis of Morbidity Risk

2.1 Volatility and Uncertainty

The Company is currently selling morbidity risk type products like health and personal accident coverage. The company has developed some approaches and methodologies to quantify the morbidity risk. The appropriateness of each approach depends on product features and availability of data.

As the company is still developing a methodology for morbidity risk measurement, further investigation for this subject is needed. This investigation requires the application of statistical methods.

Morbidity insurance products are products that cover the financial risk of sickness. Morbidity risk is the risk of variations in Claim levels and timing due to fluctuations in policyholder morbidity.

The goal of this diploma work is to study the Claim behaviour from past experience and to apply the most appropriate methods that fit the available data capturing all the volatility and uncertainty. Finally, theoretical recommendations will be made to the Company regarding the pricing of this risk type.

In order to clarify, **volatility** can be defined as the uncertainty of the Claims during the next 12 months due to the past deviation of observed Claims from the expected values. Based on previous year's data, a calculation of the distribution of Claims volumes and frequencies will be presented. This is followed by a calculation of the mean(μ), which represents the expected values. Along with the mean, a computation of the standard deviation (σ) that represents the volatility risk will be included. In the above calculations we consider that the underlying distribution and its parameters have been estimated correctly.

In addition **uncertainty** can be explained partially as the relative error in choosing the underlying distribution, as uncertainty of the distribution and the parameters of the Claims. Due to the possibility that future claims may differ in distribution and/or parameters of the distribution, G may vary from $G(a,b)$ to $G'(a',b')$

Uncertainty is divided into two components:

1. *Multi year*: We re-estimate the distribution and its parameters, and consider that the future development of the Claims will behave as the estimated distribution.
2. *One year*: Based on previous re-estimate distribution we re-estimate the parameters of the distribution for each one of the coming years.

2.2 Statistical references

2.2.1 Examination theoretical models

For an insuring organization, S denotes the random loss on the portfolio of its similar risks. Then S is the random variable for which we seek a probability distribution. In the collective risk model the basic concept is that it is a random process that generates claims for a portfolio of policies. This process is characterized in terms of the portfolio as a whole rather than in terms of the individual policies comprising the portfolio. Let N denote the number of claims produced by a portfolio of policies in a given time period. Let X_1 denote the amount of the first claim, X_2 the amount of the second claim and so on. Then

$$S = X_1 + X_2 + ... + X_N$$

represents the aggregate claims generated by the portfolio for the period under study. The number of claims N is a random variable and is associated with the frequency of the claim. In addition, the individual claim amounts $X_1, X_2,...$ are also random variables and are said to measure the severity of the claims.

<u>We make two fundamental assumptions:</u>

1. $X_1, X_2,...$ are identically distributed random variables
2. The random variables N , $X_1, X_2,...$ are mutually independent.

The first step in exploring the claim behaviour will be the study of the family distribution of N and the family distribution of the X_i 's.

The second step is to focus more upon the determination of the appropriate parameters for the distribution of N and the common distribution of the X_i 's. For N , a Poisson or a negative binomial distribution is often selected. For the Claim amount distribution, a normal, gamma or other continuous distribution may be used. These two classes of distributions provide a considerable choice for modelling the distribution of the aggregate claims S . Also X is severity and N is frequency.

Under the assumption stated earlier for the collective risk model, by conditioning N and obtaining:

$E(S) = m_1 E(N)$ and $\text{var}(S) = (m_2 - m_1^2)E(N) + m_1^2 \text{ var}(N)$, where

$m_1 = E(X)$ and $m_2 = E(X^2)$ for any claim amount X .

This leaves finding the underlying distribution for both severity and frequency.

2.2.2 Test of Appropriate distribution

Our first step is to determine which family of distributions the Claim and the Incurred Loss follow.

It may be easy to say that the Claim follows the discrete distribution and that the Incurred Loss follows a continuous distribution. However, finding the discrete distribution using the Goodness-of-Fit Test is still necessary. First, we will estimate the distribution family, which we hypothesize to be Poisson distribution. The next step will be to examine whether or not our hypothesis is valid. The general procedure consists of defining a test statistic, which is some function of the data measuring the distance between the hypothesis and the data (in fact, the badness-of-fit), and then calculating the probability of obtaining data which have a still larger value of this test statistic than the value observed, assuming the hypothesis is true. The most common tests for goodness-of-fit are the Kolmogorov-Smirnov and the chi-square test.

Below is a discussion of the Kolmogorov-Smirnov and chi-square test which is included as a reference point for the theories employed for our statistical study. It is then followed by a discussion of the quantile-quantile plot. As we discovered that the Incurred Loss follows the continuous family distribution, the quantile-quantile plot within the SPSS statistics program can help us in the estimation of the distribution.

a. Kolmogorov-Smirnov Goodness –of–Fit Test

The Kolmogorov-Smirnov (K-S) test is used to decide if a sample comes from a population with a specific distribution.

The Kolmogorov-Smirnov test is based on the empirical distribution function (ECDF). Given N ordered data points, $Y_1, Y_2, ..., Y_N$ the ECDF is defined as

$$E_N = {}^{n(i)}\!/_{N}$$

Where $n(i)$ is the number of points less than Y_i, and Y_i are ordered from smaller to largest value. This is a step function that increases by $1/N$ at the value of each ordered data point.

An attractive feature of this test is that the distribution of the K-S test statistic itself does not depend on the underlying cumulative distribution function being tested. Another advantage is that it is an exact test.

Despite these advantages the K-S test has several important limitations:

1. It only applies to continuous distributions.
2. It tends to be more sensitive near the center of the distribution than at the tails.
3. Perhaps the most serious limitation is that the distribution must be fully specified. That is, if location, scale, and shape parameters are estimated from the data, the critical region of the K-S test is no longer valid. It typically must be determined by simulation.

b. Chi-Square Goodness –of–Fit Test

The chi-square test is used to test if a sample of data came from a population with a specific distribution.

An attractive feature of the chi-square goodness–of–fit test is that it can be applied to any univariate distribution for which one can calculate the cumulative distribution function. The chi-square goodness–of–fit test is applied to binned data (i.e., data put into classes). This is actually not a restriction since for non-binned data one can simply calculate a histogram or frequency table before generating the chi-square test. However, the values of the chi-square test statistic are dependent on how the data is binned. Another disadvantage of the chi-square test is that it requires a sufficient sample size in order for the chi-square approximation to be valid.

The chi-square test is an alternative to the Anderson-Darling and Kolmogorov-Smirnov goodness–of–fit test. The chi-square goodness–of–fit test can be applied to discrete distribution such as the **Binomial** and the **Poisson**. The Kolmogorov-Smirnov and the Anderson-Darling tests are restricted to continuous distribution.

For the chi-square goodness–of–fit computation, the data are divided into k bins and the test statistic is defined as

$$\mathrm{X}^2 = \sum_{i=1}^{k} (O_i - E_i)^2 \big/ E_i$$

where O_i is the observed frequency for bin i and E_i is the expected frequency for bin i. The expected frequency is calculated by

$$E_i = N(F(Y_u) - F(Y_l))$$

Where F, the cumulative Distribution function for the distribution being tested, is Y_u, the upper limit for class and i, Y_l is the lower limit for class i, and N is the sample size.

The test statistic follows, approximately, a chi-square distribution with $(k - c)$ degrees of freedom where k is the number of non-empty cells and c is the number of estimated parameters for the distribution +1.

Therefore, the hypothesis that the data are from a population with the specified distribution is rejected if $\chi^2 > \chi^2_{(\alpha, k-c)}$ where $\chi^2_{(\alpha, k-c)}$ the chi-square percent is point function with $k - c$ degrees of freedom and a signification level of α.

c. Quantiles - Quantiles plot

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Probability plots are generally used to determine whether the distribution of a variable matches a given distribution. If the selected variable matches the test distribution, the points cluster around a straight line.

<u>The advantages of the q-q plot are:</u>

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

3. Describe of available data

3.1 Describe necessary variables of Daily Indemnity Insurance products

Before investigating the claim behaviours as mentioned in the previous paragraphs, it is necessary to determine the key variables towards this target. The accurateness and the completeness of the claim analysis depend upon the data availability described by these key variables.

1. **Gender**: Gender has two dimensions, Males and Females. This variable is necessary since pricing procedures of the Company and tariffs segregate between Males and Females.
2. **Age**: Attained age of the insured is crucial for the determination of the premium to be paid. The insurance companies provide insurance of the Daily indemnity starting from the age of "zero" up the age "65". Thus it is necessary to investigate how the claim behaviour varies in correspondence with the age. For this purpose, which is explained in more detail later in this paper, ages are groups into seventeen classes.
3. **Exposures**: Exposure is used in order to determine the probability of the risk independent of time. The maximum value is one. This value is assigned to customers who have one or more contract years. One the other hand, those who have less than 1 contract year are assigned an exposure value between zero and one. Exposure is calculated as the number of days from the contract sign date until the end the current year divided by 365 days.
4. **Incurrent Loss**: The composition of incurred losses in such is the total derived by the following formula: losses paid during the year plus loss reserves existing at the end of the year.
5. **Claim Report Year**: Essentially it is the year in which the claim is reported and is not limited by the time period of the payment of the claim, for instance a claim might incur in year 2002 but the Company may report the claim in 2005. The Claim Report Year in this example will be

2005. Also the Company uses the code **CLERPY** as an acronym for the Claim Report Year in its data files.

3.2 Summary of the original Data files

Given the key variables described above, the necessary data from the company's archives will be explored and extracted.

The Data archives combines raw data based on actual underwriting experience like Policy Number, Cover, and Gross Premium Earnings (GPE) with claims experience (i.e. Claim No, Payments and Outstanding Reserve).

Finally we arrive at the aggregated data file that shows in one row all the relevant information in respect of a particular Cover for a particular policy over a specified time period. For example during the Year 2000 for all types of coverage, each coverage's respective exposures within that year including GPE, number of claims, payments + OS, may be found and documented.

4. Application of Model

4.1 First step

In the beginning, we decided to focus on two categories of Number Claim Coverage; those are, customers who have submitted claims and those who have not submitted a claim during the claim year. Therefore, our customer population is divided by those customers who have zero claims during the year and those customers who have 1 or more than 1 claim during the same period.

The second step is to split the database based on the year of report (CLERPY) and the Gender (Gen); since, as described above, it was necessary to investigate the claim behaviour per gender and per years of report. In essence, we would like to examine the trend of the database on a year per year basis (uncertainty).

4.1.1 Fit Number Claim

With the assistance of SPSS, we can run tests that can help us fit the distribution. The first test was the Kolmogorov-Smirnov test. Within the Kolmogorov-Smirnov test, the SPSS program allows a further function which can test whether the distribution can be fitted as a Poisson distribution. From the results, appendix 1, we can say that the number of Claim Coverage follows the Poisson distribution.

However, as we know the Kolmogorv-Smirnov test is not the best test of the discrete distribution. Thus, we can select another test, which is the chi-square test. The chi-square test is another indicator of a Poisson distribution. The results, appendix 2, are almost the same as Kolmogorov-Smirnov. Therefore from the p-value results, it can be shown that the Number of Claim Coverage follows the **Poisson**

distribution. So, we can say that with 95% certainty that the Claim follows Poisson distribution.

4.1.2 *Fit the Incurred Loss coverage*

The Incurred Loss Coverage is a continuous distribution and as such we can fit the distribution employing the Q-Q plot from the SPSS-program.

The Q-Q plot in SPSS has several options in order to perform a test on distributions. Available test distributions include beta, chi-square, exponential, gamma, half-normal, Laplace, Logistic, Lognormal, normal, Pareto, Student's t, Weibull, and Uniform. Depending on the distribution selected, one can specify degrees of freedom and other parameters. These are performed for the following reasons:

- In order to obtain probability plots for transformed values. Transformation options include natural log, standardize values, difference, and seasonally difference.
- In order to specify the method for calculating expected distributions, and for resolving "ties," or multiple observations with the same value.

From the plots, appendix 3, we see that the Incurred Loss Coverage follows the **Gamma distribution**. In working with the data, we noticed two issues. The first issue was based in the distribution of categories year 2000 and the Gender Male. This category, Male 2000, follows the Laplace distribution. From the Gammas plot, we can see that one observed value is plotted too far from the other observed value. If we ignore this outlier observer and run the Q-Q plot once more, we are given a new result, appendix 4, which shows us the category, Male 2000, now also follows the Gamma distribution.

The second issue is almost the same as the issue described above. This issue is contained within the category Female 2002. This category follows the Gamma distribution, but is not very strong. We ignore the outlier observed value which is far from the last value and redo the Q-Q plot. Our new results, appendix 5, are much better and we can now say that this category too follows Gamma distribution.

4.2 Second step

As we are unsure of whether or not the Claim follows the Poisson distribution, we decided to split the Company Database once more. The key variable was the age group. We chose to process the Company age group as follows:

Years	Data name
0-4	Age 1
5-9	Age 2
10-14	Age 3
15-19	Age 4
20-24	Age 5
25-29	Age 6
30-34	Age 7
35-39	Age 8
40-44	Age 9
45-49	Age 10
50-54	Age 11
55-59	Age 12
60-64	Age 13
65-69	Age 14
70-74	Age 15
75-79	Age 16
80+	Age 17

Table 4.2

This classification was chosen because the chi-square test does not clearly show that the Claim follows the Poisson distribution. Thus it is easier to see which products must be given more care and examined more closely for each age group. Therefore, in instances where the p-value is not very strong, the

Company can change the policy value of products in this group. This is also useful from a market standpoint as we can see which age group has more claims and the Company can make adjustments to its pricing policy accordingly.

4.2.2 Fit Number Claim

Before splitting the database into the Company age groups, we were not sure if the Claim followed the Poisson distribution when we used the chi-square test.

We used the formula: $E_i = n * P_x$.

where n is the number of observers. With help of SPSS statistical program we can find the observer of the age group.

P_x is the probability of claim. We can examine if the Claim follows the Poisson distribution, and define the probability of Poisson distribution as:

$P[X = x] = \frac{e^{-\lambda}\lambda^x}{x!}$, where $x = 0{,}1$ in our case.

With help of SPSS program, we run the frequency test. The following table displays the results of this test:

Statistics

NoClm_Cov

N	Valid	963
	Missing	0
Mean		,0239
Std. Deviation		,15277
Variance		,023

Within which, we find the λ and the n. Utilizing these items, we can calculate the E_i within the Excel program. The table below from Excel shows the result as:

	AGE Group	
n	m	0,02
963	943,9313	
	18,87863	

With the results above and the help of SPSS statistical program, we have the p-value of the Claim which is summarized in appendix 6.

From this result we have a better picture of the distribution that the Claim follows. We cannot reject that the Claim follows the Poisson distribution. However, an issue is presented within the 2000-2003 years, where the p-value is not very strong.

Also we have another issue. We are concerned that we do not have an abundance of observations within a few of the Company age groups.

4.2.2 Fit the Incurred Loss coverage

In this case the Incurred Loss follows the Gamma distribution. Again, the same issue arises with the number of observations that are located as outliers and far from the quantity observed. If we take out the outlier observations, we see the incurred loss follows the Gamma distribution.

As well we have the same difficulty with the Claim and its number of observations. In many of the Company age groups, we do not have many observation points and it's difficult to say with certainty exactly which distribution each follows.

4.3 Third step

This step contains our opinion about the Company group age. We decided to process a different set of age groups than those presented previously. The decision to adapt the age groups was based on many factors.   The first was the constant issue of the amount of observations, which we have now corrected as we have more observed points within each age group. Second, we wanted to see if the results would be displayed as a Poisson distribution so that we may be clearer about which type of distribution defines the Claim.

The new age group is defined as follows:

Age	Name
0-9	Age 1
10-19	Age 2
20-29	Age 3
30-39	Age 4
40-49	Age 5
50-59	Age 6
60-69	Age 7
70-79	Age 8
80+	Age 9

Table 4.3

4.3.1 Fit Number Claim

As discussed, we performed this step as the chi-square test does not reflect that the Claim follows the Poisson distribution.

In this case, we followed the same process as described in chapter 4.2.2. The difference is only the adjustment in the Company age group. We used the same formula, which is: $E_i = n * P_x$. The tables which follow display the results of the formula.

FEMALE 2004 MALE

	AGE 1			AGE 1	
n	m	0,006	n	m	0,018
1413	1404,547		996	978,2324	
	8,427284			17,60818	

	AGE 2			AGE 2	
n	m	0,004	n	m	0,004
1809	1801,778		1374	1368,515	
	7,207114			5,47406	

	AGE 3			AGE 3	
n	m	0,025	n	m	0,018
2661	2595,3		2219	2179,415	
	64,88249			39,22948	

	AGE 4			AGE 4	
n	m	0,025	n	m	0,016
8107	7906,837		6901	6791,463	
	197,6709			108,6634	

	AGE 5			AGE 5	
n	m	0,013	n	m	0,014
8081	7976,627		9613	9479,356	
	103,6961			132,711	

	AGE 6			AGE 6	
n	m	0,013	n	m	0,022
5065	4999,581		7661	7494,298	
	64,99455			164,8746	

	AGE 7			AGE 7	
n	m	0,018	n	m	0,029
1638	1608,78		2542	2469,341	
	28,95804			71,61088	

	AGE 8			AGE 8	
n	m	0,021	n	m	0,045
243	237,9502		286	273,4153	
	4,996954			12,30369	

	AGE 9			AGE 9	
n	m	0,133	n	m	0,071
15	13,13198		14	13,04047	
	1,746553			0,925873	

Table 4.3.1 a

With the results above and the help of SPSS statistical program, we have produced the following tables regarding the p-value of the claim.

2004

FEMALE

Age1	# Clms Cov
Chi-Square	0,021
df	1
Asymp. Sig.	0,884

Age2	# Clms Cov
Chi-Square	0,088
df	1
Asymp. Sig.	0,766

Age3	# Clms Cov
Chi-Square	0,069
df	1
Asymp. Sig.	0,793

Age4	# Clms Cov
Chi-Square	0,055
df	1
Asymp. Sig.	0,815

Age5	# Clms Cov
Chi-Square	0,028
df	1
Asymp. Sig.	0,866

Age6	# Clms Cov
Chi-Square	0,140
df	1
Asymp. Sig.	0,708

Age7	# Clms Cov
Chi-Square	0,000
df	1
Asymp. Sig.	0,995

Age8	# Clms Cov
Chi-Square	0,000
df	1
Asymp. Sig.	0,994

Age9	# Clms Cov
Chi-Square	0,037
df	1
Asymp. Sig.	0,848

MALE

Age1	# Clms Cov
Chi-Square	0,009
df	1
Asymp. Sig.	0,925

Age2	# Clms Cov
Chi-Square	7,612
df	1
Asymp. Sig.	0,006

Age3	# Clms Cov
Chi-Square	0,001
df	1
Asymp. Sig.	0,970

Age4	# Clms Cov
Chi-Square	0,001
df	1
Asymp. Sig.	0,975

Age5	# Clms Cov
Chi-Square	0,212
df	1
Asymp. Sig.	0,645

Age6	# Clms Cov
Chi-Square	0,007
df	1
Asymp. Sig.	0,932

Age7	# Clms Cov
Chi-Square	0,026
df	1
Asymp. Sig.	0,871

Age8	# Clms Cov
Chi-Square	0,038
df	1
Asymp. Sig.	0,845

Age9	# Clms Cov
Chi-Square	0,006
df	1
Asymp. Sig.	0,940

Table 4.3.1 b

The original results are attached in appendix 7.

With the adjustment to the Company's age groups, the result is more accurate. Given this, clearly we can say that the Claim follows the Poisson distribution. Also we do not have large deviations in each of the other age groups.

4.3.2 Fit the Incurred Loss coverage

We processed the entire q-q test in the SPSS program. From the plot, appendix 8, it is evident that the Incurred Loss Coverage follows Gamma distribution. The observed is closer to the strong line than each of the other distributions plots, which exist in the SPSS program.

The results within Appendix 8 utilize only with the new age group described in heading 4.3.

4.4 Results

As we have completed all the possible tests that define which distribution, as discussed in headings 4.3.1 and 4.3.2, the Claim and the Incurred Loss variables follow, the next and most straight forward step is to find the parameters of each distribution.

Fortunately, we have found the distribution which satisfies our hypothesis and we have calculated the mean and the variance for each distribution. Therefore, we have computed the parameters given these items and have presented them in the tables which follow.

Yr of Report: **2000**

Gender New	NEW GROUP		Mean	Variance	Distribution	λ	α	β
MALE	1	NoClm_Cov	0,0257	0,025	Poisson	0,026		
		IL_Cov	1,6978	355,437	Gamma		0,08	209,346
	2	NoClm_Cov	0,0141	0,014	Poisson	0,014		
		IL_Cov	0,9742	228,104	Gamma		0,04	234,144
	3	NoClm_Cov	0,0305	0,030	Poisson	0,030		
		IL_Cov	5,5757	12.265,552	Gamma		0,03	2.199,830
	4	NoClm_Cov	0,0313	0,030	Poisson	0,031		
		IL_Cov	2,4777	664,284	Gamma		0,09	268,110
	5	NoClm_Cov	0,0359	0,035	Poisson	0,036		
		IL_Cov	7,9555	140.899,063	Gamma		0,05	17.710,998
	6	NoClm_Cov	0,0515	0,049	Poisson	0,052		
		IL_Cov	8,8847	10.135,042	Gamma		0,08	1.140,735
	7	NoClm_Cov	0,0581	0,055	Poisson	0,058		
		IL_Cov	8,6135	6.316,880	Gamma		0,12	733,368
	8	NoClm_Cov	0,0976	0,089	Poisson	0,098		
		IL_Cov	6,3344	558,289	Gamma		0,72	88,136
	9	NoClm_Cov	0,5000	0,333	Poisson	0,500		
		IL_Cov	13,2050	232,496	Gamma		0,750	17,607
FEMALE	1	NoClm_Cov	0,0217	0,021	Poisson	0,022		
		IL_Cov	1,3305	251,597	Gamma		0,07	189,099
	2	NoClm_Cov	0,0198	0,019	Poisson	0,020		
		IL_Cov	1,5643	294,167	Gamma		0,08	188,056
	3	NoClm_Cov	0,0775	0,072	Poisson	0,078		
		IL_Cov	7,7592	2.350,069	Gamma		0,26	302,873
	4	NoClm_Cov	0,0888	0,081	Poisson	0,089		
		IL_Cov	10,4585	4.238,556	Gamma		0,26	405,274
	5	NoClm_Cov	0,0399	0,038	Poisson	0,040		
		IL_Cov	5,7738	4.722,539	Gamma		0,07	817,929
	6	NoClm_Cov	0,0372	0,036	Poisson	0,037		
		IL_Cov	4,6090	1.301,547	Gamma		0,16	282,391
	7	NoClm_Cov	0,0559	0,053	Poisson	0,056		
		IL_Cov	5,7816	2.040,710	Gamma		0,16	352,967
	8	NoClm_Cov	0,0085	0,008	Poisson	0,008		
		IL_Cov	0,6342	47,466	Gamma		0,08	74,840

Table 4.4.a

Yr of Report: **2001**

Gender New	NEW GROUP		Mean	Variance	Distribution	λ	α	β
MALE	1	NoClm_Cov	0,0285	0,028	Poisson	0,029		
		IL_Cov	3,3464	1.373,533	Gamma		0,08	410,452
	2	NoClm_Cov	0,0199	0,020	Poisson	0,020		
		IL_Cov	1,4848	182,893	Gamma		0,12	123,175
	3	NoClm_Cov	0,0360	0,035	Poisson	0,036		
		IL_Cov	3,5396	3.195,314	Gamma		0,04	902,740
	4	NoClm_Cov	0,0356	0,034	Poisson	0,036		
		IL_Cov	3,4607	4.984,314	Gamma		0,02	1.440,261
	5	NoClm_Cov	0,0349	0,034	Poisson	0,035		
		IL_Cov	4,0135	2.236,830	Gamma		0,07	557,326
	6	NoClm_Cov	0,0565	0,053	Poisson	0,056		
		IL_Cov	9,5047	11.918,584	Gamma		0,08	1.253,969
	7	NoClm_Cov	0,0655	0,061	Poisson	0,066		
		IL_Cov	14,3298	13.381,960	Gamma		0,15	933,858
	8	NoClm_Cov	0,0829	0,076	Poisson	0,083		
		IL_Cov	26,1967	31.221,479	Gamma		0,22	1.191,810
FEMALE	1	NoClm_Cov	0,0181	0,018	Poisson	0,018		
		IL_Cov	0,6453	25,355	Gamma		0,16	39,293
	2	NoClm_Cov	0,0150	0,015	Poisson	0,015		
		IL_Cov	0,9288	131,779	Gamma		0,07	141,887
	3	NoClm_Cov	0,0690	0,064	Poisson	0,069		
		IL_Cov	9,2246	3.218,245	Gamma		0,26	348,877
	4	NoClm_Cov	0,0987	0,089	Poisson	0,099		
		IL_Cov	10,4588	2.607,951	Gamma		0,42	249,356
	5	NoClm_Cov	0,0403	0,039	Poisson	0,040		
		IL_Cov	3,4260	811,332	Gamma		0,14	236,819
	6	NoClm_Cov	0,0444	0,042	Poisson	0,044		
		IL_Cov	4,9719	2.453,391	Gamma		0,10	493,454
	7	NoClm_Cov	0,0627	0,059	Poisson	0,063		
		IL_Cov	11,1018	4.893,643	Gamma		0,25	440,796
	8	NoClm_Cov	0,0357	0,035	Poisson	0,036		
		IL_Cov	1,8079	151,609	Gamma		0,22	83,861
	9	NoClm_Cov	0,2500	0,205	Poisson	0,250		
		IL_Cov	8,8050	253,728	Gamma		0,306	28,816

Table 4.4.b

Yr of Report: **2002**

Gender New	NEW GROUP		Mean	Variance	Distribution	λ	α	β
MALE	1	NoClm_Cov	0,0340	0,033	Poisson	0,034		
		IL_Cov	1,7305	153,472	Gamma		0,20	88,687
	2	NoClm_Cov	0,0245	0,024	Poisson	0,025		
		IL_Cov	1,4758	217,524	Gamma		0,10	147,392
	3	NoClm_Cov	0,0316	0,031	Poisson	0,032		
		IL_Cov	2,5657	673,174	Gamma		0,10	262,376
	4	NoClm_Cov	0,0315	0,030	Poisson	0,031		
		IL_Cov	3,3583	2.260,360	Gamma		0,05	673,071
	5	NoClm_Cov	0,0356	0,034	Poisson	0,036		
		IL_Cov	5,2429	11.355,201	Gamma		0,02	2.165,842
	6	NoClm_Cov	0,0538	0,051	Poisson	0,054		
		IL_Cov	9,6189	10.917,949	Gamma		0,08	1.135,048
	7	NoClm_Cov	0,0538	0,051	Poisson	0,054		
		IL_Cov	12,0775	19.205,351	Gamma		0,08	1.590,178
	8	NoClm_Cov	0,0794	0,073	Poisson	0,079		
		IL_Cov	52,2621	156.115,024	Gamma		0,17	2.987,158
FEMALE	1	NoClm_Cov	0,0204	0,020	Poisson	0,020		
		IL_Cov	0,9344	107,907	Gamma		0,08	115,483
	2	NoClm_Cov	0,0134	0,013	Poisson	0,013		
		IL_Cov	0,7259	88,596	Gamma		0,06	122,046
	3	NoClm_Cov	0,0658	0,061	Poisson	0,066		
		IL_Cov	7,7530	2.316,864	Gamma		0,26	298,836
	4	NoClm_Cov	0,0958	0,087	Poisson	0,096		
		IL_Cov	11,2398	6.674,607	Gamma		0,19	593,839
	5	NoClm_Cov	0,0415	0,040	Poisson	0,042		
		IL_Cov	4,7201	1.719,223	Gamma		0,13	364,231
	6	NoClm_Cov	0,0478	0,046	Poisson	0,048		
		IL_Cov	11,3991	53.884,698	Gamma		0,02	4.727,121
	7	NoClm_Cov	0,0707	0,066	Poisson	0,071		
		IL_Cov	9,5780	3.545,528	Gamma		0,26	370,176
	8	NoClm_Cov	0,0758	0,070	Poisson	0,076		
		IL_Cov	6,2648	1.128,864	Gamma		0,35	180,190
	9	NoClm_Cov	0,3333	0,235	Poisson	0,333		
		IL_Cov	259,0233	301.210,759	Gamma		0,223	1.162,871

Table 4.4.c

Yr of Report: **2003**

Gender New	NEW GROUP		Mean	Variance	Distribution	λ	α	β
MALE	1	NoClm_Cov	0,0178	0,018	Poisson	0,018		
		IL_Cov	0,9939	89,289	Gamma		0,11	89,841
	2	NoClm_Cov	0,0082	0,008	Poisson	0,008		
		IL_Cov	0,9246	152,494	Gamma		0,06	164,933
	3	NoClm_Cov	0,0369	0,036	Poisson	0,037		
		IL_Cov	4,0883	1.712,675	Gamma		0,10	418,922
	4	NoClm_Cov	0,0312	0,030	Poisson	0,031		
		IL_Cov	4,0313	3.952,665	Gamma		0,04	980,494
	5	NoClm_Cov	0,0333	0,032	Poisson	0,033		
		IL_Cov	6,6590	23.407,519	Gamma		0,02	3.515,154
	6	NoClm_Cov	0,0542	0,051	Poisson	0,054		
		IL_Cov	20,2391	110.235,902	Gamma		0,04	5.446,690
	7	NoClm_Cov	0,0579	0,055	Poisson	0,058		
		IL_Cov	15,3712	36.247,724	Gamma		0,07	2.358,159
	8	NoClm_Cov	0,0671	0,063	Poisson	0,067		
		IL_Cov	42,6581	162.936,785	Gamma		0,11	3.819,599
	9	NoClm_Cov	0,2000	0,178	Poisson	0,200		
		IL_Cov	15,8400	1.115,136	Gamma		0,225	70,400
FEMALE	1	NoClm_Cov	0,0202	0,020	Poisson	0,020		
		IL_Cov	0,9276	81,434	Gamma		0,11	87,785
	2	NoClm_Cov	0,0071	0,007	Poisson	0,007		
		IL_Cov	0,7333	125,068	Gamma		0,04	170,561
	3	NoClm_Cov	0,0565	0,053	Poisson	0,056		
		IL_Cov	7,8377	4.383,915	Gamma		0,14	559,336
	4	NoClm_Cov	0,0856	0,078	Poisson	0,086		
		IL_Cov	10,6111	14.588,999	Gamma		0,08	1.374,884
	5	NoClm_Cov	0,0417	0,040	Poisson	0,042		
		IL_Cov	6,5514	6.143,893	Gamma		0,07	937,798
	6	NoClm_Cov	0,0451	0,043	Poisson	0,045		
		IL_Cov	12,3329	57.052,011	Gamma		0,03	4.625,985
	7	NoClm_Cov	0,0660	0,062	Poisson	0,066		
		IL_Cov	13,1508	16.423,181	Gamma		0,11	1.248,835
	8	NoClm_Cov	0,1000	0,091	Poisson	0,100		
		IL_Cov	6,8957	1.066,138	Gamma		0,45	154,610
	9	NoClm_Cov	0,3333	0,235	Poisson	0,333		
		IL_Cov	96,8450	43.667,861	Gamma		0,215	450,905

Table 4.4.d

Yr of Report: **2004**

Gender New	NEW GROUP		Mean	Variance	Distribution	λ	α	β
MALE	1	NoClm_Cov	0,0361	0,035	Poisson	0,036		
		IL_Cov	3,0036	2.261,472	Gamma		0,04	752,917
	2	NoClm_Cov	0,0087	0,009	Poisson	0,009		
		IL_Cov	1,1898	427,841	Gamma		0,03	359,601
	3	NoClm_Cov	0,0352	0,034	Poisson	0,035		
		IL_Cov	6,3721	7.375,583	Gamma		0,06	1.157,480
	4	NoClm_Cov	0,0313	0,030	Poisson	0,031		
		IL_Cov	7,1721	11.280,938	Gamma		0,05	1.572,887
	5	NoClm_Cov	0,0288	0,028	Poisson	0,029		
		IL_Cov	8,0048	17.425,625	Gamma		0,04	2.176,905
	6	NoClm_Cov	0,0427	0,041	Poisson	0,043		
		IL_Cov	26,4976	250.488,323	Gamma		0,03	9.453,228
	7	NoClm_Cov	0,0559	0,053	Poisson	0,056		
		IL_Cov	22,1750	40.296,241	Gamma		0,12	1.817,192
	8	NoClm_Cov	0,0909	0,083	Poisson	0,091		
		IL_Cov	13,9353	5.285,683	Gamma		0,37	379,301
	9	NoClm_Cov	0,1429	0,132	Poisson	0,143		
		IL_Cov	44,0200	12.520,913	Gamma		0,155	284,437
FEMALE	1	NoClm_Cov	0,0170	0,017	Poisson	0,017		
		IL_Cov	0,7441	48,829	Gamma		0,11	65,619
	2	NoClm_Cov	0,0133	0,013	Poisson	0,013		
		IL_Cov	0,6867	67,154	Gamma		0,07	97,785
	3	NoClm_Cov	0,0755	0,070	Poisson	0,076		
		IL_Cov	7,8133	1.676,192	Gamma		0,36	214,530
	4	NoClm_Cov	0,0743	0,069	Poisson	0,074		
		IL_Cov	17,8437	30.669,330	Gamma		0,10	1.718,780
	5	NoClm_Cov	0,0376	0,036	Poisson	0,038		
		IL_Cov	6,7882	7.179,452	Gamma		0,06	1.057,633
	6	NoClm_Cov	0,0397	0,038	Poisson	0,040		
		IL_Cov	12,6195	24.012,808	Gamma		0,07	1.902,828
	7	NoClm_Cov	0,0531	0,050	Poisson	0,053		
		IL_Cov	4,1449	983,826	Gamma		0,17	237,358
	8	NoClm_Cov	0,0617	0,058	Poisson	0,062		
		IL_Cov	3,1485	182,154	Gamma		0,54	57,854
	9	NoClm_Cov	0,4000	0,257	Poisson	0,400		
		IL_Cov	143,4680	41.835,208	Gamma		0,492	291,600

Table 4.4.e

When examining the results of the tables above, there is a clear difference between the Male and the Female λ . Also we see the α is less than 0.1 within the Male and Female independent age groups. However, the same cannot be said about the β, because the value has a large deviation in the Male and in the Female classifications inclusive of all age groups.

Upon more specific examination, it is very difficult to see the trend of parameters. As such, it is not easy to estimate the trend; therefore, we must continue our process using other methods to estimate the future parameters of the distribution. These other methods appear in the chapters which follow.

5. Claim Forecasting Process

As previously stated, we are in the position to estimate the parameters of the distribution. Upon which, the Company will have the capability to estimate the expectation of the total claim. For the estimation of the parameters of the distribution, which in our case is Poisson and the Gamma parameters, we can use two methods: the extrapolation method and the Linear Regression method.

5.1 Extrapolation method

Pure extrapolation of time series assumes that all we need to know is contained in the historical values of the series that is being forecasted. For cross-sectional extrapolations, it is assumed that evidence from one set of data can be generalized to another set.

Because past behavior is a good predictor of future behavior, extrapolation is appealing. It is also appealing in that it is objective, replicable, and inexpensive. This makes it a useful approach when one needs many short-term forecasts.

The primary shortcoming of time-series extrapolation is the assumption that nothing is relevant other than the prior values of a series.

We favor the use of this method only with the Gamma distribution and the estimate of the parameters.

In our case we cannot use the extrapolation method, because the parameter where a higher importance is given to the α, and α must be more 1. When we examine the tables included above, we discover that an $\alpha>1$ never appears, so we must find another method to fit the future the Incurred Loss.

5.2 Linear Regression and Results

Another method is the Linear Regression. With this method, we can estimate the future parameter for Incurred Loss and Claim.

Linear regression analyzes the relationship between two variables, X and Y. For each subject, one knows both X and Y and wants to find the best straight line through the data. In some situations, the slope and/or intercept have a scientific meaning. In other cases, the linear regression line as a standard curve to find new values of X from Y, or Y from X is used.

Prism determines and graphs the best-fit linear regression line, optionally including a 95% confidence interval or 95% prediction interval bands. One may also force the line through a particular point (usually the origin), calculates residuals, calculates a runs test, or compares the slopes and intercepts of two or more regression lines.

In general, the goal of linear regression is to find the line that best predicts Y from X. Linear regression does this by finding the line that minimizes the sum of the squares of the vertical distances of the points from the line.

Note that linear regression does not *test* whether one's data is linear (except via the runs test). It assumes that the data is linear, and finds the slope and intercept that make a straight line best fit the data.

5.2.1 Estimation of Severity

Therefore, we have computed the linear regression of the Gamma parameter and have presented it in the table which follows below.

	MALE			**FEMALE**	
0- 9 yrs	α	β	0- 9 yrs	A	β
2000	0,08	209,346	2000	0,07	189,099
2001	0,12	410,452	2001	0,16	39,293
2002	0,20	88,687	2002	0,08	115,483
2003	0,11	89,841	2003	0,11	87,785
2004	0,04	752,917	2004	0,11	65,619
2005	**0,083**	**540,2079**	**2005**	**0,115**	**39,9154**

10-19 yrs	α	β
2000	0,04	234,144
2001	0,12	123,175
2002	0,10	147,392
2003	0,06	164,933
2004	0,03	359,601
2005	**0,046**	**293,6506**

10-19 yrs	α	β
2000	0,08	188,056
2001	0,07	141,887
2002	0,06	122,046
2003	0,04	170,561
2004	0,07	97,785
2005	**0,049**	**98,5066**

20-29 yrs	α	β
2000	0,03	2199,83
2001	0,04	902,74
2002	0,10	262,376
2003	0,10	418,922
2004	0,06	1157,48
2005	**0,102**	**217,7142**

20-29 yrs	α	β
2000	0,26	302,873
2001	0,26	348,877
2002	0,26	2988,836
2003	0,14	559,336
2004	0,36	214,53
2005	**0,28**	**893,0223**

30-39 yrs	α	β
2000	0,09	268,11
2001	0,02	1440,261
2002	0,05	673,071
2003	0,04	980,494
2004	0,05	1572,887
2005	**0,0032**	**1631,901**

30-39 yrs	α	β
2000	0,26	405,274
2001	0,42	249,356
2002	0,19	593,839
2003	0,08	1374,884
2004	0,10	1718,78
2005	**0,012**	**1994,189**

40-49 yrs	α	β
2000	0,05	636,254
2001	0,07	563,422
2002	0,06	724,027
2003	0,04	1388,763
2004	0,04	1517,811
2005	**0,05**	**1479,783**

40-49 yrs	α	β
2000	0,07	817,929
2001	0,14	236,819
2002	0,13	364,231
2003	0,07	937,798
2004	0,06	1057,633
2005	**0,067**	**1036,998**

50-59 yrs	α	β
2000	0,08	1140,735
2001	0,08	1253,969
2002	0,08	1135,048
2003	0,04	5446,69
2004	0,03	9453,228
2005	**0,02**	**9931,246**

50-59 yrs	α	β
2000	0,16	817,929
2001	0,10	493,454
2002	0,02	4727,121
2003	0,03	4626,985
2004	0,07	1902,828
2005	**0,001**	**4404,662**

60-69 yrs	α	β
2000	0,12	733,368
2001	0,15	933,858
2002	0,08	1590,178
2003	0,07	2358,159
2004	0,12	1817,192
2005	**0,084**	**2564,136**

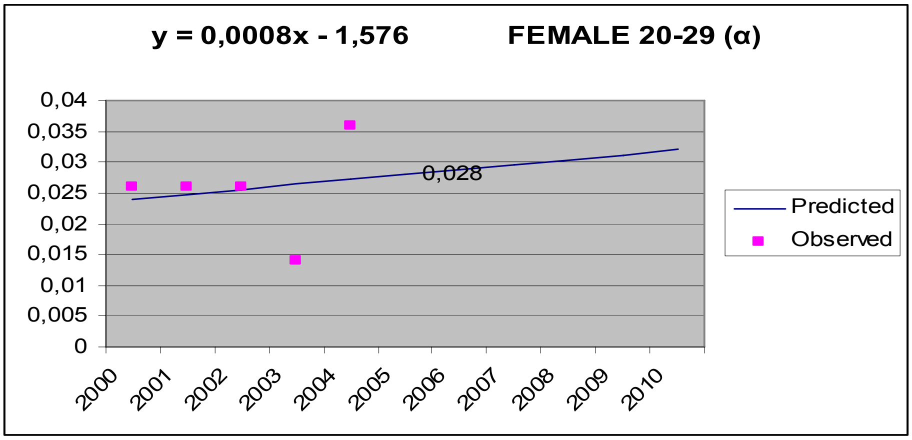
60-69 yrs	α	β
2000	0,16	282,391
2001	0,25	440,796
2002	0,26	370,176
2003	0,11	1248,835
2004	0,17	237,358
2005	**0,154**	**731,3031**

70-79 yrs	α	β
2000	0,72	88,136
2001	0,22	1191,81
2002	0,17	2987,158
2003	0,11	3819,599
2004	0,37	379,301
2005	**0,075**	**2656,237**

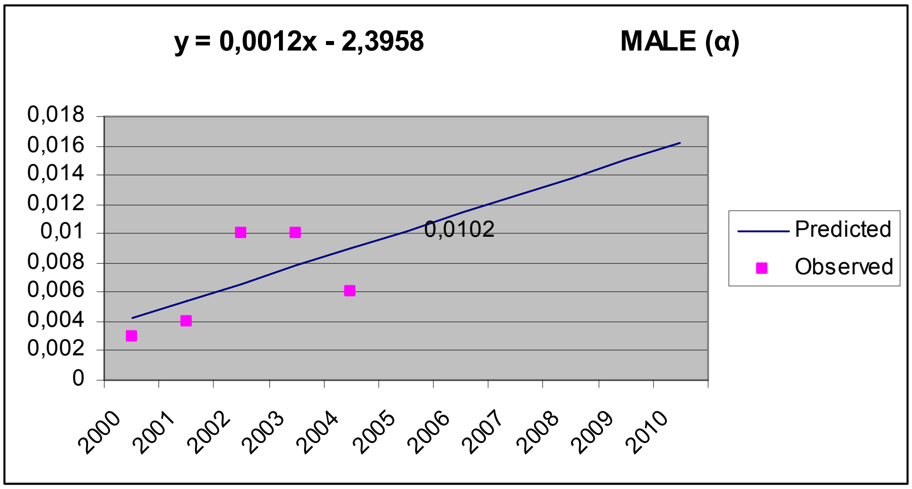
70-79 yrs	α	β
2000	0,08	352,967
2001	0,22	83,861
2002	0,35	180,19
2003	0,45	154,61
2004	0,54	57,854
2005	**0,673**	**10,0533**

Table 5.2.1.a

The results in the above tables that use linear regression are summarized in the plots which follow. Other examples, as shown in the following graphs, are illustrated as the forecast of parameter a for males and females age group 20-29.



Graph 5.2.1.a



Graph 5.2.1.a

See Appendix 9 for a representation of the analytical linear regression for each group.

5.2.2 Estimation of the Number of Claims

In conclusion, we found in this paper that the Claim follows Poisson's distribution. In order to estimate the following year, 2005, with Poisson's parameter, we take the mean of the parameters in each age group and gender classification for each year in our study. This gives us the new parameter for 2005, λ_{2005} , as it appears in the formula below:

$$\lambda_{2005} = \frac{\lambda_{2000} + \lambda_{2001} + \lambda_{2002} + \lambda_{2003} + \lambda_{2004}}{5}$$

Gender New	NEW GROUP	λ_{2000}	λ_{2001}	λ_{2002}	λ_{2003}	λ_{2004}	λ_{2005}
MALE	0-9	0,026	0,029	0,034	0,018	0,036	**0,0286**
	10-19	0,014	0,02	0,025	0,008	0,009	**0,0152**
	20-29	0,03	0,036	0,032	0,037	0,035	**0,0340**
	30-39	0,031	0,036	0,031	0,031	0,031	**0,0320**
	40-49	0,036	0,035	0,036	0,033	0,029	**0,0338**
	50-59	0,052	0,056	0,054	0,054	0,043	**0,0518**
	60-69	0,058	0,066	0,054	0,058	0,056	**0,0584**
	70-79	0,098	0,083	0,079	0,067	0,091	**0,0836**
FEMALE	0-9	0,022	0,018	0,02	0,02	0,017	**0,0194**
	10-19	0,02	0,015	0,013	0,007	0,013	**0,0136**
	20-29	0,078	0,069	0,066	0,056	0,076	**0,0690**
	30-39	0,089	0,099	0,096	0,086	0,074	**0,0888**
	40-49	0,04	0,04	0,042	0,042	0,038	**0,0404**
	50-59	0,037	0,044	0,048	0,045	0,04	**0,0428**
	60-69	0,056	0,063	0,071	0,066	0,053	**0,0618**
	70-79	0,008	0,036	0,076	0,1	0,062	**0,0564**

Table 5.2.2.a

The table above displays the λ_{2005}. In respect to the λ of male gender age groupings, it is apparent that a significant deviation is not present between each age classification. Looking at the age classes 20-49, we see that the same λ during 0.034 is calculated and the same running λ appears in age classes 50-69 as 0.055. However, a strong continuous λ among the female gender age groupings cannot be seen.

5.2.3 Estimation of Incurred Loss

As we stated in chapter 2.2.1, the estimation of incurred loss (S), which is the total number of claims times the total severity, can be made. In order to reach this result we need to calculate the expected frequency of claims and then multiply this with the expected severity of claims. By doing so, the expected claim is calculated by taking the results of S multiplied by the risk exposure.

Also we can calculate the variance of the claim that will give us in turn a more realistic pricing of the products.

Male

Group	Frequency	Severity	Incurred Loss
0-9	0,0286	156,6434	4,48
10-19	0,0152	88,81579	1,35
20-29	0,0340	65,29412	2,22
30-39	0,0320	163,125	5,22
40-49	0,0338	210,0592	7,1
50-59	0,0518	383,3977	19,86
60-69	0,0584	368,8356	21,54
70-79	0,0836	238,2775	19,92

Table 5.2.3.a

Female

Group	Frequency	Severity	Incurred Loss
0-9	0,0194	23,71134	0,46
10-19	0,0136	35,29412	0,48
20-29	0,069	362,3188	25,00
30-39	0,0888	26,91441	2,39
40-49	0,0404	172,0297	6,95
50-59	0,0428	10,28037	0,44
60-69	0,0618	182,2006	11,26
70-79	0,0564	12,05674	0,68

Table 5.2.3.b

As one can see from the above tables, the expected incurred loss, which is the product of expected frequency with expected severity, is displayed as an increasing pattern as age group is progressing for the male gender group, in general. This claim behaviour is reasonable since aging tends to bring on a higher frequency of hospitalization.

However the same can not be said for Females where someone could observe a hike in the age group of 20-29. This could be explained by maternity; however, the same should be observed also for the age group 30-39 but it is not. This could be explained by poor data experience in the latter age group. Thus, in order to apply this result for pricing, considerations should be given to the fact that the data needs to be smoothed according to the needs of the company and in order to reflect the reality of hospitalization for this age group more accurately.

Furthermore in pricing, we must be more mindful of the future parameters. Therefore, we must have a closer look at the results of the Linear Regression. We observe that the *R* (square) is poor. Given this perception, the Linear Regression is not reliable for this cover. In an effort to have a better result, our next step was to slice the outlier observations and run Linear Regression once more. If we use the

R(square) theory, we can not accept the result of the Linear Regression. However, when a comparison is made between the results discussed in this paper and the actual Company results for the years leading to 2004, the comparative results are fairly similar. This comparison makes it very difficult to reject the Linear Regression as it corresponds with the Company's past analysis. The Company will need to decide whether to accept or not to accept the Linear Regression. If the decision is made to not accept Linear Regression, we recommend that the future trend be based upon the results of the previous year, 2004. Alternatively, the Company can take the average of the value of the old parameters in order to estimate the future trend.

6. Conclusion

This thesis has described a statistical approach to determine the claim behaviour of Daily Indemnity Insurance cover. This particular piece of health insurance coverage deserves examination and was chosen for this paper as it is one of the most common components, or covers, for an insured to attach to his policy contract. As such, the Company placed interest in exploring the claim behaviours of this coverage. The available data, which was extrapolated from the raw data information in a total set of five years of experience beginning with the year 2000, was generated to fit the key variables necessary to describe the claim behaviour. The Company had established its tariffs based on a certain age group philosophy in order to comply also with other business needs. However, the analysis of this paper focused more on the theoretical approach rather than the practical approach.

While creating the various distribution models, it became clear that the results were similar or in close comparison with each other for the various classes. With the permission and guidance of the Company supervisors, the Company's age groupings were increased from a five year interval to a ten year interval for the exclusive use of this study. The new age groups, as defined on chapter 4.3, produced a clearer distribution with regard to the Claim and Incurred Loss variables. Thus, theoretically, it may be suitable to make a recommendation that the Company modify and adapt a larger age group interval where needed.

Continuing with the modified age groups and the distribution produced, the parameters of the distribution are calculated. The parameters can directly assist the Company with the pricing of the insurance coverage for the following year.

When examining the result of the linear regression, we see that there is not a large set of data present. As such, it would be best if the Company's pricing department used the result of the parameters only once for estimation purposes for the next year. Another issue with the data becomes apparent; the data is fairly recent, having been accumulated only over the past five years. Given the fact that a long term trend cannot be discovered, the company would be best served by calculating the linear regression every year and change the price of the products accordingly.

Bibliography

Bowers L. Newton. Actuarial Mathematics. USA: The Society of Actuaries, 1986

Lindgren W. Bernard. Statistical Theory Fourth Edition. Florida: Chapman & Hall, 2000

Ross M Sheldon. A First Course in Probability. Upper Saddle River, New Jersey: 2002

Ross M. Sheldon. Introduction to Probability Models. Florida: Academic Press, 2003

Retiniotis Stamatis. Statistics from the Theory to Process within SPSS 11.0. Athens: New Technology, 2004.

Tamhane, Ajit C. and Dorothy D. Dunlop. Statistical and Data Analysis from Elementary to Intermediate. Upper Saddle River, New Jersey: Prentice-Hall, 2000

Engineering Statistical Handbook. Available at http://www.itl.nist.gov

Yr of Report = 2000, Gender New = Male

One-Sample Kolmogorov-Smirnov Test[d]

		#Clms Cov
N		4[c]
Poisson Parameter[a,b]	Mean	.
Most Extreme Differences	Absolute	,002
	Positive	,001
	Negative	-,002
Kolmogorov-Smirnov Z		,006
Asymp. Sig. (2-tailed)		1,000

a. Test distribution is Poisson.

b. Calculated from data.

c. The mean was found to be ,00, but the parameter of the Poisson distribution must be positive. One-Sample Kolmogorov-Smirnov Test cannot be performed.

d. Yr of Report = 2000, Gender New = Male

Yr of Report = 2000, Gender New = Female

One-Sample Kolmogorov-Smirnov Test[d]

		#Clms Cov
N		3[c]
Poisson Parameter[a,b]	Mean	.
Most Extreme Differences	Absolute	,050
	Positive	,045
	Negative	-,050
Kolmogorov-Smirnov Z		,086
Asymp. Sig. (2-tailed)		1,000

a. Test distribution is Poisson.

b. Calculated from data.

c. The mean was found to be ,00, but the parameter of the Poisson distribution must be positive. One-Sample Kolmogorov-Smirnov Test cannot be performed.

d. Yr of Report = 2000, Gender New = Female

Yr of Report = 2001, Gender New = Male

One-Sample Kolmogorov-Smirnov Test[d]

		#Clms Cov
N		2[c]
Poisson Parameter[a,b]	Mean	.
Most Extreme Differences	Absolute	,029
	Positive	,026
	Negative	-,029
Kolmogorov-Smirnov Z		,058
Asymp. Sig. (2-tailed)		1,000

a. Test distribution is Poisson.

b. Calculated from data.

c. The mean was found to be ,00, but the parameter of the Poisson distribution must be positive. One-Sample Kolmogorov-Smirnov Test cannot be performed.

d. Yr of Report = 2001, Gender New = Male

Test Statistics Female 2004

	#Clms Cov
Chi-Square[a]	11,219
df	1
Asymp. Sig.	,001

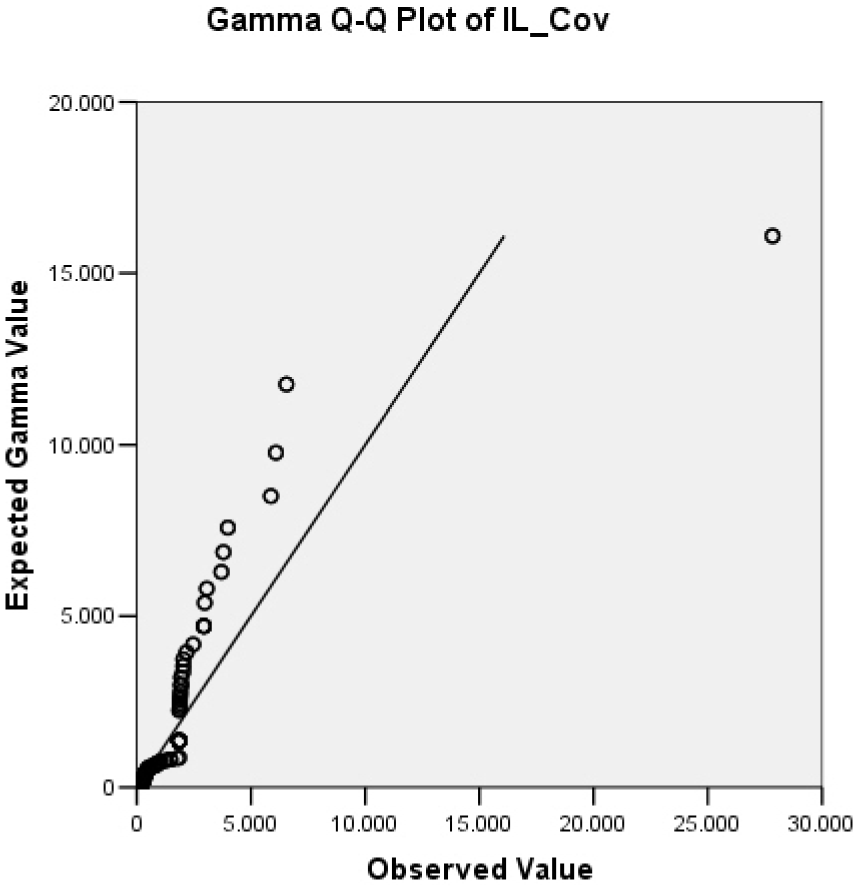
[a] 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 569,1.

Test Statistics female 2003

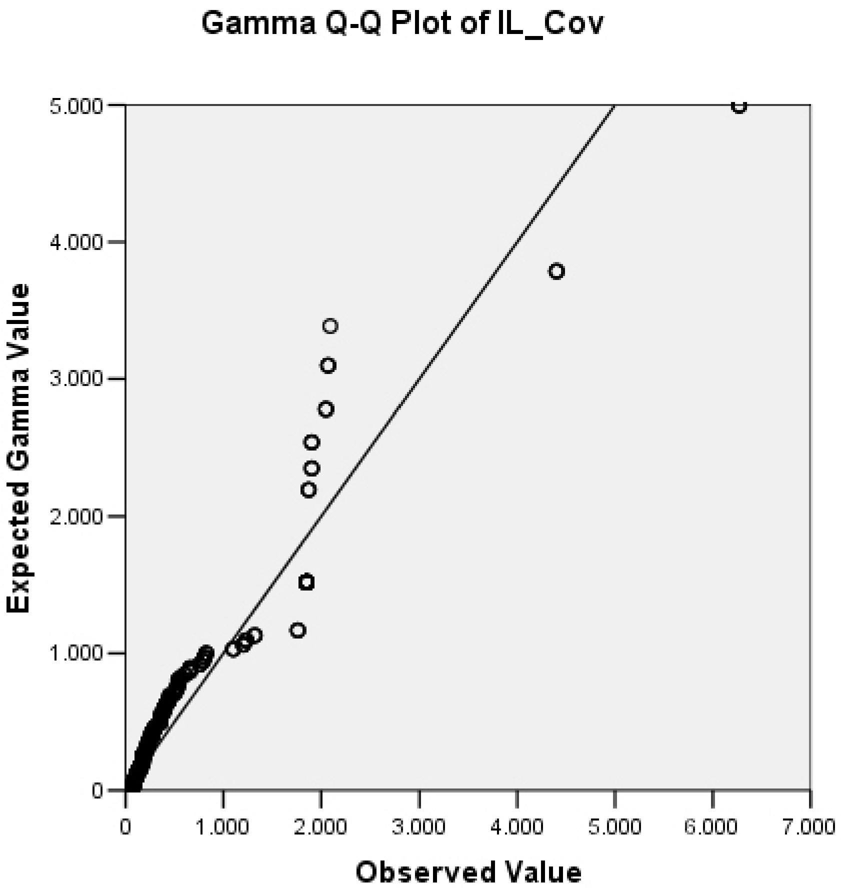
	#Clms Cov
Chi-Square[a]	1,897
df	1
Asymp. Sig.	,168

a. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 627,2.

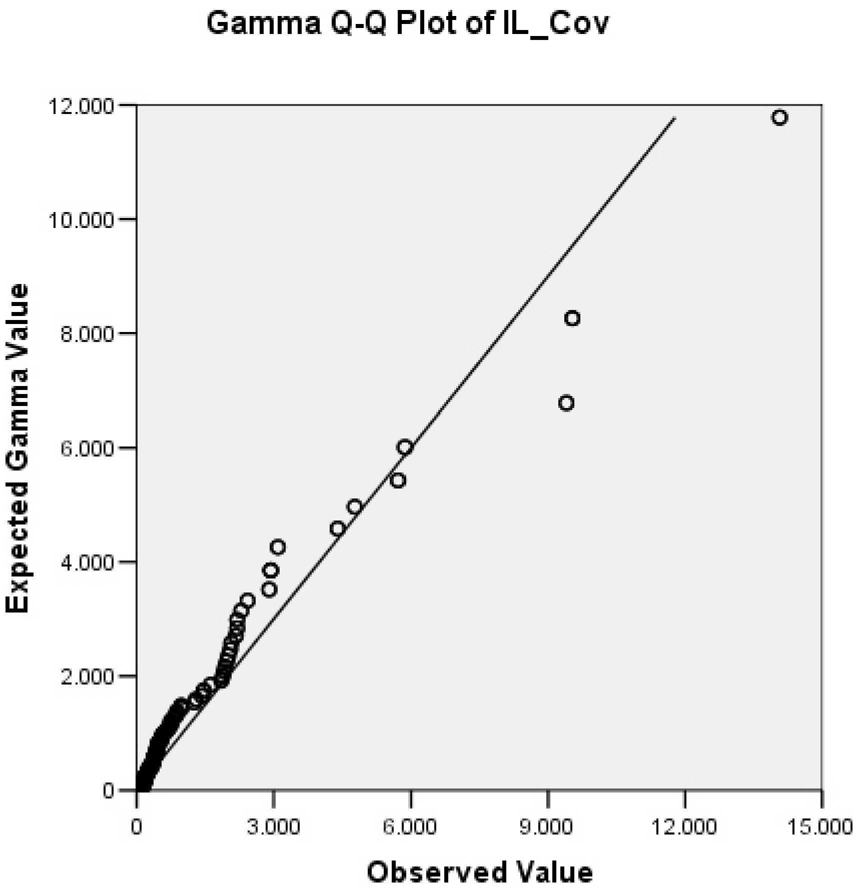
MALE 2004


Gamma Q-Q Plot of IL_Cov
Expected Gamma Value
Observed Value
20.000
15.000
10.000
5.000
0
0
5.000
10.000
15.000
20.000
25.000
30.000

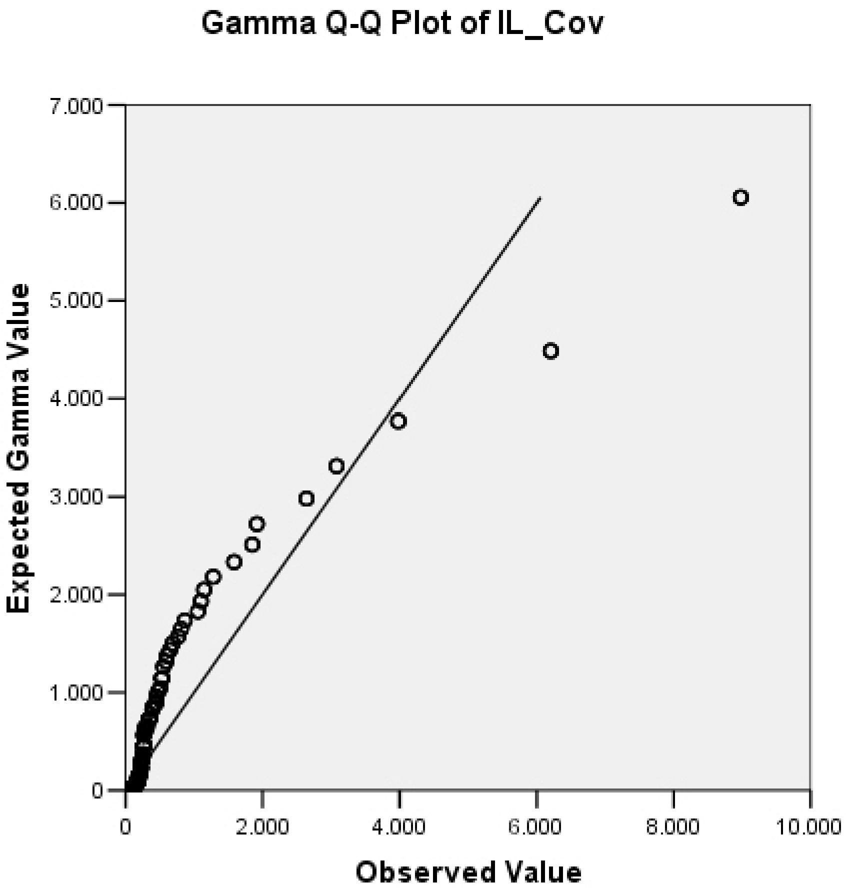
FEMALE 2004


Gamma Q-Q Plot of IL_Cov
Expected Gamma Value
5.000
4.000
3.000
2.000
1.000
0
Observed Value
0
1.000
2.000
3.000
4.000
5.000
6.000
7.000

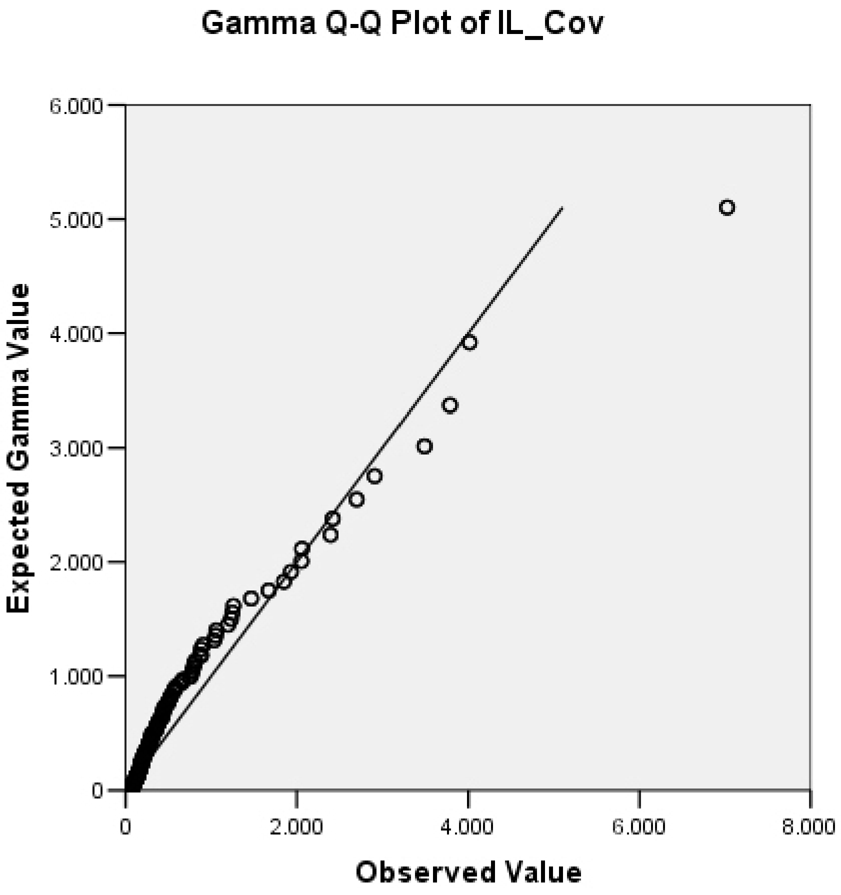
MALE 2003


Gamma Q-Q Plot of IL_Cov
Expected Gamma Value
Observed Value
12.000
10.000
8.000
6.000
4.000
2.000
0
0
3.000
6.000
9.000
12.000
15.000

FEMALE 2003

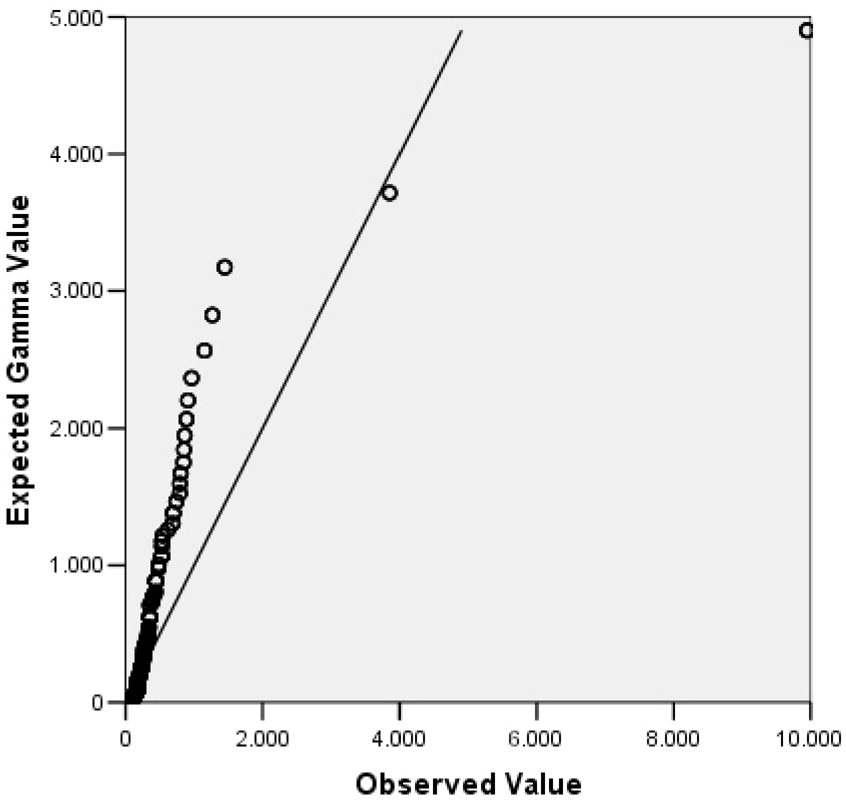

Gamma Q-Q Plot of IL_Cov
Expected Gamma Value
Observed Value
7.000
6.000
5.000
4.000
3.000
2.000
1.000
0
0
2.000
4.000
6.000
8.000
10.000

MALE 2002

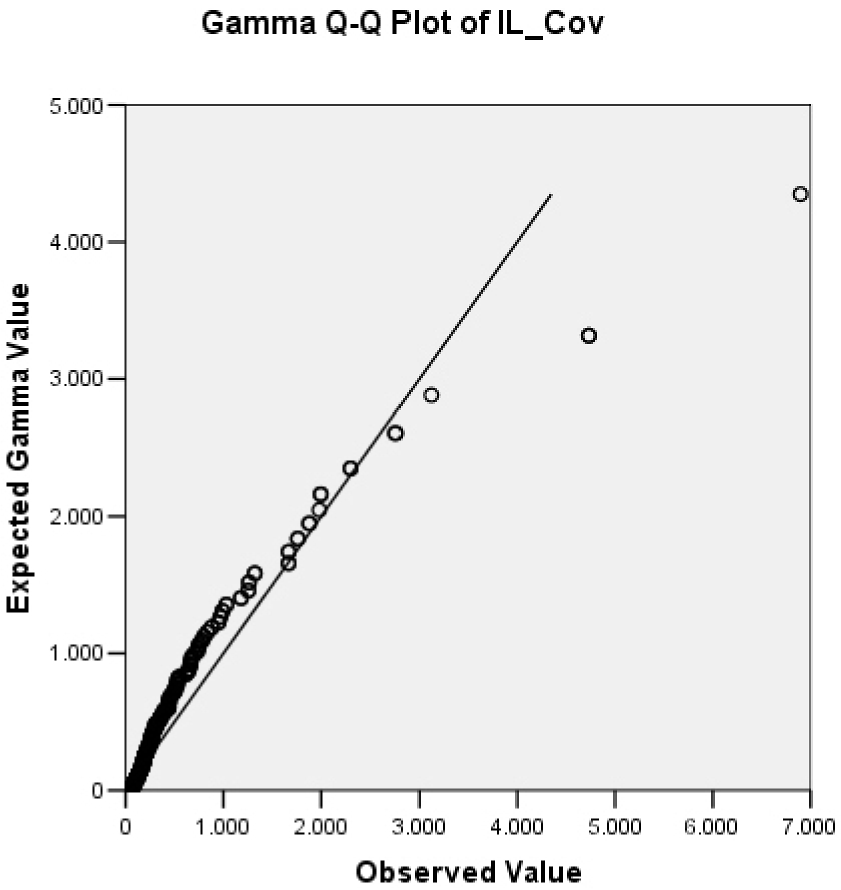

Gamma Q-Q Plot of IL_Cov
Expected Gamma Value
Observed Value
0
1.000
2.000
3.000
4.000
5.000
6.000
0
2.000
4.000
6.000
8.000

FEMALE 2002

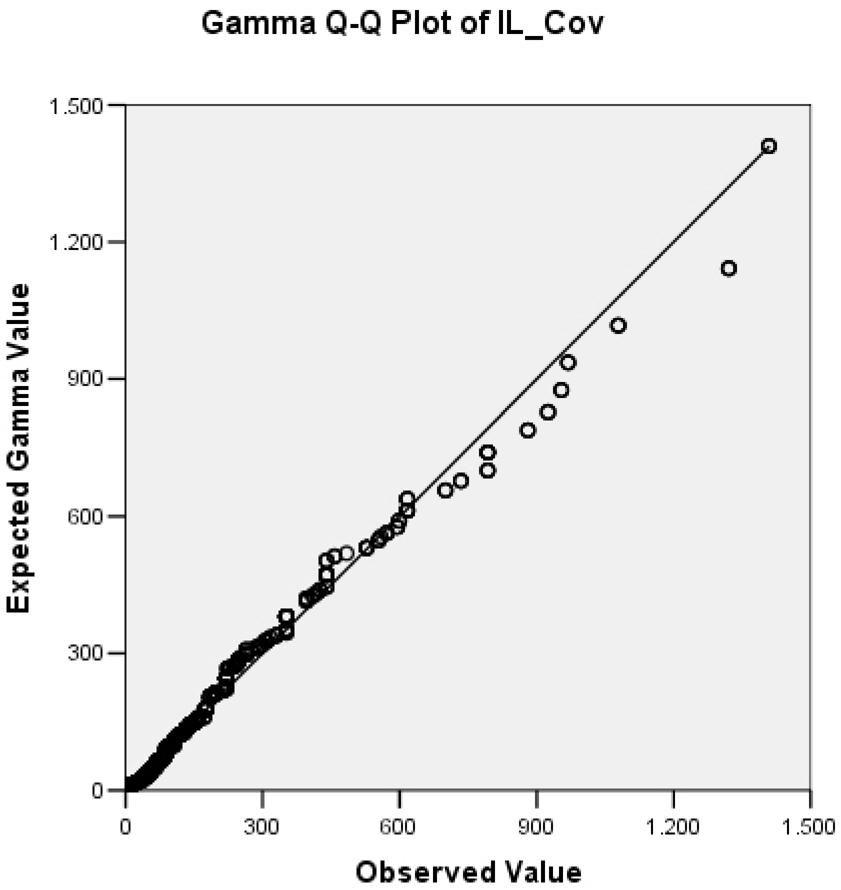
Gamma Q-Q Plot of IL_Cov



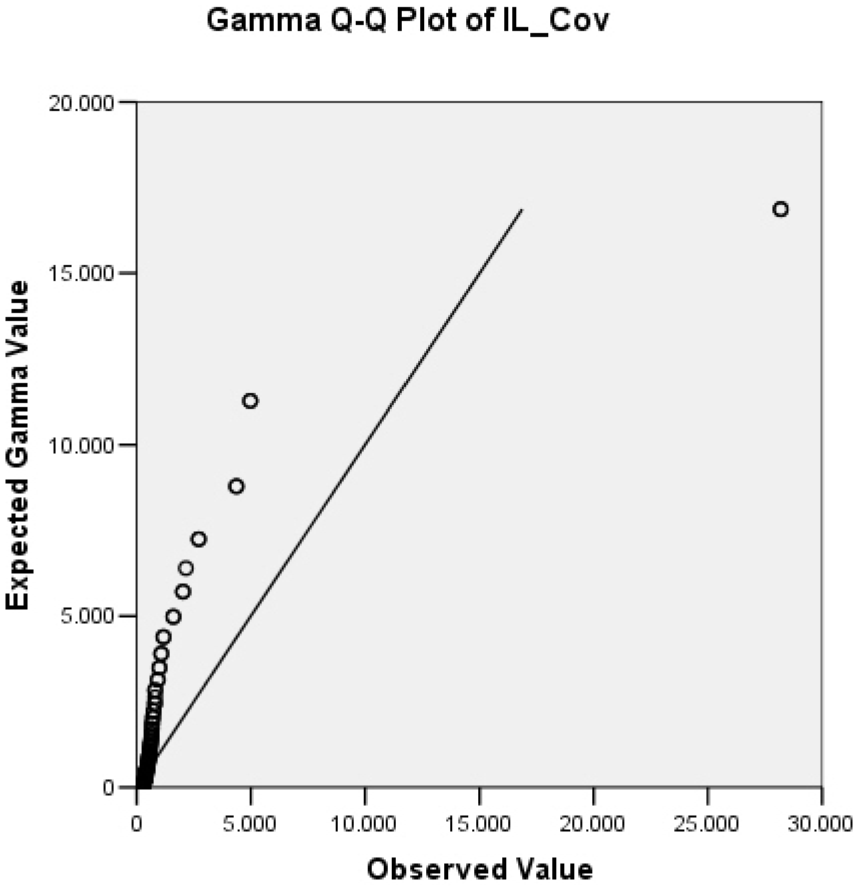
MALE 2001


Gamma Q-Q Plot of IL_Cov
Expected Gamma Value
5.000
4.000
3.000
2.000
1.000
0
0
1.000
2.000
3.000
4.000
5.000
6.000
7.000
Observed Value

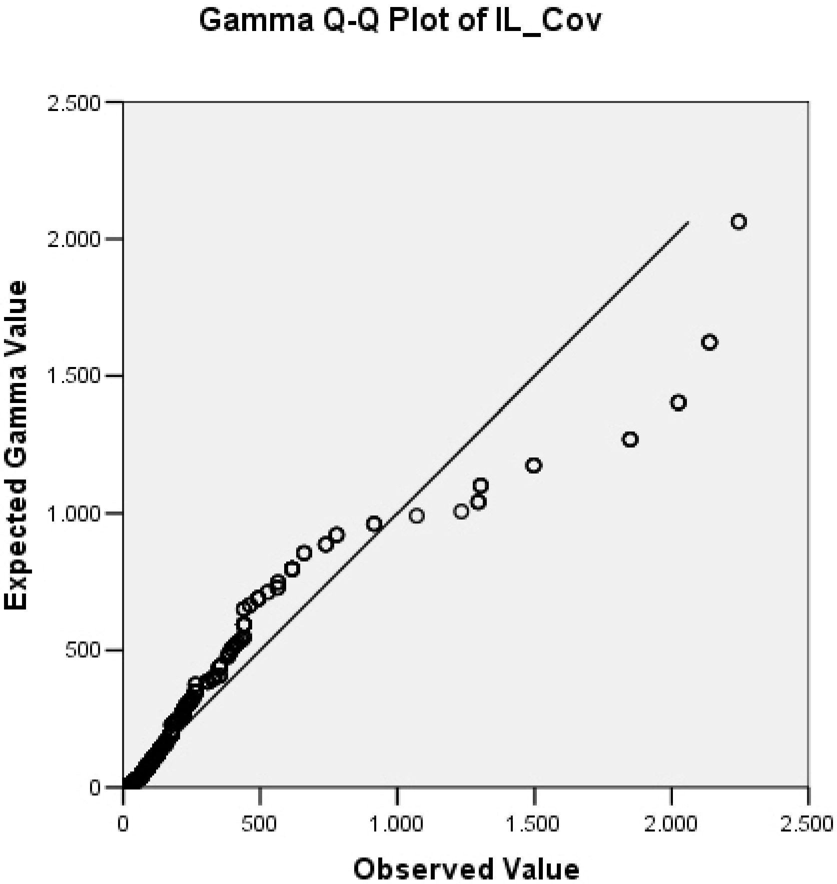
FEMALE 2001


Gamma Q-Q Plot of IL_Cov
Expected Gamma Value
1.500
1.200
900
600
300
0
0
300
600
900
1.200
1.500
Observed Value

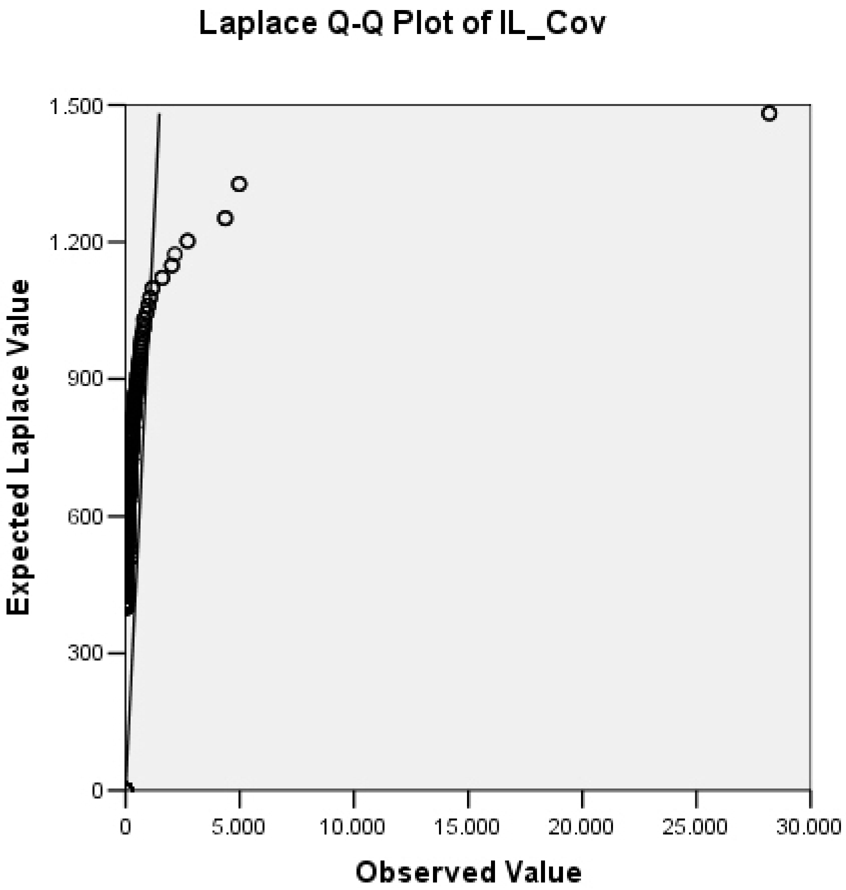
MALE 2000


Gamma Q-Q Plot of IL_Cov
Expected Gamma Value
20.000
15.000
10.000
5.000
0
0
5.000
10.000
15.000
20.000
25.000
30.000
Observed Value

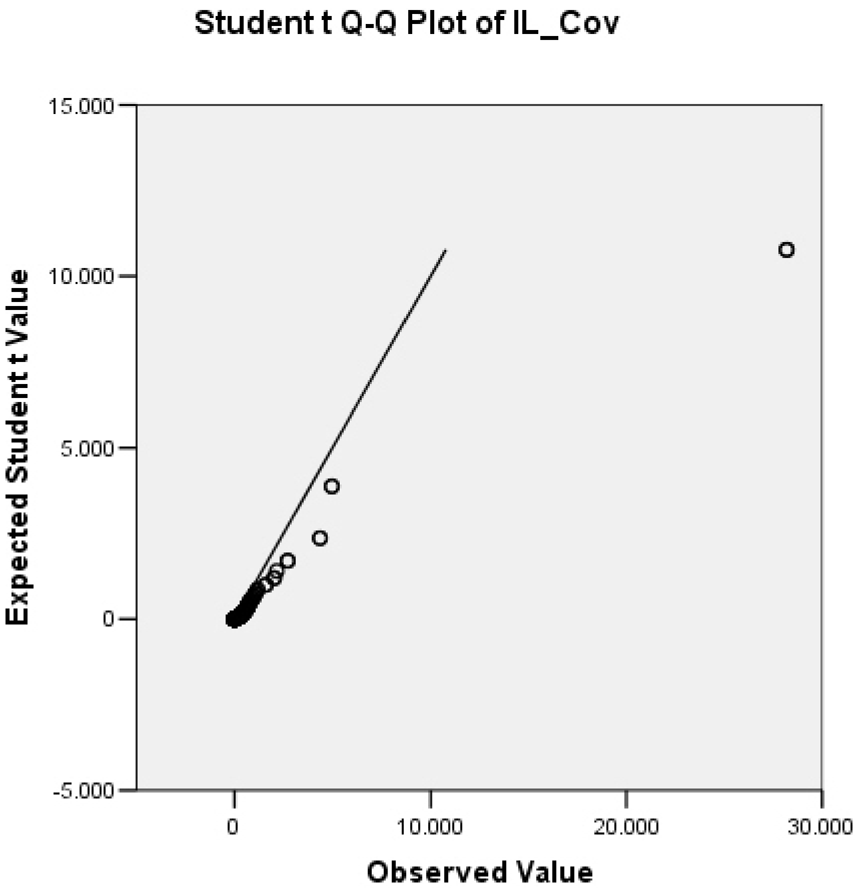
FEMALE 2000


Gamma Q-Q Plot of IL_Cov
Expected Gamma Value
2.500
2.000
1.500
1.000
500
0
0
500
1.000
1.500
2.000
2.500
Observed Value

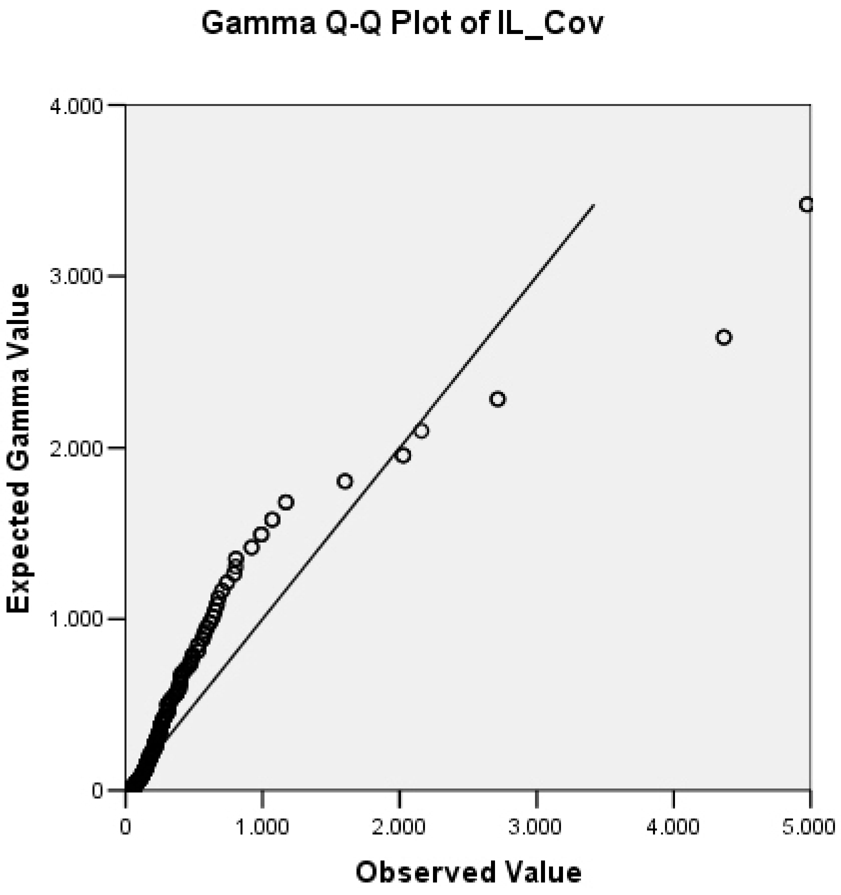
MALE 2000


Laplace Q-Q Plot of IL_Cov
Expected Laplace Value
1.500
1.200
900
600
300
0
Observed Value
0
5.000
10.000
15.000
20.000
25.000
30.000

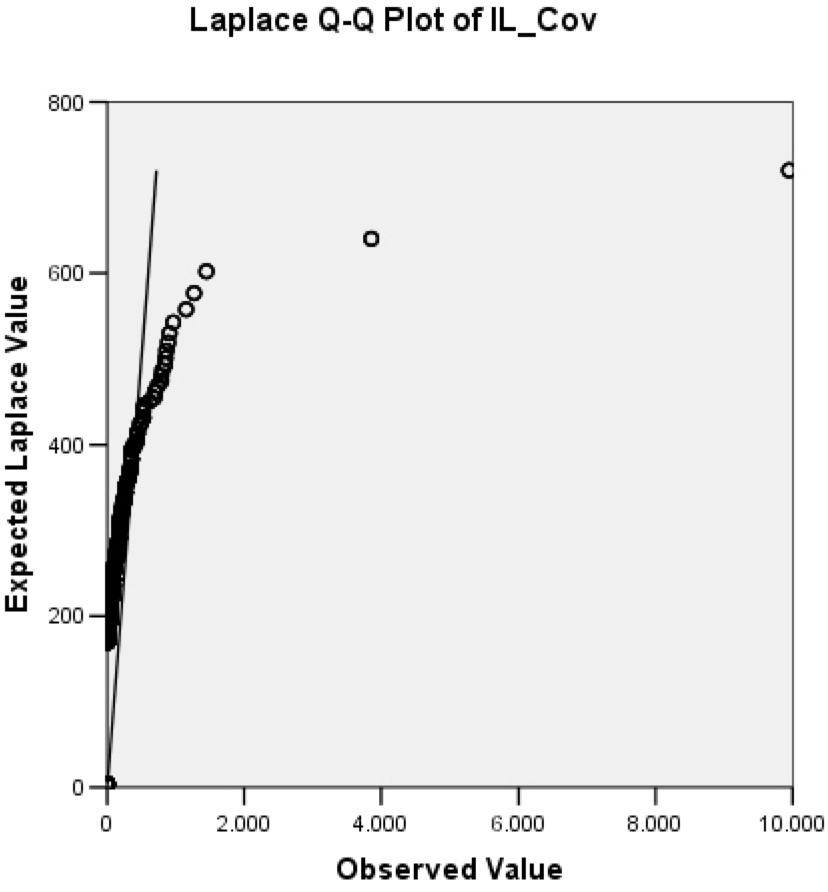
MALE 2000


Student t Q-Q Plot of IL_Cov
Expected Student t Value
15.000
10.000
5.000
0
-5.000
Observed Value
0
10.000
20.000
30.000

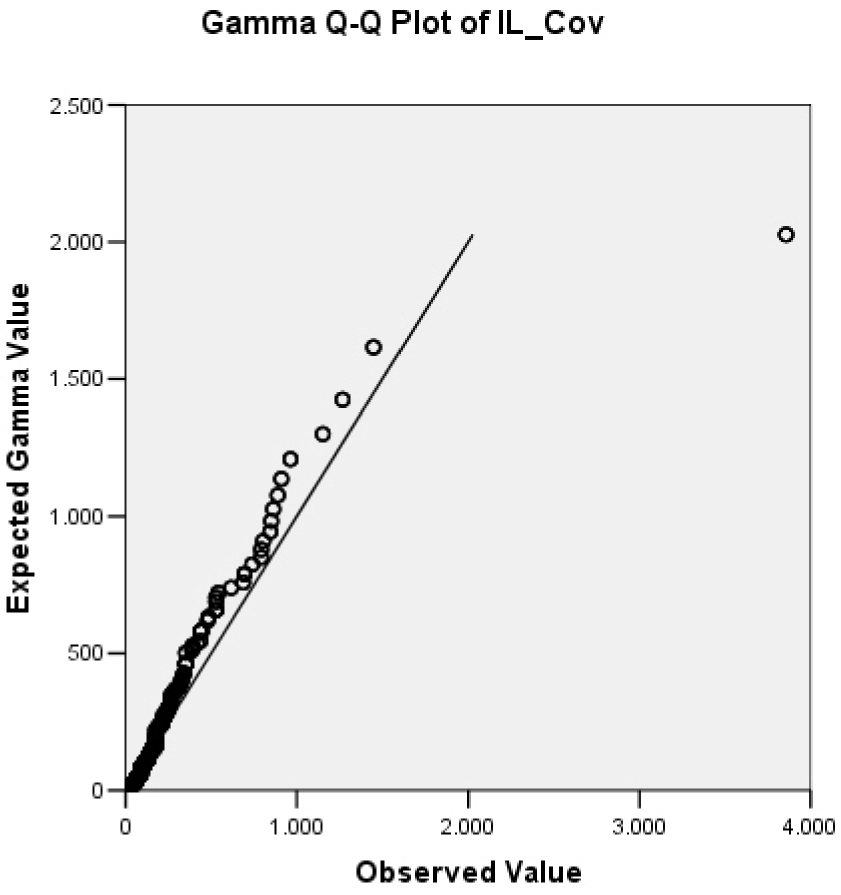
MALE 2000 (OUT)



FEMALE 2002


Laplace Q-Q Plot of IL_Cov
800
600
400
200
0
Expected Laplace Value
0
2.000
4.000
6.000
8.000
10.000
Observed Value

FEMALE 2002 OUT



2002

FEMALE

Age1	# Clms Cov
Chi-Square	0,109
df	1
Asymp. Sig.	0,741

Age2	# Clms Cov
Chi-Square	1,652
df	1
Asymp. Sig.	0,199

Age3	# Clms Cov
Chi-Square	3,621
df	1
Asymp. Sig.	0,057

Age4	# Clms Cov
Chi-Square	0,884
df	1
Asymp. Sig.	0,347

Age5	# Clms Cov
Chi-Square	3,548
df	1
Asymp. Sig.	0,060

Age6	# Clms Cov
Chi-Square	1,276
df	1
Asymp. Sig.	0,259

Age7	# Clms Cov
Chi-Square	0,219
df	1
Asymp. Sig.	0,640

Age8	# Clms Cov
Chi-Square	0,609
df	1
Asymp. Sig.	0,435

Age9	# Clms Cov
Chi-Square	1,005
df	1
Asymp. Sig.	0,316

MALE

Age1	# Clms Cov
Chi-Square	0,027
df	1
Asymp. Sig.	0,870

Age2	# Clms Cov
Chi-Square	0,433
df	1
Asymp. Sig.	0,510

Age3	# Clms Cov
Chi-Square	0,160
df	1
Asymp. Sig.	0,689

Age4	# Clms Cov
Chi-Square	0,942
df	1
Asymp. Sig.	0,332

Age5	# Clms Cov
Chi-Square	0,151
df	1
Asymp. Sig.	0,697

Age6	# Clms Cov
Chi-Square	0,000
df	1
Asymp. Sig.	0,989

Age7	# Clms Cov
Chi-Square	9,202
df	1
Asymp. Sig.	0,002

Age8	# Clms Cov
Chi-Square	2,267
df	1
Asymp. Sig.	0,132

Age9	# Clms Cov
Chi-Square	15,050
df	1
Asymp. Sig.	0,000

Appendix 7

2003

FEMALE

Age1	# Clms Cov
Chi-Square	0,010
df	1
Asymp. Sig.	0,920

Age2	# Clms Cov
Chi-Square	0,152
df	1
Asymp. Sig.	0,697

Age3	# Clms Cov
Chi-Square	0,027
df	1
Asymp. Sig.	0,870

Age4	# Clms Cov
Chi-Square	0,075
df	1
Asymp. Sig.	0,784

Age5	# Clms Cov
Chi-Square	0,067
df	1
Asymp. Sig.	0,796

Age6	# Clms Cov
Chi-Square	0,024
df	1
Asymp. Sig.	0,877

Age7	# Clms Cov
Chi-Square	0,018
df	1
Asymp. Sig.	0,894

Age8	# Clms Cov
Chi-Square	0,011
df	1
Asymp. Sig.	0,917

Age9	# Clms Cov
Chi-Square	0,025
df	1
Asymp. Sig.	0,875

MALE

Age1	# Clms Cov
Chi-Square	0,000
df	1
Asymp. Sig.	0,998

Age2	# Clms Cov
Chi-Square	0,004
df	1
Asymp. Sig.	0,951

Age3	# Clms Cov
Chi-Square	0,082
df	1
Asymp. Sig.	0,775

Age4	# Clms Cov
Chi-Square	0,014
df	1
Asymp. Sig.	0,905

Age5	# Clms Cov
Chi-Square	0,002
df	1
Asymp. Sig.	0,966

Age6	# Clms Cov
Chi-Square	0,244
df	1
Asymp. Sig.	0,621

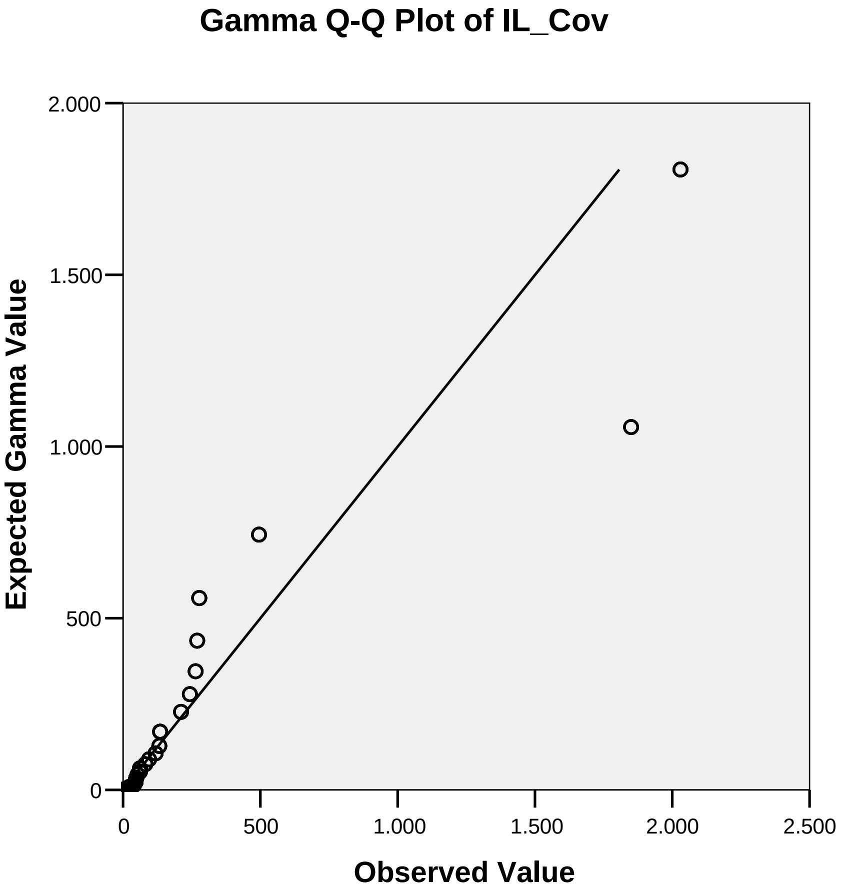
Age7	# Clms Cov
Chi-Square	0,057
df	1
Asymp. Sig.	0,812

Age8	# Clms Cov
Chi-Square	0,019
df	1
Asymp. Sig.	0,890

Age9	# Clms Cov
Chi-Square	0,010
df	1
Asymp. Sig.	0,920

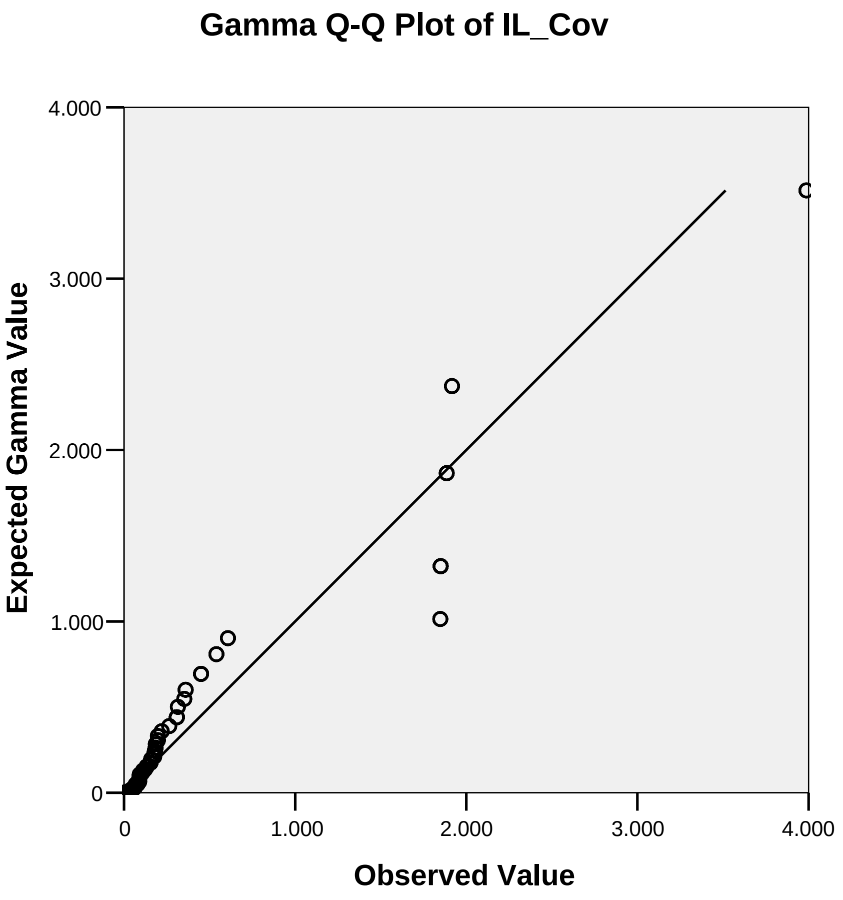
WE ARE NOW ANALYZING GROUP 3 (MALE 20-29 AGES) YEARS 2004

Appendix 8



WE ARE NOW ANALYZING GROUP 4 (MALE 30-39 AGES) YEARS 2004

Appendix 8

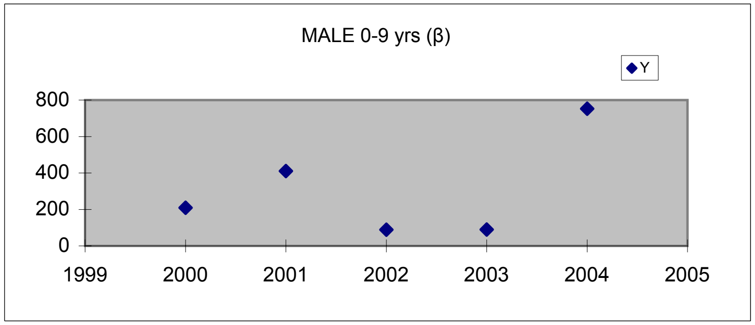


SUMMARY OUTPUT

MALE 0-9 yrs

Parameter **β**

Regression Statistics	
Multiple R	0,4326843
R Square	0,1872157
Adjusted R Square	-0,083712
Standard Error	291,59872
Observations	5


MALE 0-9 yrs (β)

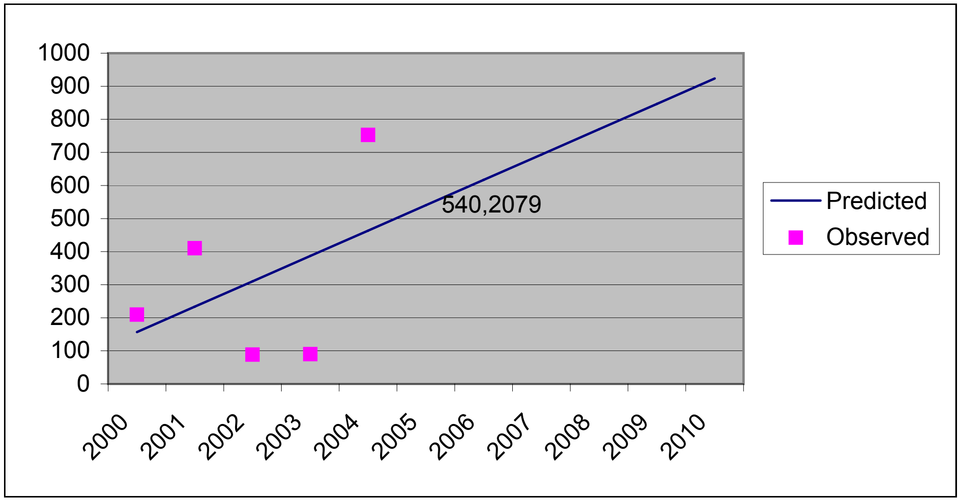
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	58756,9774	58756,98	0,691016	0,466797
Residual	3	255089,444	85029,81		
Total	4	313846,421			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	-153149,3	184607,694	-0,829593	0,467614	-740653,3	434354,8	-740653,3	434354,8
X Variable 1	76,6531	92,2116124	0,831274	0,466797	-216,8054	370,1116	-216,8054	370,1116

y = 76,6531x -153149

	Predicted	Observed
2000	156,9424	209,346
2001	233,5955	410,452
2002	310,2486	88,687
2003	386,9017	89,841
2004	463,5548	752,917
2005	540,2079	
2006	616,861	
2007	693,5141	
2008	770,1672	
2009	846,8203	
2010	923,4734	


Predicted
Observed