KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY

Actuarial Applications of Hierarchical Modeling to Health Insurance Claims

By

Michael Atta-Mensah

(BSc. Actuarial Science)

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF M.PHIL ACTUARIAL SCIENCE

October 22, 2015

Declaration

I hereby declare that this submission is my own work towards the award of the M. Phil(Actuarial Science) degree and that, to the best of my knowledge, it contains no material previously published by another person nor material which had been accepted for the award of any other degree of the university, except where due acknowledgment had been made in the text.

Michael Atta-Mensah (PG1078613)	.....	.....
Student	Signature	Date

Certified by:

Dr. A. Y. Omari Sasu	.....	.....
Supervisor	Signature	Date

Certified by:

Prof. S. K. Amponsah	.....	.....
Head of Department	Signature	Date

Dedication

Elizabeth, Mary, Janet, Rebecca, Abigail and Gabby

Abstract

This study demonstrates actuarial applications that can be performed on the Health insurance claims in the country. To achieve this, data from the CPC scheme in Accra of the NHIA in the year 2013 was employed for the study. this consisted of facility type, number of claims submitted(in-patient and out-patient) and amount submitted (in-patient,out-patient,drugs and services charges). A hierarchical model allowing for frequency, claim type and severity amount to be jointly modeled was used. Based on this hierarchical model, we proceeded to estimate premium values under various conditions, however due to lack of information from the insurer most of these estimates could not be stated categorically. Applications of the study was also made to the Value-at-Risk theory. This fact not withstanding, a case has been made for the consideration of the hierarchical modeling approach to be considered as the means of analyzing health insurance claims since this model takes into consideration not only the loss (severity) amount submitted but also considers most especially factors integral to the planning and budgeting of the insurer, and these are, the frequency and type of claim. The hierarchical modeling approach thus provided further insight which previously was overlooked.

Acknowledgements

Firstly, my gratitude goes to the God Almighty, for His mercy and love that He has showered my path during the period of writing this thesis. To my supervisor, Dr. A. Y. Omari Sasu for his guidance, patience and constructive criticisms. Also, to Mr. David Adedia and Mr. George Neurtey, you really came through for me, thank you. Pharm. Regina Esinam Abotsi, your thoughtful words of encouragement during the writing of this thesis can not be underestimated, Thank you. To all my course mates and friends for being there for me during this period. I really appreciate every help.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the current trend of insurance claims, more precisely the National Health Insurance Scheme challenges of making true its promise to health care providers in terms of payment of rendered services, it is incumbent that a more objective look is given to the scheme. It is the responsibility of actuaries to take advantage of modern statistical and computing advances to analyze claims made to the insurance authority, come up with estimates and forecasts which will better equip the insurer to face its financial demands even before they raise their heads.

1.1 Problem Statement

According to Mr. Sylvester Mensah, CEO of the National Health Authority (NHIA), Health care facilities across the country admits more than 85% of their internally generated funds come from payments from the NHIA. The health insurance scheme is the cash cow to sustaining the health care industry and the pharmaceutical supply chain in Ghana. They also confirm that about 90% of patients are health insurance subscribers. This also implies that the NHIS bares an overwhelming proportion of the cost burden of patient care in the country (todayghananews.com, 2015) From this it can be gathered that claims payment is the lifeline for service operators and any form of delay in the payment almost surely cripples services. However here is the case that payments can be delayed as long as five to six months with isolated cases of eight months (DailyGuide, 2015)

From the above, it is clearly shown how important the national insurer is to the insured, and the service providers and all needs to be done by all who matter and play a role in sustaining this social health invention. Due to this delay in payment

of claims, it has become a popular happening in recent times to hear banter, accusations and counter-accusations being traded by key stakeholders both in the electronic and print media. A recent occurrence of this unfortunate banter of accusations and counter accusation was that between the minister of finance and the CEO of the NHIA. The issue then was how many months have claims payment not being outstanding. A key stakeholder who is very much missing in this whole cloud of confusion but plays a very vital role in the sustaining of the insurance program is the actuary. The actuary possesses peculiar skills that can help develop estimates to quantify possible claim values even before a financial year begins and hence equip the national insurer for claims ahead.

1.2 Objectives of study

The overarching objective of the study is to come up with probability distribution function for claims and claim types presented to a Ghanaian health insurer.

1.2.1 Specific objectives

The overarching objective of this study can be realized through the following specific objectives:

- Claims Frequency: Claims submitted by the various service providers (pharmacies, clinics and hospitals) form a count or discrete variable and hence can be fitted to a probability distribution. This can be used to model future claims and hence equip the NHIA in its decisions.

- Claim Type: Claims submitted to the national insurer is normally a combination of various service charges by the health care provider. This aspect of claims need to be looked into and assessed for its bearing on claim payments.

- Severity (loss) component: A probability distribution of the loss/claim amount submitted is then estimated.

1.2.2 Methodology

The study will combine the various probability distributions to develop a desired model which can be used as a general model for assessing claims submitted to the insurer. The Negative Binomial regression model estimates the $r-th$ success upon which claim frequency is distributed. With this model the general probability of claims submitted being submitted and processed for payment is developed. Progressing,Claims types is modeled with the property of conditional probability theory and finally, the generalized Pareto distribution with three (3)parameters is employed.

Secondary data was obtained for NHIA in the Claims Processing and Payment Report for CPC Scheme, Accra. Data obtained comprised of various variables, however, the variables of interest to this study included Number of Claims submitted, Inpatient(Ghc), Outpatient (Ghc), Drugs(Ghc), Services(Ghc) and Total Amount Submitted(Ghc). Total amount submitted is actually a sum of all the other money amount variables. Data analysis was done with the aid of R, STATA (version 12) and Easyfit. Equation 1.1 is a mathematical sentence of the model description.

$$f(N, M, y) = f(N) \times f(M|N) \times f(y|N, M) \tag{1.1}$$

$$\text{Distribution function} = \text{Frequency} \times \text{Claim type} \times \text{Severity}$$

where N represents frequency of claims (no.claims)

M represents Kind of Claim

y signifies loss amount

Below is a graphical display of the statistical model

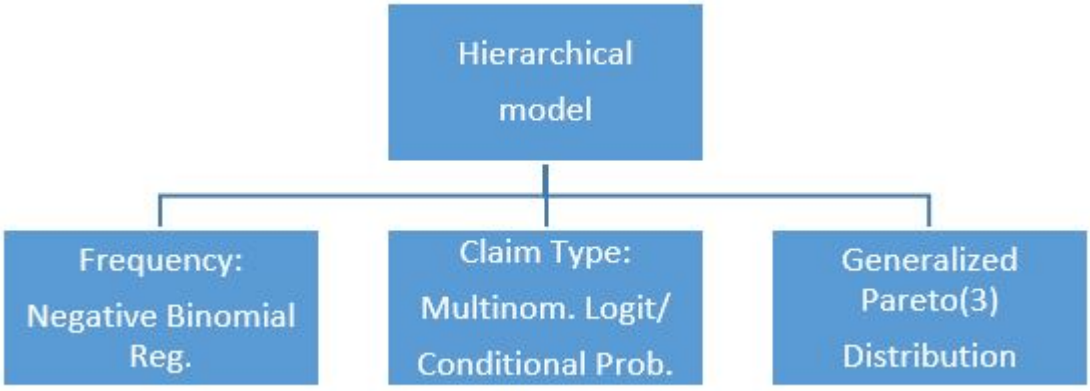


Figure 1.1: Framework of the Hierarchical Model
Source: Author's Construction

1.3 Justification of Work

At the end of this study, probability estimates of claim frequency, type and loss will be made available.These estimates will go a long way to keep the national insurer prepared for future claims submitted as this will determine their capital and Value at risk assessment. Results of this study can also go a long way to influence premiums charged and levels of coverage and other coverage modifications, should there be the need for such modifications to be introduced.

1.4 Source of Data

Secondary data was sourced from the Claims processing and payment report for CPC scheme, Accra branch of the National Health Insurance Authority for the year 2013. The decision to choose this insurer amongst the many other health insurance schemes out here in Ghana was mainly due to the level of availability and popularity of the scheme compared to the others. The National Health Insurance is a Ghana Government initiative with the aim of improving health care access to the majority of the populace. Information sourced from current literature on the subject will be used in guiding the directions and sphere of this

study.

1.5 Organization of Thesis

The Chapter one of this study contains Background of study, Problem statement, Objectives of study, Methodology, Justification of the study, and Organization of thesis. Chapter two contains the review of literature, where studies already carried out which are related to this study, methods and application of hierarchical data application and modeling were looked into. Some probability methods of interest for hierarchical modeling and application was considered under chapter three. Chapter four contains results of the data analysis, where the claims submitted to the NHIA datasets were analyzed using the methods outlined in chapter three. Finally, various findings from the analysis were discussed in Chapter five to check if the goals of this study are achieved. Recommendations are given with respect to the results obtained that are based on the methods used in the analysis.

1.6 Limitations of Study

Actuarial projections and estimates made under this study does not encompass all unforeseen occurrences or all other relevant instances; due to these reasons and also due to the fact that the analysis is not tailored for any particular insurance provider or health scheme, results as pertaining to the reality on the ground are likely to vary from that presented at the end of this study. Also, other researchers investigating the same field may come up with estimates that vary from those presented in this work due to the fact that, these researchers may operate with assumptions totally different from that utilized in this study, different data or developing models for entirely different purposes.

Chapter 2

Literature Review

In this section of the study, review of studies which had been conducted and are valid to our work are conducted. The review will be done in the areas of actuarial application to insurance data. Particular emphasis will be made on hierarchical data and loss distributions developed.

2.1 Health Insurance

According to the online investing glossary, InvestorWords.com (2015b), generally an insurance can be considered as an assurance of remuneration for a particular likely future misfortunes in return for an intermittent installment or consideration as it is called in the legal terms. Insurance in their design, are structured to protect the financial well-being of a company, an individual, or other entity in the event of an unexpected loss. Due to the nature of insurance, it may be mandatory (required by law) or optional (left at the discretion of the individual). An insurance contract is deemed entered into only upon all parties (Proposer and Acceptor) involved have gone through the various demands and are satisfied, then it is considered to be the existence between the parties, now are from that moment referred to as the insured and the insurer, respectively. The insurer takes upon itself the risk of compensating the insured for a predetermined event. This transferred liability or risk is only concluded on and remains a valid agreement upon exchange for periodic monetary payments, known as premiums to the insurer by the insured. Standard periods for payment of premiums are monthly, quarterly or annually. Due to the general nature of insurance, it can be used as a social protection tool from which health insurance is derived. Health insurance is hence defined as insuring oneself against the risk of incurring medical expenses,

thus, transferring this risk to a desired insurer. An insurer, mitigates its risk amongst a group or more precisely, a desired population of potential insureds by estimating the general health care risk and historic health expenditure amongst this group. By so doing, the insurer equips itself will the required information to develop the appropriate financial structure to sustain such a risk should it accept the risk of insuring the targeted population. Estimates developed with such an information will guide decisions on premiums and benefits or coverage limits of a health insurance policy. Health insurance contracts may also offer scope for visits to the doctor, medicine, hospital stays and any other medical expenses which may be desired and stated in the insurance contract. Various medical or health insurance policies are in existence, they however, differ in extent or inclusions of coverage, the co-payment and/or deductible size (the threshold amount below which the insurer is not liable to make payment), coverage limits (the threshold amount above which the insurer is not liable to make payment) and the treatment options accessible to the policyholder. A health insurance policy could be acquired specifically by a person as an individual, or through the procurements of an employer (InvestorWords.com, 2015a).

2.1.1 Health Insurance In Africa

Social Health Insurance (SHI) has been considered as having the potential of being a financing system in low-and average wage nations over the last two decades. SHI schemes are in existence in many Latin America countries and in recent years, have also been introduced across Asia. Despite this chalk of success, only a handful of African countries have SHI's implemented (McIntyre et al., 2003). As early as the 1930's, majority of European nations had some type of SHI introduced, and thereafter, had it implemented in various other high-wage nations (Roemer, 1991), with the same Africa expected to do same. A couple of West African nations have implemented one type or the other of social security cover aimed at improving health care services. Until recently in Southern or East Africa, it was

only Kenya which had started a type of compulsory health insurance (Kraushaar D., 1997). Some reasons given for the implementing or better still considering SHI include,it being used as a tool raking in extra income to adjust for reduction in tax-financed spending on health services (Ensor, 1999). Also, its deemed as an avenue for enhancing equity and effectiveness of health care resource use, this is done by enhancing access to health care for a broader spectrum of people. SHI is also acts as an avenue of controlling the rate of development in health care expenses.

McIntyre et al (2003), captured lessons that drawn from the creating and adjusting the SHI design in South African. This was done in connection with evidence recently acquired from other low- and average-wage nations, it was observed that a vital design prerequisite to advance value and maintainability is a common contribution and also, the risk pool over the SHI and any current private insurers. Furthermore, given the complex nature of SHI changes and the way that SHI is normally a stand out part of a more extensive bundle of health sector changes, the proper sequencing of execution of the SHI and related organizational and financing changes is key. As SHI advances in a nation, it is essential to benchmark the evolving way of its configuration against pre-determined targets keeping in mind the ultimate integrity of the policy (McIntyre et al., 2003).

Ekman (2004) in a review systematically assessing the evidence of the degree to which community-based health insurance is a reasonable choice for low-wage nations in organizing assets and offering financial security, the review contributes to the literary stock on health funding by developing and qualifying existing information, expressed that in general, the evidence base is restricted in degree and sketchy in quality. Also, there is strong proof that community-based health insurance provides some financial coverage by minimizing out-of-pocket expenditure. There is proof of moderate quality that such plans enhance cost-recuperation. There is feeble or no confirmation that plans have an impact on the nature of

consideration or the proficiency with which care is delivered. In supreme terms, the impacts are little and plans serve just a constrained area of the populace. The principal policy ramification of the review is that these forms of community financing plans are, in best case scenario, reciprocal to other more powerful frameworks of health financing. To enhance reliability and legitimacy of the proof base, analysts should concur on a more sound arrangement of result pointers and a more steady evaluation of these markers. Policy drafters should be adequately educated as to both the expenses and the advantages of actualizing different financing alternatives. The present proof base on community-based health insurance is quiet in light of this point. Another factor to be considered when contemplating Health Insurance scheme is, it's political nature. Health reform as per its characteristics, is political.

Sound technical investigation is never adequate to ensure the reception of the policy and hence financing reforms geared at advancing value are particularly prone to test personal stakes and generate resistance (Thomas and Gilson, 2004). The outline and execution of policies is about designating assets, conveying power and choosing whose needs require immediate attention (Barker, 1996). Due to its political nature, large-scale financing reforms, an example being the evolving of social health insurance (SHI), are particularly contentious since they straightforwardly impact who pays for, and who profits from the health care system. Though proponents such as these can overlooked by frontiers of reform, it is these worries that frequently bother the minds of political leaders.

Some health reforms tend to be high politically contention than others. In Olson (1965) and Nelson (1989), both authors conceded that it is redistribution to the underprivileged that is normally confronted with the most resistance. Most middle- and low-wage nations have the urban middle-class more structured with a louder voice, than the other different people-groups. As a contradiction , the rural populace are often more scattered and without the financial means to impact various policy strategies. Hence, the worries of the rich and powerful are not

hastily brushed aside (Grindle and Thomas, 1992). In a review of the Health Insurance policy development implemented in South Africa between 1994 and 1999, notwithstanding, over 10 years of debate, assessment and design, no set of Social Health Insurance (SHI) recommendations had by 1999 secured sufficient backing to end up as the premise for an execution plan. Conversely, calls to re-control the health insurance industry were quickly formulated and executed toward the end of this period. The procedures of actor engagement and administration, set against policy objectives and design details, have been fundamental to this experience (Thomas and Gilson, 2004).

An evaluation of the effect of health insurance on asset mobilization, financial security, administration use, nature of care, social incorporation and community empowerment in low-and lower-middle-wage nations in Africa and Asia by Spaan et al (2012) comprised of 159 studies- 68 in Africa and 91 in Asia. It was observed that majority of African country's studies covered on Community-Based Health Insurance (CBHI) and these were generally of greater quality; SHI studies were for the most part Asian and of mid-range quality. Of the Asian studies only one Asian study tackled Private Health Insurance (PHI). Subjects such as social inclusion, utilization and financial protection and were of prominence as compared to subjects like resource care quality, mobilization and community empowerment. There exist immense evidence which points out that CBHI and SHI offers financial protection to members by cutting down members out-of-pocket expenditure, enhances service utilization. Also CBHI further enhances resource mobilization. Despite the aforementioned, evidence also exist, though minimal, indicates a positive impact of both CBHI and SHI on social inclusion and quality of care. Hence rendering the discussion on SHI and CBHI effect on community empowerment inconclusive. Also due to insufficient studies,the discussions about PHI have also ended inconclusively in all regards. The conclusion was health insurance provides some indemnity against the detrimental impacts of client charges and a promising avenue towards general health care services coverage.

Ndiaye et al. (2007), authored an overview of the progression of Community Health Insurance (CHI) in subSaharan Africa. The study pointed out that in 2003, almost 600 CHI initiatives were enlisted in twelve nations of francophone West Africa alone. On a regional stage, systems to promote coordination have been developed in Africa which aims to bolster and maintain regular surveillance on the advancements of this commendable model for financing medical services. Also on the national stage, governments are also making ready the required legal structures and statues for the CHI implementation. CHI is also progressively seen as a procedure to meet other development objectives than just health. It consists of an intriguing model to fund health care, to pool finanacial assets fairly and to improve on the healthcare of clients. Despite the many pros, CHI development however still encounter numerous difficulties. The pertinence of more expert contribution in the administration of CHI and the requirement of vital subsidy for CHI schemes are constantly noticed. There is the additionally need to improve the relationship of CHI with alternate players in the healthcare industry and to scale-up CHI in order to pick up in viability and productivity. The blast in the quantity of schemes in Africa over recent years is a pointer of the expanding attractiveness of the model. However, in practice, enrollment rates per scheme stay low or are just increased gradually. Setting up context-specific research is required on the reasons that keep individuals from enlisting in great numbers. On that premise, adequate moves needed to be made locally can be recognized. Allegri et al (2009)accented to the above findings and contributed further stating that CHI expand access to care and provide monetary security against the expense related to sickness for needy individuals barred from formal insurance schemes. In SubSaharan Africa (SSA), experience on the field, however, shows that a sequence of operational challenges still impede the fruitful advancement of CHI, yielding negative consequences for potential advancement towards expanded access to enhanced monetary security and mind. With the aim to offer policy drafters the vital knowledge on the issues in question and with policy recom-

mendations to counter such issues, reinforcing CHI and improving its role inside SSA health schemes, Allegri et al (2009) reviewed literature which reveal that the significant challenges at present confronted by CHI in SSA are operational in nature and center around five cardinal points:

- absence of clear legislative and administrative structure;
- low enlistment rates;
- inadequate hazard management process;
- weak administrative capacity; and
- high overhead expenses.

  From the review, Allegri et al (2009) calls for suitable policy interventions, particularly:

  - More commitment towards the advancement of sufficient legislation in backing of CHI;
  - Expanding uptake of measures to grow evenhanded enlistment;
  - The acceptance and implementation of sufficient hazard management measures in all schemes;
  - Significant investments from host nations and also from supporting agencies to enhance administrative capacity; and
  - Collective measures to control overhead expenses

Again, Private health insurance is also assuming an increasing responsibility in both high-and low-wage nations, yet is ineffectively comprehended by both policy-makers and researchers (Sekhri and Savedoff, 2005). An observation from (Sekhri and Savedoff, 2005) indicates that the variance in the public and private medical insurance is most of the time overstated since well managed private insurance markets share numerous components with the public health insurance. In their closing comments made arguments that developing countries cannot overlook

private health insurance in that, it can be saddled to serve the general public interest if governments execute viable regulations and concentrate on programs for the individuals who are vulnerable and poor. Moreover, (Sekhri and Savedoff, 2005) further argued that private health insurance can be utilized as a transitional type of medical coverage to create involvement with the insurance establishments while the general public sector builds its capacity to oversee and fund coverage for health care.

2.1.2 Health Insurance In Ghana

Ghana, a middle low-income nation in sub-Saharan Africa, set out on a policy nationwide of supplanting the then out-of-pocket expenses at point of service with national health insurance module in 2003. According to Agyepong and Adjei (2008), moves at significant change need to consider and address these issues alongside moves to give confirmation to content decision-making. Without an investigation and comprehension of the legislative issues of reform and how to function inside of it, researchers and other technical players may discover their findings to bolster change might not be implemented adequately. Likewise, without an appreciation about the need for specialized or technical analysis to affirm decision making as opposed to an unpredictable utilization of political methodologies, political players may find that even with the best of motives, the desired policy targets may not be accomplished. On the perception households in held about the national health insurance scheme, it has been proven that perceptions related to plan components have the most grounded relationship with retention and voluntary enrollment choices in the National Health Insurance Scheme (NHIS). Particularly these identify with advantages, cost and convenience of NHIS. In the meantime, while household had positive perceptions with respect to the technical nature of care, advantages of NHIS, had adequate community health convictions and ease of access to NHIS management, perceptions were negative about the cost of NHIS, insurer's dispositions and peer pressure. Perception levels among

the uninsured were greatly negative than the insured concerning issues such as advantages, convenience and cost of NHIS. Perceptions linked with providers, plans and community traits assume an essential part, yet to a differing degree in household choices to voluntarily enlist and remain enlisted in insurance plans. Plan components are of key significance. Policy drafters need to identify household perceptions as potential hindrances or empowering agents to enlistment and put resources into comprehending them in their design of policies to encourage enlistment (Jehu-Appiah et al., 2012).

With regards to financing sources, (Witter, S. and Garshong, B., 2009) carried out a preliminary assessment of the NHIS up till that point in time. They observed that, the NHIS is intensely dependent on tax financing for 70-75% of its income. This has allowed speedy extension of scope, partially through the incorporation of huge exempted population subgroups. Card holders expanded from 7% of the populace in 2005 to 45% in 2008. On the other hand, just around a third of these are contributing to the plan monetarily. This brings to the fore a sustainability challenge, in that income is decoupled from the expanding enrollment numbers. Furthermore, the NHIS provides a wide range of benefits package, with no co-payments and constrained gate-keeping. The scheme also confronts cost acceleration identified with its new payment structure and the increasing access of individuals. These elements added to an increase in troubled plans and an inability to honor outstanding office claims in 2008. The NHIS has had an extensive effect on the health care all in all, assuming an increasing role in financing curative care. In 2009, it is relied upon to contribute 41% of the general resources envelope. However there is confirmation that this financing is not extra but rather has been taken from other financing channels. There are some legitimate worries about this, as the new financing source (a VAT-based duty) may be more backward. Moreover, enrollment of the NHIS at present has a skewness towards the rich, and also, a pro-urban predisposition in connection to renewals. Just a small fraction are enlisted as impoverished, and there is evi-

dence of confirmation of 'squeezing out' of non plan participants from health care usage. At last, significant hindrances remain in connection to reinforcing the purchasing aspect of the NHIS, additionally settling level headed discussions about its accountability and structure. In concluding the assessment, Witter, S. and Garshong, B. (2009), remarked that certain trade-offs will be essential between the current wide range of benefits package of the NHIS and the excellent desire to achieve an all-inclusive scope. The general resource envelope for health care is prone to be steady as opposed to expanding over the medium-term. In the more drawn out term, the investments costs in the NHIS may be defended on the off chance that it has the capacity to enhance the expense viability of purchasing and the plan's responsiveness in general.

A principal component of social health insurance, is the financial protection it offers members of the scheme, most especially, the poor and the vulnerable. A study to access this impact on the Ghanaian populace was carried out by Nguyen et al. (2011) in Nkoranza and Offinso as case studies in 2007,two years after the start of Ghana's National Health Insurance Scheme. It was observed that during the period of the study, insurance penetration was 35 percent. Despite the fact that the benefit package of insurance is liberal, the insured individuals still suffer out-of-pocket payment for services from other informal alternate sources and for medications and tests at healthcare providers which are not covered by the scheme. In any case, they paid amounts which were substantially less than that which was paid by the uninsured. Insurance has been proven to have a protective impact against the financial weight placed on healthcare, decreasing substantially the probability of suffering calamitous claim payments. The impact is especially astounding among the poorest quintile of the sample. In conclusion, Nguyen et al (2011) concluded that results emanating from the research confirmed the positive money protection impact of health insurance coverage in Ghana. The impact is more grounded among the poor community than among the general

populace. The outcomes are empowering some low wage nations who are considering a similar strategy to extend SHI. Ghana's experience likewise demonstrates that initiating insurance in itself is not sufficient to eradicate completely the out-of-pocket system of payment for medical services. Further research are required to address the quality of care and supply side's incentives, so that the insured can appreciate the full advantages of insurance (Nguyen et al., 2011).
Ultimately, the greatest stumbling block being kicked out of whenever a social health insurance is put in place is, barrier to utilization of health care. Blanchet et al. (2012) found that after implementation of the scheme,by and large people enlisted in the insurance plan are fundamentally more inclined to receive prescriptions, visit health facilities and look for orthodox medical services when ill and therefore recommended that the government's target to expand access to the formal healthcare through medical insurance has on the minimal level been partially accomplished.

With the aim to examine the relation between health insurance registration to the Ghana National Health Insurance Scheme (NHIS) and socio-economic status (SES) of inhabitants of the Asante Akim North district of the Ashanti Region, Ghana, Sarpong et al (2010) utilized information on asset variables such as lodging conditions, electricity and other variables, and on NHIS registration obtained from households in 99 villages during the span of the community survey. Principal components analysis was deployed in the survey. Households forming part of the survey were assembled into three groups according to their SES (20% high, 40% low and 40% middle). Odds ratios of NHIS registration were estimated for all SES categories, with the low category used as the benchmark group. It was brought to the fore that of the 7223 households involved in the survey, 38% registered with the NHIS, out of this, 43% middle, 21% were low and 60% high SES households. SES correlation to the NHIS registration (middle SES: OR 2.5, 95% CI 2.2-2.9; high SES: OR 4.9, 95% CI 4.3-5.7; low SES: OR 1, reference

group) was recognized to be significant. Hence the conclusion was drawn that after four years of the health insurance scheme's inception, it (the scheme) has attained registration levels of 38% within the study area. To reach this target of granting universal health care access to health facilities for the entire citizenry, especially for people who fall below average in socio-economic constraints, increasing subscription levels is a necessity (Sarpong N. et. al, 2010). Also, a study to examine the Scheme's impact on access to and usage of insurance provisions and services in the Akatsi District of the Volta region of Ghana. Both quantitative and qualitative information was gathered through vis-Ã -vis meeting with 320 people and three service suppliers utilizing organized surveys. The outcome demonstrate that level of education, age and occupation are real determinants of participation in the plan. The plan has a significant outcome on health seeking conduct and usage of health insurance provisions by eradicating the crucial financial obstructions to seeking healthcare. Absence of medical insurance serves as a noteworthy hindrance to access to advanced medical services. Expanding scope and enrollment combined with change in geographical access will enhance better and general healthcare results for the general population of Ghana (Goba and Liang, 2011).

2.2 Challenges of the Scheme

In the view of Millennium Development Goal aims for poverty reduction and health gains, there is a developing driving force towards giving all inclusive scope of medical care (World Health Organization Group, 2006), implying that the greater part of the populace has access to proper medical services when required, and at a reasonable expense. One critical move to improve affordability is to lessen the out-of-pocket payment which clients make for medical services. These are broadly perceived as a hindrance to access, particularly in developing nations, and as pushing families further into impoverishment (Xu et al., 2003).

SHI is viewed as one of the financing methodologies of healthcare with a concrete potential to spread the risk crosswise over the entire populace and time (Witter and Garshong, 2009). For this reason,it was prudent to establish the NHIS in Ghana was instituted by the National Health Insurance Act,2003 (Act 650) and National Health Insurance Regulations, 2004 (L.I. 1809) with the perspective of enhancing monetary access of Ghanaians, particularly the vulnerable and poor people, to quality essential medical insurance services and to restrict out-of-pocket payments at the point of conveying the service (Goba and Liang, 2011). Many low-and middle-wage nations depend vigorously on patients' out-of-pocket healthcare service payments to fund their medical services structure (Xu et al., 2007). According to the World Health Organisation (WHO), empirical evidence shows that out-of-pocket medical payment is the least adequate and most inequittable method for funding health insurance and keeps individuals from looking for medical care and may compound destitution (World Health Organization, 2000); (Xu et al., 2003). Health insurance plans are progressively perceived as a device to back medical services procurement in developing nations and can possibly expand use and better secure individuals against calamitous medical costs and tackle issues of equity (World Health Organization, 2000). The main features of the NHIS was developed as a required medical insurance structure, having a risk pool which draws from across regional schemes, financed from participants' contribution and a levy on the Value-Added Tax (VAT) charged on services and products, from which a wide package of benefits could be subsidized. In Ghana, the NHIA is overwhelmingly funded by taxation, which accounts for 70-75% of aggregate income, with an additional 20-25% originating from the formal sector contributions and just about 5% from the informal sector premia (as indicated by NHIA reports). This makes it less desired as the conventional subsidizing instruments (government spending plans, donor financing and client charges), at least as far as income accruing is concerned. This scenario may be exacerbated if, as guaranteed amid its election manifesto, the present government switches

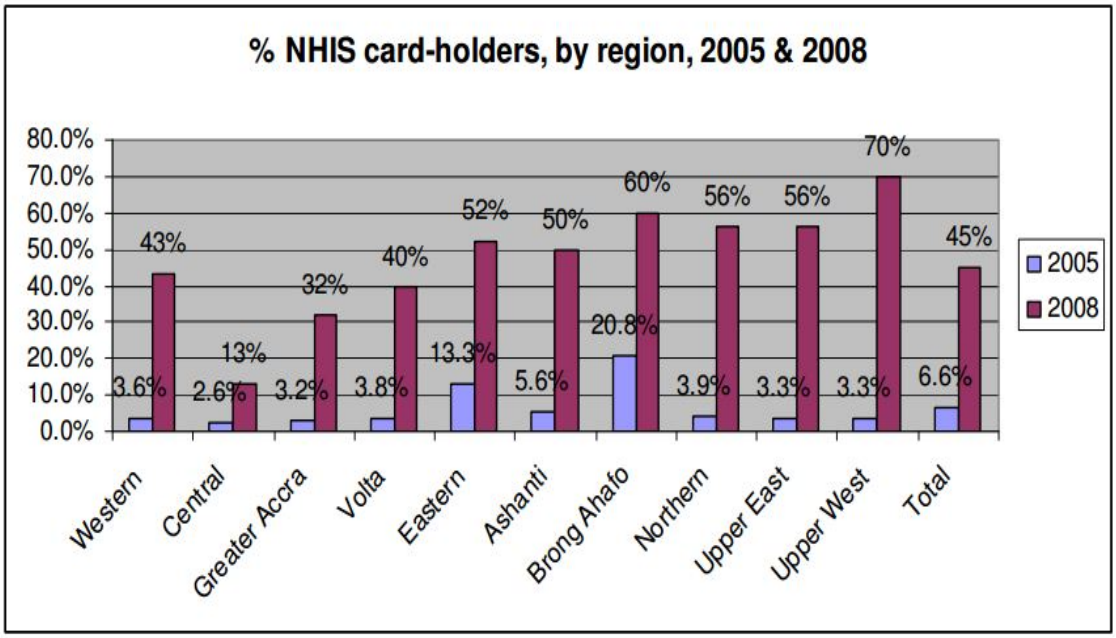
to a 'one-time premium' providing lifelong enrollment (apparently just for the informal sector). This will further disintegrate the thought that the NHIS is a contribution tied insurance framework.

In Figure 2.1 is displayed a brief pictorial summary of clients of the national insurer, (Witter and Garshong, 2009)

However, regardless of the profound benefits of the scheme, the National insurer is riddled with a lot of challenges of the most paramount is claims payments. This sterning revelation carried out in a news publication in the electronic media by Daily Guide on March 6, 2015 with the caption "NHIA IS BROKE". In this publication among the many others cited the NHIA boss, Mr. Slyvester Mensah admitting to this fact. An implication of this nonpayment of claims by the insurer is,health centers nationwide are said to be in a condition of bankruptcy as an aftereffect of obligations owed them by the Authority. This led the Ghana Medical Association (GMA) hinting the much feared cash-and-carry directive was reintroduced and NHIS card bearing patients were being dismissed (DailyGuide, 2015)

2.2.1 Actuarial Concern

Actuarial assessment is, by its inclination, a science by which uncertainty is dependably an element. Without uncertainty there is need for an actuary. Actuarial investigation is, in any case, taking into account thoroughly, exploratory routines and procedures. An essential objective, as with all science, is to give the best conceivable comprehension of reality, notwithstanding those uncertainties. Actuarial science is a branch of applied science. Consequently, the profession ought to be bothered with communicating to both within the profession and to an outside gathering of people with shifted and somewhat clashing ideologies. It is frequently important to gauge likelihood distribution to depict the loss procedures catered by the insurance contracts (Patrik, 1981). From this, the actuarial concern in the solving the challenges faced by the NHIA will be in estimating a distribution for



	2005		2008	
Membership categories	**Number of registrants**	**Proportion of total population**	**Number of registrants**	**Proportion of total population**
Formal sector	468,092	2.24%	811,567	3%
Informal sector	615,450	2.94%	3,727,454	16%
Paying members	1,083,542	5.18%	4,539,021	19.25%
Pensioners	43,208	0.21%	71,147	0.30%
Children	1,751,175	8.37%	6,305,727	27%
70+	266,421	1.27%	816,956	4%
Indigent	790,078	3.77%	302,979	1%
Pregnant women			432,728	2%
Overall exempt	2,850,882	13.62%	7,929,537	34%
Total	3,934,424	18.79%	12,468,558	54%
% of registrants paying	28%		36%	

Figure 2.1: Summary Statistics of Health Insurance subscribers in Ghana

the claims submitted to it by the health care providers. Traditionally, it will be expected that this estimation should be based on total claims submitted, however, this study will seek to develop a distribution compromised of number of claims submitted, type of claim submitted and the money amount of claim submitted. to do this, number of claims will be modeled with a negative binomial distribution, conditional probability theory used to estimate type of claims submitted and finally the severity (loss)is estimated with generalized pareto distribution.

2.3 Actuarial techniques and applications

Hierarchical models

There has been considerable interest in statistical demonstration of claims recurrence, Boucher and Denuit (2006) is an illustration. However, the literature on modeling claims amounts, particularly in conjunction with claims recurrence, is less extensive. One conceivable clarification, noted by Coutts, S.M (1984), is that the majority of the variability in general experienced may be ascribed to claim recurrence (at least when inflation was minimal) (Boucher and Denuit, 2006). Coutts, S.M, (1984) also remarks that the first paper to analyze claim frequency and severity seems to be Kahane, Y. and Levy, H. (1975), these facts are traced from Boucher and Denuit (2006). Probability models based on the hierarchical approach are mostly applied to treelike structured data and with Bayesian theory. A paramount feature of such hierarchical models is the incorporation of probability law at certain stages in the classification which is actually conditional on the results at previous stages (Dutang et al., 2008b).

Frees et al. (2009) incorporated hierarchical modeling to micro-level records of an insurance company's database from 1993-2001. The claims submitted comprised of detailed information concerning the claim type, such details included whether nature of claim was as a result of harm or damage to a third party property

or whether the claims were due to damage of the insured, corresponding claim amount relating to these information were recorded accordingly. A hierarchical model consisting of three components,namely frequency, claim type and claims severity was adopted. The frequency component was estimated with the aid of a negative binomial regression. Other variables of interest in the study were driver's gender (driver due to the fact that the database was non-life insurance-Vehicular to be precise), age, and no discount- no claim;being enjoyed by the insured as well as vehicle age and type. These variables were deemed to be relevant for predicting the event of having claim. The Claim type was developed with a multinomial logit model, this method was significant regardless of the claim being a injury to third party, property damage to third party or an insureds own damage or a combination of the aforementioned. A revelation from the method was that year, vehicle age and type, were significant predictors for this establishing a claim type. In estimating the severity component, generalized beta of the second kind was utilized for the various claim amounts. The above mentioned were put together to build the hierarchical model which proves adequate for assessing the importance of a including all available information in the data analysis. Thus the combined model enables the actuary in predicting automobile claims more appropriately and efficiently than more established traditional methods. As an application, Frees et al. (2009) demonstrated the importance of the hierarchical model by developing predictive distributions and estimating premiums under available reinsurance coverages.

The hierarchical model can also be used together with other mathematical tools to perform varied estimations. Scheel et al. (2013) assessed the impact of climate change on the industry using a Bayesian hierarchical statistical method to expatiate and forecast insurance losses because of climatic events on a local geographic scale. The quantity of climate-related insurance cases is demonstrated by combining linear models with spatially smoothed variable selection. Utilizing

Gibbs sampling and reversible jump Markov chain Monte Carlo procedures, the model is fitted on day by day climate and insurance information from each of the 319 districts which constitute central and southern Norway for the period 1997-2006. Exact out-of-sample forecasts accept the model. Scheel et al. (2013) results bring to fore intriguing provincial trends in the impact of distinctive climatic covariates. Notwithstanding being valuable for insurance pricing, the model can be utilized for transient forecasts in light of climate gauges and for long haul forecasts in view of downscaled climate models.

Also Yu (2015) proposed a statistical model for health insurance total claim amounts classified by age group, region of residence and time horizon of the insured population under Bayesian framework. The model can be used to predict future total claim amounts and thus to facilitate premium determination. The prediction is based on the past observed information and prior beliefs about the insured population, number of claims and amount of claims. The insured population growth is modeled by a generalized exponential growth model, which takes into account the random effects in age, region and time classifications. The number of claims for each classified group is assumed Poisson distributed and independent of the size of the individual claims. A simulation study was conducted to test the effectiveness of modeling and estimation, and Markov chain Monte Carlo (MCMC) used for parameter estimation. Based on the predicted values, the premiums are estimated using four premium principles and two risk measures.

Armed with the R statistical software, Dutang et al. (2008a), defined a hierarchical model as one which satisfies meet the listed criteria below:

- Unsophisticated and easily understood from the mathematical theory of the model to the R derivation and vice versa
- Unlimited to any number of stages and nodes;

A hierarchical model is fully defined by the quantity of nodes at each stage ($I, J_1, \ldots, J_I$ and $n_{11}, \ldots, n_{IJ}$, above) and by the likelihood laws at each stage. An example of a hierarchical model is given by:

$X_t \mid \wedge\Theta \sim Poisson(\wedge)$

$\wedge \mid \Theta \sim Gamma(3, \Theta)$

$\Theta \sim Gamma(2, 2)$

According to Guszcza (2010), hierarchical modeling offers a "third way" modeling grouped data. In this model parameters reflecting group enrollment enter one's model through properly determined likelihood sub-models. An essential special instance of hierarchical models includes different perceptions through time of each unit. An earlier overview of how statistical modeling of claims and severity can be helpful for pricing automobile coverage was discussed by Brockman and Wright (1992). An integral part of hierarchical modeling is statistical software, as frees et al, current computing hardware, researchers can promptly get access to information at the individual policyholder level that is term "micro-level". This essentially because actuaries use statistical models to abridge smaller micro-level information that subsequently should be translated appropriately for monetary decision-making. (Frees et al., 2009) The actuar project Dutang et al. (2008a) is a package of Actuarial Science function for the R statistical. Albeit different packages on CRAN which offer functions that may be useful to statisticians, actuar expects to serve as a focal area for all the more particularly actuarial use and information sets. The task was formally outdoored in 2005 and is under dynamic advancement. The variant of actuar accessible on CRAN is 0.9-3. The list of capabilities of the package could be categorized into three primary classes: loss appropriations modeling, risk hypothesis and credibility hypothesis. This and many more statistical programs out there help the actuary to confidently analyze and develop appropriate estimates for given loss datasets.

Chapter 3

Methodology

3.1 Introduction

In this chapter, the various components of the hierarchical statistical model is outlined and discussed. Attention will be on the model methods considered in this study and their properties. It explains in details the steps that were utilized in the modeling process which includes the data processing and data analysis that were used.

3.2 Negative Binomial Regression

The frequency component of the statistical method employed is estimated with Negative Binomial Regression. The negative Binomial model is written by Zwilling (2013) as:

$$\ln \mu = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{3.1}$$

with the indicator variables $x_1, x_2 ... x_p$ provided, and the population of regression coefficients $\beta_0, \beta_1, \beta_2 ... \beta_p$ are estimated. Negative binomial regression is a kind of generalized linear model in which the dependent variable is the count of times an occasion happens. A helpful parametrization of the negative binomial distribution is formulated as Hilbe (2011):

$$p(y) = P(Y = y) = \frac{\Gamma(y + 1/\alpha)}{\Gamma(y + 1)\Gamma(1/\alpha)} \left(\frac{1}{1 + \alpha\mu}\right)^{1/\alpha} \left(\frac{\alpha\mu}{1 + \alpha\mu}\right)^{y} \tag{3.2}$$

where $\mu > 0$ represents the mean of Y and $\alpha > 0$ is the heterogeneity parameter. Hilbe (2011) in the above equation derives this parameterization as a Poisson-

gamma mixture, then again on the other hand as the quantity of failures before the $(1/\alpha)^{th}$ success, however one does not require $1/\alpha$ to be an integer. Given a random sample made up of n elements, for element i the dependent variable y_i and the predictor variables$x_{1i}, x_{2i}...x_{pi}$. Employing matrix and vector notation, denote $\beta = (\beta_0, \beta_1, \beta_2...\beta_p)^T$, and assemble the explanatory variable data into the matrix X as follows:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\ ; & : & : & ... & : \\ . & . & . & ... & . \\ 1 & x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix} \tag{3.3}$$

Assigning the i^{th} row of X to be x_i and exponentiating (3.1) then write the distribution (3.2)

$$p(y_i) = \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i + 1)\Gamma(1/\alpha)} \left(\frac{1}{1 + \alpha e^{x_i\beta}}\right)^{1/\alpha} \left(\frac{\alpha e^{x_i\beta}}{1 + \alpha e^{x_i\beta}}\right)^{y_i} \tag{3.4}$$

Then estimate α and β using maximum likelihood estimation. The likelihood function is given by

$$l(\alpha, \beta) = \prod_{i=1}^{n} p(y_i) = \prod_{i=1}^{n} \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i + 1)\Gamma(1/\alpha)} \left(\frac{1}{1 + \alpha e^{x_i\beta}}\right)^{1/\alpha} \left(\frac{\alpha e^{x_i\beta}}{1 + \alpha e^{x_i\beta}}\right)^{y_i} \tag{3.5}$$

and the log-likelihood function is

$$\begin{aligned} lnL(\alpha, \beta) = \sum_{i=1}^{n} (y_i ln\alpha + y_i(x_i\beta) - (y_i + \frac{1}{\alpha}) ln(1 + \alpha e^{x_i\beta}) + ln\Gamma\left(y_i + \frac{1}{\alpha}\right) \\ - ln\Gamma\left(y_i + 1\right) - ln\Gamma\left(\frac{1}{\alpha}\right)) \end{aligned} \tag{3.6}$$

The values of α and β that maximize $lnL(\alpha, \beta)$ will be the maximum likelihood estimates sought after, and the estimate variance-covariance matrix of the estimators is $\sum = -H^{-1}$ where H is the Hessian matrix of second derivatives of the log-likelihood function. At that point, the variance-covariance matrix can be utilized to locate the standard Wald confidence interval and $p - values$ of the coefficient estimates.

3.3 Negative Binomial Distribution

For a given sequence of Bernoulli trials which are independent of each other, with every trial having outcomes of a "success" or "failure". The chances or likelihood of success is represented by p and failure $(1-p)$. The sequence is allowed to continue till a number r denoting failures occur. r is predetermined before the start of the experiment. Thus, the random number of successes observed denoted by X, is said to have negative binomial distribution. An example is the probability of having to fail a professional exam three times before passing it on the fourth try. Rationale for the choice of negative binomial distribution is the correction for over-dispersion in the other count distribution (Poisson) which was initially employed in the analysis used in this research. The negative binomial distribution under probability theory and in statistics theory, is classified as a discrete likelihood distribution based on the quantity of success in a sequence of independent and identically distributed (iid) Bernoulli trials before a predefined (non-random) count of failures (denoted r) occurs. $X \sim$ NB(r; p) The negative binomial distribution is mathematically represented by mass function as:

$$\Pr(X = k) = \binom{k+r-1}{k} p^k (1-p)^r \quad \text{for } k = 0, 1, 2, \ldots \tag{3.7}$$

The parameters in the brackets depicts a binomial coefficient, and is represented by:

$$\binom{k+r-1}{k} = \frac{(k+r-1)!}{k!\,(r-1)!} = \frac{(k+r-1)(k+r-2)\cdots(r)}{k!}. \tag{3.8}$$

This can then again be represented in the form below when the term "negative binomial" is taken into account,

$$\frac{(k+r-1)\cdots(r)}{k!} = (-1)^k \frac{(-r)(-r-1)(-r-2)\cdots(-r-k+1)}{k!} = (-1)^k \binom{-r}{k}. \tag{3.9}$$

To better appreciate the equation above, consider the chances that for every given sequence of r failures and k successes denoted by $(1-p)^r p_k$, since all results of the $k+r$ trials are independent,that is, their occurrence is not contingent on any other. This reasoning is attributable to the fact that the rth failure is the last to come, and as it stands the k trials containing the successes is from the remaining $k+r-1$ trials. Due to the combinatorial interpretation for the above binomial coefficient, offers exactly the count of all these sequences of length $k+r-1$.

Recurrence relation

$$\{(k+1)\Pr(k+1) - p\Pr(k)(k+r) = 0, \Pr(0) = (1-p)^r\} \tag{3.10}$$

The Negative Binomial Expectation

The negative binomial distribution with parameters (r, p) has its average number of trials $k+r$ as:

$$\frac{pr}{1-p} \tag{3.11}$$

Extension to real-valued r

In this subsection of the discussion, consider r as a real, positive number. Using a multiplicative formula, the binomial coefficient is then defined and rewritten with the gamma function as:

$$\binom{k+r-1}{k} = \frac{(k+r-1)(k+r-2)\cdots(r)}{k!} = \frac{\Gamma(k+r)}{k!\,\Gamma(r)}. \tag{3.12}$$

By the binomial series and (3.8) above, for every $0 \leq p < 1$

$$(1-p)^{-r} = \sum_{k=0}^{\infty} \binom{-r}{k} (-p)^k = \sum_{k=0}^{\infty} \binom{k+r-1}{k} p^k, \tag{3.13}$$

and thus, the various components of the probability mass function actually amounts to one.

Other formulations

- X can be defined as the aggregate count of trials required to obtain r failures, and not just the count of successes. This is so because the aggregate count of trials is equivalent to the count of successes with the count of failures added, this definition varies from the initial definition used in this discuss, by including a constant r. To convert formulations having this new definition into the definition utilized in this study, one can replace every "k" with "$k-r$" wherever it occurs in the material, and further reduce the median, mean, and mode by r. Likewise to reformulate the formulas in this research to this new definition, one should supplant "k" with "$k+r$" and r added to the median, mean and mode. Doing these changes will essentially imply using a probability mass function of the kind below:

$$f(k;r,p) \equiv \Pr(X=k) = \binom{k-1}{k-r}(1-p)^r p^{k-r} \quad \text{for } k = r, r+1, r+2, \ldots, \tag{3.14}$$

  this distribution probably mimics the binomial distribution more closely than the definition used above. A point worth noting is, the parameters forming the binomial coefficient are decremented with respect to order,that is: the last "failure" occurs last, and thus the other events are one position short when potential orderings are being counted. Again this description of negative binomial distribution, however doesn't readily approach a positive real parameter r.

- With regards to p denoting the chances of failure and not of a success, is another alternate definition. To convert formulations between this definition and that already established, as in the initial definition this study, one ought to replace "p" with "$1-p$" everywhere it appears in the text.

- Another definition is where the support X refers to the count of failures,instead of the count of successes. In the definition in which X numbers

failures and p being the chances of success- has just the same formulations as in the situation where X denotes successes and p represents the probability of failure. That notwithstanding, the contributory text still contain the wording "failure" and "success" interchanged when compared with the previous case.

- The two definitions discussed previously can be employed simultaneously, that is, p depicts the probability of failure and X depicts or numbers total trials.

- With regards to negative binomial regression, the mean, m of the distribution is specified, it is then related to explanatory variables just like that of linear regression or any other generalized linear models Hilbe (2011). Thus, the likelihood mass function then is formulated as

$$\Pr(X = k) = \left(\frac{r}{r+m}\right)^r \frac{\Gamma(r+k)}{k!\,\Gamma(r)} \left(\frac{m}{r+m}\right)^k \quad \text{for } k = 0, 1, 2, \ldots. \quad (3.15)$$

$m + \frac{m^2}{r}$ represents the variance, the parameter r denotes the "shape parameter", "dispersion parameter" or "clustering coefficient" Lloyd-Smith (2007) or "heterogeneity" Hilbe (2011) or "aggregation" parameter Crawley (2012). The "aggregation" term is mostly employed in ecology when tallies of individual species are being described. When the aggregation parameter r decreases towards zero results in an increase in organisms aggregation; when r approaches infinity there is a corresponding absence of aggregation, this could be demonstrated by Poisson regression. The reciprocal of r, in some applications of the negative binomial regression is referred to as the "dispersion parameter"

- Sometimes the distribution is parameterized in terms of its mean μ and variance σ^2. In that case,

$$p = \frac{\sigma^2 - \mu}{\sigma^2} r = \frac{\mu^2}{\sigma^2 - \mu} \, and \Pr(X = k) = \binom{k + \frac{\mu^2}{\sigma^2 - \mu} - 1}{k} \left(\frac{\mu}{\sigma^2}\right)^{\left(\frac{\mu^2}{\sigma^2 - \mu}\right)} \left(\frac{\sigma^2 - \mu}{\sigma^2}\right)^k . \tag{3.16}$$

3.3.1 Occurrence

Waiting time in a Bernoulli process

Given a situation in which r is an integer, then the negative binomial distribution can be referred to as the Pascal distribution. The Pascal Distribution is a likelihood distribution which consists of a series with a certain count of failures and successes which are iid Bernoulli trials. The Bernoulli distribution has $k + r$ trials, probability of success denoted by p, the negative binomial has the likelihood of failures as r and k successes, with the last trial ending in a failure. Stated differently, the negative binomial distribution describes the likelihood distribution with the count of successes occurring before the rth failure in a Bernoulli sequence of events, with a likelihood p of successes on each trial. A Bernoulli process by nature being a discrete time process, and thus having integers for the count of trials, failures, and successes are integers.

As an Alternative for overdispersed Poisson

The negative binomial distribution, given its optional reformulation discussed early on, can be a substitute to the Poisson distribution. This use of the Negative Binomial distribution is most useful in discrete data analysis whose sample variance exceeds the sample mean over an unbounded positive range. In such scenarios, the observed outcomes are said to be overdispersed in regards to the Poisson distribution whose mean is commensurate to its variance. Therefore such a Poisson distribution is considered as an inappropriate model. However,the negative binomial distribution possess an extra parameter more than the Poisson, this parameter can act as a means of adjusting the variance independent of the mean.

Relation to other probability distributions

- A geometric distribution with sequence (0, 1, 2, 3, . . .) is deemed an exceptional instance of the negative binomial distribution, $\text{Geom}(p) = \text{NB}(1,\, 1-p)$.
- Also as a special case, is the discrete phase-type distribution.
- Discrete Compound Poisson distribution also has the negative binomial distribution as an exceptional case.

Relation to Poisson distribution

Given a negative binomial distributions sequence in which the terminating parameter r approaches infinity, where p denotes the success probability within atrial, approaches zero in a manner so that the distribution's mean remains constant. Representing the mean with λ, then p is given by $p = \lambda/(\lambda + r)$
$\lambda = r\,\frac{p}{1-p} \quad \Rightarrow \quad p = \frac{\lambda}{r+\lambda}$. Under this formulation the probability function is represented by
$f(k; r, p) = \frac{\Gamma(k+r)}{k!\cdot\Gamma(r)}p^k(1-p)^r = \frac{\lambda^k}{k!}\cdot\frac{\Gamma(r+k)}{\Gamma(r)\,(r+\lambda)^k}\cdot\frac{1}{\left(1+\frac{\lambda}{r}\right)^r}$
In the case where the limit $r \longrightarrow \infty$, the second parameter converges to one, and the third converges to the exponent function:
$\lim_{r\to\infty} f(k; r, p) = \frac{\lambda^k}{k!}\cdot 1\cdot\frac{1}{e^\lambda}$, which is the mass function of a Poisson-distributed random variable with expected value λ. Stated differently, the optional parameterized negative binomial distribution approaches the Poisson distribution and r controls the deviation from the Poisson. This makes the negative binomial distribution suitable as a strong option for the Poisson, which converges to the Poisson for large r, however which has bigger fluctuation than the Poisson for less large r. $\text{Poisson}(\lambda) = \lim_{r\to\infty}\text{NB}\left(r,\ \frac{\lambda}{\lambda+r}\right)$.

Relation to Gamma-Poisson mixture

The negative binomial distribution also is a compound probability distribution formulated as a product of a continuous mixture of Poisson distributions where the combination of the Poisson distribution rate is a gamma distribution. That is, the negative binomial can be viewed as a Poisson(λ) distribution, in which λ is considered a random variable,with a gamma distribution with parameters defined by; scale $\sigma = p/(1-p)$ or correspondingly a rate $\beta = (1-p)/p$ and shape, r. Formally stated, this interprets as the negative binomial distribution's probability mass function is written as

$$f(k;r,p) = \int_0^\infty f_{\text{Poisson}(\lambda)}(k) \cdot f_{\text{Gamma}\left(r,\,\frac{1-p}{p}\right)}(\lambda)\,\mathrm{d}\lambda \tag{3.17}$$

$$= \int_0^\infty \frac{\lambda^k}{k!} e^{-\lambda} \cdot \lambda^{r-1} \frac{e^{-\lambda(1-p)/p}}{\left(\frac{p}{1-p}\right)^r \Gamma(r)}\,\mathrm{d}\lambda \tag{3.18}$$

$$= \frac{(1-p)^r p^{-r}}{k!\,\Gamma(r)} \int_0^\infty \lambda^{r+k-1} e^{-\lambda/p}\,\mathrm{d}\lambda \tag{3.19}$$

$$= \frac{(1-p)^r p^{-r}}{k!\,\Gamma(r)}\; p^{r+k}\,\Gamma(r+k) \tag{3.20}$$

$$= \frac{\Gamma(r+k)}{k!\,\Gamma(r)}\; p^k (1-p)^r. \tag{3.21}$$

Hence the negative binomial distribution can be termed as a Gamma-Poisson (mixture) distribution.

Relation to a Geometric distribution sum

Given Y_r, a random variable which follows the negative binomial distribution, having parameters p and r, and domain 0, 1, 2,..., then Y_r is said to be a summation of r non-dependent variables which follows the geometric distribution (on 0, 1, 2,...) with parameter $1-p$. Due to the central limit theorem, Y_r (shifted and scaled) for large r, is approximately normal. Hence for a random variable B_{s+r}, which is binomial distributed with parameters $1-p$ and $s+r$, then

$$\Pr(Y_r \leq s) = 1 - I_p(s+1, r) \tag{3.22}$$

$$= 1 - I_p((s+r) - (r-1), (r-1) + 1) \tag{3.23}$$

$$= 1 - \Pr(B_{s+r} \leq r - 1) \tag{3.24}$$

$$= \Pr(B_{s+r} \geq r) \tag{3.25}$$

$$= \Pr(\text{after } s + r \text{ trials, there are in the least event} r \text{ successes}). \tag{3.26}$$

With this case, the negative binomial distribution acts as an "inverse" of the binomial distribution. When independent and identical negative-binomially distributed random va,,riables r_1 and r_2 having their parameter p to be of equal value is summed, the result also has a negative-binomial distribution with the same p however, the "r-value" of the result is a sum of the r-values of the initial random variables, ie., $r_1 + r_2$. With regards to divisibility, the negative binomial distribution is infinitely divisible, that is, given Y has a negative binomially distributed, for a positive integer n, there exist $Y_1, \ldots, Y_n$ which are independent and identically distributed random variables with their sum having the same distribution as that of Y.

3.3.2 Formulating as a compound Poisson distribution

A negative binomial distribution NB(r, p), could be presented inn the form of a compound Poisson distribution:$Y_n, n \in N_0$ represents an iid random variables, where each variable depicts a logarithmic distribution Log(p), having the accompanying probability mass function:
$f(k; r, p) = \frac{-p^k}{k \ln(1-p)}, \qquad k \in \mathbb{N}$. Let N be a random variable, which is not dependent on the sequence, and assume that N is distributed by a Poisson distribution having mean $\lambda = -rln(1-p)$. Hence NB(r, p) represents the distribution of the random sum given by: $X = \sum_{n=1}^{N} Y_n$. To ascertain this, the probability

generating function G_X of X is derived, which makes up the constituents of the probability generating functions G_N and G_{Y1}. Using

$G_N(z) = \exp(\lambda(z-1)), \qquad z \in \mathbb{R}$, and

$$G_{Y_1}(z) = \frac{\ln(1-pz)}{\ln(1-p)}, \qquad |z| < \frac{1}{p}, \tag{3.27}$$

Resulting in

$$\begin{aligned} G_X(z) &= G_N(G_{Y_1}(z)) && (3.28)\\ &= \exp\left(\lambda\left(\frac{\ln(1-pz)}{\ln(1-p)} - 1\right)\right) && (3.29)\\ &= \exp\big(-r(\ln(1-pz) - \ln(1-p))\big) && (3.30)\\ &= \left(\frac{1-p}{1-pz}\right)^r, \qquad |z| < \frac{1}{p}, && (3.31) \end{aligned}$$

this, which serves as the probability generating function of the NB (r, p) distribution.

3.3.3 Features

Cumulative distribution function

The cumulative function of the Neg. Binom. is expressed in the regularized form as the incomplete beta function.

$$f(k; r, p) \equiv \Pr(X \leq k) = 1 - I_p(k+1, r) = I_{1-p}(r, k+1). \tag{3.32}$$

Sampling and point estimation of p

Assume p is not unknown and a trial is carried out in which it is chosen at the onset that testing will proceed until r successes are obtained. An adequate statistic for the trial is k, the count of failures. In evaluating p, the minimum variance which is not a biased estimator is

$$\hat{p} = \frac{r-1}{r+k-1}. \tag{3.33}$$

The maximum likelihood estimate of p is

$$\tilde{p} = \frac{r}{r+k}, \tag{3.34}$$

however, this estimate is a one-sided or stated differently, biased one. The inverse $(r+k)/r$, is the unbiased estimate of $1/p$, however Haldane (1945).

Linkage between Binomial theorem & the Negative Binomial

Consider a random variable Y, which is binomially distributed with parameters p and n. Assuming $p+q=1$, having $q, p \geq 0$. Implies that the binomial theorem is

$$1 = 1^n = (p+q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}. \tag{3.35}$$

Employing the Newton's binomial theorem, the above equation is equally written as:

$$(p+q)^n = \sum_{k=0}^{\infty} \binom{n}{k} p^k q^{n-k}, \tag{3.36}$$

with the upper bound of the summation being infinite. Thus, the binomial coefficient

$$\binom{n}{k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}. \tag{3.37}$$

instead of just a positive integer, the binomial coefficient is defined when n is a real number. However in this case of the binomial distribution, has a value of zero if $k > n$. Assume $r > 0$ and utilize a negative exponent:

$$1 = p^r \cdot p^{-r} = p^r (1-q)^{-r} = p^r \sum_{k=0}^{\infty} \binom{-r}{k} (-q)^k. \tag{3.38}$$

The terms are positive at that point, and the term
$p^r \binom{-r}{k} (-q)^k$ is only the likelihood that the count of failures before the rth success

is equivalent to k, if r is an integer. (Should it be the case that r is a negative non-integer, so that the exponent is a positive non-integer, then some of the terms in the summation above are negative, hence not having a likelihood distribution on the set for all non-negative integers.) Additionally, it is permissible for non-integer estimations of r. At that point one has a legitimate negative binomial distribution, which generally is a Pascal distribution, which harmonizes with the Pascal distribution when r happens to be a positive integer.

3.3.4 Estimation of Parameters

The maximum likelihood estimation approach

Given iid observations $(k_1, \ldots, k_N)$, the likelihood function is given by

$$L(r, p) = \prod_{i=1}^{N} f(k_i; r, p) \tag{3.39}$$

the log-likelihood function is then calculated as

$$\ell(r, p) = \sum_{i=1}^{N} \ln\left(\Gamma(k_i + r)\right) - \sum_{i=1}^{N} \ln(k_i!) - N \ln\left(\Gamma(r)\right) + \sum_{i=1}^{N} k_i \ln(p) + Nr \ln(1-p). \tag{3.40}$$

To obtain the maximum, partial derivatives in terms of p and r are taken and equated to zero:

$$\frac{\partial \ell(r, p)}{\partial p} = \sum_{i=1}^{N} k_i \frac{1}{p} - Nr \frac{1}{1-p} = 0 \tag{3.41}$$

$$\frac{\partial \ell(r, p)}{\partial r} = \sum_{i=1}^{N} \psi(k_i + r) - N\psi(r) + N \ln(1-p) = 0 \tag{3.42}$$

where

$\psi(k) = \frac{\Gamma'(k)}{\Gamma(k)}$ is known as the di-gamma function. Solving for p in the first equation gives:

$$p = \frac{\sum_{i=1}^{N} k_i}{Nr + \sum_{i=1}^{N} k_i} \tag{3.43}$$

Placing this result in the second equation gives:

$$\frac{\partial \ell(r,p)}{\partial r} = \sum_{i=1}^{N} \psi(k_i + r) - N\psi(r) + N \ln \left(\frac{r}{r + \sum_{i=1}^{N} k_i / N} \right) = 0 \qquad (3.44)$$

In a closed form, r in this equation cannot be solved for. An iterative method, such as Newton's can be used if a numerical solution is required.

3.4 Multinomial logistic regression

In statistics, multinomial logistic regression is a classification method that generalizes logistic regression to multiclass problems, i.e. with more than two possible discrete outcomes (Greene, 1993). That is, it is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables (which may be real-valued, binary-valued, categorical-valued, etc.). Multinomial logistic regression is used when the dependent variable in question is nominal (equivalently categorical, meaning that it falls into any one of a set of categories which cannot be ordered in any meaningful way) and for which there are more than two categories. The multinomial logit regression acts as a unique classification problem solution which presumes a linear combination for the observed parameters in which some other case-specific features could be utilized in determining the dependent variables probability in each unique outcome.

3.4.1 Multinomial Logit. Assumptions

An assumption of the multinomial logit model is, data are specified according to cases; that is, for each case possess a unique value for each independent variable. Again multinomial logit model also makes the assumption that, for any case, the given independent variables cannot perfectly predicted the dependent variable. Like the other types of regression models, there is not a need for statistically independence among the independent variable. That notwithstanding, collinearity

is presumed to be comparatively low, since due to high correlation, there is a difficulty in differentiating between the influence of several variables.

When the multinomial logit is utilized in modeling choices, it depends on the Independence Irrelevant Alternatives (IIA) assumption, although this is not what may be deemed desirable always. The IIA assumption specifies that, the odds of desiring a state or class over another does not rely on the absence or presence of other "non-relevant" options.
When the multinomial logit is employed to formulate choices in some situations, tend to impose a lot of constraint on the relative preferences between the various options. This fact especially is important to consider if the analysis aims to forecast how choices would differ if one option was to disappear. Other models such as the nested logit or the multinomial probit may be resorted to in such scenarios since these do not violate the IIA, (Baltas and Doyle, 2001).

3.4.2 The Multinomial Logit Model

Various descriptions for the mathematical model underlying multinomial logistic regression are in existence, although they are all equivalent. Hence, there exist some difficulty in comparing the various treatments of the subject under various texts. The underlying principle behind all of these, just like any statistical grouping technique is to, linearly develop a predictor function that builds a score from a given collection of weights that are combined linearly with the explanatory features of a given observation using a dot product:

$$\text{score}(\mathbf{X}_i, k) = \boldsymbol{\beta}_k \cdot \mathbf{X}_i, \tag{3.45}$$

X_i indicates the illustrative variables vector which depicts perception i, k represents a vector of regression coefficients ascribed to the outcome k and (X_i, k) speaks to the score connected with coordinating perception i to classification k.

As per discrete decision hypothesis, where perceptions signify individuals and results with regards to decisions, the score is regarded to be the utility joined with individual i deciding on the k elective. The most elevated score demonstrates the anticipated result. That which changes between the multinomial logit and alternate systems, with the same central theme, is the procedure in deciding the best coefficients and the way in which scores are deciphered. Under the multinomial logit model, it is conceivable to straightforwardly change a score to a likelihood estimate, subsequently expressing the likelihood of perception i deciding on result k. This goes about as a key point for including multinomial logit model into different strategies that may incorporate different techniques in an examinations methodology. Not having such a method, have a tendency to expand blunder in results.

Basically, the setup of the multinomial logit has the same setup as any other logistic regression model. The categorical other than binary nature of dependent variables is the only difference in the setups.
The beginning presumption is that there is a progression of N observed information focuses with every information point i $(1, \cdots, N)$ comprised of M illustrative variables $x_{1,i} \ldots x_{M,i}$ and an accompanying categorical outcome Y_i, which explains one of the K conceivable values. The K conceivable values denotes unique groupings. Multinomial logit is intended for building a model that clarifies the relationship existing between the illustrative variables and the result. It seeks to do this such that the turnouts of other trials with the same underlying principles can be predicted accurately predicted when new data point are made available.

Like alternate regressions of the linear type, multinomial logistic regression utilizes a straight line function in predicting the likelihood that perception i has

result k. It does this using the formulation below:

$$f(k,i) = \beta_{0,k} + \beta_{1,k}x_{1,i} + \beta_{2,k}x_{2,i} + \cdots + \beta_{M,k}x_{M,i}, \tag{3.46}$$

where $\beta_{m,k}$ represents the regression coefficient linking the mth explanatory variable with the kth outcome. The regression coefficients and explanatory variables normally,can be written as a vector of size $M+1$, so that the function above can be written as:

$$f(k,i) = \boldsymbol{\beta}_k \cdot \mathbf{x}_i, \tag{3.47}$$

where $\boldsymbol{\beta}_k$ is the collection of regression coefficients associated with the $k-th$ outcome, and $\mathbf{x}_i$ (which is a row vector) is the collection of explanatory variables linked with the $i-th$ observation.

To build a multinomial logit model, one way of doing this,is assume that, for K conceivable turnouts, running $K-1$ autonomous binary logistic regression models, one of these possible turnouts is selected to act as a fulcrum on which the other $K-1$ outcomes are separately regressed. Below is how this how this process is done.

$$\ln \frac{\Pr(Y_i = 1)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_1 \cdot \mathbf{X}_i \tag{3.48}$$

$$\ln \frac{\Pr(Y_i = 2)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_2 \cdot \mathbf{X}_i \tag{3.49}$$

$$\cdots\cdots \tag{3.50}$$

$$\ln \frac{\Pr(Y_i = K-1)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_{K-1} \cdot \mathbf{X}_i \tag{3.51}$$

$$\tag{3.52}$$

Different sets of coefficients were introduced. One representing each possible outcome. Exponentiating both sides, and solving for the probabilities, the result

is:

$$\Pr(Y_i = 1) = \Pr(Y_i = K)e^{\boldsymbol{\beta}_1 \cdot \mathbf{X}_i} \tag{3.53}$$

$$\Pr(Y_i = 2) = \Pr(Y_i = K)e^{\boldsymbol{\beta}_2 \cdot \mathbf{X}_i} \tag{3.54}$$

$$\cdots\cdots \tag{3.55}$$

$$\Pr(Y_i = K - 1) = \Pr(Y_i = K)e^{\boldsymbol{\beta}_{K-1} \cdot \mathbf{X}_i} \tag{3.56}$$

$$\tag{3.57}$$

Considering the way that the total of the probabilities must accumulate to one, the probability of a K outcome is then given by:

$$\Pr(Y_i = K) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}} \tag{3.58}$$

And the other probabilities:

$$\Pr(Y_i = 1) = \frac{e^{\boldsymbol{\beta}_1 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}} \tag{3.59}$$

$$\Pr(Y_i = 2) = \frac{e^{\boldsymbol{\beta}_2 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}} \tag{3.60}$$

$$\cdots\cdots \tag{3.61}$$

$$\Pr(Y_i = K - 1) = \frac{e^{\boldsymbol{\beta}_{K-1} \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}} \tag{3.62}$$

$$\tag{3.63}$$

Considering the fact that regressions are multiplied uncovers why the model is built on the presumption of IIA.

The β_k parameters (though known) in every vector are mutually evaluated by maximum a posteriori (MAP) estimation, an augmentation of maximum likelihood. This is done by regularization of the weights to avoid obsessive arrangements. An iterative strategy, for example, the generalized iterative scaling

(Darroch and Ratcliff, 1972) or the iteratively reweighted least squares (IRLS) (Bishop, 2006),or by means of gradient-based optimization algorithms such as L-BFGS (Malouf, 2002), or by specialized coordinate descent algorithms (Yu et al., 2011) are employed.

The multi-way regression can directly be inferred from the binary logistic regression formulated as a log-linear model. The logarithm of the likelihood of observing a given turnout utilizing the linear predictor in conjuction with additional normalization factor:

$$\ln \Pr(Y_i = 1) = \boldsymbol{\beta}_1 \cdot \mathbf{X}_i - \ln Z \tag{3.64}$$

$$\ln \Pr(Y_i = 2) = \boldsymbol{\beta}_2 \cdot \mathbf{X}_i - \ln Z \tag{3.65}$$

$$\cdots\cdots \tag{3.66}$$

$$\ln \Pr(Y_i = K) = \boldsymbol{\beta}_K \cdot \mathbf{X}_i - \ln Z \tag{3.67}$$

$$\tag{3.68}$$

Like it is in the binary case, an additional term $-\ln Z$ is introduced to make sure that the entire probabilities set structures a likelihood distribution, that is, ensure the set sums up to 1:

$$\sum_{k=1}^{K} \Pr(Y_i = k) = 1 \tag{3.69}$$

The extra term is added to guarantee standardization, other than multiply as usual, mainly because the logarithm of the likelihood was taken. Exponentiating both sides converts the additive term to a multiplicative variable, and as such

simultaneously additional term was written in the form $-\ln Z$ other than $+Z$:

$$\Pr(Y_i = 1) = \frac{1}{Z} e^{\boldsymbol{\beta}_1 \cdot \mathbf{X}_i} \tag{3.70}$$

$$\Pr(Y_i = 2) = \frac{1}{Z} e^{\boldsymbol{\beta}_2 \cdot \mathbf{X}_i} \tag{3.71}$$

$$\cdots\cdots \tag{3.72}$$

$$\Pr(Y_i = K) = \frac{1}{Z} e^{\boldsymbol{\beta}_K \cdot \mathbf{X}_i} \tag{3.73}$$

The value of Z can be computed by applying the above limitation which requires that all likelihoods ought to total to 1:

$$1 = \sum_{k=1}^{K} \Pr(Y_i = k) = \sum_{k=1}^{K} \frac{1}{Z} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i} \tag{3.74}$$

$$= \frac{1}{Z} \sum_{k=1}^{K} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i} \tag{3.75}$$

Therefore:

$$Z = \sum_{k=1}^{K} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i} \tag{3.76}$$

The factor is considered as "constant" since it is not a component of Y_i, the variable over which the likelihood dissemination is characterized. In relation to the explanatory variables it is absolutely not constant, most especially, with regards to the unknown coefficients β_k, will be determined through some optimization procedure. The resulting probabilities equations are given as:

$$\Pr(Y_i = 1) = \frac{e^{\boldsymbol{\beta}_1 \cdot \mathbf{X}_i}}{\sum_{k=1}^{K} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}} \tag{3.77}$$

$$\Pr(Y_i = 2) = \frac{e^{\boldsymbol{\beta}_2 \cdot \mathbf{X}_i}}{\sum_{k=1}^{K} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}} \tag{3.78}$$

$$\cdots\cdots \tag{3.79}$$

$$\Pr(Y_i = K) = \frac{e^{\boldsymbol{\beta}_K \cdot \mathbf{X}_i}}{\sum_{k=1}^{K} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}} \tag{3.80}$$

$$\tag{3.81}$$

Or generally:

$$\Pr(Y_i = c) = \frac{e^{\boldsymbol{\beta}_c \cdot \mathbf{X}_i}}{\sum_{k=1}^{K} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}} \tag{3.82}$$

The following function: $\text{softmax}(k, x_1, \ldots, x_n) = \frac{e^{x_k}}{\sum_{i=1}^{n} e^{x_i}}$ is alluded to as the softmax function. The impact serves as the reason for exponentiating the values $x_1, \ldots, x_n$ since it exaggerates the contrasts between them. Therefore, $\text{softmax}(k, x_1, \ldots, x_n)$ returns a value approaching 0 anytime x_k is fundamentally not exactly the most extreme of all the values, and returns a value approaching 1 whenever it is used on the max value, however, the result may not be so when it is very close to the next-max value. Therefore, the softmax function is utilized in the construction of weighted average which acts as a smooth function and approximates the indicator function as: $f(k) = \begin{cases} 1 \text{ if } k = \arg\max(x_1, \ldots, x_n), \\ 0 \text{ otherwise.} \end{cases}$

Hence the probability equations can be written as

$$\Pr(Y_i = c) = \text{softmax}(c, \boldsymbol{\beta}_1 \cdot \mathbf{X}_i, \ldots, \boldsymbol{\beta}_K \cdot \mathbf{X}_i) \tag{3.83}$$

Therefore in binary logistic regression, the softmax function serves as the equivalent of the logistic function. A fact worth noting is the fact that, since all probabilities ought to sum up to 1, not all the vector of coefficients, β_k, are particularly identifiable, hence making one completely determined when all the other

coefficients are known. Due to this reason, only $k-1$ separately identifiable probabilities are in existence, implying $k-1$ specific vectors of coefficients. At the point when a constant vector is introduced to all of the coefficient vectors, the comparison of the equations are indistinguishable. By this the point discussed above is demonstrated:

$$\frac{e^{(\boldsymbol{\beta}_c+C)\cdot\mathbf{X}_i}}{\sum_{k=1}^{K} e^{(\boldsymbol{\beta}_k+C)\cdot\mathbf{X}_i}} = \frac{e^{\boldsymbol{\beta}_c\cdot\mathbf{X}_i}e^{C\cdot\mathbf{X}_i}}{\sum_{k=1}^{K} e^{\boldsymbol{\beta}_k\cdot\mathbf{X}_i}e^{C\cdot\mathbf{X}_i}} \tag{3.84}$$

$$= \frac{e^{C\cdot\mathbf{X}_i}e^{\boldsymbol{\beta}_c\cdot\mathbf{X}_i}}{e^{C\cdot\mathbf{X}_i}\sum_{k=1}^{K} e^{\boldsymbol{\beta}_k\cdot\mathbf{X}_i}} \tag{3.85}$$

$$= \frac{e^{\boldsymbol{\beta}_c\cdot\mathbf{X}_i}}{\sum_{k=1}^{K} e^{\boldsymbol{\beta}_k\cdot\mathbf{X}_i}} \tag{3.86}$$

Sterning from the above, traditionally C is set as, $C = -\boldsymbol{\beta}_K$. Basically, the constant is set in a manner which so as to convert one vector to 0, the other vectors are converted to the distinction between these vectors and the vector converted to 0. This process is the same as choosing one of the K options and pivoting on it, and assessing how much the other $K-1$ options fare (this can either be better or worse). Mathematically, transformation of the coefficients as below:

$$\boldsymbol{\beta}'_1 = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_K \tag{3.87}$$

$$\cdots\cdots \tag{3.88}$$

$$\boldsymbol{\beta}'_{K-1} = \boldsymbol{\beta}_{K-1} - \boldsymbol{\beta}_K \tag{3.89}$$

$$\boldsymbol{\beta}'_K = 0 \tag{3.90}$$

Leading to:

$$\Pr(Y_i = 1) = \frac{e^{\boldsymbol{\beta}'_1 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}'_k \cdot \mathbf{X}_i}} \tag{3.91}$$

$$\cdots\cdots \tag{3.92}$$

$$\Pr(Y_i = K - 1) = \frac{e^{\boldsymbol{\beta}'_{K-1} \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}'_k \cdot \mathbf{X}_i}} \tag{3.93}$$

$$\Pr(Y_i = K) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}'_k \cdot \mathbf{X}_i}} \tag{3.94}$$

$$\tag{3.95}$$

This has the same form as that of the model above, as far as $K-1$ non-dependent two-way regressions, notwithstanding the (') on the coefficients.

Taking after the two-way latent variable model formulated for binary logistic regression, the multinomial logistic regression can also be defined as a latent variable model. The latent variable definition is mostly found in the hypothesis of discrete choice models, this makes it fairly easy for comparing the multinomial logit to the multinomial probit model, and by extension to more complicated models. Consider that, for each given information point i with a probable result k, there exist an unobserved random variable(continuous latent variable $Y_{i,k}$*-) with distribution:

$$Y^*_{i,1} = \boldsymbol{\beta}_1 \cdot \mathbf{X}_i + \varepsilon_1 \tag{3.96}$$

$$Y^*_{i,2} = \boldsymbol{\beta}_2 \cdot \mathbf{X}_i + \varepsilon_2 \tag{3.97}$$

$$\cdots \tag{3.98}$$

$$Y^*_{i,K} = \boldsymbol{\beta}_K \cdot \mathbf{X}_i + \varepsilon_K \tag{3.99}$$

$$\tag{3.100}$$

where $\varepsilon_k \sim \text{EV}_1(0,1)$, -standard type-1 extreme value distribution. A non-random procedure developed from the latent variables is employed in determining the value of variable Y_i (i.e. this signifies that the observed outcomes have randomness taken from it and into the latent variables), whereby result k is considered picked if and if only the accompanying utility ($Y^*_{i,k}$) exceeds the utilities of all available options. Due to the continuous nature of the latent variables, the probability that two latent variables wi;; have the same value is 0, hence no need worrying about such a situation occurring. Stated differently:

$$\Pr(Y_i = 1) = \Pr(Y^*_{i,1} > Y^*_{i,2} \text{ and } Y^*_{i,1} > Y^*_{i,3} \text{ and } \cdots \text{ and } Y^*_{i,1} > Y^*_{i,K}) \tag{3.101}$$

$$\Pr(Y_i = 2) = \Pr(Y^*_{i,2} > Y^*_{i,1} \text{ and } Y^*_{i,2} > Y^*_{i,3} \text{ and } \cdots \text{ and } Y^*_{i,2} > Y^*_{i,K}) \tag{3.102}$$

$$\cdots \tag{3.103}$$

$$\Pr(Y_i = K) = \Pr(Y^*_{i,K} > Y^*_{i,1} \text{ and } Y^*_{i,K} > Y^*_{i,2} \text{ and } \cdots \text{ and } Y^*_{i,K} > Y^*_{i,K-1}) \tag{3.104}$$

Or equivalently:

$$\Pr(Y_i = 1) = \Pr(\max(Y^*_{i,1}, Y^*_{i,2}, \ldots, Y^*_{i,K}) = Y^*_{i,1}) \tag{3.105}$$

$$\Pr(Y_i = 2) = \Pr(\max(Y^*_{i,1}, Y^*_{i,2}, \ldots, Y^*_{i,K}) = Y^*_{i,2}) \tag{3.106}$$

$$\cdots \tag{3.107}$$

$$\Pr(Y_i = K) = \Pr(\max(Y^*_{i,1}, Y^*_{i,2}, \ldots, Y^*_{i,K}) = Y^*_{i,K}) \tag{3.108}$$

$$\tag{3.109}$$

A more careful look at the first equation reveals that it can be re-written as:

$$\Pr(Y_i = 1) = \Pr(Y_{i,1}^* > Y_{i,k}^* \ \forall \ k = 2, \ldots, K) \tag{3.110}$$

$$= \Pr(Y_{i,1}^* - Y_{i,k}^* > 0 \ \forall \ k = 2, \ldots, K) \tag{3.111}$$

$$= \Pr(\boldsymbol{\beta}_1 \cdot \mathbf{X}_i + \varepsilon_1 - (\boldsymbol{\beta}_k \cdot \mathbf{X}_i + \varepsilon_k) > 0 \ \forall \ k = 2, \ldots, K) \tag{3.112}$$

$$= \Pr((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_k) \cdot \mathbf{X}_i > \varepsilon_k - \varepsilon_1 \ \forall \ k = 2, \ldots, K) \tag{3.113}$$

Some notes worth considering:

- Generally, given $X \sim \text{EV}_1(a, b)$ and $Y \sim \text{EV}_1(a, b)$ then $X - Y \sim \text{Logistic}(0, b)$. This implication of this is that the difference between two iid extreme-value-distributed variables takes after the logistic distribution, given that the first variable is considered irrelevant. Considering that the first parameter is a location parameter,in that the mean is varied by a fixed amount by it, moreover when two variables are varied by the same amount, their difference does not change. It still remains the same. Therefore all statements related to the basic likelihood of a said decision incorporate the logistic distribution, thereby making the decision of extreme-value distribution seem subjective.

- The second parameter (scale) in the extreme-value or logistic distribution is such that when $X \sim \text{Logistic}(0, 1)$ then $bX \sim \text{Logistic}(0, b)$. This interprets as replacing scale 1 with an error variable having an arbitrary scale parameter's effect can be augmented easily by multiplying the regression vectors with the same scale. From these (preceding point with that just made), it is gathered that, using a standard extreme-value distribution (with parameters; scale 1, location 0) for error variables results in no loss of generality as against the use of an arbitrary extreme-value distribution. Stated more profoundly, the model is non-identifiable that is, has no unique collection of ideal coefficients, when the more broad distribution is utilized.

- Including any subjective constant to the coefficient vectors imposes no bear-

ing on the model. This is attributable to the fact that the differences in regression coefficients vectors are utilized. This implies that, like the log-linear model scenario, just $K-1$ of the coefficient vectors are unique, and the $K-th$ set to an arbitrary value.

3.4.3 Intercept Estimation

In multinomial logistic regression, a dependent variable is made the benchmark category. Odds ratios for the various non-dependent variables with respect to the dependent variable category are determined. However, the exception of the reference category is made during the analysis. The exponential beta coefficient depicts changes in the dependent variable odds of being in a particular category with regards to the reference category, as in the change linked with a unit change inf the relating variable.

3.4.4 Natural language processing application of Multinomial Logit.

Multinomial LR classifiers are usually employed as an option to Naive Bayes classifiers in natural language processing. This is due to the fact that, no assumption of statistical independence of the random variables which plays the role of predictors, are made. Despite this, the Multinomial logit not be appropriate as learning in such a model, relatively to a naive Bayes classifier is slower and hence not be considered as proper given an expansive number of classes to learn.

3.5 General Probability

Many are the events bound to occur at one point or the other, however of relevant importance to have a realistic or near-actual measures of the time of such an event happening. To this reasoning is the concept of the probability theory founded. The probability concept estimates the likeliness of an event occurring

under circumstances purely random. Classic examples include number popping up on a die throw, side of a coin facing upwards after a toss, horse winning a race, to-mention-but-a few. Probability of events occurring or vice-versa is quantified as a numerical value between 0 and 1(the endpoints inclusive), with 0 signifying an impossible outcome and certainty of an outcome signified by 1 Stuart and Ord (2009). From this it can be concluded that the closer the probability of an event is to 1, the higher the level of certainty that the said event will occur. Probability theory hinges on some guiding principles or axioms which acts as the benchmark for the discipline.

3.5.1 Axioms of Probability Theory

These probability axioms are also known as the Kolmogorov axioms. Given the following parameters Ω, F, P these together form a measure space denoted by (Ω, F, P) with $P(\Omega) = 1$, where Ω is the sample space, F the event space and P is the probability measure.

Axiom 1

The likelihood of an event is a non-negative real number:

$P(E) \in \mathbb{R}, P(E) \geq 0 \qquad \forall E \in F$

However there are theories in which there exist negative probability. In such theories, this first axiom is relaxed.

Axiom 2

This axiom is the Unit measure assumption. This assumption states that the likelihood of a rudimentary event in the whole sample space will happen is 1. Stated all the more particularly, no elementary events are outside the sample space.

$P(\Omega) = 1.$

Hence should a calculation of an event results in a value greater than 1 indicates that the probability calculation is erroneous. Most of such errors is due to the inability of one performing the calculation to clearly define the entire sample space.

Axiom 3

The supposition of σ-additivity. That is for any countable sequence of mutually exclusive events E_1, E_2, ... satisfies the condition below
$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$.

3.5.2 Some other Useful Probability laws

Based on the Kolmogorov axioms, some other useful laws governing the Probability theory can be derived. These are:

- **Monotonicity**

$$\text{if} \quad A \subseteq B \quad \text{then} \quad P(A) \leq P(B).$$

- **Numeric bound**

  this property sterns directly from the above and quite intuitive too.

$$0 \leq P(E) \leq 1 \qquad \forall E \in F.$$

- **Null set probability**

  This is given as:

$$P(\varnothing) = 0.$$

- The Probability Addition law.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{3.114}$$

This law is also known as the sum rule. The proof of which is as follows. $P(A \cup B) = P(A) + P(B \setminus (A \cap B))$ (by Axiom 3) now, $P(B) = P(B \setminus (A \cap B)) + P(A \cap B)$. Eliminating $P(B \setminus (A \cap B))$ from both equations provides the desired result.

Proof of Properties

Proofs for the above axioms exist to both provide insightful and insightful openings and supplications. Most especially is the connection made between the numeric bound and the other two axioms.

Starting with the monotonicity axiom, let $E_1 = A$, $E_2 = B \backslash A$, given that $A \subseteq B$ and $E_i = \varnothing$ such that $i \geq 3$,making it obvious that the E_i are pairwise disjoint. Hence, $E_1 \cup E_2 \cup \ldots = B$. Therefore,

$$P(A) + P(B \backslash A) + \sum_{i=3}^{\infty} P(\varnothing) = P(B) \tag{3.115}$$

, which satisfies the numeric bound axiom. This is so because a series of non-negative numbers are formed by the left hand side of the equation which eventually converges $P(B)$, which actually is limited and hence $P(A) \leq P(B)$ and $P(\varnothing) = 0$. Moving further, by contradiction if $P(\varnothing) = a$, then the left hand side of the is not less than

$$\sum_{i=3}^{\infty} P(E_i) = \sum_{i=3}^{\infty} P(\varnothing) = \sum_{i=3}^{\infty} a = \begin{cases} 0 & \text{if } a = 0, \\ \infty & \text{if } a > 0. \end{cases} \tag{3.116}$$

Should $a > 0$, then a contradiction occurs, since the aggregate does not surpass the finite $P(B)$. And thus, $a = 0$. Hence as a side effect of the monotonicity axiom Proof, its been proved that $P(\varnothing) = 0$.

3.5.3 Some Useful Definitions

- Independent Events: Events A and B are said to be autonomous if their joint likelihood is given by:

$$P(A \text{ and } B) = P(A \cap B) = P(A)P(B), \tag{3.117}$$

- Mutually exclusive events: Given that either of events A or B happens in a trial is termed as the union of the the events A and B and is denoted by $P(A \cup B)$. However when two events are mutually exclusive, then their likelihood of occurring is;

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B). \tag{3.118}$$

- Conditional probability; This is the likelihood that some event B occurs given that some other event event A has occurred. Conditional probability is denoted by $P(A \mid B)$. $P(A \mid B)$ is defined as;

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{3.119}$$

  The conditional probability is undefined if the event serving as the denominator or the precursor equals 0, that is, $P(B) = 0$

3.6 Generalized Pareto Distribution

The generalized Pareto distribution belongs to the family of continuous probability distributions. It's main application is in the modeling the tails of other distributions. The GPD is characterized by three parameters: scale σ ,location μ, and shape ξ (Coles, 2001), (Dargahi-Noubary, 1989). However, it is sometimes defined by only scale and shape (Hosking and Wallis, 1987) and on other occasions only by the shape parameter. In some other literature, the shape parameter

is denoted as $\kappa = -\xi$ (Davison, 1984). The cumulative density function of the generalized distribution is given by (Embrechts et al., 1997):

$$F_\xi(z) = \begin{cases} 1 - (1+\xi z)^{-1/\xi} & \text{for } \xi \neq 0, \\ 1 - e^{-z} & \text{for } \xi = 0. \end{cases} \tag{3.120}$$

where $z = \dfrac{x-\mu}{\sigma}$. z ≥ 0 for $\xi \geq 0$ and $0 \leq z \leq -1/\xi$ for $\xi < 0$. $\xi \in \mathbb{R}$ and the probability density function (pdf) is also given by:

$$\left\{ f_{(\xi,\mu,\sigma)}(x) = \tfrac{1}{\sigma}\left(1 + \tfrac{\xi(x-\mu)}{\sigma}\right)^{\left(-\frac{1}{\xi}-1\right)} \right. \tag{3.121}$$

or equivalently

$$f_{(\xi,\mu,\sigma)}(x) = \frac{\sigma^{\frac{1}{\xi}}}{(\sigma + \xi(x-\mu))^{\frac{1}{\xi}+1}} \tag{3.122}$$

The GPD actually has its cdf as a solution of the following differential equation:

$$\left\{ \begin{array}{l} (\xi z + 1) f'_\xi(z) + (\xi + 1) f_\xi(z) = 0, \\ f_\xi(0) = 1 \end{array} \right\}$$

for x $\geqslant \mu$ when $\xi \geqslant 0$, and $\mu \leqslant$ x $\leqslant \mu - \sigma/\xi$ when $\xi < 0$. And its pdf is also a solution of the differential equation below:

$$\left\{ \begin{array}{l} f'(x)(-\mu\xi + \sigma + \xi x) + (\xi + 1) f(x) = 0, \\ f(0) = \frac{\left(1 - \frac{\mu\xi}{\sigma}\right)^{-\frac{1}{\xi}-1}}{\sigma} \end{array} \right\} \tag{3.123}$$

Given that the shape ξ and location μ are both zero, then the GPD is identical to the exponential distribution. Also if the shape $\xi > 0$ and location $\mu = \sigma/\xi$, then GPD is identical to the Pareto distribution with scale $x_m = \sigma/\xi$ and shape $\alpha = 1/\xi$.

If U is uniformly distributed on (0, 1], then

$$X = \mu + \tfrac{\sigma(U^{-\xi}-1)}{\xi} \sim \text{GPD}(\mu, \sigma, \xi \neq 0)$$

and

$X = \mu - \sigma \ln(U) \sim \text{GPD}(\mu, \sigma, \xi = 0)$.

The cdf's inversion resulted in the above formulation. The generalized Pareto distribution permits a continuous scope of conceivable shapes that incorporates both the exponential and Pareto distributions as uncommon cases. The generalized Pareto distribution permits one to let the available information "decide" which distribution is proper. The generalized Pareto distribution has three essential structures, each corresponding to a constraining distribution of exceedence information from an alternate class of basic distributions.

- Distributions whose tails diminish exponentially, for example the normal, lead to a generalized Pareto shape parameter of zero.
- Distributions whose tails diminish as a polynomial, for example the Student's t, lead to a positive shape parameter.
- Distributions whose tails are limited, for example the beta, lead to a negative shape parameter.

3.7 Distribution Selection Test

in this section, various model selection test on which models and distributions were selected or deemed appropriate for given dataset are discussed.

3.7.1 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test also called K-S test or KS test, is a nonparametric test employed in the assessing the equality of continuous, one-dimensional likelihood distributions which can be utilized to contrast a sample with a reference likelihood distribution such as one-sample K-S test, or also compare to two samples such as two-sample K-S test. The distance from the empirical distribution of the sample to the aggregate of the reference distribution is estimated by the K-S statistic. For two given samples, the K-S statistic estimates the separation

between their empirical distribution. The null hypothesis under which the null distribution is calculated, does this on the premise that samples(ie. two-sample case) are coming from the same distribution or in the one-sample case,the sample is drawn from the benchmark distribution (in the one-sample case). Whatever the scenario maybe, only continuous distributions treated under the null hypothesis, however in the situations other than the null hypothesis, there are no such restrictions in place. Due to the sensitive nature of the two-sample K-S, that is, much attention is paid to both the difference in shape and location of the empirical CDF of the two samples, this renders the two-sample K-S of great use amongst the the general non-parametric methods for comparing two samples. The K-S test can be changed to function as a goodness of fit test, after some adjustments though. In assessing for normality of the distribution (a special case), samples ought to be standardized and then contrasted with a standard normal distribution. This procedure is the same as having the reference distribution's mean and variance set equal to that of the sample estimates. It is known that implementing these adjustments in defining the particular reference distribution alters the test statistic null distribution.

The $K - S$ statistic

F_n denoting the empirical distribution function for n *iid* observations X_i, is given as:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[-\infty,x]}(X_i) \tag{3.124}$$

where $I_{[-\infty,x]}(X_i)$ denotes the indicator function. The indicator function equal 1 if $X_i \leq x$ and equals 0 otherwise. The $K - S$ statistic for a given CDF, $F(x)$ is:

$$D_n = \sup_x |F_n(x) - F(x)| \tag{3.125}$$

where sup_x represents the supremum of the set of distances.

Kolmogorov distribution

The Kolmogorov distribution is defined by the random variable

$$K = \sup_{t\in[0,1]} |B(t)| \tag{3.126}$$

where $B(t)$ denotes the Brownian bridge. The CDF of K defined as:

$$\Pr(K \leq x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2} = \frac{\sqrt{2\pi}}{x}\sum_{k=1}^{\infty}e^{-(2k-1)^2\pi^2/(8x^2)}. \tag{3.127}$$

As indicated above, under the null hypothesis the sample originates from the hypothesized distribution $F(x)$, $\sqrt{n}D_n \xrightarrow{n\to\infty} \sup_t |B(F(t))|$ in distribution, where $B(t)$ is the Brownian bridge. If F is continuous then under the null hypothesis $\sqrt{n}D_n$ converges to the Kolmogorov distribution, which does not rely on F. This outcome might likewise be known as the Kolmogorov theorem. The goodness-of-fit test or the Kolmogorov-Smirnov test is developed by utilizing critical values of the Kolmogorov distribution. The null hypothesis is rejected at level α if $\sqrt{n}D_n > K_\alpha,$ where $K\alpha$ is found from $\Pr(K \leq K_\alpha) = 1 - \alpha..$ The asymptotic power of this test is 1.

Discrete null distribution

The $K - S$ test in order for it to be used on discrete variables calls for some alteration, however the test statistic form is exactly as it was as in the continuous case, the calculation of its value is the only area of slight change as it is more subtle. This can be noticed when the test statistic is computed between a continuous distribution $f(x)$ and a step function $g(x)$ which has a discontinuity at point x_{i}. Stated differently, no such limit such as $\lim_{x\to x_{\mathrm{i}}} g(x)$. Therefore the statistic is computed as:

$$\sup_x |g(x) - f(x)| = \max_{\mathrm{i}} \left[\max\left(|g(x_{\mathrm{i}}) - f(x_{\mathrm{i}})|, \lim_{x\to x_{\mathrm{i}}} |g(x) - f(x_{\mathrm{i}-1})\right)\right], \tag{3.128}$$

Unless the limiting value of the underlying distribution is known, it is going to be unclear how the limit will be replaced.

Two-sample $K-S$ test

The Kolmogorov-Smirnov test may additionally be utilized to test whether two underlying one-dimensional likelihood distributions contrast. For this situation, the K-S statistic is:

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|, \tag{3.129}$$

where $F_{1,n}$ and $F_{2,n'}$ are the empirical distribution functions of the first and the second sample respectively, and *sup* is the supremum function. The null hypothesis is rejected at level α if: $D_{n,n'} > c(\alpha)\sqrt{\frac{n+n'}{nn'}}$.
While the $K-S$ test can also be used to assess if a given $F(x)$ is the underlying likelihood distribution of $F_n(x)$, this process can also be rearranged to provide confidence limits for $F(x)$ itself. Given a critical value for test statistic D_α such that $P(D_n\, D_\alpha) = \alpha$, then a bandwidth of $\pm D\alpha$ around $F_n(x)$ will totally contain $F(x)$ with likelihood $1-\alpha$.

3.7.2 Anderson-Darling test

The Anderson-Darling test is a statistical test of whether a given sample of data is drawn from a given likelihood distribution. In its fundamental frame, the test supposes that there are no parameters to be assessed in the distribution being tested, in which case the test and its collection of critical values is distribution-free. On the other hand, the test is regularly utilized as a part of settings where a group of distributions is being tested, in which case the parameters of that family need to be estimated and account must be taken of this in adjusting either the test-statistic or its critical values. At the point of testing if a normal distribution satisfactorily portrays a collection of data, it is a standout amongst the most effective statistical tools for detecting most variations from normality. K-sample Anderson-Darling tests are accessible for testing whether several collections of

observations can be modeled as originating from a solitary populace, where the distribution function does not have to be specified. Notwithstanding its utilization as a test of fit for distributions, it can be utilized in parameter estimation as the premise for a type of least separation estimation methodology.

Single-sample test

The Anderson-Darling test is part of the category of quadratic EDF statistics, that is, empirical distribution function based test. Given F is the hypothesized distribution, and the empirical CDF denoted by F_n, then the distance from F and F_n is estimated by the quadratic EDF statistics. This is done by the formulation below

$$n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \, w(x) \, dF(x), \tag{3.130}$$

$w(x)$ represents a loading function. When $w(x) = 1$, the statistic becomes what is known as the Cramèr-von Mises statistic. Anderson-Darling (1954) hinges on the distance

$$A = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x) \, (1 - F(x))} \, dF(x), \tag{3.131}$$

this is acquired when $w(x) = [F(x) \, (1 - F(x))]^{-1}$. in this manner, when contrasted against the Cramér-von Mises distance, the Anderson-Darling distance pays more attention to observations in the tails of the distribution. For a given sample, the Anderson-Darling test assess if the sample is coming from a specified distribution. The Anderson-Darling test uses the fact that, when presented a hypothesized underlying distribution and with the assumption that the data originates from this distribution, transforming the data into a Uniform distribution is possible. The test statistic A to examine if sample data $\{Y_1 < \cdots < Y_n\}$ (the data ought to be in order) originates from a distribution having the CDF, Φ, is formulated as: ‘

$$A^2 = -n - S \tag{3.132}$$

where

$$S = \sum_{i=1}^{n} \frac{2i-1}{n} \left[\ln(\Phi(Y_i)) + \ln\left(1 - \Phi(Y_{n+1-i})\right) \right] \tag{3.133}$$

The theoretical distribution's critical values can be contrasted against that of the test statistic. It must be noted that, in this situation no parameters are evaluated in connection to the distribution function Φ. The test statistic can also be utilized in the test of fit of a family of distributions, however when doing this, the statistic ought to be compared against the critical values akin to the family of theoretical distributions in question and also dependent on the parameter estimation used.

The Anderson-Darling test has some shortcomings which casts a dent on its relevance. Both literature and testing has found that A^2 is a standout amongst the most efficient EDF statistics in spotting deviations from normality. The difference in the computation is attributed to information available about the distribution.Such information include:

- 1: The mean μ and variance σ^2 defined.
- 2: The variance σ^2 finite, however mean μ is not.
- 3: The variance σ^2 is undefined but mean μ is known.
- 4: Both the variance σ^2 and the mean μ are unknown.

The observations X_i, for $i = 1, \ldots n$, of the variable X that are to be assessed are ordered from lowest to the highest and the assumption is made to indicate the ordered observations. The formulation below is used to represent that X_i. Let

$$\hat{\mu} = \begin{cases} \mu, & \text{if the mean is defined.} \\ \bar{X}, = \frac{1}{n}\sum_{i=1}^{n} X_i & \text{otherwise.} \end{cases}$$

$$\hat{\sigma}^2 = \begin{cases} \sigma^2, & \text{if the variance is defined.} \\ \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2, & \text{if the variance is undefined, but the mean is.} \\ \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, & \text{otherwise.} \end{cases}$$

The X_i values are standardized to get new values Y_i, which is denoted by $Y_i = \frac{X_i - \hat{\mu}}{\hat{\sigma}}$. The standard normal CDF Φ, A^2 is formulated as

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)(\ln \Phi(Y_i) + \ln(1 - \Phi(Y_{n+1-i}))) \tag{3.134}$$

An optional formulation in which just one observation is considered at each stage in the summation is

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\left[(2i-1)\ln \Phi(Y_i) + (2(n-i)+1)\ln(1 - \Phi(Y_i))\right]. \tag{3.135}$$

An altered statistic can be formulated by:

$$A^{*2} = \begin{cases} A^2\left(1 + \frac{4}{n} - \frac{25}{n^2}\right), & \text{the case where both variance and mean are unknown.} \\ A^2, & \text{otherwise.} \end{cases}$$

Should A^{*2} surpass the given critical value, it implies the hypothesis of normality is to be rejected with some level of significance.

- Note: If $\hat{\sigma} = 0$ or any $\Phi(Y_i)$=(0 or 1) then A^2 is undefined

3.7.3 Chi-squared test

An alternative notation for the chi-squared test is χ^2. The chi-square test refers to any statistical hypothesis test whose sampling distribution of the test statistic is hinged on a chi-squared distribution given the null hypothesis is true. Alternatively, a chi-squared test is a test in which this assertion is asymptotically true, implying that the sampling distribution could be approximated to a chi-squared distribution closely to that which is required through increasing the sample size enough. An application of the chi-squared test is the determining whether there exist a significant variance between the anticipated frequencies and the observed

frequencies in one or more classifications.

3.7.4 Akaike information criterion, AIC

The AIC represents a measure of the relative adequacy of a statistical model for a collection of data points, that is, for a given set of models for the data, AIC assess the adequacy or appropriateness of every one of the models, in contrast to the other models under consideration. Therefore the AIC serves as a means by which model selection is made. AIC has the information theory as its foundation: providing a comparative assessment on the data unaccounted for when a given model represents the process generating the data. Thus, it tends to focus on the trade-off between the goodness of fit of model and the many-sided nature of the model. However, the AIC lacks to test of a model under the null hypothesis; i.e. AIC cannot provide any information on the adequacy of the model in an exact nature. In a scenario where all models under consideration are inappropriate, AIC does not indicate in no sense of it.

Given a statistical model with accompanying data. Let L denotes the model's optimized value from the probability function; let k represent the estimated count of parameters in the model. Therefore the value of the AIC model formulated as:

$$\mathrm{AIC} = 2k - 2\ln(L) \tag{3.136}$$

For a collection of possible models for the data, the model of choice is the model having thr least AIC value. Based on the assessment of the likelihood function, the AIC appreciates goodness of fit however, it includes also a penalty which actually is an increasing function of the number of estimated parameters. This penalty discourages overfitting.

3.7.5 AICc

AICc serves as an amendment for the AIC with a limited sample sizes. The formulation for the AICc is contingent on the statistical model. Supposing the given model is linear, univariate and the residuals normally-distribute, however its conditioned on the regressors. AICc is formulated as:

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{n-k-1} \tag{3.137}$$

where n represents the drawn population size and k, the count of parameters. Should the univariate linear model with normal residuals assumption be contravened, it implies then that the formulation for the AICc will normally change. AICc is fundamentally an AIC with more penalty for the extra parameters. Employing AIC, in the stead of AICc in the scenario where n is not greater than k^2, builds up the likelihood of choosing models that have a lot of parameters, that is, overfitting. Its substantial to note that the likelihood of AIC overfitting is very considerable, sometimes.

The AICc is used rather than AIC, in the case where n is small or in the case where k is large. However due to the fact that AICc approaches the AIC when n becomes vast, AICc ought to be used generally in any case. Should all models under consideration possess identical k, then both the AICc and AIC will have same valuations; therefore, no disadvantage will exist in utilizing the AIC in place of the AICc. Moreover, should n be much more than k^2, at that point the amendment will be immaterial; thereby making negligible the disadvantage in using AIC in the place of the AICc.

3.7.6 Comparisons with other model selection methods

Comparing with Bayesian Information Criterion

The Bayesian Information Criterion (BIC) unlike the AIC, does penalize more effectively for the parameters number than the AIC does. Burnham and Anderson (2002) did a comparison on the AIC/AICc and BIC and it was shown that the AIC and AICc could be formulated exactly in the Bayesian structure as the BIC, simply by varying the prior employed in the Bayesian framework. Also it an argument was made that the AIC/AICc possess a theoretical advantages over the BIC. The first of these was that, since the AIC/AICc is formulated from the information principles; though the name of the BIC suggest otherwise was not appropriate. Also, due to the fact that the Bayesian-structure formulation of BIC has a prior represented by $1/R$ (where R denotes the count of models under consideration), this is considered as not being "prudent", because the prior ought to be a diminishing function of k. Moreover, these present a couple of simulation studies in literature that tend to recommend that the AICc tends to possess realistic advantages ahead of the BIC.

Given the context of regression, comparison between AIC and BIC, reveals that the AIC is asymptotically ideal in selecting the model whose mean squared error is minimum, on premise that the assumption of same "true" model is not amongst the set of models under consideration; in relation to this assumption, the BIC is not asymptotically optimal. The author further revealed that the rate of convergence of the AIC to the optimum is to certain extents, the best conceivable.

Comparing with the Chi-squared test

Most at times when presented with a set of models where all the likelihood functions are assumed to be normally distributed (i.e. have mean to be zero) and

independent. This assumption results in the chi-squared tests, hinged upon the χ^2 distribution. Employing the chi-squared test is related to using the AIC test. Under this presumption, the maximum likelihood is represented as:

$L = \prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma_i^2}\right)^{1/2} exp\left(-\sum_{i=1}^{n} \frac{(y_i - f(x_i;\sigma))^2}{2\sigma_i^2}\right)$

$\therefore \ln(L) = \ln\left(\prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma_i^2}\right)^{1/2}\right) - \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - f(x_i;\sigma))^2}{\sigma_i^2}$

$\therefore \ln(L) = C - \chi^2/2$

C represents a constant which is not dependent on the model in use, however only dependent on the particular data points in use, that is, the test does not alter if the information points does not alter. Thus AIC $= 2k - 2\ln(L) = 2k - 2\,(C - \chi^2/2) = 2k - 2C + \chi^2$, since the contrasts in AIC are significant, C the constant, could be overlooked permitting the AIC to be taken as$= 2k + \chi^2$ when comparing models.

3.7.7 Bayesian Information Criterion, BIC

The Bayesian information criterion (BIC) also referred to as Schwarz criterion (denoted SBC, SBIC) is a model selection criterion with a limited collection of models; the model of choice is the model whose BIC value is the lowest. BIC is built partly on the likelihood function and also bears a close relation to the AIC. During models fitting, the likelihood can be increased by adding extra parameters, however by so doing it may result in overfitting. The issue of overfitting is resolved by both AIC and BIC with the introduction of a penalty term which compensates for the count of parameters contained in the model; the BIC penalty term or value is greater than of the AIC. BIC is formally characterized as

$$\text{BIC} = -2 \cdot \ln \hat{L} + k \cdot \ln(n). \tag{3.138}$$

where

x= represents observed data

θ= represents model parameter

n = data points contained in x, or simply put the drawn populace

k = the free parameters to be evaluated.

If the given model is a linear regression, then k refers to the count of regressors, which includes the intercept; P(x|M) = the model's marginal likelihood of the observed data M

$\hat{L}$ = the value of maximized value of the likelihood function of the model M, i.e. $\hat{L}$=P(x| $\hat{\theta}$,M), where $\hat{\theta}$ represents the likelihood function maximization parameter values. The BIC can also be viewed as an asymptotic result based on assumptions that the data under consideration has a distribution belonging to the exponential family. Stated differently,this implies that when the the integral of the likelihood function p(x| θ,M) multiplied by the prior probability distribution,denoted P(θ |M), over the parameters θ of the model M for fixed observed data x is formulated as

$$-2 \cdot \ln p(x \mid M) \approx \text{BIC} = -2 \cdot \ln \hat{L} + k \cdot (\ln(n) - \ln(2\pi)). \tag{3.139}$$

Given the assumption that the model errors otherwise referred to as disturbances are independent and identically distributed as stated or required by normal distribution and the limiting state of the derivative of the log probability regarding the actual variance is zero, results in:

$$\text{BIC} = n \cdot \ln(\widehat{\sigma_e^2}) + k \cdot \ln(n) \tag{3.140}$$

$\widehat{\sigma_e^2}$ represents the error variance; and this situation is characterized as

$$\widehat{\sigma_e^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x_i})^2. \tag{3.141}$$

and is a one-sided estimator for the real variance. The residual sum of squares (RSS) of the BIC given by

$$\text{BIC} = n \cdot \ln(RSS/n) + k \cdot \ln(n) \tag{3.142}$$

During the testing of numerous linear models against a saturated model, the BIC can be formulated in regards of the deviance χ^2 as

$$\text{BIC} = \chi^2 + df \cdot \ln(n) \tag{3.143}$$

the *df* represents the count of degrees of freedom in the test. When selecting from numerous models, the model having the least BIC is the model of choice. BIC is an increasing function of the error variance σ_e^2 and an increasing function of k. That is, unexplained variation in the dependent variable and the count of explanatory variables increase the value of BIC. Subsequently, lower BIC suggests either less explanatory variables, better fit, or both. The BIC by and large penalizes free parameters all the more firmly than the AIC, despite the fact that it relies on upon the extent of n and relative size of n and k. It is essential to remember that the BIC can be utilized to compare estimated models only when the numerical estimations of the dependent variable are indistinguishable for all evaluations being analyzed. The models being looked at need not be nested, dissimilar to the scenario when models are being contrasted by the utilization of an F-test or a likelihood ratio test.

Attributes of the Bayesian information criterion

1. It is not dependent on the prior or the prior is "obscure" (a constant).

2. It can gauge the proficiency of the parameterized model in terms of predicting the data.

3. It penalizes the multifaceted nature of the model where multifaceted nature

refers to the quantity of parameters in the model.

4. It is approximately equivalent to the minimum description length measure however with negative sign.

5. It can be utilized to pick the quantity of groups as indicated by the inherent multifaceted nature present in a specific dataset.

6. It is closely identified with other penalized likelihood criteria such as BIC and the AIC.

3.8 Premium Principle

Speaking of it loosely, a premium principle is a guideline for matching a premium value to an insurance risk. It's these principle actuaries use to come up with insurance premiums. According to (Young, 2004), χ signify the collection of nonnegative random variables on the likelihood space (Ω, F, P); this is the accumulation of insurance-loss random variables- likewise known as insurance risks. Let X, Y, Z, etc. signify typical members of χ. Concluding, let H signify the premium function, from χ to the set of (expanded) non-negative real numbers. Along these lines, it is conceivable that $H[X]$ takes the value ∞.

Independence

$H[X]$ depends just on the aggregate distribution function of X, in particular S_χ, in which $S_\chi(t)= P\omega \in \Omega : X(\omega) > t$. That is, the premium of X relies just on the tail probabilities of X. This property expresses that the premium relies just on the money related loss of the insurable event and the likelihood that a said financial misfortune happens, not the reason of the financial loss.

Risk loading

$H[X] \geq EX$ for every $X \in \chi$. Stacking for risk is attractive on the grounds that one for the most part obliges a premium rule to charge at least the normal payout

of the risk X, in particular EX, in return for guaranteeing the risk. Otherwise, the guarantor will lose money on average.

Chapter 4

Analysis

4.1 Introduction

This chapter contains the results of the study on actuarial applications of a hierarchical health insurance claim data. Estimates for the Frequency, Claim Type and Severity models were developed. The analysis was carried out with Total number of claims (noofclaims) as the response variable (ie Frequency component) and two independent variables: In-patient (inpat) and Out-patient (outpat). For the severity component, the response variable was total amount submitted (totalamtsub.ghc) and the independent variables: In-patient (inpat.ghc), Out-patient (outpat.ghc), Service (service.ghc) and Drugs (drugs.ghc) charges were chosen. The results were presented in a series of tables.

4.2 Original dataset

Table 4.1 contains the summary statistics of the original dataset. This includes Mean, Maximum and minimum values of the dataset. From the table, it is pointed out that out-patient number and out-patient amount in Ghana cedis is the topmost variable affecting health insurance claims. This can be explained or attributed to the fact that with country situation in the tropics disease conditions such malaria and other ailments which are rampant but does not require in-patient services will be on the high. With values being as high as 9295 and Ghc. 222854 depicting frequency and amount submitted, respectively.

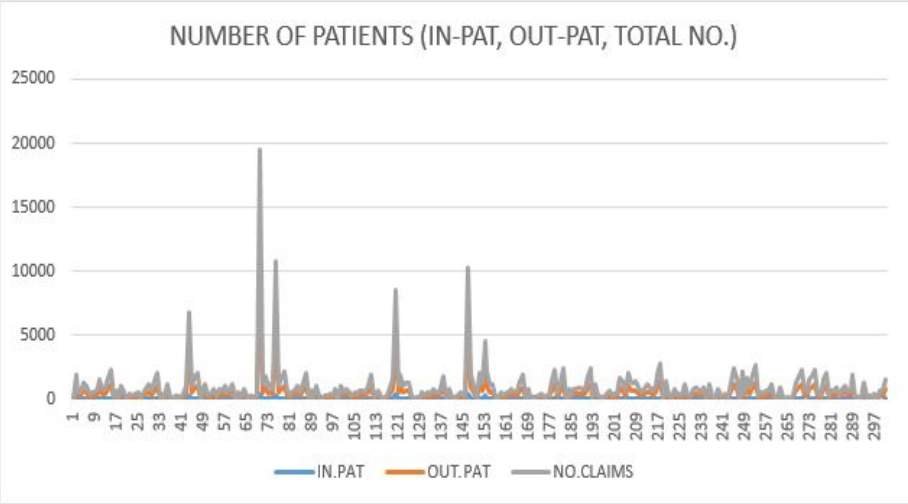


Figure 4.1: Frequency of Claims

From the diagram above it with many observations hovering about the mean, there were six(6) observations that were far off from the mean, with the farthest being seventieth (70th) data point which had an observation of nine thousand seven hundred and twenty-seven (9727) an observation such as this gives reason for such deviations off the mean.

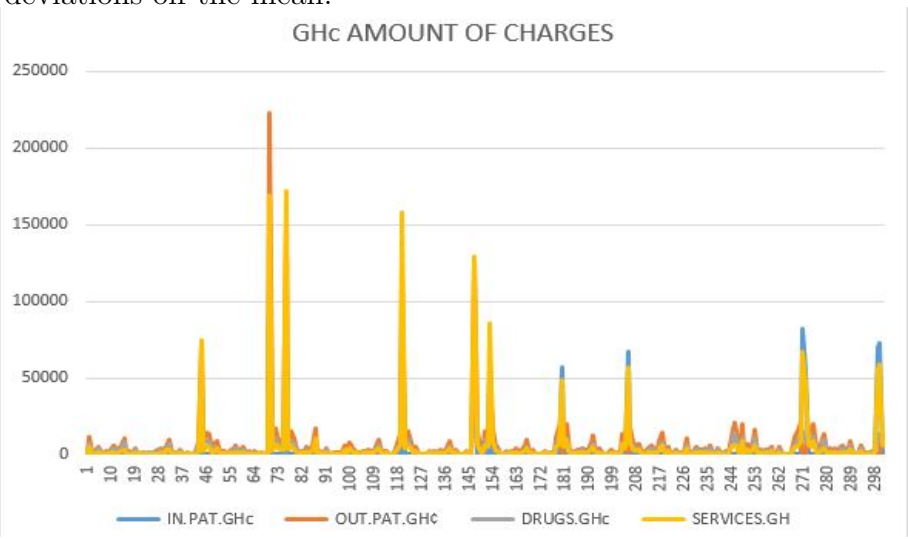


Figure 4.2: Claim Amounts Submitted

When amounts of for the various services charged by providers is plotted the outcome is no different from that of the number of claims submitted. Points of interest which were 180, 205, 271 and 300, at these point it was observed that the

various money amounts submitted for Inpatient charges were higher than that of outpatient charges.

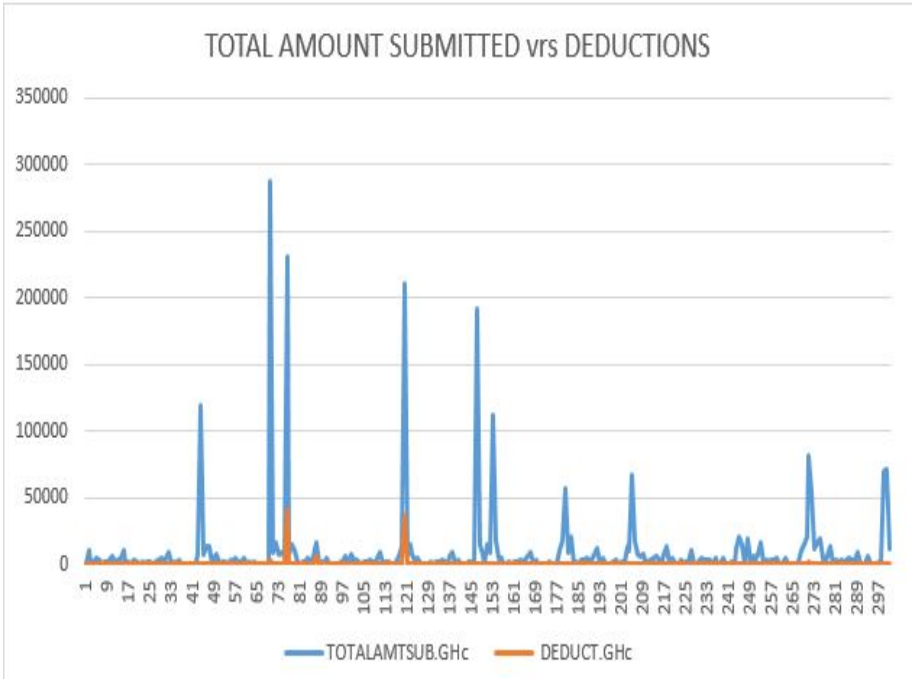


Figure 4.3: Claim amounts submitted against Deductibles

Total money amounts submitted to the NHIA as claims requesting payments were found to be subjected to some payment deductions (information regarding the composition deductions were not made available).From the graphical point, it suggested that amounts (claims) size submitted are far more greater than the deductions and deductions seemed not to be even or lacked respect for claim size.

4.3 Correlation between Claim groups

Table 4.1 brings to the fore correlation values amongst the independent variables affecting the total amount of claim charges. Its evident that services charges are strongly correlated to with all the other independent variables. From this could it be asked if a walk into any health facility will automatically register a billable charge for the insurer? From intuition, one can deduce that since an inpatient call requires patient being on admission, definitely services which wouldn't be ordinarily offered, such as feeding, bed, etc will be made available, hence the

strong correlation between services amount and inpatient charges.

Table 4.1: Correlation between variables

	IN.PAT	OUT.PAT	NO.CLAIMS	IN.PAT.GHc	OUT.PAT.GHc	DRUGS.GHc	SERVICES.GH	TOTALAMTSUB.GHc	DEDUCT.GHc
IN.PAT	1.00								
OUT.PAT	0.58	1.00							
NO.CLAIMS	0.64	1.00	1.00						
IN.PAT.GHc	0.99	0.51	0.57	1.00					
OUT.PAT.GHc	0.67	0.97	0.98	0.60	1.00				
DRUGS.GHc	0.76	0.96	0.97	0.69	0.97	1.00			
SERVICES.GH	0.92	0.80	0.83	0.90	0.88	0.90	1.00		
TOTALAMTSUB.GHc	0.89	0.87	0.90	0.85	0.93	0.95	0.99	1.00	
DEDUCT.GHc	0.48	0.50	0.51	0.53	0.62	0.51	0.69	0.65	1.00

4.4 Covariancce between claim groups

Table 4.2: Covariance between variables

	IN.PAT.GHc	OUT.PAT.GHc	DRUGS.GHc	SERVICES.GH	TOTALAMTSUB.GHc
IN.PAT.GHc	177651856.6				
OUT.PAT.GHc	154917457.9	370060476.7			
DRUGS.GHc	87731071.63	179165644.6	91522715.84		
SERVICES.GH	244838242.8	345812289.9	175374000.4	415276532.4	
TOTALAMTSUB.GHc	332569314.5	524977934.5	266896716.2	590650532.8	857547249

A brief comment on the covariance table is that, the high covariance between services and inpatient amounts is due to frequency. It was evident from the dataset, inpatient were but a few.

4.5 Hierarchical Health insurance Claims data

The overall analysis plan of this study is shown below in a model description table below

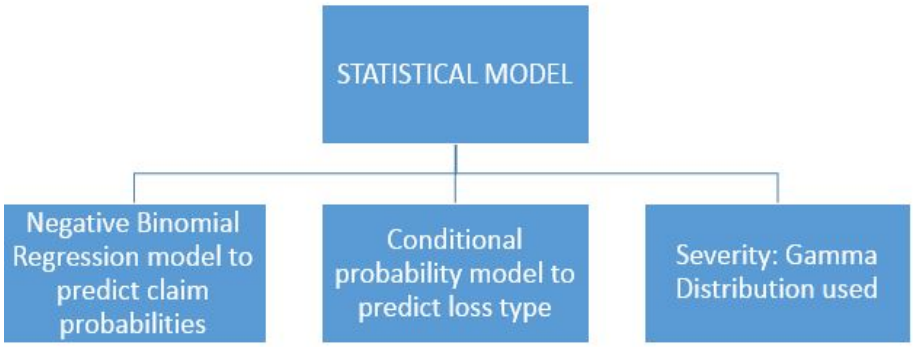


Figure 4.4: Hierarchical model

The statistical model as shown above is a joint distribution made up of a frequency, Claim type and severity components. Mathematically, this is:

$$f(N, M, y) = f(N) \times f(M|N) \times f(y|N, M) \tag{4.1}$$

$Joint function = frequency \times Conditional Claim type \times Conditional Severity$

Value of M —– numeric notation for claim type

$Y_i - - - - - - - - - - - - - - Claim Type filed i = 1, 2, ..., 14$

$Y_1 - - - - - - - - IN.PAT.GHc(In-patient\ amount\ submitted\ in\ Ghana\ Cedis)$

$Y_2 - - - - - - - - OUT.PAT.GHc(Out-patient\ amount\ submitted\ in Ghana cedis)$

$Y_3 - - - - - - - - DRUGS.GHc\ (Drugs\ amount\ submitted\ in\ Ghana\ Cedis)$

$Y_4 - - - - - - - - SERVICES.GHc(Service\ charges\ amount\ submitted\ in\ Ghana\ Cedis)$

These four were the main categories by which all other combinations were made.

4.5.1 Frequency Component

Modeled by Negative Binomial Regression. The general Negative binomial distribution is defined as a random variable k signifying the number of failures observed in a series of Bernoulli trails until r successes have occurred. The general distribution is given below

$$f(k; r, p) \equiv \Pr(X = k) = \binom{k + r - 1}{k} p^r (1 - p)^k \quad \text{for } k = 0, 1, 2, \ldots \tag{4.2}$$

Table 4.3: Summary Statistics frequency parameters

Variable	Obs	Mean	Std. Dev.	Min	Max
noclaims	301	483.9801	791.3498	4	9727
Inpat	301	13.35216	66.31452	0	432
Outpat	301	470.6279	750.8928	0	9295

These denote the conditional means and variances. These distinctions recommend that over-dispersion is available and that a count model would be suitable. A

Poisson model is initially fitted to the data. Below is the outcome of the model estimation

Table 4.4: Statistics of Poisson regression

Poisson regression					
Number of obs	301				
LR chi2(2)	106813.3				
Prob >chi2	0				
Log likelihood	-31381.138				
Pseudo R2	0.6299				
No. Of Claims	Coef.	Std. Err.	z	$P > \|z\|$	[95% Conf. Interval]
Inpat	.0014276	3.84E-05	37.15	0.000	.0013523 0.001503
Outpat	.0003474	2.14E-06	162.6	0.000	.0003432 0.000352
_cons	5.843828	0.003038	1923.92	0.000	5.837875 5.849782

Table 4.5: Measures of Fit of Poisson model

Measures of Fit for poison of noclaims		
Log-Lik Intercept Only: -84787.768	Log-Lik Full Model:	-31381.138
D(298): 62762.275	LR(2):	106813.261
	Prob >LR:	0
McFadden's R2: 0.630	McFadden's Adj R2:	0.63
Maximum Likelihood R2: 1.000	Cragg & Uhler's R2:	1
AIC: 208.532	AIC*n:	62768.275
BIC: 61061.556	BIC':	-106801.846

Estimating the Negative Binomial

Fitting Poisson model:

Table 4.6: Negative Binomial parameters

Negative binomial regression	
Number of obs	301
LR chi2(2)	324.25
Dispersion = mean	Prob >chi2 = 0.0000

Table 4.7: Continuation of Negative Binomial parameters

Log likelihood = -1999.7736		Pseudo R2 = 0.0750			
Noclaims	Coef.	Std. Err.	z	$P > \|z\|$	[95% Conf. Interval]
inpat	-0.0002112	0.0005686	-0.37	0.71	-.0013256 .0009031
Outpat	0.0015218	0.0001095	13.9	0	.0013072 .0017364
_cons	5.128114	0.064243	79.82	0	5.0022 5.254028
/lnalpha	-0.967968	0.0787552			-1.122325 -.8136107
Alpha	0.3798541	0.0299155			.325522 .4432547
Likelihood-ratio test of alpha=0: chibar2(01) = 5.9e+04					Prob>=chibar2 = 0.000

The output begins with an iteration log. The model fitting starts off by first fitting a Poisson model, then a null model (intercept only model) and lastly the negative binomial model. Since it employs maximum likelihood estimate, iterations are made until the adjustment in the log likelihood is adequately little. The last value in the iteration log is the last estimation of the log likelihood for the full model. The log likelihood can be utilized as a tool for contrasting models. The header data is next to be displayed. On the right-hand side, the count of observations utilized as a part of the investigation (301) is given, alongside the Wald chi-square statistic with three degrees of freedom for the full model, trailed by the p-value for the chi-square. This is a test of the model all in all. From the p-value, it can see that the model is factually relevant. The header likewise incorporates a pseudo-R2, which is 0.075. Below the header, the negative binomial regression coefficients for each of the variables, alongside standard errors, z-scores, p-values and 95% confidence intervals for the coefficients. The variable inpat has a coefficient of -.0002112, which is statistically insignificant, however,

the variable outpat is. This implies that for each one-unit increment on inpat, the anticipatory log count of the noclaims diminishes by -0.0002112, however the model increased by 0.0015218 per one unit increase in outpat. Furthermore, the log-transformed over-dispersion parameter (/lnalpha) is evaluated and shown alongside the untransformed value. A Poisson model is a model in which the value of alpha is restricted to zero. Stata finds the maximum likelihood estimate of the log of alpha and after that ascertains alpha from this. This implies that alpha is constantly more than zero and that Stata's nbreg takes into account over dispersion, that is, the mean being lesser than the variance. Beneath the coefficients table, a likelihood ratio test that alpha is equal to zero–the likelihood ratio test contrasting this model to a Poisson model. In this model the related chi-squared value is 5.9e+04 having a single degree of freedom. This unequivocally recommends that alpha is non-zero and the negative binomial model is more adequate for the analysis than the Poisson model. Again the measures of fit values indicate that the negative binomial regression model is superior to the poisson model.

Additional information about the fitted model is provided below

Table 4.8: Negative Binomial Regression Measures of Fit

Measures of Fit for nbreg of noclaims			
Log-Lik Intercept Only: 2161.899	Log-Lik Full Model:	-1999.774	
D(297): 3999.547	LR(2):	324.252	Prob >LR: 0.000
McFadden's R2: 0.075	McFadden's Adj R2:	0.073	
Maximum Likelihood R2:	0.659	Cragg & Uhler's R2:	0.659
AIC:	13.314	AIC*n:	4007.547
BIC:	2304.535	BIC':	-312.837

4.6 The Negative Binomial regression model

This therefore implies that

$$\log(noclaims) = \text{Intercept} + b_1(inpat) + b_2(outpat)$$

This implies:

$$noclaims = \exp(\text{Intercept} + b_1(\text{inpat}) + b_2(\text{outpat}))$$
$$= \exp(\text{Intercept}) * \exp(b_1(\text{inpat})) * \exp(b_2(\text{outpat}))$$

$$log_{noclaims} = 5.128114 - 0.0002_{inpat} + 0.00152_{outpat} \tag{4.3}$$

λ_i = Conditional mean of variable i with i being outpat or inpat

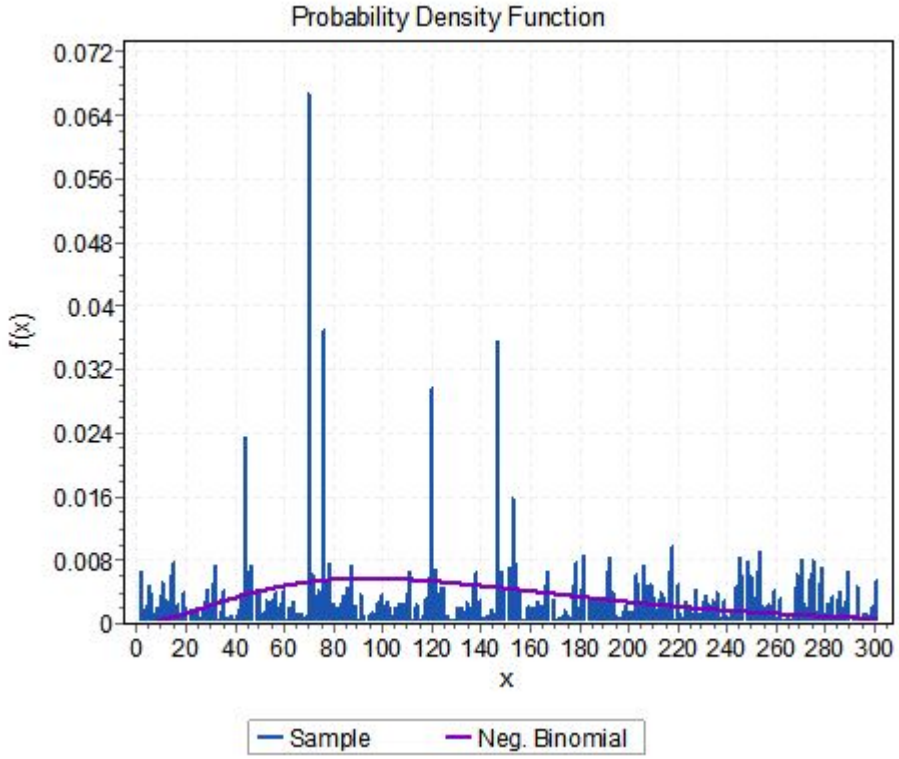
$\sigma = \frac{1}{r}$, Dispersion parameter, alpha which is .3798541

This therefore implies that

$$r = 2.632579 \approx 3$$

From analysis, it is gathered from the original dataset that Pr(Success) is 0.006.

Figure 4.5: Negative Binomial regression Density Function



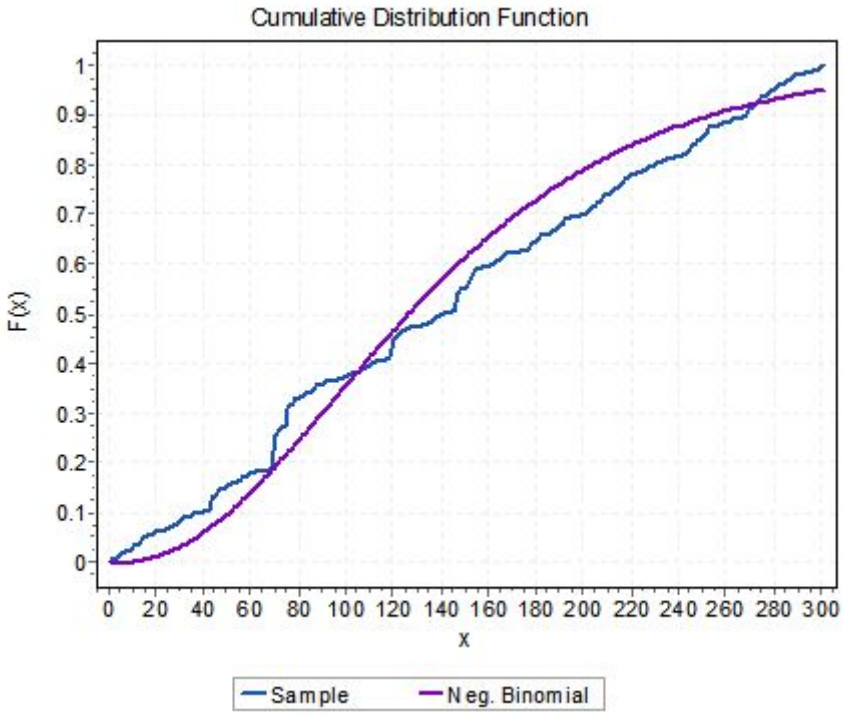


Figure 4.6: Cum. Dist. of Negative Binomial

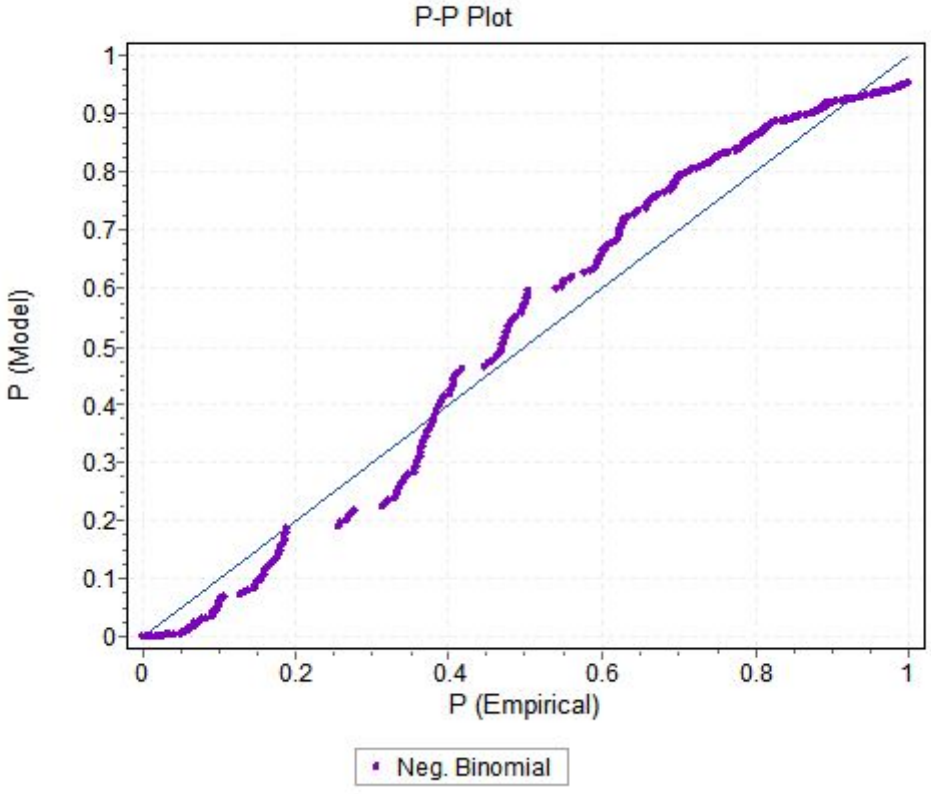


Figure 4.7: Negative Binomial P-P plot

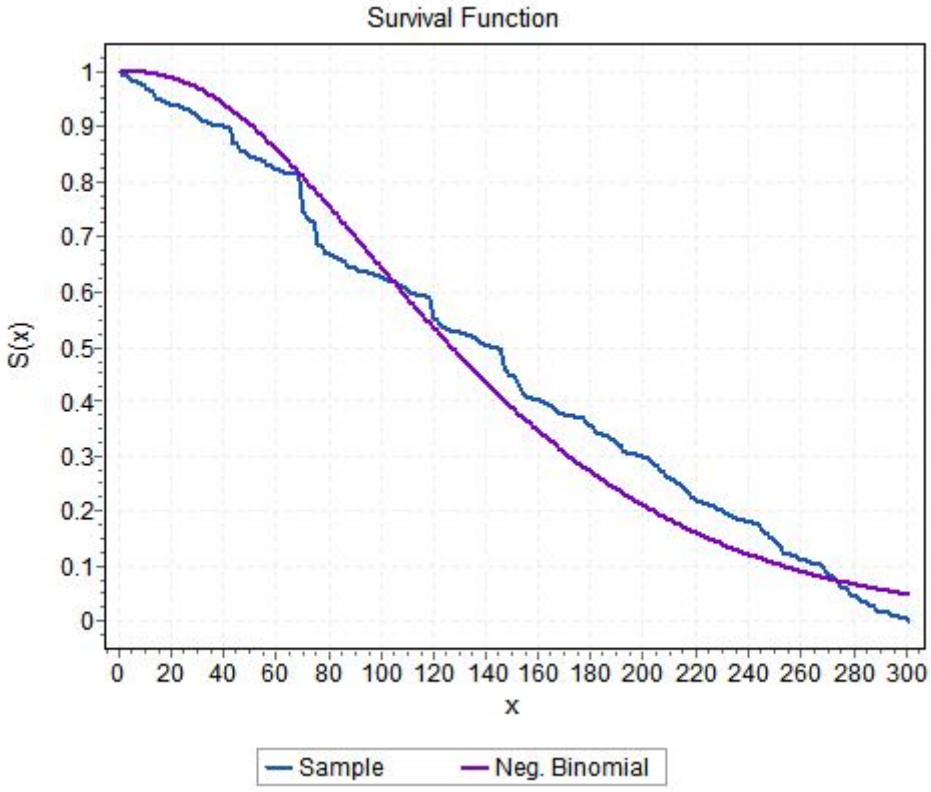


Figure 4.8: Negative Binomial Survival Func. Plot

From the above model, an estimate of claim frequency given random values of in-patient frequency and outpatient frequency can be made. Hence:

$$\binom{k+3-1}{k}(0.006)^3(0.994)^k \tag{4.4}$$

4.7 Multinomial claim type

4.7.1 Distribution of Claims

We now proceed to perform the multinomial logistic regression to ascertain to impact of the various claims types influence on number of claims submitted.

Multinomial logistic regression

Table 4.9: Multinomial Regression Parameters

Log likelihood= -28.865643		Number of obs			301
		LR chi2(3)			125.67
		Prob >chi2			0.0000
		Pseudo R2			0.6852
cla	Coef.	Std. Err.	z	$P > \|z\|$	[95% Conf. Interval]
Y2+Y3 noclaims _cons	-1.946208 42.99249	196.4911 4819.785	-0.01 0.01	0.992 0.993	-387.0617 383.1693 -9403.612 9489.597
Y1+Y3+Y4 ,noclaims, _cons	-.0011831 -3.431665	.001643 .6583364	-0.72 -5.21	0.471 0.000	-.0044033 .0020371 -4.72198 -2.141349
Y2+Y3+Y4	(base outcome)				
Y_1, Y_2, Y_3, Y_4 noclaims, _cons	-.0011831 -3.431665	.001643 .6583364	-0.72 -5.21	0.471 0.000	-.0044033 .0020371 -591267.7 591072.6

After fourteen (14) iterations,the iteration number tells how quick the model converged. One role of the log likelihood of -28.865 is that, it is also employed as means of comparing with other models. The likelihood ratio chi-square of 48.23 with a p-value¡0.0001 signifies that the model in its entirety better fits significantly than an empty model (that is, a no-predictor model) The model estimates are displayed in the table able. Thus it can be stated that a unit increment in the variable *noclaims* is linked with a -1.946208 decrement in the relative log odds of being in claim type Y_2, Y_3 against Y_2, Y_3, Y_4. Likewise a -.0011831 decrease in relative log odds of being in Claim type Y_1, Y_3, Y_4 against Y_2, Y_3, Y_4, however there was a positive change of .052674 Y_1, Y_2, Y_3, Y_4 against Y_2, Y_3, Y_4. Also evident from the table are $P > |z|$ values which suggest s that the developed model parameters are highly insignificant and hence not adequate to be utilized for the desired intentions. in view of this development one tends to fall onto conditional probability theory to aid in estimating probabilities associated with Claim types submitted to the insurer.

With the above observations, Probability of Observing a particular claim type is

Table 4.10: Claim Type Statistics

	Claim Type (M=m)				
Type of Facility	Y_2, Y_3	Y_1, Y_3, Y_4	Y_2, Y_3, Y_4	Y_1, Y_2, Y_3, Y_4	Total
Pharmacy	5	0	0	0	5
Clinic/ Health Centre	2	6	282	6	296
Total	7	6	282	6	301

estimated by:

$$Pr(M = m) = \frac{V_m}{\sum_{s=1}^{14} V_s} \tag{4.5}$$

The table below gives the values of Various m and their corresponding Probability values

Table 4.11: Claim Type Probability

Claim Type	V(M=m)
Y_2, Y_3	0.023
Y_1, Y_3, Y_4	0.020
Y_1, Y_2, Y_3, Y_4	0.020
Y_2, Y_3, Y_4	0.937

4.8 Severity Component

This is modelled by the Generalized Pareto distribution with three (3) parameters which is given by:

$$\mathrm{f(x)}=\begin{cases} \frac{1}{\sigma}\left(1+k\frac{(x-\mu)}{\sigma}\right)^{-1-1/k} & \mathrm{k}\neq 0 \\ \\ \frac{1}{\sigma}\exp\left(-\frac{(x-\mu)}{\sigma}\right) & \mathrm{k}=0 \end{cases}$$

From analysis $k = 0.70324, \sigma = 2712.3$ and $\mu = 187.54$. Below is the graphical display of the fitted model.

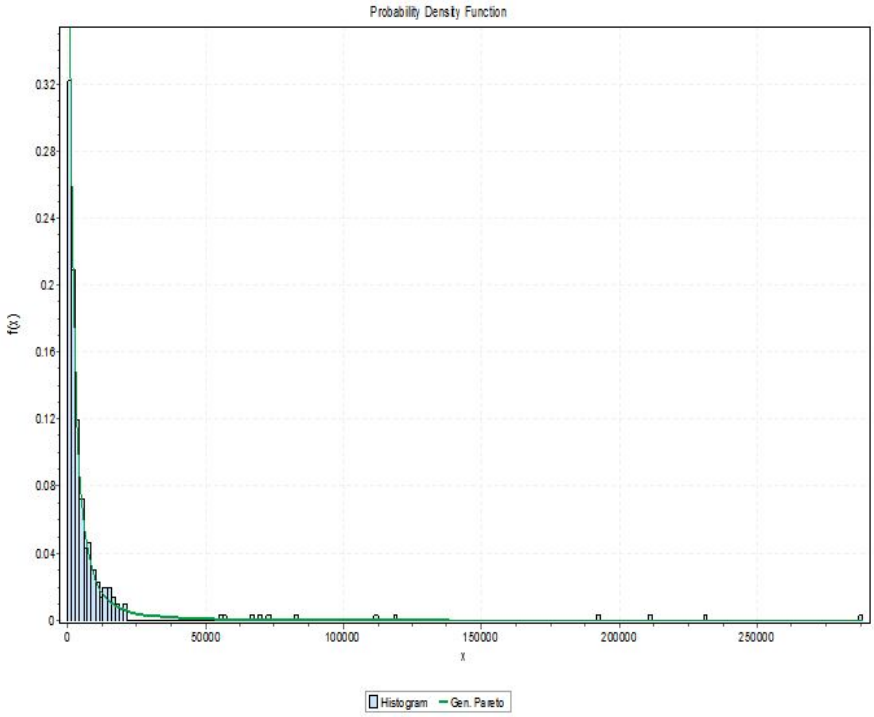


Figure 4.9: Density Func. of Severity model

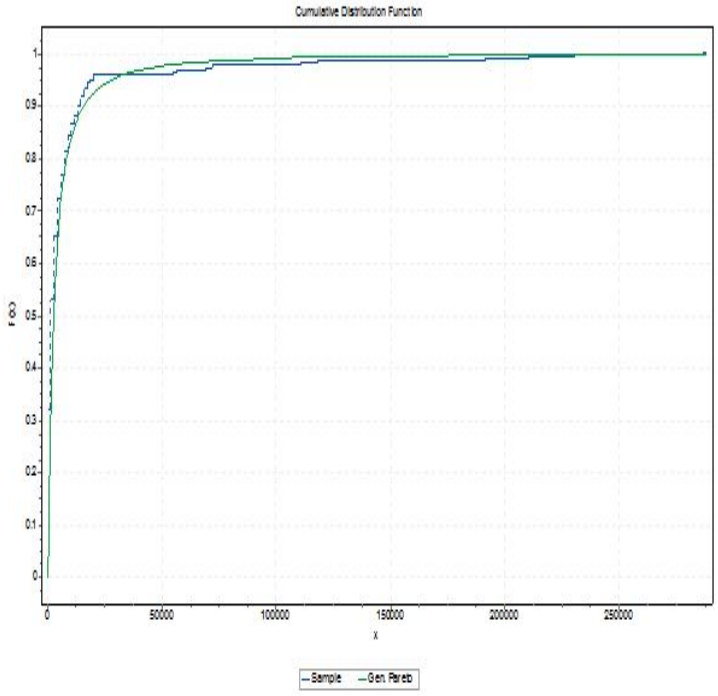


Figure 4.10: Cumulative Density Func. of Severity model

Figure 4.11: P-P Plot of Severity Component

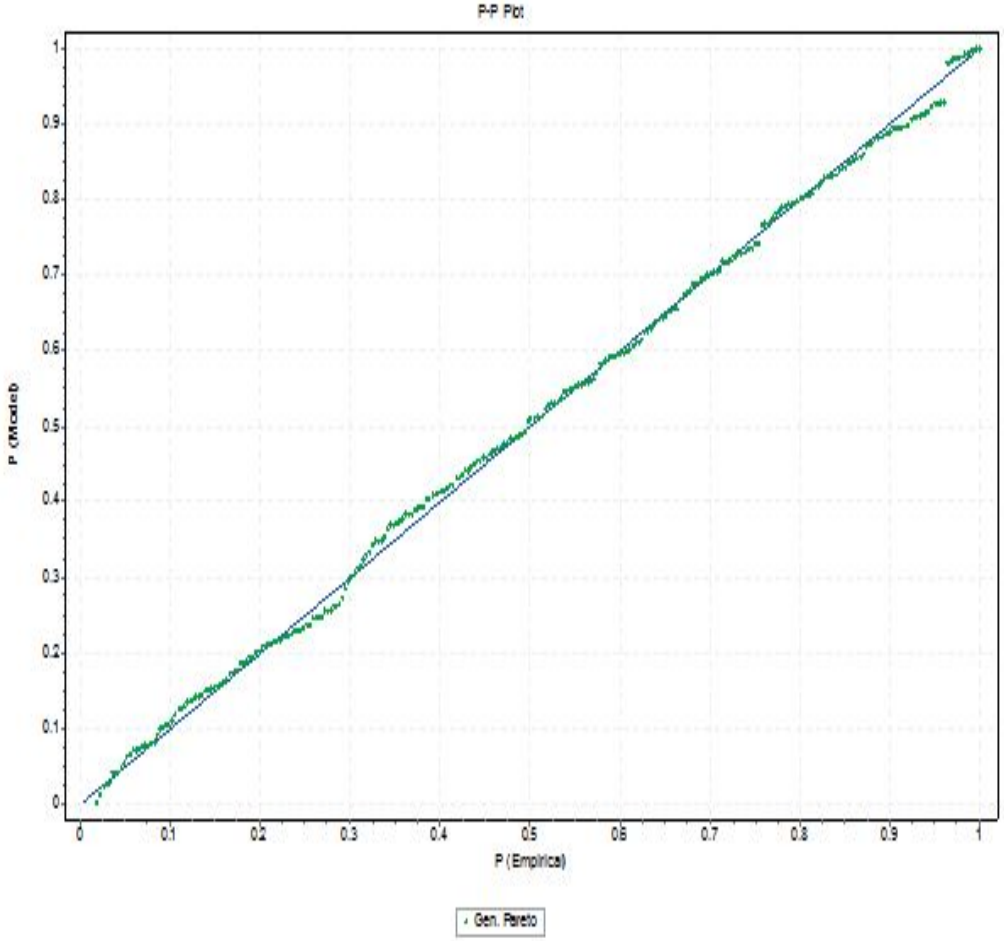
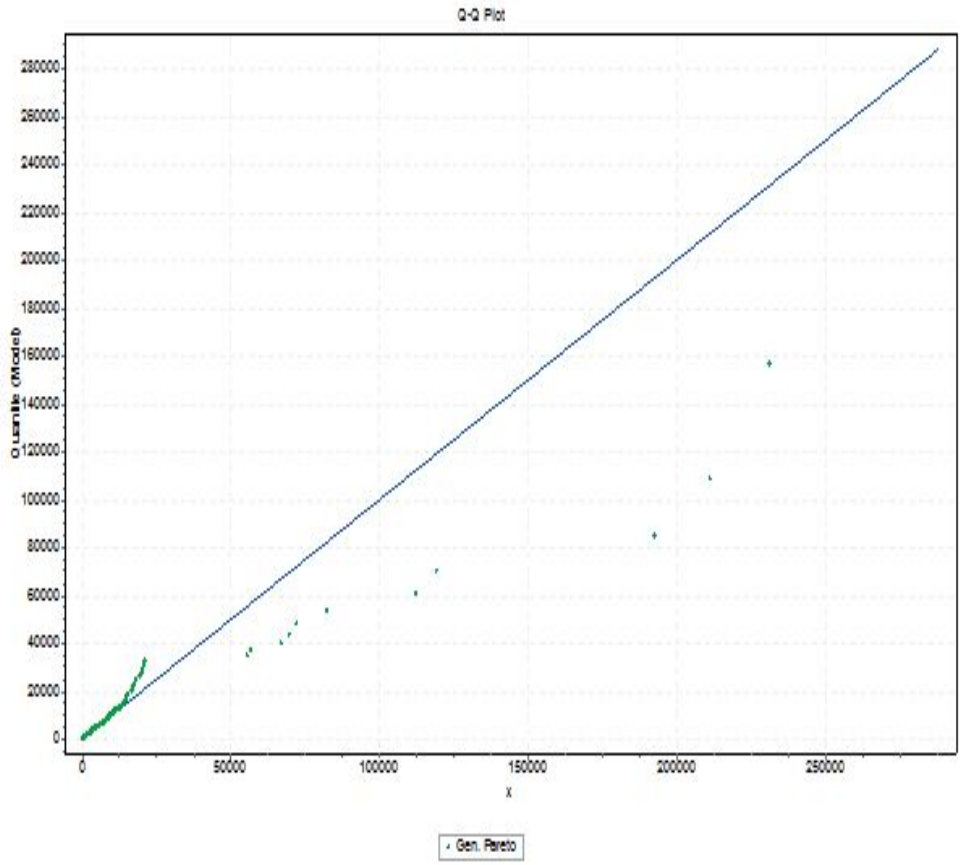


Figure 4.12: Q-Q Plot of Severity model



From the above analysis, its now convenient to finally give the distribution of the statistical model, which now is,

$f(N, M, y) = f(N) \times f(M|N) \times f(y|M, N)$

$$f(N, M, y = \binom{k+3-1}{k}(0.006)^3(0.994)^k \times \frac{V_m}{\sum_{s=1}^{14} V_s} \times \frac{1}{\sigma}\left(1 + k\frac{(x-\mu)}{\sigma}\right)^{-1-\frac{1}{k}} \tag{4.6}$$

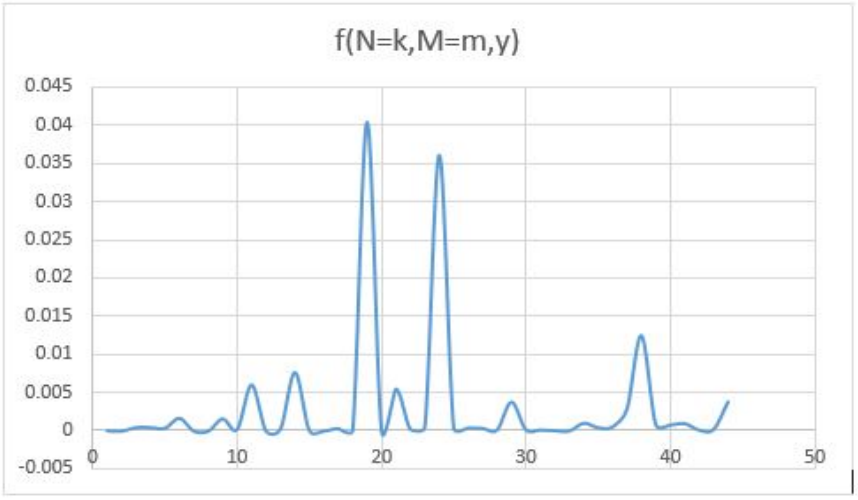
$$f(N, M, y = \binom{k+3-1}{k}(0.006)^3(0.994)^k \times \frac{V_m}{\sum_{s=1}^{14} V_s} \times \frac{1}{2712.3}\left(1 + 0.70324\frac{(x-187.54)}{2712.3}\right)^{-1-\frac{1}{0.70324}} \tag{4.7}$$

We then proceeded to randomly generate some values for k, m and y. Below is a table displaying the results of this randomly generated data to the model developed,

Table 4.12: Randomly Generated Claim Numbers, Claim Type and Severity levels

N=k,M=m,y	f(N=k,M=m,y)
5,8,46	4.08623E-05
936,12,11393	9.24113E-06
248,12,1812	0.000475728
301,12,1956	0.000477388
327,12,2462	0.000390786
460,13,11958.2	0.001683328
7,8,137.9	6.52344E-06
600,13,100000	1.2592E-05
750,13,7900	0.001614373
900,14,600	0.000224866
1000,13,900	0.006042416
2000,12,5982.79	2.04976E-07
204,12,2834.64	0.000277889
661,13,3163.75	0.007673949
686,12,5688.01	7.08686E-05
919,8,11066.69	1.19673E-05
802,8,689.45	0.000352359
911,8,2497.36	0.000101773
235,13,568.19	0.040390854
950,12,3863.69	4.68644E-05
934,13,1619.46	0.005504136
976,8,260.47	0.000235209
380,14,1531.91	0.000570665
461,13,736.92	0.036065991
144,8,1657.32	0.000372876
482,14,1786.63	0.000441882
738,14,702.51	0.000378719
797,12,2200.73	0.000151644
828,13,3574.72	0.00382272
204,8,4479.62	0.000185111
745,14,2462	0.000163094
726,14,4816.31	7.90472E-05
106,14,10096.9	2.23119E-05
479,8,392.29	0.001044407
196,14,1514.04	0.000466709
154,14,661.25	0.000575735
915,13,3452.94	0.002880403
851,13,621.1	0.012501779
166,12,179.97	0.0008267
238,8,960.7	0.00080575
365,14,480.98	0.000995781
997,12,789.78	0.00013849
788,8,1687.6	0.000224603
322,13,8018.19	0.003835572

Figure 4.13: Density Func. Joint Hierarchical model



The above graph is very reflective of the distribution expected of a loss data. Again it is very descriptive of the general trend line of the number of claims variable. The developed model was tested to find an established model which is used in modeling loss data and the findings made was very interesting. It was established that the developed model sits perfectly with the Pareto (Second Kind) Distribution with parameters α=0.69265 and β=2.0489E-4, Pearson Type 6 Distribution with parameters $\alpha 1$=0.64174,$\alpha 2$=0.82053,β=5.6707E-4 and γ=0 and the Burr Distribution with parameters k=1.4019, α=0.70705, β=7.3546E-4 and γ=0. The Pareto distribution for the most part is utilized as a part of the depiction of social, scientific, geophysical, actuarial, and different sorts of noticeable phenomena. The density function and cumulative density function of the Pareto (Second Kind) Distribution is given by

$$f(x) = \frac{\alpha\beta^{\alpha}}{(x+\beta)^{\alpha+1}} \tag{4.8}$$

$$F(x) = 1 - \left(\frac{\beta}{x+\beta}\right)^{\alpha} \tag{4.9}$$

The Pearson system was derived as an effort to model visibly skewed observations. The Type 6 Distribution of Pearson is also known as a beta prime distribution

or the F-distribution. It has for its density and cumulative density;

$$f(x) = \frac{((x-\gamma)/\beta)^{\alpha_1 - 1}}{\beta B(\alpha_1, \alpha_2)(1 + (x-\gamma)/\beta)^{\alpha_1+\alpha_2}} \tag{4.10}$$

$$F(x) = I_{(x-\gamma)/(x-\gamma+\beta)}(\alpha_1, \alpha_(2)) \tag{4.11}$$

where where B indicates the Beta Function, $I_z$and is the Regularized Incomplete Beta Function. Also in statistics, econometrics and probability theory, the Burr Type XII distribution or simply the Burr distribution is defined as a continuous probability distribution for a non-negative random variable. The Burr distribution is also referred to as the Singh-Maddala distribution and also a variant of the distributions sometimes called the "generalized log-logistic distribution". The distribution is mostly employed in modeling household income. However for consistency with the research and literature, the decision is to select the Pareto (Second Kind) Distribution with parameters α=0.69265 and β=2.0489E-4 Hence the research can be concluded that the estimated distribution model is appropriate for modeling loss distribution The acceptance criteria for the model fit was the Kolmogorov Smirnov, Anderson Darling and Chi-Squared. The table below displays other information concerning the model fit.

Table 4.13: Model Fitness- Kolmogorov-Smirnov

Kolmogorov-Smirnov					
Sample Size	44				
Statistic	0.07317				
P-Value	0.95887				
Rank	1				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	0.15796	0.18053	0.20056	0.22426	0.2406
Reject?	No	No	No	No	No

4.9 Actuarial Applications of study

Insurance is rapidly becoming more of a commodity, with customers often choosing their insurer purely on the basis of price. As a result, accurate ratemaking

Table 4.14: Model Fitness- Anderson-Darling

Anderson-Darling					
Sample Size	44				
Statistic	0.39416				
α	0.2				
Critical Value	1.3749	1.9286	2.5018	3.2892	3.9074
Reject?	No	No	No	No	No

Table 4.15: Model Fitness- Chi-Squared

Chi-Squared					
Deg. of freedom	5				
Statistic	2.3117				
P-Value	0.80455				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	7.2893	9.2364	11.07	13.388	15.086
Reject?	No	No	No	No	No

has become more important than ever (SAS Institute Inc, 2011). By this, an important application of the distribution developed by the hierarchical model is the pricing of the insurance (premium). Stated differently, Rate making (insurance pricing) is the determination of what rates, or premiums, to charge for insurance. A rate is the price per unit of insurance for each exposure unit, which is a unit of liability or property with similar characteristics (Spaulding C. W, 2014).

4.9.1 Net Premium

According to Yu (2015), amongst the many premium principles, the net premium principle is one of the commonly applied principles in the literature. It is feasible and simple in application and satisfies many preferred properties. The underlying principle is that the risk is eventually eliminated after selling a great many identical and independently distributed policies. Thus the premium would just to cover the claims only. It does not encompass any load for expenses or profit. This principle is defined as:

$$P(X) = E(X) \tag{4.12}$$

Applying this to the results (the expected loss amount will be 661992.3446).That is $E(X) = 661992.3446$. To obtain the actual premium per head client one divides the total loss expected by the expected number of clients (mean) who accessed the various health care providers. This premium calculation can be described as the pure premium.

$$P(X) = \frac{E(X)}{ExpectedNumberofClients} \tag{4.13}$$

$$P(X) = \frac{9327}{484} = 19.27 \tag{4.14}$$

The advantage of the net premium principle is that it requires the least amount of information from the predicted posterior distribution with a handy calculation process. It is a crude method of providing estimation when there is no sophisticated analysis of the predicted variables. At the same time the disadvantages are too remarkable to be neglected. In reality it is almost impossible to sell infinitely many independent and identical policies. Bearing no risk loading makes the premiums exposed to extreme events and fluctuations such as very large claim amounts. Hence it is not recommended to apply the net premium principle in practice, but to treat it as an estimated measure (Yu, 2015).

4.9.2 Expected Value Premium Principle

The expected value premium principle, often regarded as the extension of the net premium principle, expresses as

$$P(X) = (1 + \xi)E(X); \xi > 0; \tag{4.15}$$

where ξ is the loading factor. If $\xi = 0$, it is the same as the net premium principle. Clearly the premium under this principle is larger than the expected loss. The difference between the expected loss and the premium can be referred as the premium loading which provides protection against unexpected losses. Loading is

an extra cost incorporated into the insurance strategy to cover misfortunes which are high above the expected for the organization originating as a result of insuring a person who is inclined to a type of risk. It can also be defined as a sum that is incorporated into the insurance cost (Sanjeev Sinha, 2013). This sum takes care of the working expense of the insurer, and additionally the chance that the insurer's misfortunes for the duration will be greater than expected, and the adjustments in the interest earned from the insurer's ventures. This is added to the sum needed to cover losses, known as the pure insurance cost (BusinessDictionary.com, 2015). Various factors influence the loading factor, some of which are the insurer's administrative costs, costs of capital,and the ability of the insurer to pass along higher premiums to the employer (in this case the government) and the consumer (client or risk exposure unit). The loading factor can be determined based on the risk tolerance level of the insurers. A big value of ξ produces large protection margin while less attraction to the potential buyers. Therefore, it is suggested to pay attention to the loading factor and do constant testing to ensure that the factor is set at a right level Yu (2015).

4.9.3 Variance Premium Principle

The variance premium principle can be expressed as:

$$P(X) = E(X) + \omega V(X), \omega > 0 \tag{4.16}$$

If $\omega = 0$,Variance Premium Principle is the same as the net premium principle. The premium depends not only on the expected value but also the variance of the loss. Unlike the other premium principles, the variance premium principle considers the the variability of the loss; the more variability the loss, the higher the premium. In contrast to the previous case that the risk loading is proportional to the expected loss, here it is proportional to the variance of the loss Yu (2015). From the results the variance of the randomly generated dataset is 1114320680.

Hence per the variance premium principle, the premiums will be pegged per head (risk unit/ clients) at:

$$P(X) = \frac{9327 + \omega(1114320680)}{484} \tag{4.17}$$

In this principle, just like the methods stipulated above, the insurer has the sole right in determining the risk load to the premium and again, it is strictly linked to the risk tolerance of the insurer. However there is some ambiguity with regards to the interpretation of the empirical indication of the variance and the expectation since both parameters different units.

4.9.4 Standard Deviation Premium Principle

This is expressed as:

$$P(X) = E(X) + v\sqrt{V(X)}, v > 0 \tag{4.18}$$

From the equation above, it can be said that the structure of the standard deviation premium is the same as variance principle.As the standard deviation and the expectation of the loss share the same unit, it is more convenient to interpret the underlying reasoning of the principle. Each of the aforementioned principles have its pros and cons, however as this does not have any linkage to the objectives of this research, it would not be discussed.

4.10 Value-at-Risk

In finance,be it mathematics or risk administration, value at risk (VaR) is an application mostly employed as a risk measure of loss on a particular portfolio of money related resources. Given a probability$p$, time horizon and portfolio, at this point $100p\%$ VaR can be explained as the threshold loss value, such that the likelihood that the loss on the portfolio over the given time period surpasses

this value is p. In the insurance and banking industry,VaR is also known as the quantile risk measure or quantile premium principle Yu (2015). Given a confidence level $\alpha \in (0, 1)$, the VaR of the portfolio at the confidence level α is defined by the least number l such that the likelihood that the loss L surpasses l is at maximum $(1 - \alpha)$. Mathematically, given L is the loss of a portfolio, at that point $\mathrm{VaR}_\alpha(L)$ is the level α-quantile, i.e.

$$\mathrm{VaR}_\alpha(L) = \inf\{l \in \mathbb{R} : P(L > l) \leq 1 - \alpha\} = \inf\{l \in \mathbb{R} : F_L(l) \geq \alpha\}. \quad (4.19)$$

where l represent the loss

Chapter 5

Conclusion

5.1 Introduction

In this study, actuarial applications were applied to insurance claims data submitted by of three hundred and one (301) heath facilities to the NHIA. Claims submitted was total charges for service rendered to people who called upon the various health facilities. Claims submitted consisted of number of inpatient and outpatient visits, inpatient, outpatient, drugs and other services rendered charges. We sought out to develop a probability distribution for claims submitted to the national insurer. In the quest to establish this, the analysis designed so that a hierarchical model approach was feasible. We proceeded by breaking the estimation procedure into frequency component, claim type and severity(loss or claim amount) submitted. Frequency or number of claims submitted was modeled by the a negative binomial distribution. The negative binomial model was deemed appropriate after having various measures of fit gave enough proof of its appropriateness. Furthermore, with the aid of conditional probability the type of claim submitted was also estimated. Then finally, the claim amount was estimated with the generalized pareto distribution with three parameters. Softwares used in the analysis, consisted of Stata (version 12), R and EasyFit. Stata were used in the actual estimation estimation procedure whereas EasyFit was used to assess the model developed by comparing actual data distributions against that which was estimated or developed.

5.2 Findings and Conclusions

The results reported in chapter four showed that hierarchical modeling was efficient in developing claim distribution just as traditional loss distributions. The number of inpatient visits to the health care provider was proven by the negative binomial model to be insignificant in determining the number of claims submitted to the national insurer. Also it was found that though the number of number of claims submitted is a count variable, the negative binomial regression model was a superior model at estimating the claim frequency than the Poisson regression model.

5.3 Recommendations

In respect of the findings, the following recommendations are given on the use and application of hierarchical methods. Interestingly, another finding of this study was the fact that the estimated probability distribution was in perfect alignment with the traditionally more accepted distribution models such as the General Pareto (II) Distribution which is also accepted to which are distributions normally used to estimate severity.

- We recommend the use of hierarchical methods because of the effects of truly bringing to bear all aspects of a claim submitted to the national. Hierarchical methods help to uncover variables which may be relevant in the determining claim distribution but are not taken into consideration by more traditional severity estimation distributions.

- The central limit theorem hinges on large population drawn theory and it is not in all situations that this law holds, therefore it is advisable to use hierarchical methods, since it does not impose strict distributional assumptions on the datasets.

- Finally, many statisticians and actuaries do not use hierarchical methods

because, they believe these methods are computationally complex with less information on how they are used. However, the research recommends the use of these methods because,there are statistical packages which now have functions for the application of robust methods

REFERENCES

Agyepong, Irene Akua and Adjei, Sam (2008) Public social policy development and implementation: A case study of the Ghana National Health Insurance scheme. *Health Policy and Planning*, 23:150-160

Baltas, G. and Doyle, P. (2001). Random utility models in marketing research: A survey. *Journal of Business Research*, 51:115–125.

Barker, C. (1996). *The health care policy process*. SAGE Publications Ltd

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Blanchet, N. J., Fink, G., and Osei-Akoto, I. (2012). The effect of Ghana's National Health Insurance Scheme on health care utilization. *Ghana medical journal*, 46:76–84.

Boucher, J.-P. and Denuit, M. (2006). Fixed versus random effects in poisson regression models for claim counts: A case study with motor insurance. *ASTIN Bulletin*, 36:285–301.

Brockman, M. and Wright, T. (1992). Statistical motor rating: making effective use of your data. *Journal of the Institute of Actuaries*, 119:457–543.

BusinessDictionary.com, (2015).
http : *//www.businessdictionary.com/definition/loading.html*
Retrieved May 10, 2015, from BusinessDictionary.com

Coles, S. (2001). An introduction to statistical modeling of extreme values. Springer-Verlag London

Coutts, S.M., (1984). Motor Insurance: An Actuarial Approach. *JIA*, III:1

Crawley, M. J. (2012). *The R Book*. Wiley.

Daily Guide (2015). NHIA is broke. *http : //www.dailyguideghana.com/nhia − is − broke/*

Dargahi-Noubary, G. R. (1989). On tail estimation: An improved method. *Mathematical Geology*, 21(8):829–842.

Darroch, J. and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics (Institute of Mathematical Statistics)*, 43((5)):1470–1480.

Davison, A. C. (1984). Modelling excesses over high thresholds, with an application. In *Statistical Extremes and Applications*, page p. 462. Kluwer. In de Oliveira, J. Tiago.

Dutang, C., Goulet, V., and Pigeon, M. (2008a). actuar: An r package for actuarial science. *Journal of Statistical Software*, 25(7):1–37.

Dutang, C., Goulet, V., and Pouliot, L.-P. (2008b). Simulation of compound hierarchical models.

Ekman, B. (2004). Community-based health insurance in low-income countries: asystematic review of the evidence. *Health policy and planning*, 19(5):249–270.

Embrechts, Paul and Kluppelberg, Claudia and Mikosch, Thomas (1997). Modelling extremal events for insurance and finance. *Springer*

Ensor, T. (1999). Developing health insurance in transitional asia. *Social Science and Medicine*, 48:871–879.

Frees, E. W., Shi, P., and Valdez, E. A. (2009). Actuarial applications of a hierarchical insurance claims model. *ASTIN Bulletin*, 39:165–197.

Goba, F. F. K. and Liang, Z. (2011). The national health insurance scheme in ghana: Prospects and challenges: a cross-sectional evidence. *Global Journal of Health Science*, Vo l. 3(No. 2):90–101.

Greene, W. H. (1993). *Econometric Analysis*. Prentice Hall, fifth edition:.720-723. edition.

Grindle, M. and Thomas, J.W. (1992.). Public choices and policy change: the political economy of reform in developing countries. Baltimore, MD: The Johns Hopkins University Press.

Guszcza, J. (2010). Actuarial applications of hierarchical modeling. In *CAS RPM Seminar*.

Haldane, J. B. S. (1945). On a method of estimating frequencies. *Biometrika*, 33(No. 3):222–225.

Hilbe, J. M. (2011.). *Negative Binomial Regression, 2nd ed., ,*. New York: Cambridge University Press.

Hosking, J. R. M. and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29((3)):339–349.

SAS Institute Inc, (2011) How can we price our insurance products more accurately? SAS Institute Inc.

InvestorWords.com (2015a). Health insurance.
http : *//www.investorwords.com*/2289/*healthinsurance.html*

InvestorWords.com (2015b). Insurance.
http : *//www.investorwords.com*/2510/*insurance.html*

Jehu-Appiah, C., Aryeetey, G., Agyepong, I., Spaan, E., and Baltussen, R. (2012). Household perceptions and their implications for enrolment in the National Health Insurance Scheme in Ghana. *Health Policy and Planning*, 27:222–233.

Kahane, Y. and Levy, H. (1975) Regulation in Insurance Industry: Determination of Premiums in Automobile Insurance. *Journal of Risk and Insurance*, 42:117-132

Kraushaar , D. (1997). The Kenya National Hospital Insurance Fund: what can we learn from thirty years experience.
Boston: Management Sciences for Health

Lloyd-Smith, J. (2007). Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. PLoS ONE 2(2): e180.

Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. In *Sixth Conf. on Natural Language Learning (CoNLL)., pp. 49-55.*

McIntyre, D., Doherty, J., and Gilson, L. (2003). A tale of two visions: the changing fortunes of social health insurance in south africa. *Health policy and planning*, 18(1):47–58.

Ndiaye P., Soors W. and Crie B. (2007). Editorial:a view from beneath: Community health insurance in africa. *Tropical Medicine and International Health*, 12(2):157-161.

Nelson, J. (1989). *Fragile coalitions: the politics of economic adjustment.*, chapter The politics of pro-poor adjustment, pages 95–113. New Brunswick: Transaction Books.

Nguyen, H. T., Rajkotia, Y., and Wang, H. (2011). The financial protection effect of ghana national health insurance scheme: evidence from a study in two rural districts. *International Journal for Equity in Health*, 10(1):4.

Olson, M. (1965). The logic of collective action: public goods and the theory of groups. Cambridge, MA: Harvard University Press.

Patrik, G. (1981). Estimating casualty insurance loss amount distributions. In *Proceedings of Casuallty Actuarial Society.*

Roemer, M. I. (1991). The countries National health systems of the world. volume I

Sarpong, N. and Loag, W. and Fobil, J. and Meyer, C. G and Adu-Sarkodie, Y. and May, J. and Schwarz, N. G. (2010). National health insurance coverage and socio-economic status in a rural district of ghana. *Tropical Medicine and International Health*, 15(2):191–197.

Scheel, I., Ferkingstad, E., Haug, A. F. O., Hinnerichsen, M., and Meze-Hausken, E. (2013). A bayesian hierarchical model with spatial variable selection: the effect of weather on insurance claims. *Royal Statistical Society*, 62(Part1,):pp.85–100.

Sekhri, N. and Savedoff, W. (2005). Policy and practice private health insurance : implications for developing countries. *Bulletin of the World Health Organization*, 83(2):127–134.

Sanjeev Sinha, (2013). What is loading in insurance & how does it affect your premium $http://articles.economictimes.indiatimes.com/2013-11-27/news/44520212-1-claims-loading-insurance-policy-insurance-company$

Spaulding William C., (2014). Rate Making: How Insurance Premiums Are Set. $http://thismatter.com/money/insurance/rate-making.htm$

Stuart, A. and Ord, K. (2009). Kendall's advanced theory of statistics, distribution theory (volume 1) 6th ed. ISBN 9780534243128.

Thomas, S. and Gilson, L. (2004). Actor management in the development of health financing reform: health insurance in south africa 1994 -1999. *Health policy and planning*, 19(5):279–291.

todayghananews.com (2015). NHIA boss speaks on challenges facing the scheme. $http://todayghananews.com/2015/03/27/nhia-boss-speaks-on-challenges-facing-the-scheme/$

Witter, S. and Garshong, B. (2009). Something old or something new? Social health insurance in Ghana.

World Health Organization Group (2000). The world health report 2000. Health systems: Improving performance. In *The World Health Report 2000. Health systems: Improving performance.*

World Health Organization Group (2006). Moving towards universal coverage series.geneva. In *Moving Towards Universal Coverage Series.*

Xu, K., Evans, D., Carrin, G., Aguilar-Rivera, A., Musgrove, P., and Evans, T. (2007). Protecting households from catastrophic health spending. *Health Affairs (Millwood)*, 26(http://dx.doi.org/10.1377/hlthaff.26.4.972):972–983.

Xu, K., Evans, D., Kawabata, K., Zeramdini, R., Klavus, J., and Murray, C. (2003). Household catastrophic health expenditure: a multicountry analysis. *Lancet*, 362:111–117.

Young, R. V. (2004). Premium principles. In *Encyclopedia of Actuarial Science.* John Wiley & Sons, Ltd.

Yu, G. Q. (2015). Hierarchical bayesian modeling of health insurance claims. Master's thesis, SIMON FRASER UNIVERSITY; Faculty of Science , Department of Statistics and Actuarial Science.

Yu, H.-F., Huang, F.-L., and Lin, C.-J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85:41–75.

Zwilling, M. L. (2013). Negative binomial regression. The Mathematica Journal Volume 15 $http://www.mathematica-journal.com/2013/06/negative-binomial-regression/$