Kabarak Journal of Research & Innovation
www.kabarak.ac.ke

MORTALITY FORECASTING USING NEGATIVE BINOMIAL MODEL

Markson Rugut

Department of, School of Science, Engineering & Technology, Kabarak University

rugutmarkson@gmail.com

Submitted 26th March, 2021, Accepted 11th March, 2022 and Published 27th June, 2022

ABSTRACT

People living in America today are living longer compared to yester years; this is due to reduced mortality rates and increase in life expectancy. Over the years mortality rates have continuously been declining. However, other courses of death like suicide rate have been soaring high for the last 25 years which is a worrying trend for both the state and the insurance providers. According to study report conducted between 2011 and 2012, life expectancy improved by 0.1 years from 78.7 to 78.8 years. This was attributed to improvement of services in the health sector in the country. This constant change in mortality rates is proving to be a challenge to the insurance industry and pension providers in designing the right products for their consumers. In this paper Negative Binomial Regression model is used to forecast male mortality of the United States of America (1980-220). The results of this analysis demonstrate that Negative Binomial with simple polynomial functions give better mortality forecast over the given period.

Key words: Negative Binomial, Mortality, Forecasting

I. INTRODUCTION

Longevity risk management has proved to be a challenge to many life insurance companies and pension service providers over the years. Improvement in the medical filed and good personal hygiene has led to an increase in life expectancy. A mortality forecasting model which can fit well to historical data and gives a better forecast needs to be developed. In the USA life expectancy varies according to states, ethnicity, race and gender. Mortality forecasting is very essential both to the state and pension provider, annuity providers and insurance companies. The state uses these forecasted rates to plan health care programs and also in planning the social security of the nation. Forecasting rate of death is basically based on subjective judgments, or an expert opinion. However, reduction in mortality rates in recent decades have proven to be a challenge even to experts, and in this regard more methods use in classifying risk have been employed. Bailey and Simon introduced the minimum bias models, with the Generalized Linear Models; these provided a better statistical explanation for the minimum bias models which give way for selection of various parametric models. Feldblum and Brosius in there analysis later pave way for several bias functions, which include zero bias, least squares, minimum chi-squares and maximum likelihood. Thus, the purpose of this paper is to assess the viability of Negative Regression models in forecasting mortality

A. Statement problem

Mortality forecasting remains a major challenge for many governments, pension institution and annuity managers. Therefore, there is need for a more accurate forecasting model which gives good fit to the historical mortality data for consistent predictive performance. A Negative Binomial Regression model is used to forecast male mortality of the United States of America.

B. Objective of the study

The main aim of this study is to use negative binomial regression model to forecast mortality rates using male mortality of USA (1980-2020).

II. LITERATURE REVIEW

The forecasting of uncertainty in mortality improvement is especially important to actuarial risk management. More specifically, longevity risk resembles investment risk in that it is non-diversifiable, since any change in the overall mortality level is likely to affect all policyholders of the same cohort. As a result, it cannot be mitigated by the traditional insurance mechanism of selling large number of policies. However, it is different in that there is a lack of mortality-linked securities that could possibly be used for hedging. In other words, capital is often required to cushion longevity risk, and such capital is, of course, determined by measures of uncertainty associated with mortality projections. Tuljarpurkar (1997) suggested that methods for forecasting uncertainty can be segregated into two categories, namely, static and dynamic. In a static approach, the forecaster makes assertions about demographic characteristics, such as life expectancy at birth, and the trajectories along which those characteristics will change between the start and end. These trajectories are then assigned a probability, usually determined subjectively. A shortcoming of the static approach is that the dynamics of getting from “here” to “there” are largely ignored. In contrast, dynamic forecasts employ a stochastic model that is fitted to historical data. The resulting models have uncertainty embedded within them, as reflected in historical change, and yield trajectories in the form of sample paths, which are particularly valuable in assessing financial liabilities that are sensitive to the timing and pattern of demographic change. According to Polder et al., (2006) there are changing of death patterns hence there is need for research on causes of death and forecasts. This should be on understanding of health patterns, costs of social care and what drives overall mortality change. Li and Chan (2007) summed up the authentic period death rates to age 150 by the social model proposed by Himes et al. (1994). Given this grid of extrapolated passing rates, they got a Lee-Carter

mortality projection from which accomplice life tables for various birth partners were figured. At long last, they inferred the dissemination capacity of omega for every accomplice utilizing traditional extraordinary esteem theory. Notwithstanding, the dissemination capacity of must be computed numerically as the extrapolation of death rates was performed in a non-parametric way.

A. Life expectancy

Life expectancy is the average number of complete years a person is expected to live a selected group or state. Studies conducted in the USA shows that women have a higher expectancy than men by about 5 years while Hispanics have the highest life expectancy than non-Hispanics among the races. According to doctor Xu the author of the report, the reason as to why there is this variation in life expectancy among the sexes is majorly due to behavioral activities. Jiaquan Xu said "Men usually take more risks, and they participate in risky outdoor activities like climbing and scuba diving," he says. "Also, teenage boys do more high-risk activities, and they get in more car wrecks, than girls".

B. Mortality rates

Mortality rate is the probability that a life aged x at time t is expected to die in the next t+1 years. A study conducted in the year 2002 by Parikh, Guttenman, England and Pokorski found an approximately 1.2 months per year improvement in life expectancy in the developed world and a global average of about 4.5 months every year. Cousin-Frankel(2011) studied the cost implications of these improvements in life expectancy and he found out that it cost the US social security administration about $50 billion annually. Insurance firms and define benefit pension plans uses mortality forecast in determining the amount of cash reserves to be held in order to meet future liabilities of the organization. According to Halonen (2007), increments amount needed to cushion pension inform of reserves in the US will lead to an increase in pension liability by about 5-10%. These findings shows that it is utmost important for the insurance, pension firms and the government to fully understand how mortality trends will flow in the future. For us to elucidate these mortality changes we must have a forecasting model which best captures these trends.

III. METHODOLOGY

A. Negative binomial regression model

Negative binomial is a mixture of two distribution the Poisson and Gamma distribution and was first derived by Greenwood and Yule (1920) to adjust over-dispersion in discrete data. Poisson-Gamma function is a type of Poisson regression model in which the dependent variable y_{ij} is the number of times

death occurs. The number of death variable $y_{ij}$is modeled as a Poisson variable with a mean λ_{ij} where the model error is assumed to be Gamma distribution. According to Cameron and Travedi (1998), if Poisson mean assumed to have a random intercept term which enters the conditional mean in a multiplicative manner, then we get the following equations.

$$\lambda_{ij}= exp(\beta_0+\sum_{j=1}^{k} x_{ij}\beta_j + \varepsilon_{ij})$$

$$\lambda_{ij} = e^{\sum_{j=1}^{k} x_{ij}\beta_j}\ e^{(\beta_0+\varepsilon_{ij})}$$

$$\lambda_{ij}= e^{\left(\beta_0 +\sum_{j=1}^{k} x'_{ij}\beta_j\right)}e^{\varepsilon_{ij}}$$

$$\lambda_{ij} = \mu_{ij}v_{ij}$$

Where, $e^{(\beta_0+\varepsilon_{ij})}$ is defined as the random intercept μ_{ij}=$e^{\left(\beta_0 +\sum_{j=1}^{k} x'_{ij}\beta_j\right)}$ is the log-link between the mean $\lambda_{ij}$and the independent variables x's, while β_s are the regression coefficients which follows the lee-carter specification and v_{ij} is the error term.

The marginal distribution of y_{ij} is obtained by integrating the error term v_{ij}.

$f(y_{ij};\mu_{ij})=\int_0^{\infty} g(y_{ij};\mu_{ij},v_{ij})\,h(v_{ij})dv_{ij}$

Where,

$h(v_{ij})$ is the mixing distribution.in this case $g(y_{ij};\mu_{ij},v_{ij})$ is the Poisson distribution while the error term function $h(v_{ij})$ is the gamma distribution.

Letting v_{ij} be a two-parameter gamma distribution then its distribution will be given by;

$$z(v_{ij};\beta,\delta)=\frac{\delta^{\beta}}{\Gamma\beta}v_{ij}^{\beta-1}e^{-v_{ij}\delta}\ ,\quad \beta>0,\delta>0,v_{ij}>0$$

The mean of this distribution is;

$E[v_{ij}]=\beta/\delta$, while it's $VAR[v_{ij}]=\beta/\delta^2$.

By setting $\beta=\delta$ gives a one parameter gamma with $E[v_{ij}]$ =1 and $1/\beta$.as the variance

By substituting v_{ij} in the gamma distribution above it will be transforms to a function of Poisson mean given by;

$$z(\lambda_{ij};\beta,\mu_{ij})=\frac{(\beta/u_{ij})^{\beta}}{\Gamma\beta}\lambda_{ij}^{\beta-1}e^{-\frac{\lambda_{ij}}{\mu_{ij}}\delta}$$

The joint distribution of y_{ij} and u_{ij} is then given by

$$f(y_{ij};\mu_{ij}\beta)=\int_0^{\infty}\frac{exp(-\lambda_{ij})\lambda_{ij}{}^{y_{ij}}}{y_{ij}!}\frac{(\beta/u_{ij})^{\beta}}{\Gamma\beta}\lambda_{ij}^{\beta-1}e^{-\frac{\lambda_{ij}}{\mu_{ij}}\delta}d\lambda_{ij}$$

The unconditional distribution of death is thus obtained by summing out λ_{ij} in the above function.

Kabarak Journal of Research & Innovation
www.kabarak.ac.ke

$$f(y_{ij};\mu_{ij}\beta)=\frac{\left(\beta/u_{ij}\right)^{\beta}}{\Gamma\beta\Gamma(y_{ij}+1)}\int_0^{\infty} exp\left(-\lambda_{ij}\left(1+\frac{\beta}{u_{ij}}\right)\right)\lambda_{ij}^{y_{ij}+\beta-1}d\lambda_{ij}$$

$$f(y_{ij};\mu_{ij}\beta)=\frac{\left(\beta/u_{ij}\right)^{\beta}\left(1+\frac{\beta}{u_{ij}}\right)^{-(y_{ij}+\beta)}\Gamma(y_{ij}+\beta)}{\Gamma\beta\Gamma(y_{ij}+1)}$$

$$f(y_{ij};\mu_{ij}\beta)=\frac{\Gamma(y_{ij}+\beta)}{\Gamma\beta\Gamma(y_{ij}+1)}\left(\frac{\beta}{\mu_{ij}+\beta}\right)^{\beta}\left(\frac{u_{ij}}{u_{ij}+\beta}\right)^{y_{ij}}$$

This is the probability density function of negative binomial distribution

Where $u_{ij} > 0$ is the mean incident rate of y_{ij} per unit of exposure.

By setting the dispersion parameter

$\beta = 1/k > 0$

The distribution of death then becomes

$$f(y_{ij})=\frac{\Gamma\left(y_{ij}+\frac{1}{k}\right)}{y_{ij}!\,\Gamma\left(\frac{1}{k}\right)}\left(\frac{1}{1+ku_{ij}}\right)^{\frac{1}{k}}\left(\frac{ku_{ij}}{1+ku_{ij}}\right)^{y_{ij}}$$

While the mean and variance of negative binomial regression model will be given by:

$E[y_{ij}] = u_{ij}$ and $\text{VAR}[y_{ij}] = u_{ij}(1 + \kappa u_{ij})$

Thus if, $y_{ij} \sim negb(u_{ij})$ then the log-link will be given by,

$u_{ij} = exp\{\beta_0 + (x_{ij})T\beta\}$

β_0 is the intercept

Substituting u_{ij}, the probability mass function of Negative binomial regression model will then be:

$$f(y_{ij})=\frac{\Gamma\left(y_{ij}+\frac{1}{k}\right)}{y_{ij}!\,\Gamma\left(\frac{1}{k}\right)}\left(\frac{1}{1+k(\exp\{\beta_0+x_{ij}{}^t\beta\}}\right)^{\frac{1}{k}}\left(\frac{k(\exp\{\beta_0+x_{ij}{}^t\beta\}}{1+k(\exp\{\beta_0+x_{ij}{}^t\beta\}}\right)^{y_{ij}}$$

B. Estimating of parameters of negative binomial

We estimate κ and β using the maximum likelihood method in order to fit negative binomial equation into US male mortality data.

The likelihood function is given by;

$$L(k,\beta)=\prod_{ij}\frac{\Gamma\left(y_{ij}+\frac{1}{k}\right)}{y_{ij}!\Gamma\left(\frac{1}{k}\right)}\left(\frac{1}{1+k(\exp\{\beta_0+x_{ij}{}^t\beta\}}\right)^{\frac{1}{k}}\left(\frac{k(\exp\{\beta_0+x_{ij}{}^t\beta\}}{1+k(\exp\{\beta_0+x_{ij}{}^t\beta\}}\right)^{y_{ij}}$$

But,

$$\ln\left(\frac{\Gamma\left(y_{ij}+\frac{1}{k}\right)}{\Gamma\left(\frac{1}{k}\right)}\right)=\sum_{z=0}^{y_{ij}-1}\ln\left(z+\frac{1}{k}\right)$$

The log-likelihood function then becomes

$$L=\sum_{1}^{n}\left\{\sum_{z=1}^{y_{ij}-1}\ln\left(z+\frac{1}{k}\right)-\ln\left(\Gamma\left(y_{ij}+1\right)\right)-\left(y_{ij}+\frac{1}{k}\right)\ln\left(1+k\mu_{ij}\right)+y_{ij}\ln\left(\mu_{ij}\right)+y_{ij}\ln(k)\right\}$$

The coefficients of this model are then estimated by taking the first-order conditions and then equate them to zero. Cameron (1998) and lawless (1987) gave the following first-order conditions.

$\frac{\partial L}{\partial \beta_j}=\sum_{i=1}^{n}\frac{x_{ij}\left(y_{ij}-\mu_{ij}\right)}{1+k\mu_{ij}}=0$ j=1,2,3….,k

$\frac{\partial L}{\partial k}=\sum_{i=0}^{n}\left\{k^{-2}\left(\ln\left(1+k\mu_{ij}\right)-\sum_{z=0}^{y_{ij}-1}\frac{1}{z+\frac{1}{k}}\right)+\frac{y_{ij}-\mu_{ij}}{k\left(1+k\mu_{ij}\right)}\right\}=0$

The series of equations can also be solved using Newton-Raphson method.

To obtain the confidence interval of β we take Hessian matrix of the second derivatives of the log-likelihood function.

$\frac{-\partial^2 L}{\partial \beta_r \partial \beta_s}=\sum_{i=1}^{n}\frac{\mu_{ij}\left(1+ky_{ij}\right)x_{ir}x_{is}}{\left(1+k\mu_{ij}\right)^2}$, r,s=1,2,3…..p

$\frac{-\partial^2 L}{\partial \beta_r \partial k}=\sum_{i=1}^{n}\frac{\mu_{ij}\left(y_{ij}-\mu_{ij}\right)x_{ir}}{\left(1+k\mu_{ij}\right)^2}$, r=1,2,3…..p

$$\frac{-\partial^2 L}{\partial k^2}=\sum_{i=1}^{n}\left\{\sum_{z=0}^{y_{ij}-1}\left(\frac{z}{1+kz}\right)^2+2k^{-3}\ln\left(1+k\mu_{ij}\right)-\frac{2k^{-2}\mu_{ij}}{1+k\mu_{ij}}+\frac{\mu_{ij}^2\left(y_{ij}+\frac{1}{k}\right)}{\left(1+k\mu_{ij}\right)^2}\right\}$$

Assuming that the covariance matrix is normally distribution, then the confidence interval of β and k is thus given by;

$$\begin{bmatrix}\hat{\beta}\\ k\end{bmatrix}\sim \mathrm{N}\left(\begin{bmatrix}\beta\\ k\end{bmatrix},\begin{bmatrix}VAR[\beta] & 0\\ 0 & VAR[k]\end{bmatrix}\right),$$

Where,

$\mathrm{VAR}[\beta]=\left(\sum_{i=1}^{n}\frac{\mu_{ij}}{1+k\mu_{ij}}x_i\acute{x}_\iota\right)^{-1}$

$\mathrm{VAR}[k]=\sum_{i=1}^{n}\left\{k^{-4}\left(\ln\left(1+k\mu_{ij}\right)-\sum_{z=0}^{y_{ij}-1}\frac{1}{z+\frac{1}{k}}\right)^2+\frac{\mu_{ij}}{k^2\left(1+k\mu_{ij}\right)}\right\}^{-1}$

C. Description of data

The study focus to establish male mortality rates in the United States using the datasets over a span of 27 years from the 1980 to 2006. Validation set will be from 2008 to 2020 .This datasets will be classified according to sex, age, and time. One-year age group of 0-109 with an open interval for 110+ will be the population size starting from 1980 to 2020; the exposure to risk will be organized in the same manner.

D. Observed males exposed to risk

The study will use male to observe the mortality rate count from 1980 to 2007. These datasets will be grouped into 1x1 age-time interval with some changes that reflect the deaths during the year.

E. Estimation of observed mortality rates

We will estimate rate of mortality by taking ratio between death counts of male over their size of exposure from the year 1980 to 2020. These ratios will be present as log mortality as in the mortality tables.

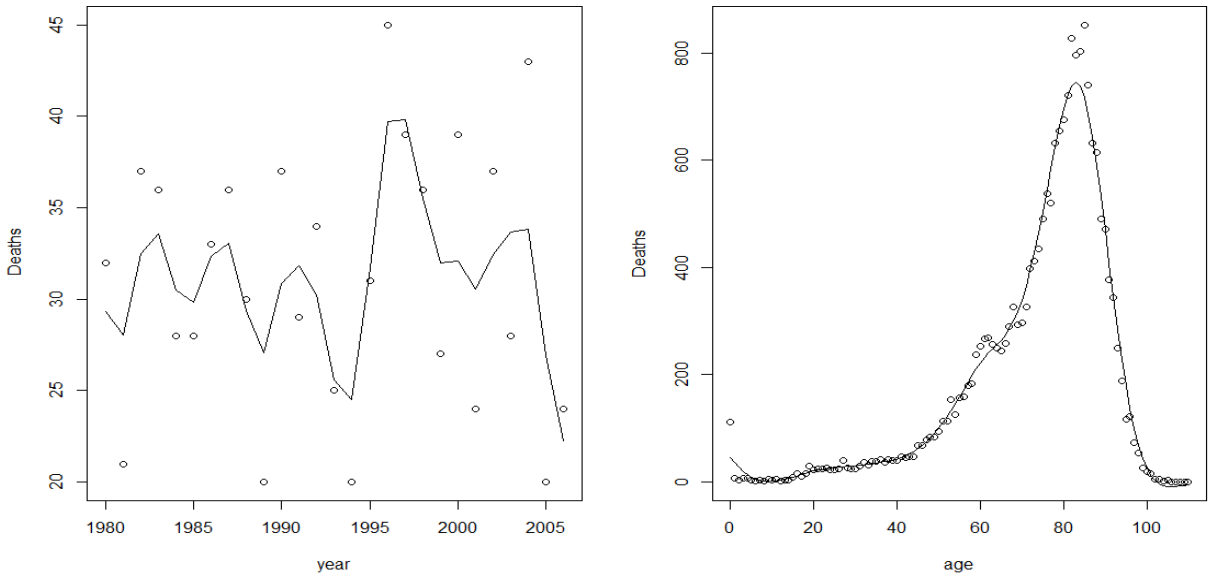
F. Sources of the data

This study will source its data from the Human Mortality Database (maintained by the University of California, Berkeley (U.S.A) www.mortality.org.

IV. RESULTS AND DISCUSSION

Figure 1:

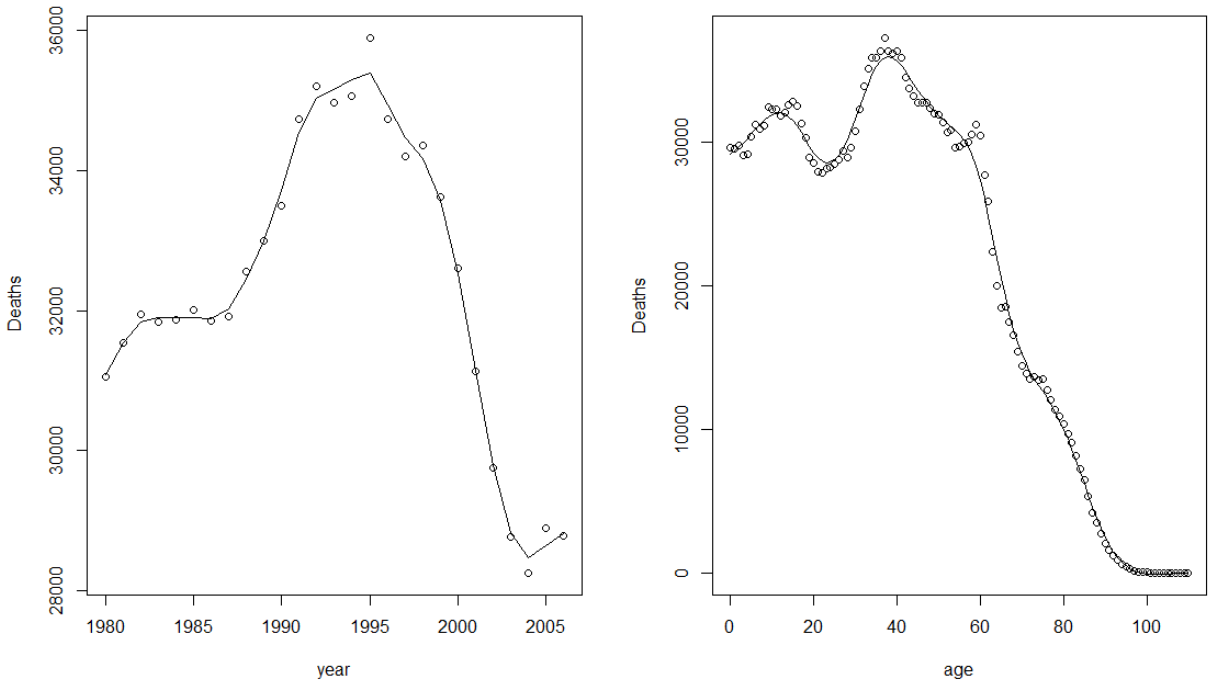
Observed Male Deaths at Age 27 and Year 2006



The figure above shows that the observed count of male deaths in the USA. The highest deaths aged 27 was recorded between year 1995 and 2000 while death observation in calendar year 2006, age group 80 recorded the highest deaths.

Figure 2:

Observed Male Exposure at Age 27 and Year 2006



From these two-dimension plot of male exposure from 1980-2006, 1995 recorded the highest male exposure. It can be seen that the exposure rises at age 20 till age 40 then it steadily drops to zero at age 100.

A. Goodness of fit of the model

In order to fit negative binomial model we fist check normality within the residual terms and the randomness of these elements.

Figure 3:

Goodness of Fit for Negative Binomial (Deviance of Log Number of Death)

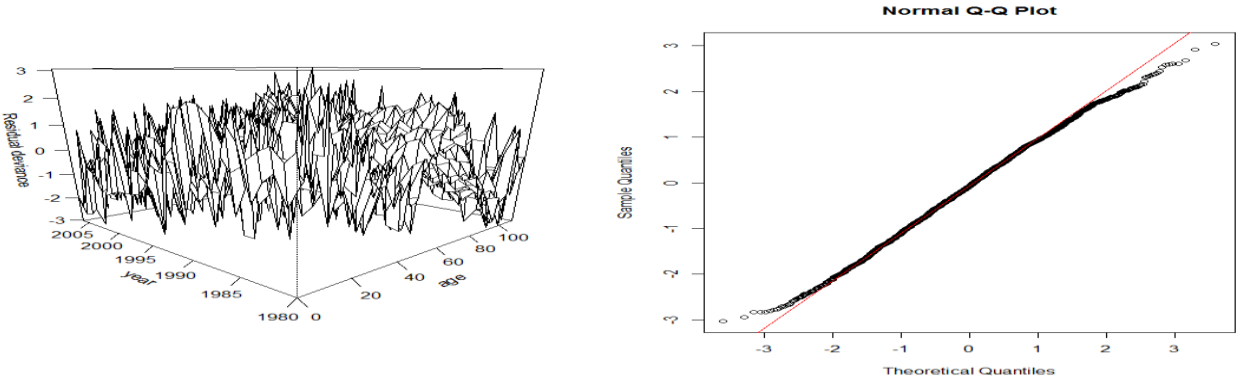
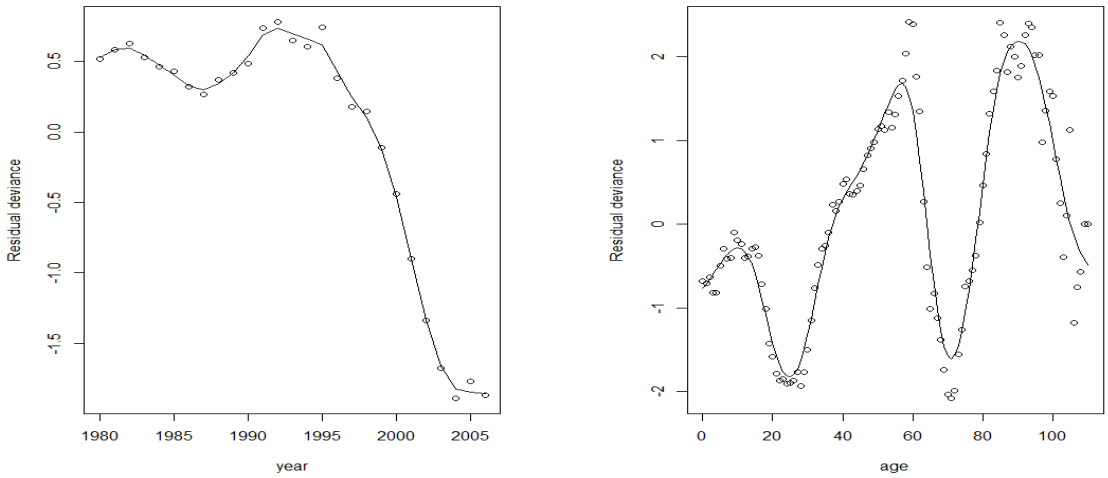


Figure 4:

Deviance Plots of Log Mortality at Age 27 and Year 2006



The above two-dimension plots of deviance shows that the deviance reduced from year 1980 to 1994 then it drastically drops to year 2006 while across age the deviance is random.

Figure 5:

Fitted Against Observed Number of Deaths

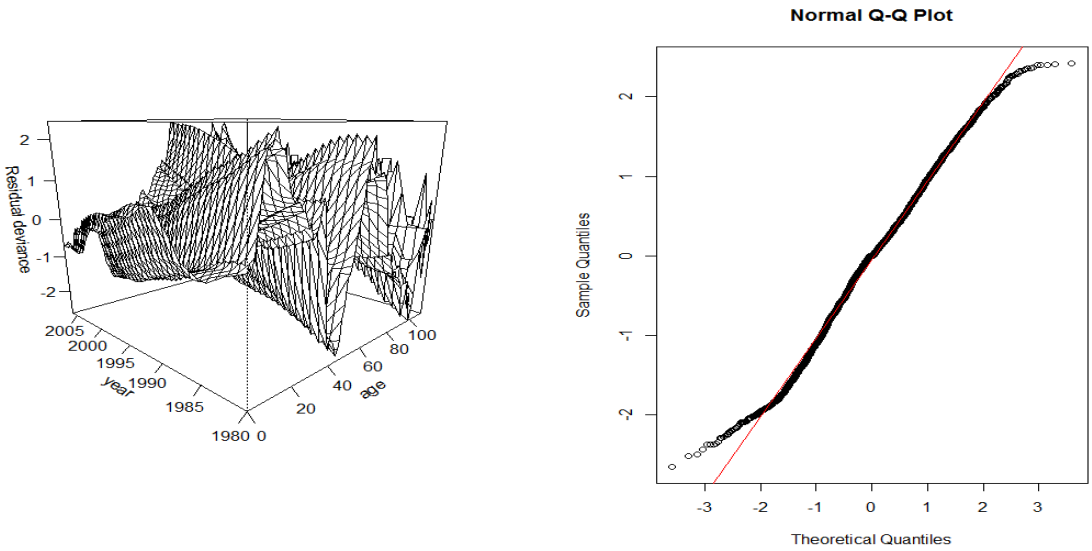
From the above graphs it is clear that negative binomial does not capture well historical trends over time but the model fits well through age.

B. Goodness of risk of exposure

In this section we validate the risk of exposure by again checking the randomness of the deviance and also its normality.

Figure 6:

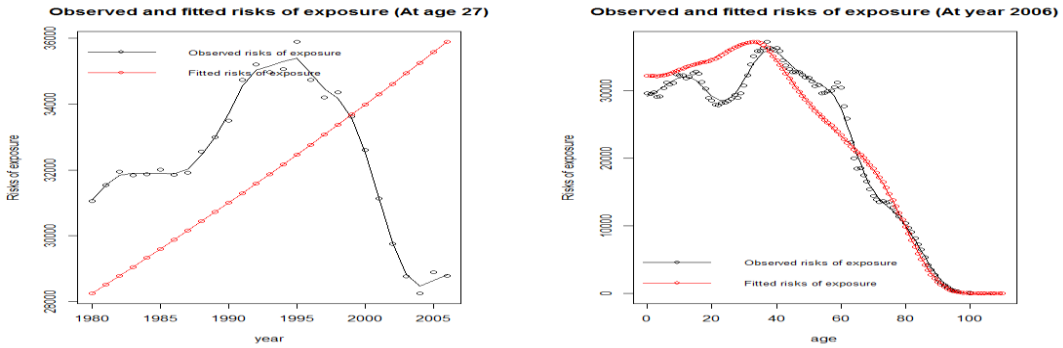
Goodness of Fit for Risk of Exposure



The three-dimension plot above shows that the residual deviance is random. From the Q-Q plots, there is some variation at the beginning and at the end of the plot while most of the deviance is normal.

Figure 7:

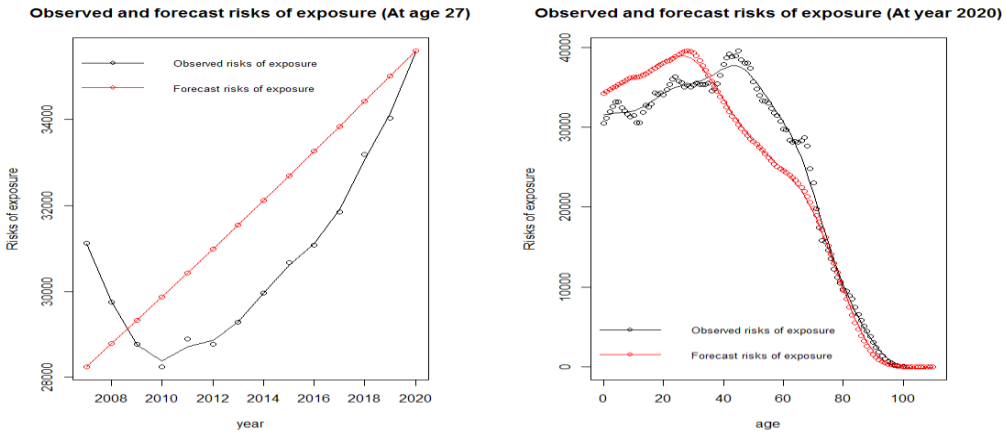
Observed Against Fitted Risk of Exposure in Year 2006 and Age 27



These plots shows that there is lots of variation between fitted and observed risk of exposure over the tome but there is some fit in year 2006 through age. Thus, the model does not capture well the trends of risk of exposure for age group 27 in year 2006.

Figure 8:

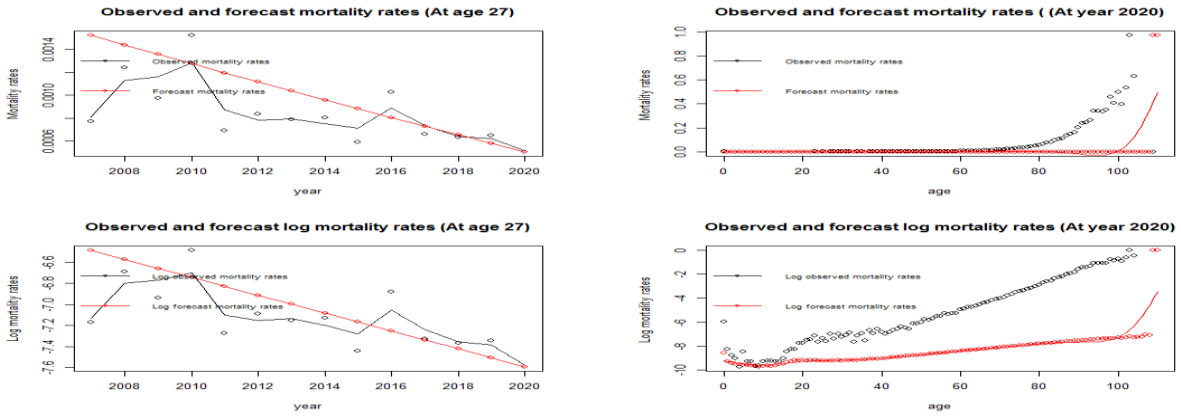
Goodness of Forecast of Exposure with Negative Binomial



From the above graphs we can see that forecasted risk of exposure depart from the historical trends over the years but it considerably fits with age.

Figure 9:

Goodness of Forecast of Mortality Rates



The forecasted mortality throughout the years does not seem to model the historical mortality rate. Fitted mortality decreases linearly while observed mortality increases from year 2008 to 2011 then it gradually drops throughout the years. However, on close inspection, the curve shows that between the year 2016

and 2020 the model seems to be a perfect match. In the year 2020, the model fits well between age 0 and 70 then it departs considerably. The model also gives a poor fit with lo-mortality after age 20.

Conclusion(s)

In the analysis, Negative Binomial gives low mean square error and a better fit with sample data compared with fitted data. The goodness of fit is also reflected in the forecasted data compared with observed data. Negative Binomial with simple polynomial functions gave better forecast over the given period. Thus, we recommend the use of simple polynomial function to model mortality in Negative Binomial

Recommendation(s)

From the above analysis it is evident that Negative Binomial with simple polynomial functions gave a better forecast over the given period. Thus, we recommend the use of simple polynomial function to model mortality.

V. REFERENCES

Bell, W. R. (1997). Comparing and assessing time series methods for forecasting age specific fertility and mortality rates. *Journal of Official Statistics 13(3)* pp. 279-303.

Cameron, A.C., & Travedi, P. K. (1998). *Regression Analysis of Count Data.* Cambridge University Press.

Delwarde, A. (2007) Negative binomial version of Lee-Carter model for mortality forecasting. *Applied Stochastic Models in Business and Industry*. https://doi.org/10.1002/asmb.679.

Di Cesare, M. (2009) Forecasting Mortality. Different approaches for different cause of deaths? The cases of Lung Cancer; Influenza, Pneumonia, and Bronchitis; and Motor Vehicle Accidents. *British Actuarial Journal*.15(S1):185 – 211. https://doi.org/10.1017/S1357321700005560.

Hilbe, J. M. (2007). *Negative binomial regression.* Cambridge University Press.

Hinde, J. and C. G. B. Demetrio (1998). Overdispersion: models and estimation, Computational Statistics and Data Analysis, *27 (2),* pp. 151-170. https://doi.org/10.1016/S0167-9473(98)00007-3

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics, 15 (3)* pp. 209-225.

Valeria, D., Piscopo, G., & Russolillo, M. (2014) Adaptive Neuro-Fuzzy Inference Systems vs. Stochastic Models for Mortality Data. *Smart Innovation, Systems and Technologies*.