Ranking System for Ordinal Longevity Risk Factors using Proportional-Odds Logistic Regression

Nur Haidar Hanafi[1], Puteri Nor Ellyza Nohuddin[2]
Institute of Visual Informatics[1,2]
Universiti Kebangsaan Malaysia
43650 UKM, Bangi
Selangor, Malaysia
Faculty of Computer and Mathematical Sciences[1]
Universiti Teknologi MARA
70300 Seremban
Negeri Sembilan, Malaysia

***Abstract*—Longevity improvements have traditionally been analysed and extrapolated for future actuarial projections of longevity risk by using a range of statistical methods with different combinations of statistical data types. These methods have shown great performances in explaining the trend movements of the longevity rate. However, actuaries believe that knowing the trend movements is not enough, especially in controlling the impact of the longevity risk. Accessing the effects of each level of the risk factors, especially ordinal risk factors, towards the improvements of the longevity rate would provide significant additional knowledge to the trend movements. Therefore, this study was conducted to determine the potentiality of Proportional-Odds Logistics Regression in ranking the levels of the ordinal risk factors based on their effects on longevity improvements. Based on the results, this method has successfully reordered the levels of each risk factor to be according to their effects in improving longevity rate. Hence, a more meaningful ranking system has been developed based on these new ordered risk factors. This new ranking system will help in improving the ability of any statistical methods in projecting the longevity risk when handling ordinal variables.**

***Keywords*—Longevity risk; ordinal risk factors; risk ranking; proportional-odds ratios; effect analysis**

I. Introduction

Longevity improvements indicate a good sign that people are enjoying better fitness and better health conditions than the previous generations. A high life expectancy is considered as a success story to the enforcement of public health policies and socio-economic development systems. The increasing number of older people due to the improvement in life expectancy has undeniably changed the landscapes of every society and economic activity across various industries. These older people have become an increasing consumer group with specific needs and significant spending patterns, especially on health care and mortality protection policies.

Providing readily accessible and available health care and insurance policies can significantly contribute to equal social, economic, political, and cultural participation of these older people. There are increasing demands for insurance companies and private pension managers to bring to the market more products that are suited to the needs and wants of this consumer group. However, according to Hanafi and Nohuddin [1], the policymakers are more concerned about the impact that the life expectancy improvement may have on their financial position and reserve status due to the exposure of longevity risk.

This concern arises due to various negative impacts of longevity risk which include higher financial responsibilities for governments and annuity policy providers [2], risk of outliving resources during old ages for individuals [3][4]; and reduction in the ability of younger members of a family to take care of the older ones due to the extended mobility of the workforce [5]. To address this concern, it is essential for the policymakers to accurately assess the size of longevity risk in ensuring that product pricing, risk management, and asset allocation can be done smoothly.

Several statistical models have been developed to accurately predict longevity risk including cause-of-death specific models, disease-based models, and population-based models. These models were generated based on a diverse spectrum of mortality rate projections tools. Some of the models involved application of life tables [6][7], stochastic mortality models [8], generalized dynamic factor models with vine-copulae simulations [9]; and various data mining techniques including logistic regression technique and decision tree technique [10][11].

Policymakers find these statistical models to be more appealing to them because they help in improving the underwriting process by providing analytics-based approaches using readily accessible customers information to yield a more accurate, consistent, and efficient decision. This helps in simplifying policy applications for smaller face amounts, reducing data acquisition and storing cost, refining underwriting requirements, and processing data via automated software packages. However, actuaries believe that predicting the lonvegity risk movements is not enough, especially in controlling the impact of the longevity risk.

Accessing the effects of each level of the risk factors, especially ordinal risk factors, towards the improvements of the longevity rate would provide significant additional knowledge. Therefore, this study was conducted to determine the potentiality of Proportional-Odds Logistics Regression in ranking the levels of the ordinal risk factors based on their effects on longevity improvements. This new ranking system will help in improving the ability of any statistical methods in projecting

the longevity risk when dealing with ordinal risk factors.

This paper is divided into different important sections to ease the discussion process. The first section discussed the motivation behind this research which include the different risk factors influencing the longevity improvements and the gap analysis on the existing risk ranking systems, especially when using ordinal risk factors. The second section discussed the nature of Proportional-odds Logistic Regression and its potentiality as an alternative to the current risk ranking system. The third section showed the empirical illustration using data on death records with a combination of different risk factors data types. The fourth section combined all the information from the third section into a meaningful risk ranking system. Conclusion and recommendations are then included in the last section.

II. Motivation

A. Risk Factors

Identifying significant risk factors is the most important phase in developing statistical models to predict longevity risk. Various studies have been done to identify significant risk factors in influencing life expectancy improvements. These studies were done using historical data; either non-disease data or genomic data or a combination of both. Non-disease data is readily accessible demographic data from the customers' database while genomic data is medically obtained database such as complete sets of DNA data including all of the customer's gene maps. Different combination of the risk factors would change how the models perform in predicting the longevity risk.

In the underwriting process, selecting a method that is cheaper and producing faster results would benefit the policymakers. Genomic data would require a large space to store and complex purpose-built software to analyse as compared to non-disease data. Therefore, most policymakers would prefer to use non-disease data when developing the models for future projections of longevity risk. Based on past studies using non-disease data, there are five most significant risk factors which include gender, race, residential status, marital status and education level [12][13][14][15][16][4][17][18][2].

B. Risk Ranking System

A major drawback of using statistical models is that they produced very complex methods which are difficult to be explained to the stakeholders and non-statistical practitioners. Therefore, multiple attempts have been done to transform them into a meaningful risk ranking system. A risk ranking system is a phase in the risk management process whereby the identified risks are assessed either using quantitative or qualitative analysis to determine which risks have the highest consequence of occurrence in order of importance. This method is considered to be a simpler mechanism as compared to the more complex statistical methods. It simplifies the estimation of risks, increases the visibility of the risks and assists the policymakers in decision making.

One notable method among a limited number of risk ranking systems that are available is a risk matrix model. A risk matrix model is an m by n risk matrix which provides the assessment of the size of individual risks versus the assessment of group risks along with the amount of overall exposure of the insurance company, particularly with regard to general and life insurance products [19]. One major advantage of this model comes from its relatively simple and transparent characteristics as well as its ability to trace risk trends over time. However, the simplicity of this method leads to inconsistent possible results.

Using qualitative risk parameters in the risk matrix model is a subjective process of numerical interpretation of the risk parameters by means of crisp intervals. This type of interpretation violates the real-life gradual transition between intervals. This problem has long been pointed out and a fuzzy risk ranking approach has been introduced in place of the crisp framework in order to overcome it [20].

A fuzzy model risk ranking system is a system developed by defining risk factors as fuzzy sets [21]. By doing so, an insurance company can utilize multiple prognostic factors that are imprecise and vague. This model provides a more realistic way of modelling longevity risks since it allows for interactions between multiple risk factors at once. Furthermore, it captures expert knowledge and allows for expertise descriptions to be done in a more intuitive and human-like manner.

When dealing with ordinal risk factors, most of the existing risk ranking models have failed to provide a thorough knowledge of the effects of each level of the risk factors towards the improvements of life expectancy. Analysing, extrapolating and scoring the degree of longevity risk while ignoring the substantive features of the risk factors will produce biased results. Requirements for models with the potential to accurately represent the actual characteristics of each of the risk factors should be met.

This study is conducted to assess the potentiality of Proportional-Odds Logistics Regression (POLR) in ranking the ordinal risk factors' levels based on their individual effects on longevity improvements. Hence, a more meaningful ranking system can be developed based on these new ordered risk factors. This new ranking system will help in improving the ability of any statistical methods in projecting the longevity risk when handling ordinal variables.

III. Proportional-Odds Logistic Regression

An ordinal regression model is a regression model specifically developed for ordinal dependent variables based on discrete or continuous covariates. It is similar to binary and multinomial logistic regression whereby it uses the same iterative procedure called maximum likelihood estimation. There are several types of ordinal logistic regression models. The most frequently used in practice is the proportional odds model [22][23][24].

$$logit[P(Y \leq j)] = \alpha_j - \Sigma\beta_i X_i \tag{1}$$

The above equation represents the proportional-odds models where $j = 1, \cdots, J-1$ and $i = 1, \cdots, M$. Let Y denotes the response category in the range $1, 2, \cdots, j$ such that $j \geq 2$. On the right side of the equation is a simple linear model with one slope, β, and an intercept α_j that changes depending on the category j in which the intercepts depend on j, but the

slopes are all equal. According to this equation, the model is basically generating the probability of being in one category lower level versus being in categories above it.

Some advantages of the logit link are worth to be mentioned here. The model yields constant odds ratios across each split with interpretations very similar to logistic regression. It represents both orderings as well as categorical nature without any substantial increase in the difficulty of interpretation, such that it decreases variability and increases interpretability of the subject matter. Thus, the model has fewer terms than a multinomial regression model.

IV. Empirical Illustration

The analysis procedures illustrated in this section play a major purpose in highlighting the potentiality of POLR in ranking the ordinal risk factors' levels based on their effects on longevity improvements. The POLR has the ability to capture the strength of the effects that the independent variables have on a given dependent variable, thus it can be used to understand how much the dependent variable changes when the independent variables are changed.

A. Data Descriptions

The death records of Americans who had died at the age of 70 and above from 2013 until 2015 were used in this study with a total of 3,765,210 recorded deaths. Only deaths that occurred at the age of 70 and above were selected for this study because the size of longevity risk is more significant amongst those within this age group. Such datasets were used because they represent real-life risk exposures of people being able to live and die beyond their life expectancy; including both typical and extreme ones.

Any unnatural deaths would disturb the accuracy of this study, thus only deaths caused by natural events were included; while those caused by suicide, homicide or other unnatural events were removed. Only six non-disease risk factors were included in this study which is the age at death, gender, race, residential status, marital status and education level. These variables are a combination of nominal and ordinal variables without any range variables. Age at death was selected as the dependent variable where each age group represents the level of longevity risk exposure.

The descriptions of these non-disease risk factors along with their variable types and codes as coded in the RStudio software are presented in Table I. The R software version 1.1.383 installed in Windows 10 with processor Intel(R) Core (TM) i7-6500U CPU at 2.5GHz was used in the analysis process. Different R packages have been used for various purposes which included readr package for reading rectangular data [25], plyr package for splitting, applying and combining data [26], ggplot2 package for data visualisation [27], ordinal package for ordinal regression modelling [28], stats package for running a Kruskal-Wallis Test [29]; and effects package for effect displays [30][31][32].

B. Descriptive Analysis

The basic features of each risk factor with respect to each category of the dependent variable based on the percentage of deaths was visualised using bar chart plots as shown in Figure 1. Separating each risk factor based on their item composition is an important process in getting to know the structure of this dataset, thus highlighting potential relationships between variables and giving extra information for some early presumptions of the longevity risk exposure.

TABLE I. Data descriptions.

Variable	Description	Variable Type
Age_Death	The age at death; recorded as a single age, was then discretized into 5-year age groups so that it became categorical data. {Age_Death = "70-74", "75-79", ..., "100+"}	Ordinal
Residential	{Residential = "Residents", "Intrastate NR", "Interstate NR", "Foreign R"}	Nominal
Education	The education status of the deceased; recoded using the revised 2003 education codes. {Education = "Primary", "Secondary", "Diploma/GED", "Degree", "Master", "PHD/Professional"}	Ordinal
Gender	{Gender = "Female", "Male"}	Nominal
Marital_Status	{Marital_Status = "Single", "Married", "Widowed", "Divorced"}	Nominal
Race	{Race = "White", "Black", "American Indian", "Asian/Pacific Islander"}	Nominal

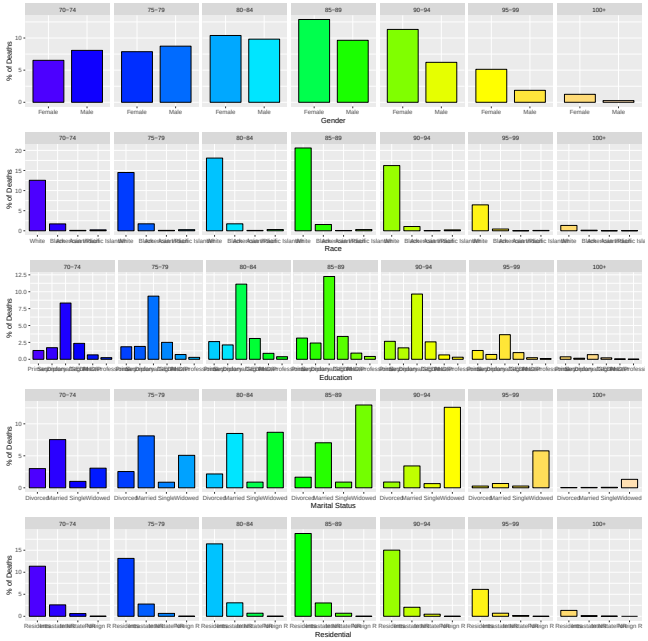

Fig. 1. Data composition of each risk factor with respect to age at death.

According to Figure 1, the majority of the deaths among citizens aged 70 up to age 79 were male before being dominated by female starting from age 80 onwards. White coloured people, having diploma/GED as their highest education level and residents of the United States of America have recorded the highest percentage of deaths for all age groups. From ages 70 to 79, the majority of the deaths occurred among married people before being dominated by widowers starting from age 80 onwards.

The correlation between the dependent variable and each risk factor is represented by mosaic plots as shown in Figure 2. The size of the boxes corresponds to the size of death records within each level of each risk factor as in Figure 1. The

shading density of each box is based on the Pearson residual values. The blue and red coloured boxes represent the level of the residual for that category. More specifically, a blue coloured box means that more observations in that box that would have been expected under the null model; whereas a red coloured box means that there are fewer observations in that box than the one would have been expected. This is known as positive and negative relationships between each category of the dependent variable with each category of the risk factors.

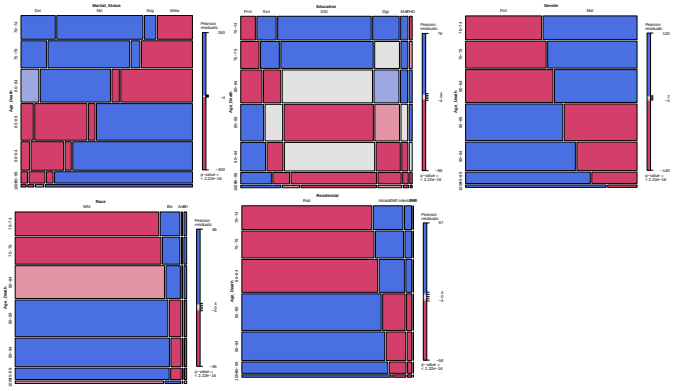

Fig. 2. Classic mosaic plot for correlation between age at death and the risk factors.

The p-value of the Pearson Residuals indicates the significant departure from the independence of the association between any given two risk factors. Therefore, when a p-value is greater than 0.05 then there is no association between the two risk factors. From Figure 2, the values of Pearson residuals for all the risk factors used in this study are less than 0.05. This result shows that all five risk factors are significantly affecting the deceased's life expectancy.

There are a few interesting findings that could be discussed further with respect to Figure 2. Even though all risk factors are significantly associated with age at death as proven by the p-values, there exist some levels within the Education risk factor that are not associated with age at death. This condition can be seen from grey coloured boxes. For example, owning a diploma/GED certificate did not affect those who died between ages 80 to 84 and between ages 90 to 94.

C. Proportional-Odds Assumption

One of the assumptions underlying POLR is that the relationship between each pair of the dependent variable's categories is the same, thus it is called a parallel regression assumption. A POLR study can only be done if and only if this assumption is checked to ensure that the coefficients of the relationship between the lowest category versus all higher categories of the dependent variable are the same as those that describe the relationship between the next lowest category and all higher categories.

Figure 3 shows the graphical tool used in assessing the parallel slopes assumption using all observations. Since the relationship between all pairs of the dependent variable's category is the same, there is only one set of coefficients. The plots as displayed in this graph are predictions from the logit model used to model the probability that ages at death is greater than or equal to a given value using one risk factor at a time. The normalisation of all the first set of coefficients by setting them to be zero was done so that there is a common reference point for all risk factors.

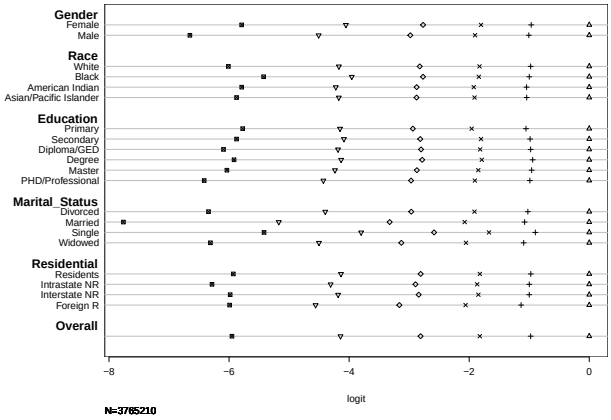

Fig. 3. Parallel slopes assumption.

If the proportional-odds assumption holds, for each risk factor, the distance between the symbols for each set of categories of the dependent variable should remain similar. Looking at the coefficients for the variables paired in Figure 3, it can be seen that the distance between all sets of coefficients are almost similar which indicate that this dataset met the requirement for the proportional-odds assumption. In contrast, the markers are much further apart on the line for the married category under the marital status risk factor which suggests that this assumption may not hold only for this particular item.

D. POLR Model Fitting

The standard interpretation of the ordered logit coefficient is that for a one-unit increase in a particular risk factor, the dependent variable level is expected to change by its respective regression coefficient in the ordered log-odds scale given that the other risk factors in the model are held constant. The coefficients from the model can be somewhat difficult to interpret because they are scaled in terms of logs. Therefore, the coefficients are converted into proportional-odds ratios. The proportional-odds ratios are interpreted pretty much as would be for the odds ratios from a binary logistic regression.

The results of the POLR model fitting for this study as produced by RStudio can be found in Figure 4. The p-value for all the risk factors is less than 0.05, hence they are statistically significant in influencing age at death at a 95% confidence interval. These findings are similar to the findings for the Pearson correlation between the dependent variable and each risk factor in Figure 2.

The noticeable difference between the results produced by the multinomial regression model as compared to the results produced by the POLR model can be seen from the generation of intercept values between two categories of the dependent variable. These values represent the relationship between the lowest category versus all higher categories of the dependent

variable. Mathematically, the intercept $70-74|75-79$ corresponds to $logit[P(Y \leq 1)]$ and can be interpreted as the log of odds of occurrence that a person is going to die between age 70 to 74 versus the log of odds of occurrence that a person is going to die within other age groups.

```
Call:
  polr(formula = Age_Death ~ Gender + Race + Education + Marital_Status
       + Residential, data = trainingOLR, Hess = TRUE)

Coefficients:
                                  Value  Std. Error     t value       p value
GenderMale                -0.170377795 0.002411575  -70.650013  0.000000e+00
RaceBlack                 -0.500173926 0.004041636 -123.755325  0.000000e+00
RaceAmerican Indian       -0.738564500 0.016678078  -44.283550  0.000000e+00
RaceAsian/Pacific Islander -0.247583738 0.009142014  -27.081968 1.605887e-161
Education.L                0.116151395 0.005744923   20.218096  6.785178e-91
Education.Q                0.372753990 0.005014475   74.335598  0.000000e+00
Education.C               -0.127579365 0.004732022  -26.960855 4.255488e-160
Education^4                0.137140806 0.004138242   33.139869 7.925691e-241
Education^5                0.004914821 0.002882278    1.705186  8.815970e-02
Marital_StatusMarried      0.286213236 0.003920787   72.998927  0.000000e+00
Marital_StatusSingle       0.570099259 0.006275987   90.838189  0.000000e+00
Marital_StatusWidowed      1.679247582 0.003894953  431.134273  0.000000e+00
ResidentialIntrastate NR  -0.347892683 0.003130040 -111.146414  0.000000e+00
ResidentialInterstate NR  -0.312632246 0.006095587  -51.288289  0.000000e+00
ResidentialForeign R      -1.048515742 0.038922680  -26.938426 7.795082e-160

  Intercepts:
                                  Value  Std. Error     t value       p value
70-74|75-79               -1.305594694 0.004179443 -312.384846  0.000000e+00
75-79|80-84               -0.232103229 0.004076751  -56.933377  0.000000e+00
80-84|85-89                0.739988309 0.004110147  180.039395  0.000000e+00
85-89|90-94                1.863261075 0.004232425  440.234849  0.000000e+00
90-94|95-99                3.300080066 0.004603971  716.789995  0.000000e+00
95-99|100+                 5.150912337 0.006497250  792.783490  0.000000e+00
```

Fig. 4. The POLR model fitting.

The coefficients were then converted into proportional-odds ratios for ease of interpretation of the results since they are interpreted pretty much the same as the odds ratios from a binary logistic regression. The values of the proportional-odds ratios for each category of each risk factor can be found in Figure 5 below. These proportional-odds ratios were calculated based on a pre-selected baseline category from each risk factor, thus all interpretations must be done based on this pre-selected baseline. For example, for male, the odds of being more likely to live longer is 15.7% lower than female, given that all other risk factors are constant. Another example of interpretation, for widowers, the odds of being more likely to live longer is 5.36 times that of single people, given that all other risk factors are constant.

```
                                   OR     2.5 %    97.5 %
GenderMale                  0.8433461 0.8394620 0.8472413
RaceBlack                   0.6064252 0.6017385 0.6111592
RaceAmerican Indian         0.4777993 0.4624863 0.4936622
RaceAsian/Pacific Islander 0.7806848 0.7668627 0.7946764
Education.L                 1.1231659 1.1108266 1.1355783
Education.Q                 1.4517272 1.4376506 1.4658213
Education.C                 0.8802236 0.8732524 0.8882488
Education^4                 1.1469896 1.1380117 1.1560739
Education^5                 1.0049269 0.9994453 1.0104417
Marital_StatusMarried       1.3313763 1.3214369 1.3414066
Marital_StatusSingle        1.7684426 1.7487420 1.7882863
Marital_StatusWidowed       5.3615203 5.3218538 5.4013225
ResidentialIntrastate NR    0.7061747 0.7019582 0.7104095
ResidentialInterstate NR    0.7315189 0.7229358 0.7402352
ResidentialForeign R        0.3504575 0.3248096 0.3783115
```

Fig. 5. The proportional-odds ratios and their 95% confidence intervals.

Some interesting findings can be discussed further from the results in Figure 5. For gender risk factor, a female American is more likely to live longer than a male American. For race risk factor, Black, American Indian and Asian/Pacific Islander people are less likely to live longer than White people; with American Indian to have more than 50% fewer odds compared to White. For education risk factor, a person with degree qualification is less likely to live longer than those with primary school qualification. For the other education categories, they are more likely to live longer than those who have primary school qualification. For marital status risk factor, being married, single or widowed are more likely to live longer compared to being divorced. For residential risk factor, intrastate non-resident, interstate non-resident or foreign resident is less likely to live longer than those who are a permanent resident of the United States.

The values of these proportional-odds ratios were then used to generate the ranking of longevity risk based on the likeliness of each category of each risk factor in influencing the longevity risk with respect to the pre-selected baseline category. However, depending only on these values would only generate a one-way ranking because the arrangements of the ranking were done based on only one category. A good risk ranking must also be able to explain the interaction between each category of each risk factor with each other. Thus, an effect analysis was carried out to overcome this situation.

E. Effect Analysis with Visualisation

The main purpose of doing an effect analysis is to determine the interactions between each category of each risk factor with each other. The first step in doing such analysis is to check if there are statistically significant differences between the independent variables. The most common statistical tests for analysing such relationships is the ANOVA test. However, using an ordinal data type as the dependent variable means that the assumption that the data follows a normal distribution will be violated. Given that the assumption of normality is violated, a typical ANOVA test in this situation would at best lack sensitivity, and at worst provide spurious estimates.

Fortunately, there are non-parametric versions of the ANOVA test which do not depend on the assumption of normality, and so are quite suitable for the ordinal data type. One of them is the Kruskal-Wallis test which is most appropriate for statistical testing between an ordinal level dependent variable and nominal level independent variables. This test assumes that the population has the same distribution, except for a possible difference in the population medians. The cross-sectional test for the dataset used in this study produced p-values of $2.2e^{-}16$ for each combination of risk factors. This shows that all of the independent variables are significantly different from one another with the p-value of each interaction is less than 0.05. Thus, the outcome of the dependent variable is influenced by each independent variable without any disturbance from the relationships among them.

Visualizing how the dependent variable responses with the changes across different level of the independent variables through effect display would help in determining the interactions between each category of each risk factor with each other. The effect display is carried out by allowing the independent variables to range over their combinations of values while holding other independent variables at fixed values. Figure 6 until Figure 15 show the effect of changing the value of the independent variables on the probability of being included

into each level of the dependent variable based on a different combination of the independent variables while holding the other independent variables at fixed values.

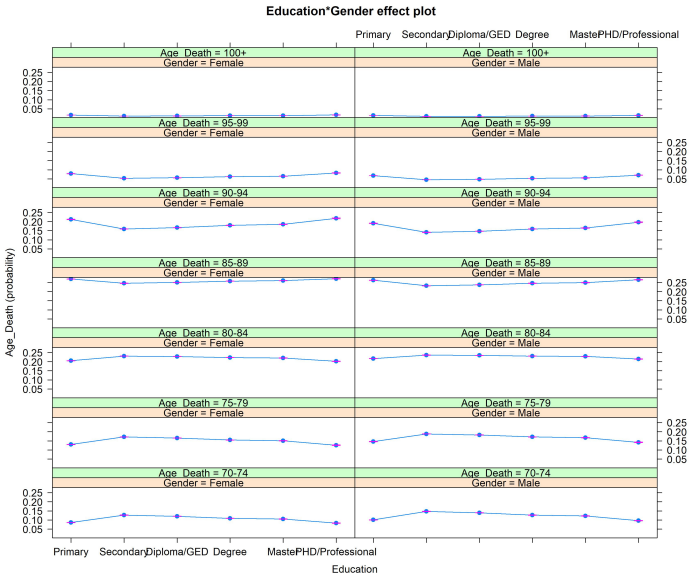

Fig. 6. Gender versus Race effect plot.

Based on Figure 6, it was found that, regardless of gender and race, the probability of dying at the age of 100 and above stays low and constant. Being a White person would increase the probability of dying within the age group of 85 to 89, regardless of gender. Changing the race value from White to Black would reduce the probability of dying between the age of 85 to 94, regardless of gender, with only a slight reduction in probability for the age group of 95 to 99. Being a male American would have more significant in the probability of dying between the age of 70 to 79, with a slight increment in probability for age 80 to 84. No dramatic differences between female and male categories.

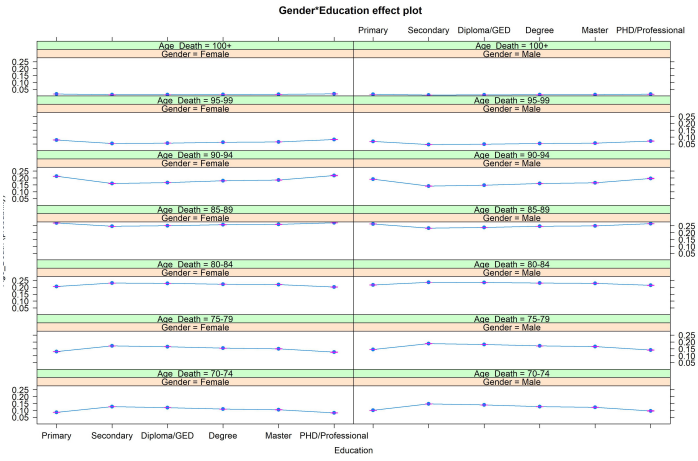

Fig. 7. Gender versus Education effect plot.

The effect of gender and education on the age of death also reveals some interesting information, as in Figure 7. The effect of changing the values of education level shows only small differences in the probability of dying across all ages for both genders. The probability of dying within the age group of 85 to 89 is dramatically high, regardless of gender and education level. The probability of dying at the age of 100 and above stays low and constant, regardless of gender and education level. There is a major drop in the probability of dying between ages 90 to 94 given that the level of education is changed from primary to secondary for both genders.

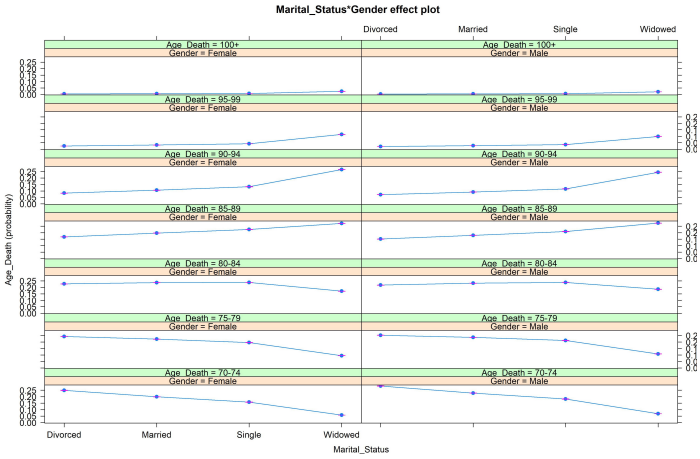

Fig. 8. Gender versus Marital Status effect plot.

Figure 8 shows the effect of gender and marital status on the age of death. The probability of dying at the age of 100 and above stays low and constant, regardless of gender and marital status. No significant differences in the effect of changing marital status on the age of death for both genders. There are some dramatic changes in the probability of dying for all age groups when the marital status is changed from single to widowed for both genders, especially for the age group of 70 to 74 and 90 to 94. There exist some reversed effects between these two age groups across all levels of marital status, for both genders.

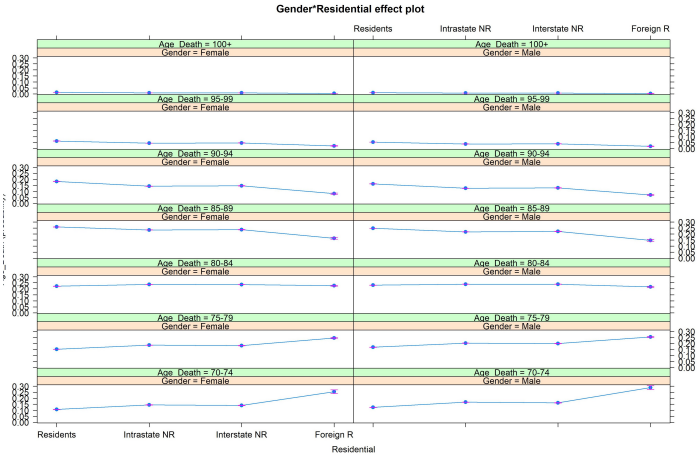

Fig. 9. Gender versus Residential effect plot.

The probability of dying at the age of 100 and above stays low and constant, regardless of gender and residential status, as shown in Figure 9. The effect of changing the values for

residential status, for both genders, on the probability of dying within the age group of 95 to 99 and 80 to 84, is almost not visible. The effect of intrastate non-resident and interstate non-resident is almost the same for all age groups. However, dramatic changes can be seen between residents and foreign residents for all age groups regardless of gender. The most obvious effect could be found for the age group of 70 to 74.

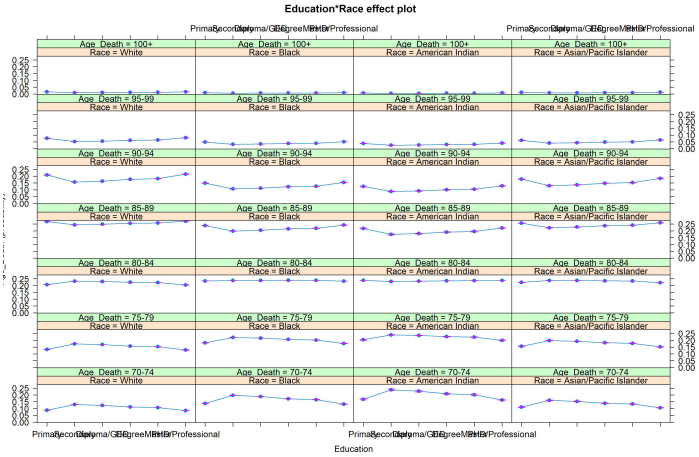

Fig. 10. Race versus Education effect plot.

Figure 10 shows the effect analysis for race and education level on the probability of dying for all age groups. The probability of dying at the age of 100 and above stays low and constant regardless of race and education level. The effect of all levels of education is the same across race on the probability of dying within the age group of 80 to 84 and 95 to 99. The probabilities of dying within the age group of 70 to 74 and 75 to 79 are highly affected by the changes in the education level of Black and American Indian people. The same thing can be said for the age group of 90 to 94, but with a reversed effect. The probability of a White person to die within the age group of 85 to 89 is constantly high, regardless of the level of education.

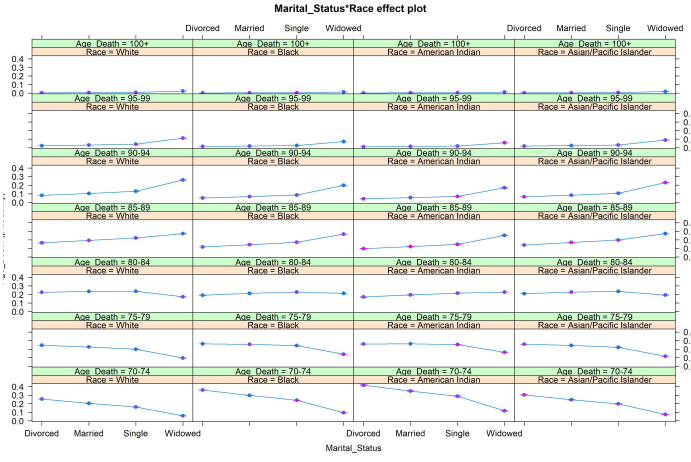

Fig. 11. Race versus Marital Status effect plot.

Figure 11 shows the effect of changing the values in the race and marital status on the probability of dying within each age group. The probability of dying at the age of 100 and above stays low and constant, regardless of race and marital status. Dramatic changes in the probability of dying for every age group are visible if there are changes in the marital status for all races. Being a divorced American Indian has the highest probability of dying within the age group of 70 to 74, whereby being a widowed person would have a higher probability of dying within higher age groups, regardless of race.

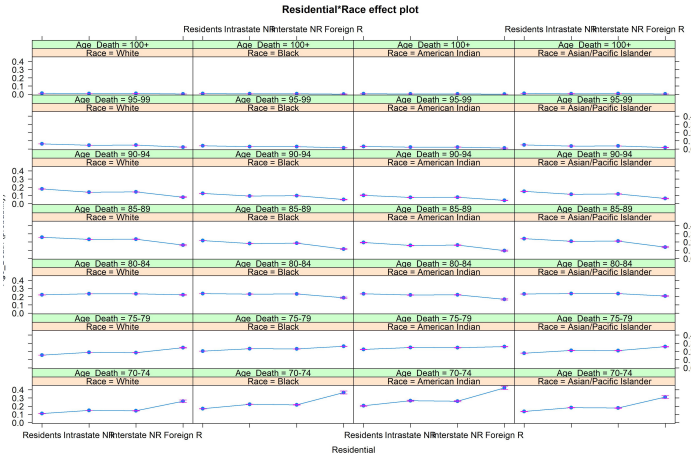

Fig. 12. Race versus Residential effect plot.

Figure 12 shows the effect of changing the values in the race and residential status on the probability of dying within each age group. The probability of dying at the age of 95 and above stays low and constant, regardless of race and residential status. The probability of dying within the age group of 80 to 84 stays moderate and constant, regardless of race and residential status. Dramatic changes in the probability of dying for the other age groups are visible, given that there are changes in the residential status for all races. The most dramatic changes can be seen in the probability of dying within the age group of 70 to 74 across all races. Based on this, foreign residents would have a higher probability of dying within this age group as compared to residents across all races, with the most dramatic increment for American Indian.

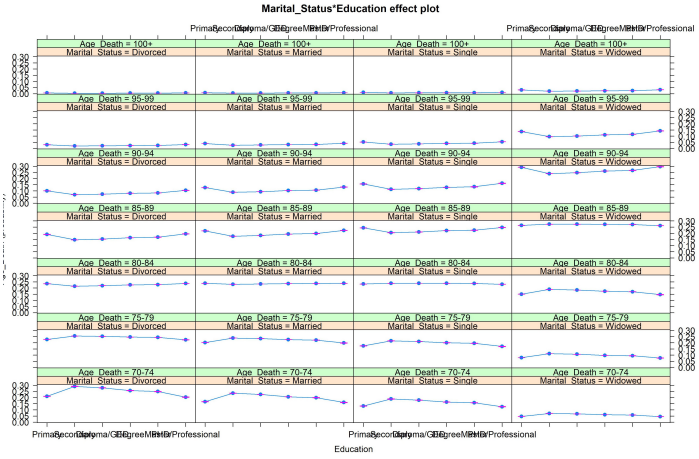

Fig. 13. Marital Status versus Education effect plot.

Figure 13 shows the effect of changing the values in marital status and education level on the probability of dying within

each age group. The probability of dying at the age of 100 and above stays low and constant, regardless of marital status and education level. The effect of education levels stays low for age group 95 to 99 for divorced, married and single people. The most dramatic changes can be seen in the probability of dying within the age group of 70 to 74 for all education levels. Changing the marital status from divorced to widowed would dramatically reduce the probability, regardless of the education level. Being widowed will also increase the probability of dying within the age group of 90 to 94 which has recorded a very high increment.

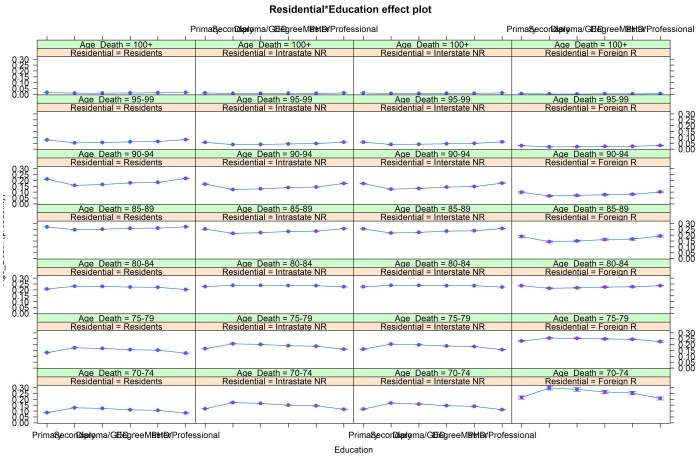

Fig. 14. Residential versus Education effect plot.

Figure 14 shows the effect of changing the values in residential status and education level on the probability of dying within each age group. The probability of dying at the age of 95 and above stays low and constant, regardless of residential status and education level. A reversed situation can be found in the age group of 80 to 84, whereby the probability of dying within this age group stays high and constant. The effect of changing the education level among foreign residents for the age group of 70 to 74 is very significant, whereby the probability of dying within this age group is higher than the other residential statuses.

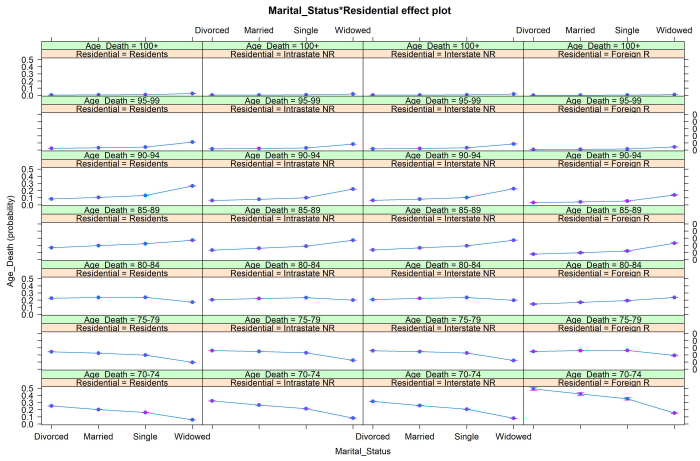

Fig. 15. Residential versus Marital Status effect plot.

Figure 15 shows the effect of changing the values in marital status and residential status on the probability of dying within each age group. The probability of dying at the age of 100 and above stays low and constant regardless of marital status and residential status. The probability of dying within the age group of 95 to 99 shows a slight increment between single and widowed people for all residential statuses. Dramatic changes in the probability of dying can be found for the age group of 70 to 74 across all marital statuses with the highest changes could be seen between residents and foreign residents. Changing the marital status of divorced to married would significantly reduce the probability of dying within this age group.

V. Longevity Risk Ranking

The findings from the odds-ratio analysis, along with the information abstracted from the effect display plots, were transformed into a meaningful risk ranking system, as shown in Figure 16. All levels of the risk factors were rearranged according to the effect level that they have on the probability of dying for each age group, whereby a higher age group represents a higher level of longevity risk. Based on the risk ranking diagram as shown in Figure 16, a high level of longevity risk is predicted to exist in those who are a White female widowed and the resident of the United States with the highest education level is diploma/GED. Those with such profile are predicted to live longer as compared to the other profiles.

One interesting characteristic of this risk ranking system is on the nature of each risk factor. At the beginning of this study, all the risk factors were treated as nominal variables. Rearranging them according to their level of effect on the probability of dying for all age groups has transformed them to be ordinal risk factors, thus improved their ability in providing more knowledge on the longevity risk. This ranking system is considered as a simpler mechanism compared to the more complex statistical methods. It simplifies the estimation of risks, increases the visibility of the risks and assists the policymakers in decision making.

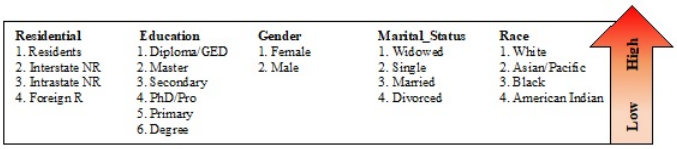

Fig. 16. Level of longevity risk according to the rank of each level for each variable.

VI. Conclusion and Recommendations

The POLR model has been proven to have good potential as an alternative model in ranking the ordinal risk factors' levels based on their individual effects on longevity improvements. A more meaningful ranking system has been developed based on a set of ordinal non-disease risk factors. This new ranking system has the potentiality of improving the ability of any statistical methods in projecting the longevity risk when using ordinal variables. Thus, it increases the visibility of the longevity risk and could be used to assist the policymakers in decision making.

However, some limitations need to be highlighted in this study. There is no means of classifying this ranking system and comparing it with the other countries, thus it's generalisation cannot be proved further and can only be applied for US longevity risk projections. As a recommendation for future researchers, the same methods used in this study could be applied to data from other countries and a comparative study should be conducted in comparing the findings from this study with other countries. It would be better if the procedures in this study could be replicated using data from various insurance companies and pension providers to see the impact that this method has in projecting the longevity risk within these industries.

REFERENCES

[1] N. H. Hanafi and P. N. E. Nohuddin, "Data Mining Approach in Mortality Projection: A Review Study," Advanced Science Letters, vol. 23, no. 3, pp. 1612-1615, 2018.

[2] A. Bozikas and G. Pitselis, "An empirical study on stochastic mortality modelling under the age-period-cohort framework: The case of Greece with applications to insurance pricing," Risks, vol. 6, no. 2, p. 44, 2018.

[3] R. Zugic, G. Jones, C. Yiasoumi, K. McMullan, A. Tacke, M. Held and B. Moreau, "Longevity CRObriefing Emerging Risks Initiative Position Paper," CRO Forum, Amsterdam, 2010.

[4] S. Haberman, V. Kaishev, P. Millossovich, A. Villegas, S. Baxter, A. Gaches, S. Gunnlaugsson and M. Sison, "Longevity basis risk: A methodology for assessing basis risk," The Institute and Faculty of Actuaries & The Life and Longevity Markets Association, London, 2014.

[5] J. Bravo, P. Real and C. Silva, "Participating life annuities incorporating longevity risk-sharing arrangements," Portugal, 2009.

[6] A. D. Lopez, J. Salomon, O. Ahmad, C. J. Murray and D. Mafat, "Life tables for 191 countries: data, methods and results," World Health Organization, Geneva, 2001.

[7] R. I. Ibrahim, "Expanding an abridged life table using the Heligman-Pollard model," MATEMATIKA: Malaysian Journal of Industrial and Applied Mathematics, vol. 24, pp. 1-10, 2008.

[8] M. Denuit and J. Trufin, "From regulatory life tables to stochastic mortality projections: The exponential decline model," Insurance: Mathematics and Economics, vol. 71, pp. 295-303, 2016.

[9] H. Chulia, M. Guillen and J. M. Uribe, "Modelling longevity risk with generalized dynamic factor models and vine-copulae," ASTIN Bulletin: The Journal of the IAA, vol. 46, no. 1, pp. 165-190, 2016.

[10] L. Guo and M. C. Wang, "Data Mining Techniques for Mortality at Advanced Age," 2007.

[11] L. Guo, "Predictive Modeling for Advanced Age Mortality," in Living to 100 and Beyond Symposium, Florida, 2008.

[12] E. Daly, A. Mason and M. J. Goldcare, "Using mortality rates as a health outcome indicator: A literature review. Report to the Department of Health," National Centre for Health Outcomes Development, Oxford, 2000.

[13] S. Vinnakota and N. S. Lam, "Socioeconomic inequality of cancer mortality in the United States: a spatial data mining approach," International journal of health geographics, vol. 5, no. 1, p. 9, 2006.

[14] M. Heron, "National vital statistics reports," National Center for Health Statistics, 2007.

[15] World Health Organization, "Global health risks: mortality and burden of disease attributable to selected major risks," World Health Organization, Geneva, 2009.

[16] P. Berry, L. Tsui and G. Jones, "Our New 'Old'Problem–Pricing Longevity Risk in Australia," in 6th International Longevity Risk and Capital Markets Solutions Conference, Sydney, 2010.

[17] Office for National Statistics, "Mortality Statistics: Metadata," Office for National Statistics, South Wales, 2015.

[18] D. Allen and S. Lee, "Modelling Life Insurance Risk Prudential Insurance Data Set," in SAS Student Symposium Forum, 2018.

[19] D. Drljača, "Risk assessment through matrix model in insurance companies," Poslovna ekonomija, vol. 10, no. 2, pp. 43-65, 2016.

[20] O. Abul-Haggag and W. Barakat, "Application of fuzzy logic for risk assessment using risk matrix," International Journal of Emerging Technology and Advanced Engineering, vol. 3, no. 1, pp. 49-54, 2013.

[21] P. Horgby, "Risk classification by fuzzy inference," The Geneva Papers on Risk and Insurance Theory, vol. 23, no. 1, pp. 63-82, 1998.

[22] R. Brant, "Assessing proportionality in the proportional odds model for ordinal logistic regression," Biometrics, pp. 1171-1178, 1990.

[23] R. Bender and U. Grouven, "Using binary logistic regression models for ordinal data with non-proportional odds," Journal of clinical epidemiology, vol. 51, no. 10, pp. 809-816, 1998.

[24] R. Williams, "Generalized ordered logit/partial proportional odds models for ordinal dependent variables," The Stata Journal, vol. 6, no. 1, pp. 58-82, 2006.

[25] H. Wickham, J. Hester and R. Francois, "readr: Read Rectangular Text Data," 2018.

[26] H. Wickham, "The Split-Apply-Combine Strategy for Data Analysis," Journal of Statistical Software, vol. 40, no. 1, pp. 1-29, 2011.

[27] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, New York: Springer-Verlag New York, 2016.

[28] R. H. B. Christensen, ordinal: Regression Models for Ordinal Data, 2019.

[29] R Core Team, R: A Language and Environment for Statistical Computing, Vienna: R Foundation for Statistical Computing, 2019.

[30] J. Fox and J. Hong, "Effect Displays in R for Multinomial and Proportional-Odds Logit Models: Extensions to the effects Package," Journal of Statistical Software, vol. 32, no. 1, pp. 1-24, 2009.

[31] J. Fox and S. Weisberg, "Visualizing Fit and Lack of Fit in Complex Regression Models with Predictor Effect Plots and Partial Residuals," Journal of Statistical Software, vol. 87, no. 9, pp. 1-27, 2018.

[32] J. Fox and S. Weisberg, An R Companion to Applied Regression, 3rd ed., CA: Thousand Oaks, 2019.