# Extreme value analyses of US P&I mortality data under consideration of demographic effects

Huey Chiy LEE
Hans WACKERNAGEL

Juillet 2007

Centre de Géosciences
Ecole des mines de Paris, Fontainebleau, France

# Extreme value analyses of US P&I mortality data under consideration of demographic effects

Huey Chiy LEE
Hans WACKERNAGEL

**CONFIDENTIEL**

No. R071113HLEE

Centre de Géosciences
Ecole des mines de Paris, Fontainebleau, France

Ecole des Mines de Paris – Centre de Géosciences
Equipe Géostatistique
35, rue Saint Honoré
77300 Fontainebleau, France

Tél. 01 64 69 47 81
Fax 01 64 69 47 05

Huey Chyi LEE, Hans WACKERNAGEL (2007) Extreme value analyses of US P&I mortality data under consideration of demographic effects. Rapport technique. Juillet 2007. 25 p.

| | |
|---|---|
| Equipe | Géostatistique |
| Visa | J.P. Chilès |
| | |

# Extreme value analyses of US P&I mortality data under consideration of demographic effects

Huey Chyi Lee, National University of Singapore
Hans Wackernagel, Centre de Géosciences Equipe Géostatistique

July 2007

# Acknowledgements

This study is performed during an internship at Ecole Nationale Supérieure des Mines, Centre de Geosciences. I would like to express my gratitude to all who gave me the opportunity to take part in this study. I am deeply indebted to my supervisor Hans Wackernagel for his continuous guidance and support. I would also like to extend my appreciation to Mark Wilson, visiting professor from Michigan State University, who provided me with good advice and stimulating suggestions. Furthermore, this work has also benefitted from discussions with Dr Fabrice Carrat and Magali Lemaître from unit 707 of Institut National de la Santé et de la Recherche Médicale (INSERM). Last but not least, I would also like to thank Professor Wong Yan Loi from National University of Singapore who interviewed and offered me this internship, together with the director of Centre de Geosciences, Jean-Paul Chiles and also the Embassy of France in Singapore who provided subsidy for my course of internship.

# Introduction

Past influenza pandemics, notably the Spanish influenza (1918-1919) with a death toll of more than 40 million people, and other pandemics such as Asian influenza (1957), Hong Kong influenza (1968), Russian influenza (1977) [3], have greatly impacted humans all around the world. The psychological and economic repercussions from these pandemics have raised concerns on potential of a future influenza pandemic. However, with the current detection technologies, early diagnosis of the continuously mutating influenza strains lacks the required sensitivity and specificity. Influenza leads to pneumonia in the more serious cases, and most influenza deaths result from secondary bacterial pneumonia. This occurs more often in the >65 age group compared to the other age groups. The combined cause-of-death category pneumonia and influenza (P&I) ranks as the seventh leading cause of death in the United States, only to be preceded by heart disease, cancer, stroke, chronic lower respiratory diseases, unintentional injuries and diabetes[1]. Thus, the knowledge of future outbreaks of pneumonia and influenza is essential for prevention and control of the magnitude of outbreak.

In this study, we have applied the extreme value theory to predict the distribution of extremes of mortality due to pneumonia and influenza for the >65 age group from the data of 1968-1998. Epidemiology is intimately linked to demography: therefore the extreme value analysis of mortality data is preceeded by a detailed statistical analysis of the demographic changes in age structure of the american population over a period of 30 years.

This report is organized into 2 chapters, in the first chapter, analyzes have been performed on the demographic structures of the four most populated states in the United States, namely, California, Texas, New York and Florida. In the second chapter, we performed basic analyzes on the mortality of California and Texas and fitted the generalized extreme value distribution to California and Texas. We also improved on the fit by introducing a covariate in a non-stationary context.

[1] http://www.cdc.gov/nchs/products/pubs/pubd/hestats/leadingdeaths03/leadingdeaths03.htm

# Contents

# Chapter 1

# Demography

We are interested in a general analysis of the demographics of selected states in the United States (US) as a background study to identify trends and changes in the population.

The 1969-2004 US population data for each state is obtained from the National Cancer Institute (NCI) website under the category County-Level Population Files[1]. These datasets include the modifications made by the NCI to the Census Bureau estimates. In each dataset, there is information on resident population along with year, state postal abbreviation, state FIPS code, county FIPS code, registry, race, origin, sex and age group. For our study purposes, we extracted 5 variables from the datasets, namely year, state FIPS code, sex, age group and population.

As census was actually carried out every 10 years (for our dataset the years 1970, 1980, 1990 and 2000), the population sizes for other years were estimated and published under the Population Estimates Program. The size of dataset decreased the speed of operations in R substantially, and thus state specific data was extracted independently and used in our study instead of the combined states data.

Table 1.1: Age groups classifications

| Group | Ages (years) | Group | Ages (years) |
|---|---|---|---|
| 00 | 0 | 10 | 45-49 |
| 01 | 1-4 | 11 | 50-54 |
| 02 | 5-9 | 12 | 55-59 |
| 03 | 10-14 | 13 | 60-64 |
| 04 | 15-19 | 14 | 65-69 |
| 05 | 20-24 | 15 | 70-74 |
| 06 | 25-29 | 16 | 75-79 |
| 07 | 30-34 | 17 | 80-84 |
| 08 | 35-39 | 18 | 85+ |
| 09 | 40-44 | | |

## 1.1 California state population

Firstly, we studied the demographic evolution of California as the population size of California is the largest among all other states, it constitutes 12% of the total population of

[1] http://seer.cancer.gov/popdata/download.html

US. Understanding the changes in the diverse and dynamic population of California over the years could provide us clues to a fragment of the larger picture of US population.

According to the 19 age groups as defined in the dataset, bar-plots of each age group are generated for 1970, 1980, 1990 and 2000 (Fig.1.1). The first mode that appeared in the age groups 2-5 in 1970 is due to the Baby Boomers (persons born between 1946 and 1964)[1], it shifts to the right in the years as aging occurs. Different age groups increase at different rates throughout the years but there is an evident overall increase in all the age groups as seen in the bar-plots. The primary cause of population growth in California is migration, the number of immigrants in California being more than twice as many as the next leading state New York, and the leading sources of immigrants are Latin America and Asia, as suggested by 2000 Census Supplemental Survey.

For visualization of the same data from a different perspective, the line graphs are generated in Fig.1.2. It is observed that each cluster of age groups behaves differently. For instance, age groups 1-4 show an increase from year 1990 onwards, age groups 5-8 show increase throughout the years and reach a peak in 1990 then decline gradually, age groups 9-12 shows an apparent increase from 1980 onwards, and age groups 13-18 remains relatively stable throughout the years. The individual behaviors of each single age group can likewise be read from the graphs, according to their color codes.

Figure 1.1: Bar-plots of California population by age groups

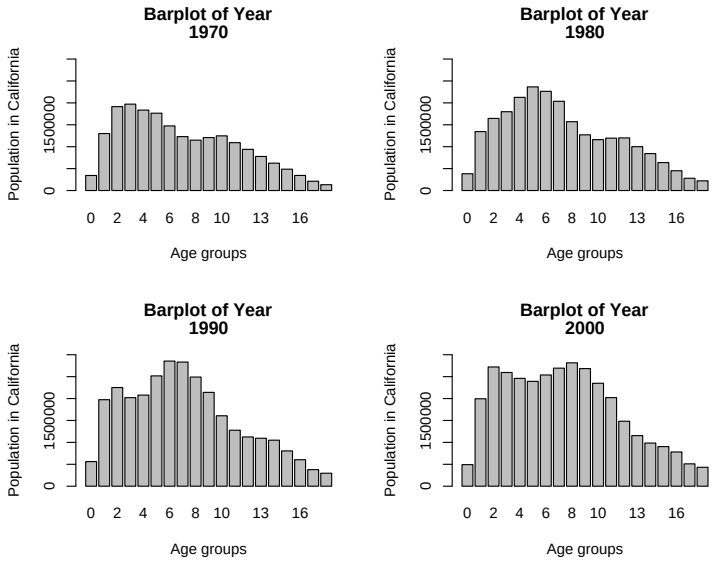
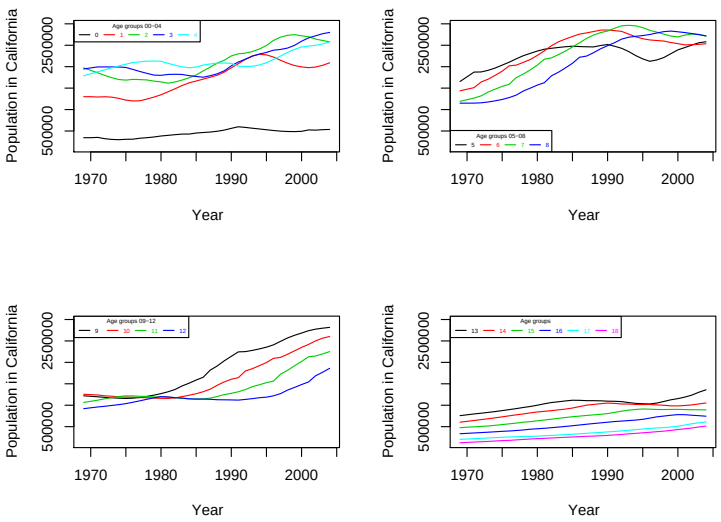


Figure 1.2: Line plots of California population by age groups




## 1.2 Texas state population

Texas is the second most populated state in the US and thus is also of interest to us. The same types of graphs are generated for Texas so that comparison could be made. From the barplots, it is evident that the population size of Texas is smaller than that of California in general, as can be seen in Fig.1.3. In year 2004, the population of Texas is 22.5 million compared to 35.8 million in California. However, the overall distribution and shape of the barplots are almost similar to California. Small differences can only be observed in the diminished magnitude of the second peaks (age groups 9-11 in 1970, age groups 11-12 in 1980, age groups 5-8 in 1990) in Texas, and a slightly leveled shape of age groups 2-8 compared to California.

From the line plots of Texas population by age groups on Fig.1.4, we observe increase for all 19 age groups which indicates increase in the Texas population as a whole. The increase in population is accounted for by three factors: natural increase, net immigration and net migration. The line plots for age groups 0-8 are slightly different from the corresponding age groups of California, with the most evident difference being in 1995. Age groups 9-18 of Texas have a similar trend as California.

Figure 1.3: Bar-plots of Texas population by age groups

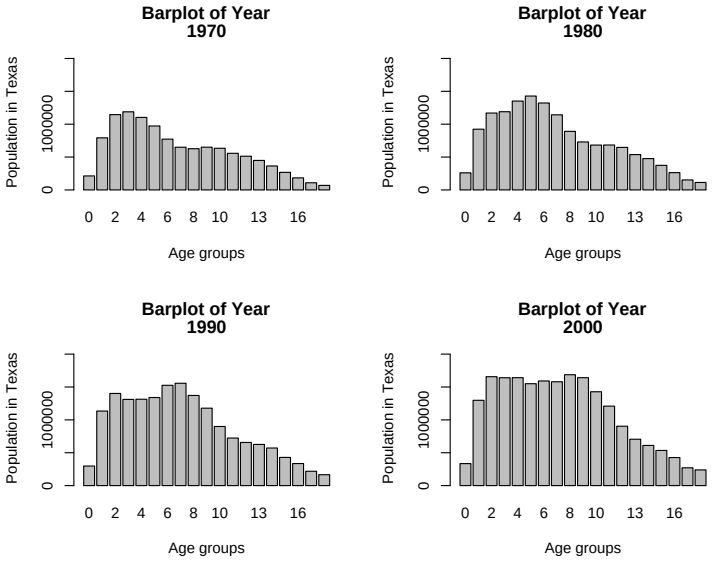
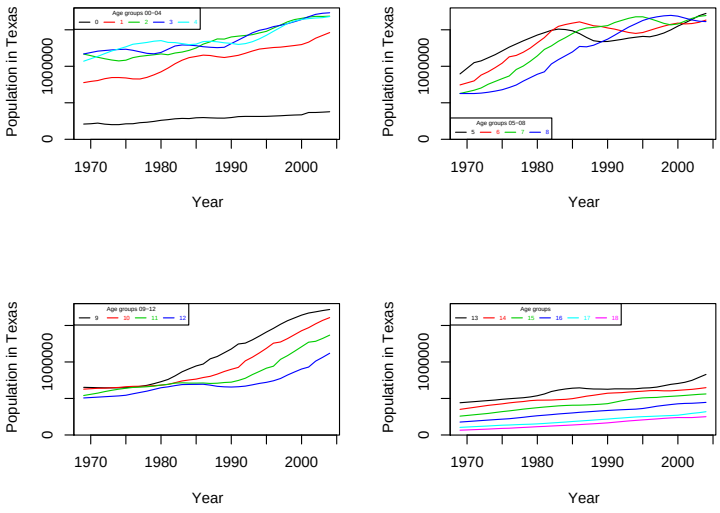



Figure 1.4: Line plots of Texas population by age groups




## 1.3 New York state population

New York state is the third largest state in population, with a total population of 19.3 million in 2004. The barplots of New York state show that the distribution of the population for each age groups are slightly different from California and Texas on Fig.1.5. A noticeable trend in the bar-plots is the aging of the population in New York state. The population of the oldest (18th) age group increases from 116064 in 1969 to 354790 in 2004.

The line plots of New York indicate that there is a decrease in population for age groups 0-8 on Fig.1.6. The population increase in age groups 9-18 is at a lower rate compared to California. This is due to a higher rate of emigration to other states, especially Florida and

Arizona, as these other states provide housing at a lower cost and work opportunities[4].

Figure 1.5: Bar-plots of New York population by age groups

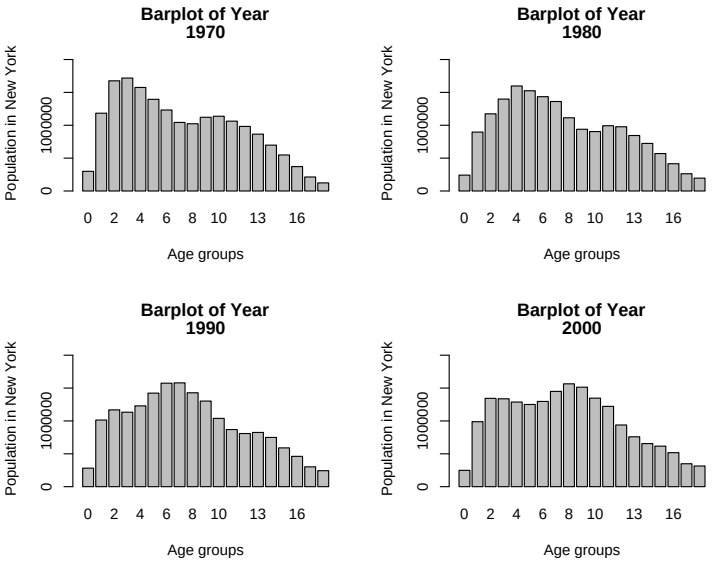
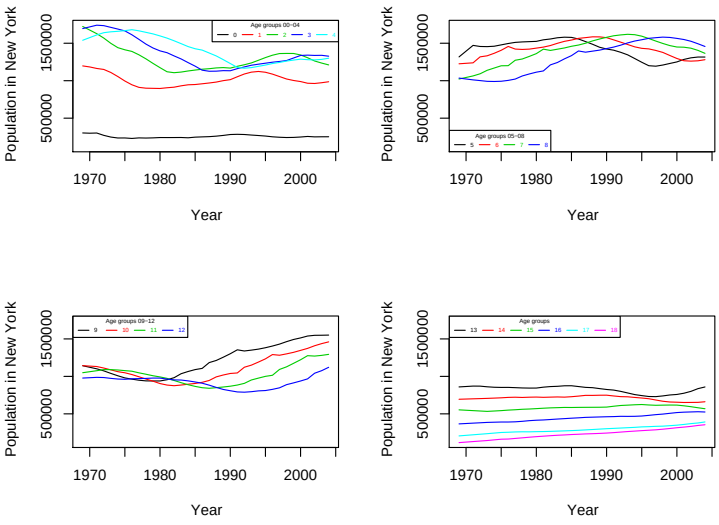


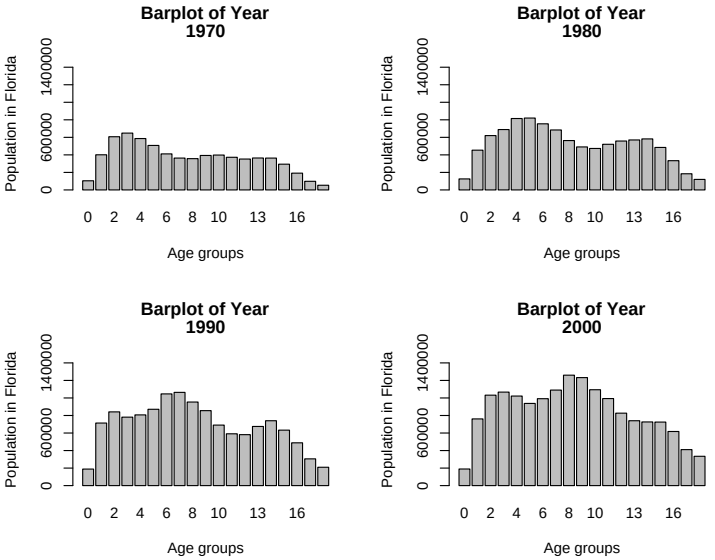

Figure 1.6: Line plots of New York population by age clusters




## 1.4 Florida state population

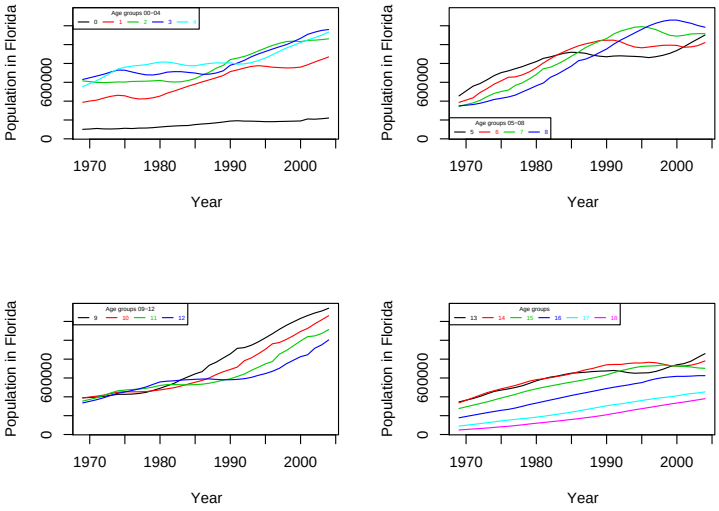
Florida is the nation's third fastest-growing state and the fourth largest state by population. The bar-plots indicate that in year 1990, there is an evident increase in the population for age groups 13-15 whereas no such increase is observed in California (Fig.1.7). The population of age groups 13-18 is relatively high compared to California, which also suggest aging of population in Florida.

Figure 1.7: Line plots of Florida population by age groups




The population growth is shown in the line plots of each age groups, and is dependent on both natural increase and migration. The age groups 13-18 is observed to be increasing steadily throughout the years. All other age groups also show increasing trends.

Figure 1.8: Line plots of Florida population by age clusters



# Chapter 2

# Mortality

Our dataset on mortality rates was obtained by dividing Pneumonia & Influenza deaths for each month by the corresponding year's total population count, with the 1970 population applied to deaths occurring in years 1968-1970, among people aged 65 or more years.

## 2.1 Basic analysis on mortality data

For fundamental understanding of the mortality data for California and Texas, graphical representations of the mortality data of each states were generated. As the data was provided in months of each year, the maximum mortality rate of the year is taken. Then, the time series plot and histogram are plotted.

### 2.1.1 California mortality maxima

From the graphs plotted below, we observed that there is an obvious increasing trend of maximum mortality across years 1968-1998 (Fig.2.1). Generally, the maximum mortality rate tripled in 30 years. This is due to the aging population as influenza mortality occurs primarily among the elderly. This indicates that there is a characteristic of the California mortality that change systematically through time. From the histogram generated in Fig.2.2, we observe a bimodal distribution of the mortality rates, with the first peak at 0.04 and second peak at 0.08.

Figure 2.1: Time-series Plot for California

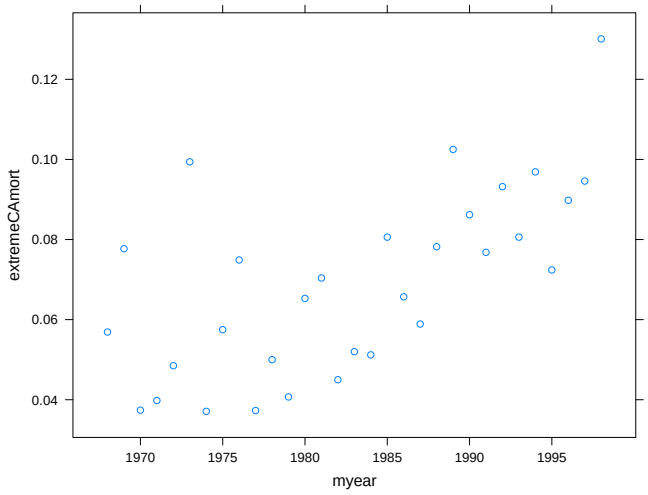
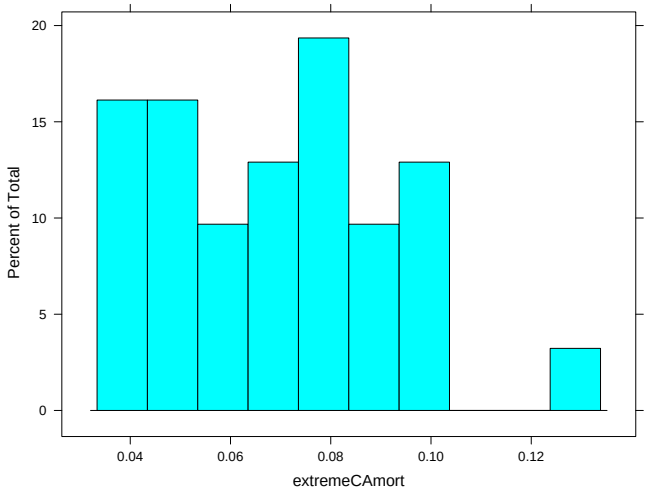



Figure 2.2: Histogram for California




### 2.1.2 Texas mortality maxima

The mortality across the years is scattered randomly, however, there is an outlier in year 1968 (Fig.2.3). Compared to the mortality in California, the maximum mortality of Texas are lower. The histogram of Texas mortality is symmetric as seen in Fig.2.4, except for a few low values which generate a second mode.

Figure 2.3: Time-series Plot for Texas

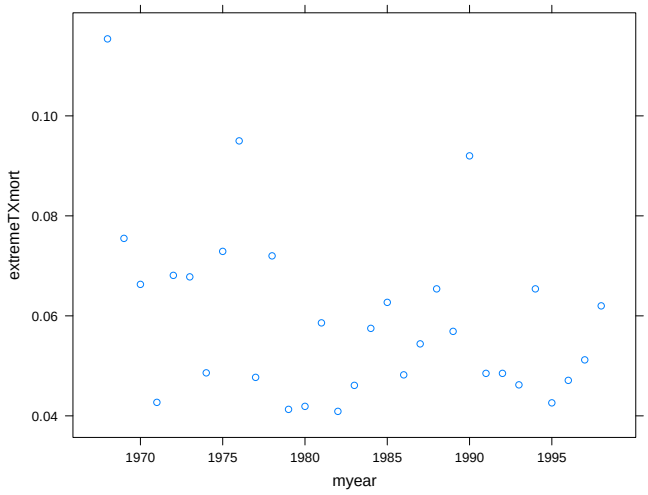
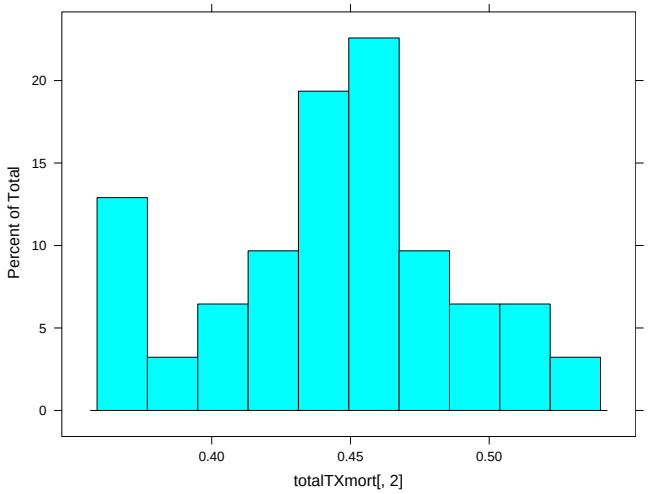


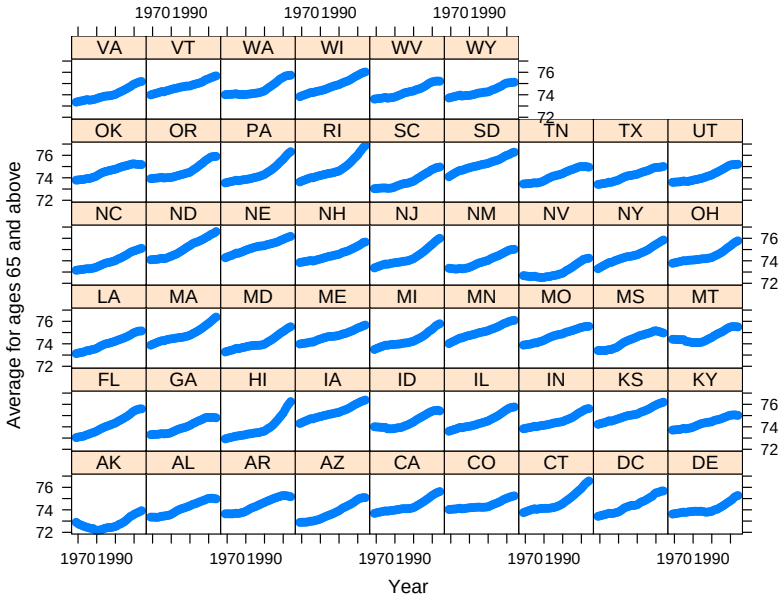

Figure 2.4: Histogram for Texas




### 2.1.3 Weighted average age for all states

To study the aging of the population, the average age weighted by population size, for the age groups >65 is calculated for each year from 1968-2004, for all states. The graphical output indicates that the aging of population in each state is occurring at a different rate (Fig.2.5). The weighted average age of Arkansas decreases from 1970-1980, unlike all other states which showed increase. In the overall, the weighted average age in Arkansas is the lowest. There is a similar pattern of slight increase till 1990 followed by a steep increase through 2004 in states like Hawaii, Connecticut, New Jersey, Rhode Island and Pennsylvania. California and Washington show similar trends over the years, the weighted average age showed minimal increase up till 1990 and an obvious increase thereafter.

Figure 2.5: Weighted average line plots for all states




## 2.2 Classical extreme value model

Extreme value analysis is used to estimate the probability of events that are more extreme than any that have already been observed. Due to inadequacies of adopting one of the three classes of extreme value distributions: Gumbel, Fréchet and Weibull families, the generalized extreme value (GEV) family of distributions which unifies three families of extreme value distributions into a single family is chosen[2]. After fitting a GEV model to the mortality data, we generated return level plots and also diagnostic plots to assess the suitability of the GEV model.

### 2.2.1 California GEV fit

Maximization of the GEV log-likelihood for the mortality data of California returns the estimates ($\mu$,$\sigma$,$\xi$) = (0.06, 0.02, -0.10) with standard errors (0.004, 0.003, 0.160). The maximized log-likelihood for this fit is -74.11. To assess the quality of the fitted model, a set of four graphical diagnostics is used (Fig.2.7). From the probability plot and quantile plot, we found that the data fits acceptably well, the plotted points follow approximately the straight line and have a near linear behavior. The return level curve shows asymptotes to a finite level and so provide a satisfactory representation of the empirical estimates. The corresponding density plot is consistent with the histogram of the data. We conclude from these diagnostic plots that the fitted model is suitable for California mortality data.

Figure 2.6: Profile likelihood for $\xi$ in the California mortality data

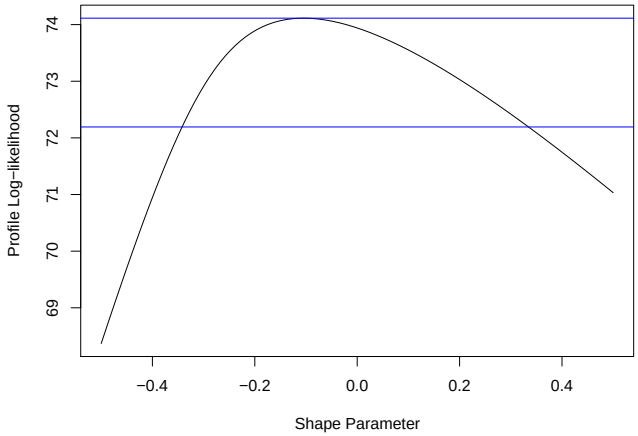
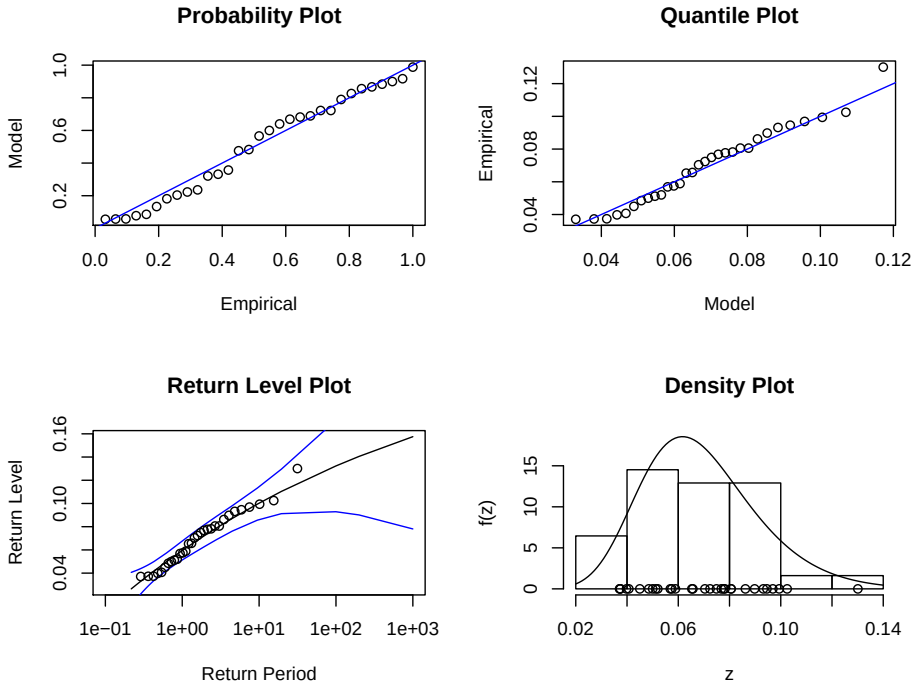



Figure 2.7: Diagnostic plots for GEV fit to California mortality




As better accuracy comes with profile likelihood, the profile likelihood for 10-year and 100-year return level in California mortality are plotted respectively on Fig.2.8 and Fig.2.9. The the 10-year return level plot is almost symmetrical, in contrast, the asymmetry in the 100-year return level plot indicates that the data provides weak information about the return level. We concluded that the GEV fit is adequate for the California mortality, in the next section, we will introduce the weighted average age as a covariate as an attempt to obtain a better fit.

Figure 2.8: Profile likelihood of return level for 10-year return return period

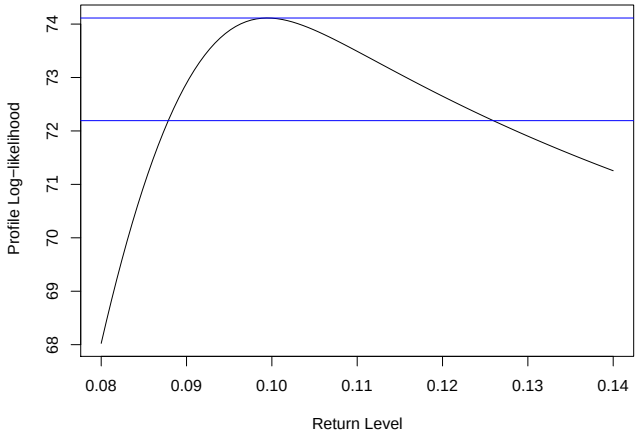
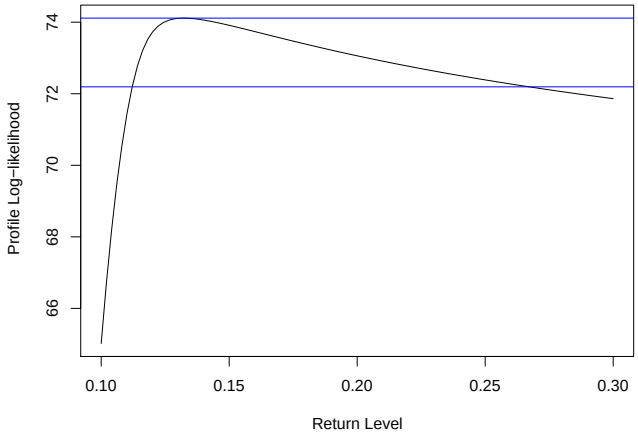



Figure 2.9: Profile likelihood of return level for 100-year return return period

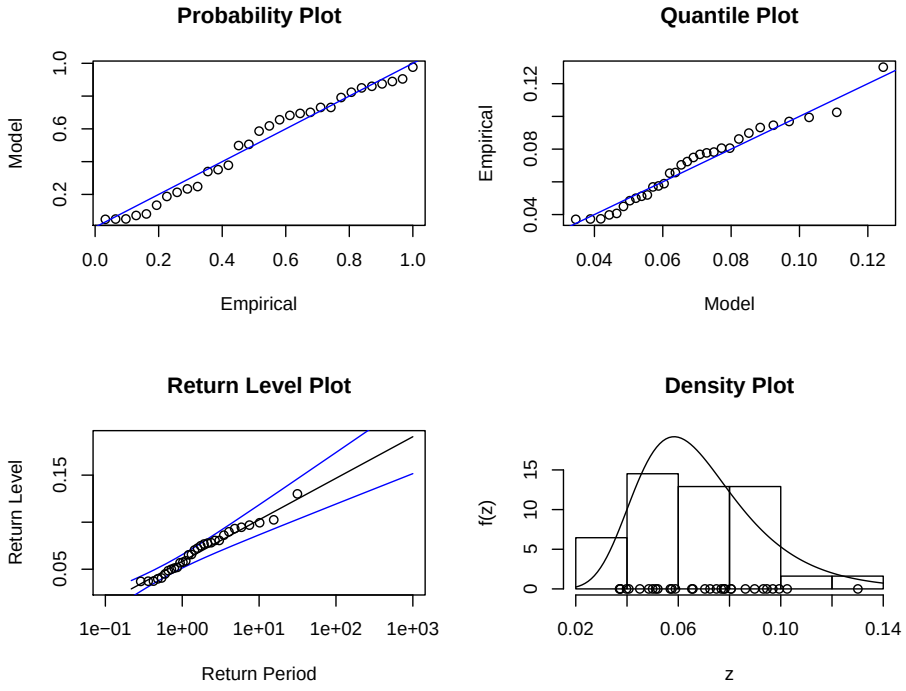



### 2.2.2 California Gumbel fit

As the 95% confidence interval for profile likelihood for $\xi$includes 0 (Fig.2.6), and the set of points in the return level plot is near-linear as seen in Fig.2.7, a Gumbel model might be appropriate in replacing the GEV family. Maximum likelihood in the Gumbel case corresponds to the estimates ($\mu$,$\sigma$) = (0.06, 0.02) with standard errors (0.004, 0.003). The maximized log-likelihood for this fit is -73.9. However, the likelihood ratio test statistic for the reduction to the Gumbel model is D=2{-73.9-(-74.11)}=0.42, this value is small when compared to the $\chi^2_1$ distribution, suggesting that the Gumbel model is adequate for California mortality data. From the diagnostic plots for Gumbel fit (Fig.2.10), we observe that the probability plot and quantile plot for Gumbel model are similar to that of GEV model. However, the goodness-of-fit is comparable with the GEV model as the estimated parameters in the two models are very similar. The difference between the GEV model and Gumbel model is the precision of estimation, the model parameters and return levels have estimates with smaller

confidence intervals in the Gumbel model compared to the GEV model.

Figure 2.10: Diagnostic plots for Gumbel fit to California mortality




### 2.2.3 Texas GEV fit

We obtained the maximum likelihood estimates ($\mu$,$\sigma$,$\xi$) = (0.05, 0.01, 0.34) with standard errors (0.002, 0.002, 0.24). The maximized log-likelihood for this fit is -89.0, considerably lower than that of California. From the diagnostic plots generated for the Texas GEV fit, we observe slight deviations of plotted points for probability plot and quantile plot. The return level curve shows an extremely large confidence interval and thus is not a good representation of the empirical estimates. Finally, the corresponding density plot is generally consistent with the histogram of data. Consequently, we conclude that the fitted model is not suitable for California mortality data.

Figure 2.11: Profile likelihood for $\xi$ in the Texas mortality data

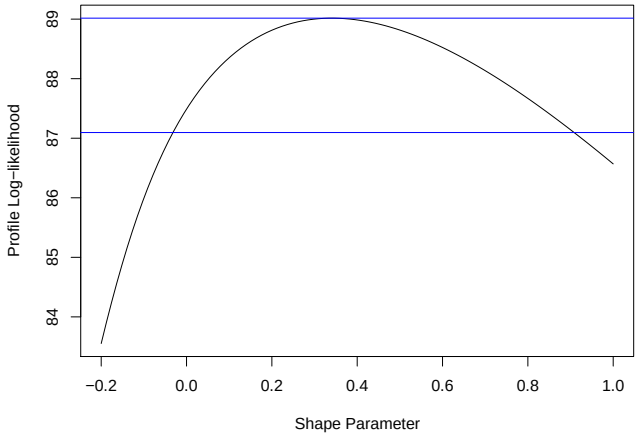
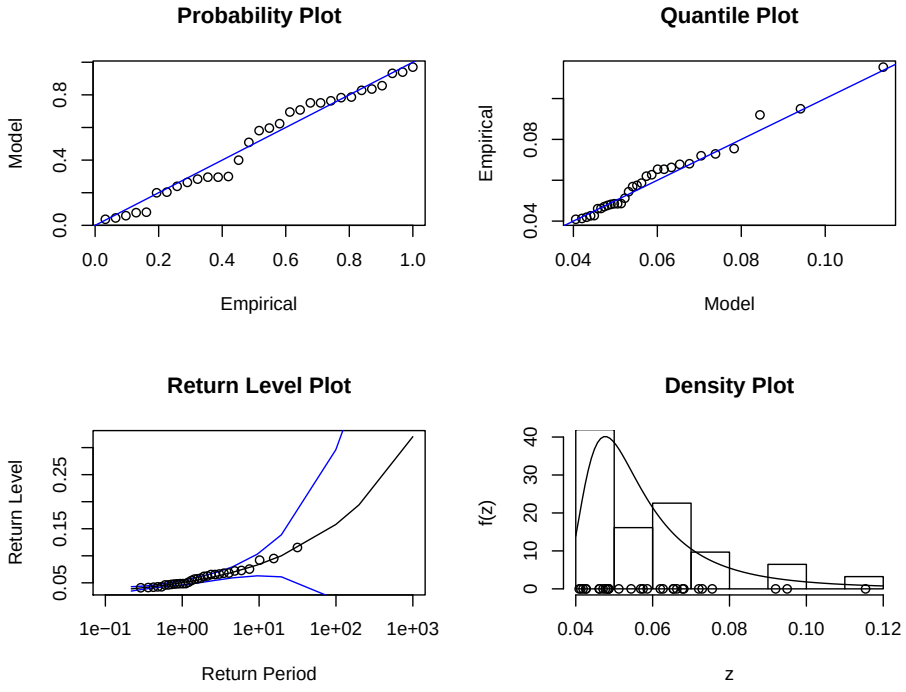



Figure 2.12: Diagnostic plots for GEV fit to Texas mortality




The profile likelihood for 10-year return level in Texas mortality (Fig.2.13) shows slight asymmetry whereas the asymmetry in the 100-year return level plot (Fig.2.14) is more evident as the data provide increasingly weaker information about high levels of the process. The GEV fit is less satisfactory for the Texas mortality data compared to that of California.

Figure 2.13: Profile likelihood of return level for 10-year return return period

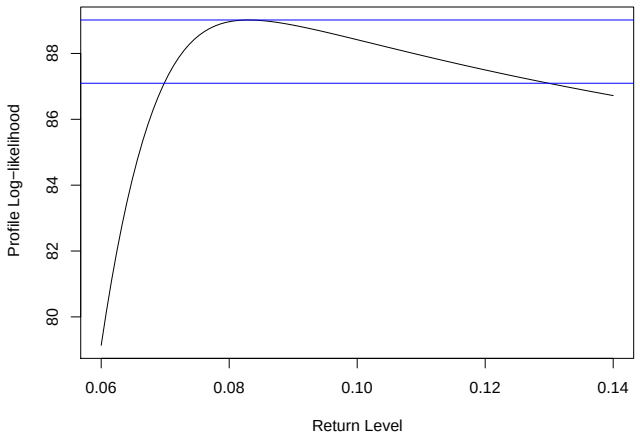
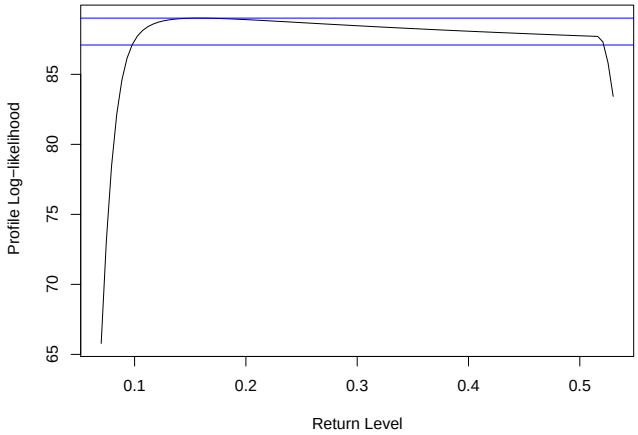



Figure 2.14: Profile likelihood of return level for 100-year return return period




The likelihood ratio test statistic for the reduction to the Gumbel model is D=2{-87.5-(-89.0)}=1.5, this value is small when compared to the $\chi_1^2$ distribution, suggesting that the Gumbel model is adequate for Texas mortality data. The 95% confidence interval of profile likelihood for $\xi$ in Fig.2.11 excludes 0 and the set of plotted points in return level plot is not linear (Fig.2.12), therefore, the Gumbel model is unlikely.

## 2.3 Non-stationary modeling

From the weighted average plots, we found that the weighted average ages for both California and Texas change systematically through time. Thus, there are limitations of fitting a GEV model to the mortality data which assumes a constant distribution through time. The extremal behavior of the mortality can be related to a covariate, the weighted average age in this case. To introduce the covariate, we have fitted a new model to the mortality of

California and Texas. A scatterplot of California mortality against weighted average age is shown below in Fig.2.15. The similar plot for Texas mortality is shown for comparison in Fig.2.16. From the plots, we observe a trend in California mortality data, it deviates slightly from our prediction that mortality is exponentially related to weighted average age. Conversely in Texas mortality data there is no such trend, the mortality is distributed randomly across all ages, but there are a few outliers in the plot (Fig.2.16).

Figure 2.15: Scatterplot of California mortality data against covariate

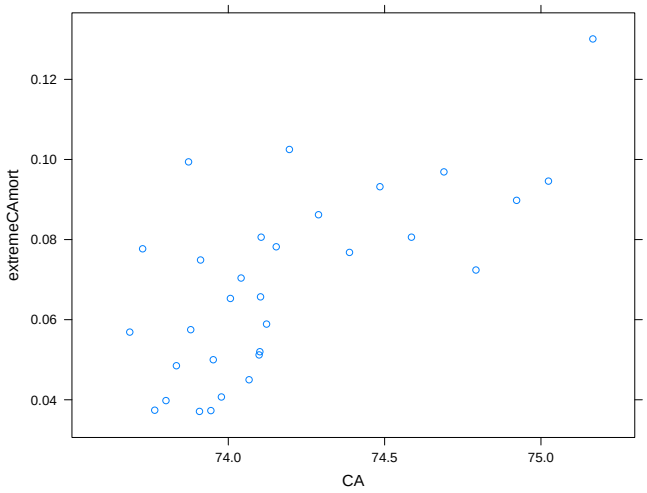
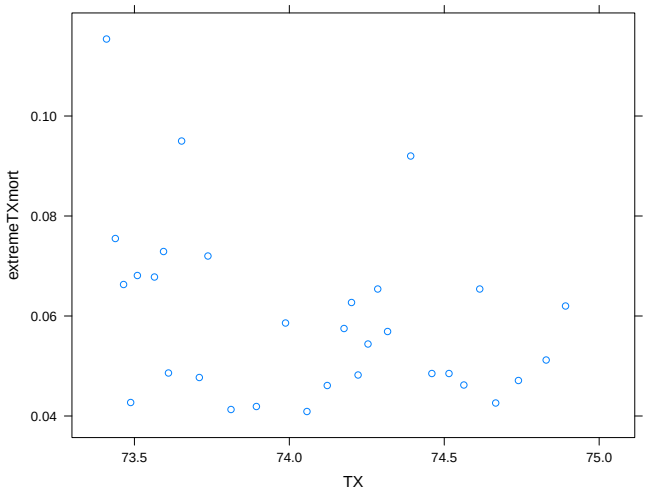



Figure 2.16: Scatterplot of Texas mortality data against covariate




### 2.3.1 California non-stationary fit

After introduction of covariate, we have fitted three different models to the California mortality data, Model 1 ($\mu$ depends linearly on year), Model 2 ($\sigma$ depends exponentially on year) and Model 3 (Linear dependence of $\mu$on year and weighted average age). The results of fit are shown in the table below Tab.2.1. From the table, we see that the negative log-likelihoods for all three models are less than that of the GEV fit, this suggests that the 3

models have improved accuracy. Among these three models, Model 3 is the best suited as it has the lowest negative log-likelihood.

The deviance statistics for comparing these three models with GEV model is calculated as shown in the table. All deviance statistics values are large when compared to a $\chi^2_1$ distribution at 95% level, more precisely, $D_1$=21.6 $> \chi^2_{1,0.95}$=3.84, $D_2$=20.2 $> \chi^2_{2,0.95}$=5.99 and $D_3$=22.6 $> \chi^2_{2,0.95}$=5.99. It implies that these models explain a substantial amount of the variation in the data, and are likely to be a genuine effect in the mortality rather than a chance feature in the data.The diagnostic plots show satisfactory fit of all three models.

Table 2.1: Maximized log-likelihoods and parameter estimates and standard errors of three models for California

| | Model 1: $\mu$depends linearly on year | Model 2: $\sigma$depends exponentially on year | Model 3: Linear dependence of $\mu$on year and weighted average age |
|---|---|---|---|
| Maximum likelihood estimates | 0.06 0.19 0.01 0.1 | 0.06 0.2 -4.4 1.1 0.2 | -1.66 0.1 0.02 0.01 0.08 |
| Standard errors of estimates | 0.003 0.03 0.002 0.2 | 0.003 0.03 0.2 2.1 0.2 | 0.002 0.03 NA 0.002 0.04 |
| Calculated Maximum likelihood estimates | $\mu$=0.06+0.19* year<br>$\sigma$=0.01<br>$\xi$=0.1 | $\mu$=0.06 + 0.2* year<br>$\sigma$=exp[-4.4+ 1.1*year]<br>$\xi$=0.2 | $\mu$=-1.66 + 0.1* year -0.02*w.avr<br>$\sigma$=0.01<br>$\xi$=0.08 |
| Negative log-likelihood | -84.9 | -84.2 | -85.4 |
| Deviance statistics from GEV model | $D_1$=2{84.9-74.1}=21.6 | $D_2$=2{84.2-74.1}=20.2 | $D_3$=2{85.4-74.1}=22.6 |

Figure 2.17: Residual diagnostic plots for Model 1

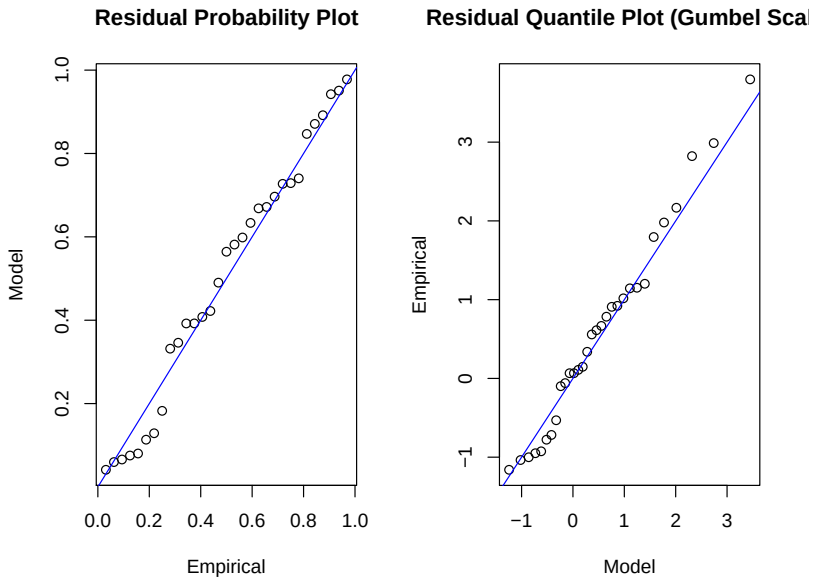


Figure 2.18: Residual diagnostic plots for Model 2

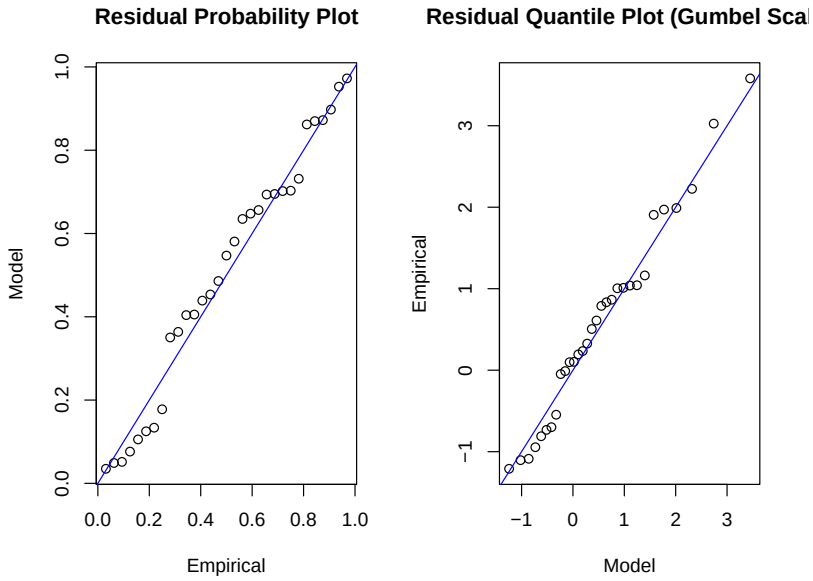
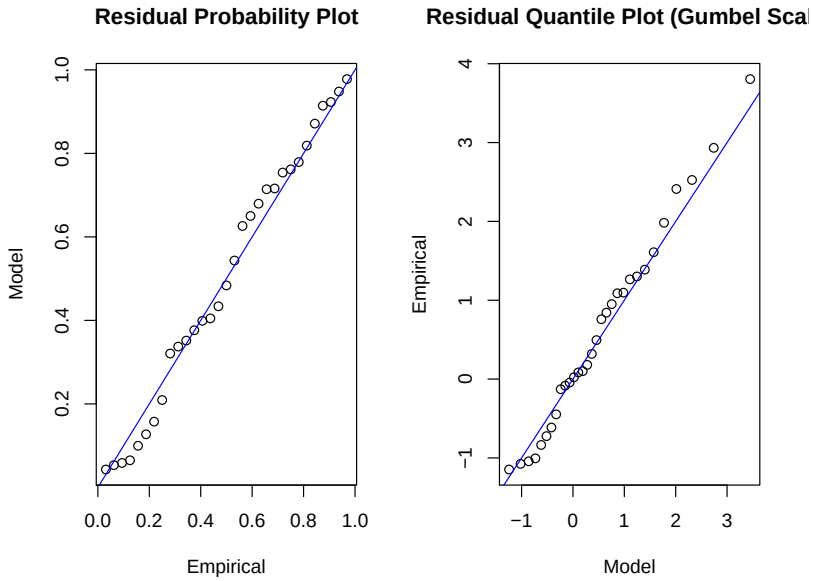



Figure 2.19: Residual diagnostic plots for Model 3




### 2.3.2 Texas non-stationary fit

The same models are fitted to Texas mortality data, and the results are tabulated in Tab.2.2. The negative log-likelihoods of all three models are close to that of the GEV fit for Texas mortality. The deviance statistics are small compared to the $\chi^2_1$ distribution at 95% level, $D_1$=0 $< \chi^2_{1,0.95}$=3.84, $D_2$=4.8 $< \chi^2_{2,0.95}$=5.99 and $D_3$=0.4 $< \chi^2_{2,0.95}$=5.99. This implies that the increase in model size does not bring worthwhile improvements in the model's capacity to explain the data. It follows that the constant GEV model provides an adequate description

Table 2.2: Maximized log-likelihoods and parameter estimates and standard errors of three models for Texas

| | Model 1: $\mu$depends linearly on year | Model 2: $\sigma$depends exponentially on year | Model 3: Linear dependence of $\mu$on year and weighted average age |
|---|---|---|---|
| Maximum likelihood estimates | 0.05 0.001 0.01 0.35 | 0.05 -0.01 -4.7 -2.6 0.2 | -0.3 -0.01 0.005 0.009 0.4 |
| Standard errors of estimates | 0.002 0.013 0.002 0.28 | 0.002 0.01 0.2 1.1 0.2 | 2.3e-3 1.2e-2 2e-6 1.8e-3 2.9e-1 |
| Calculated Maximum likelihood estimates | $\mu$=0.002+0.013* year<br>$\sigma$=0.002<br>$\xi$=0.28 | $\mu$=0.002 + 0.01* year<br>$\sigma$=exp[0.2+ 1.1*year]<br>$\xi$=0.2 | $\mu$=-0.3 - 0.01* year -0.005*w.avr<br>$\sigma$=0.009<br>$\xi$=0.4 |
| Negative log-likelihood | -89.0 | -91.4 | -89.2 |
| Deviance statistics from GEV model | $D_1$=2{89.0-89}=0 | $D_2$=2{91.4-89}=4.8 | $D_3$=2{89.2-89}=0.4 |

of the process. From the diagnostic plots we observe reasonable linearity of the plotted points, but imperfection occurs at the center part of the plots.

Figure 2.20: Residual diagnostic plots for Model 1

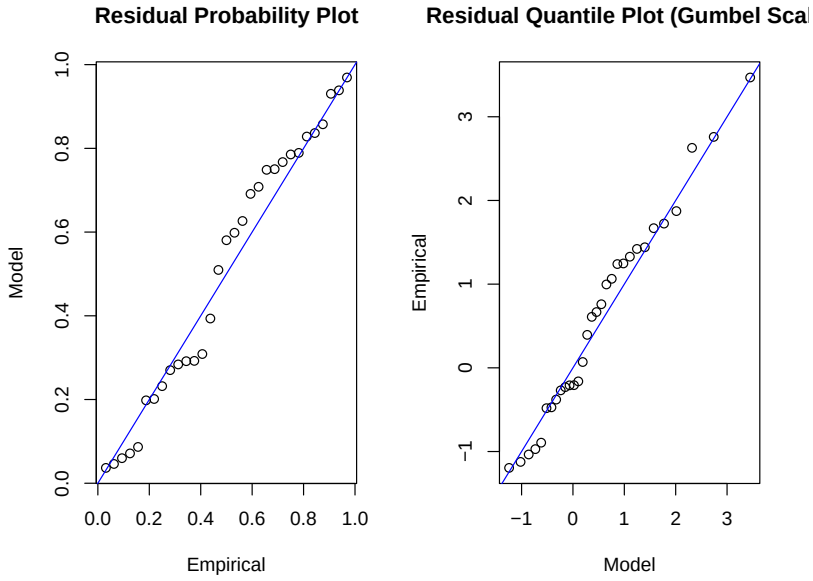


Figure 2.21: Residual diagnostic plots for Model 2

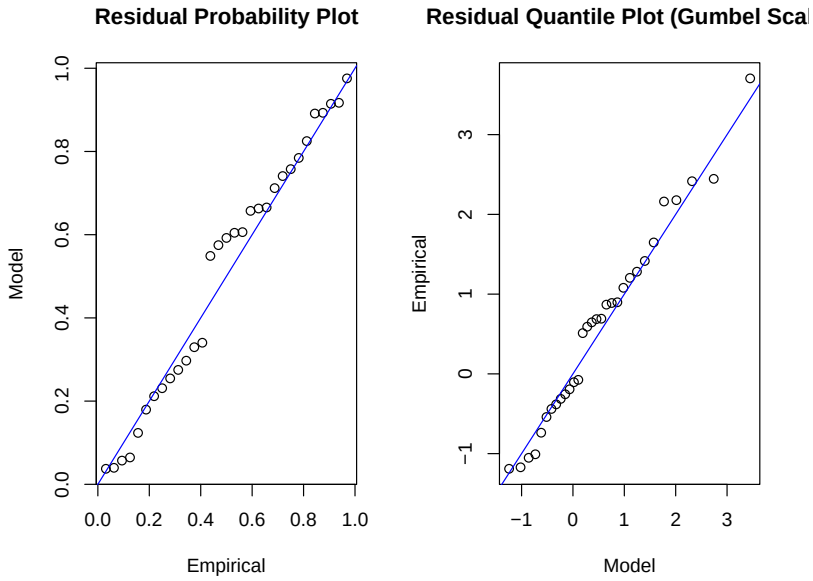
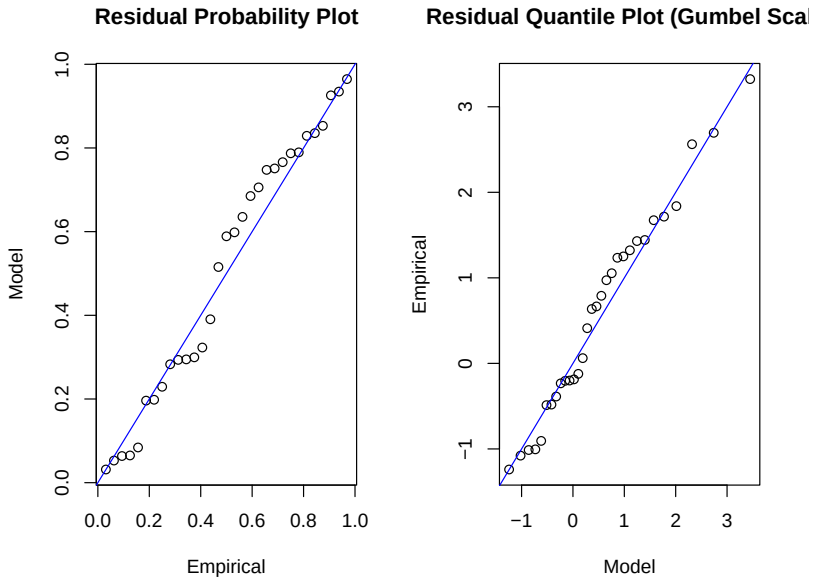



Figure 2.22: Residual diagnostic plots for Model 3



# Bibliography

[1] Population Projections 1993 California Department of Finance. Progression of aging: The impact of baby boomers.

[2] S Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

[3] Taisuke Horimoto and Yoshihiro Kawaoka. Influenza: Lessons from past pandemics, warnings from current incidents. *Nature Reviews*, 2005.

[4] Roberts S. State-to-state migration is linked to cost of housing. The New York Times, April 2006.