# Living to 100 and Beyond:
# An Extreme Value Study

Zhongxian (Jerry) Han
University of Central Florida, Orlando, Florida

Presented at The Living to 100 and Beyond Symposium
Sponsored by the Society of Actuaries

Orlando, Fla.

January 12–14, 2005

## Abstract

Extreme value theory has recently emerged as a newly developed statistical modeling tool for the analysis of extreme observations. This research paper focuses on parametric modeling of mortality rate for the elderly and the oldest population, together with the limiting age.

A classic threshold model is fitted to the data of each year using maximum likelihood methodology, separated by categories of year and sex. Then a model with transformed generalized Pareto distribution is fitted using a hyperbolic transformation, where the limiting age is introduced as a new parameter. The third model, a transformed exponential distribution, fits the data best and has good explanation. Log-likelihood functions for all models are given to find parameter estimations together with their confidence intervals. Last-$k$-years thresholds are specifically used to do a time series analysis of the limiting age in the 20th century. As a direct application, continuous mortality rates functions above the threshold can be derived from the model.

## 1. Introduction

With the improvements in general living conditions and the advancement of medical technology, the average human life span in the United States has increased over 20 years on average during the last century. As a result, the demographic structure has changed significantly. In many developing countries, the fast-growing aging population is having more and more impact on many economic and social fields.

Modeling of general mortality and survival functions has being going on for centuries. The earliest one dates back to 1729, when De Moivre proposed a uniform distribution of time-until-death variable $T(x)$ on the interval from current age $x$ to the limiting age $\omega$. Gompertz (1825) used a growing exponential function to model the mortality function, and Makeham (1860) modified it by adding a constant term. Weibull proposed a power function (1939). Even today, these models are still widely used in making life tables or approximations. In the latter three models, there is no upper bound on the survival age; the limiting age is infinite.

As a basic underlying distribution of future lifetime variable, life tables are popularly adopted in premium and reserve calculations and in risk management. However, the nature of the faster acceleration in mortality rates for the elderly population makes it less accurate in the end-of-life tables. For instance, the limiting age ω varies from 100 to 120 in different life tables. As we seek parametric models that could describe this behavior, the newly developed extreme value theory emerges as a promising solution.

## 2. About the Data

The data used in the analysis is from life tables of the United States during 1901–1999 provided by the Human Life-Table Database (HLD) [5]. We mainly use the complete life table columns of $d_x$ as the recorded number of deaths among the survivors' group, broken down by year and sex.

## 3. General Pareto Distribution (GPD) Model

### 3.1 Theoretical background

With either a detailed record of survivorship group in each year or general life table structure, peaks-over-threshold model can be applied in an elegant manner. The following theorem serves as a foundation of the asymptotic conditional distribution of excess over high threshold.

Let $X_1, X_2, ..., X_n$ be a sequence of independent random variables with common distribution $F$, and let $M_n = \max\{X_1, ..., X_n\}$. If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n > 0\}$ such that

$$\Pr\{(M_n - b_n)/a_n \le z\} \to G(z) \quad as\ n \to \infty$$

for a non-degenerated distribution function $G$, then $G$ is a member of the GEV family. Furthermore, for high threshold $u$, the conditional distribution of Y = (X – u | X > u) follows a Generalized Pareto Distribution (GPD) with distribution function:

$$H(y) = 1 - (1 + \frac{\xi y}{\tilde{\sigma}})^{-1/\xi} \tag{1}$$

defined on {y: y > 0 and $(1 + \xi y / \tilde{\sigma}) > 0$}, where $\tilde{\sigma} = \sigma + \xi(u - \mu)$. Details of the proof can be found in Coles [2] and Embrechs [3]. The corresponding survival function $s(y) = 1 - H(y) = (1 + \frac{\xi y}{\tilde{\sigma}})^{-1/\xi}$ implies a mortality function $\mu(y) = \frac{\tilde{\sigma}/\xi}{\tilde{\sigma} + \xi y}$ whose reciprocal is linear in $y$.

### 3.2 The log-likelihood function

The life table data, which contains $d_x$ at each age x, can be interpreted as the interval data, i.e., the number of observations for the age-at-death variable X on interval [x, x+1] is $d_x$. Suppose the largest age in the life table is $\omega_l$, then for $\xi \neq 0$, the log-likelihood can be derived from equation (1):

$$l(\tilde{\sigma},\xi) = \sum_{y=0}^{\omega_l-u-1}\left\{\log[(1+\frac{\xi y}{\tilde{\sigma}})^{-1/\xi} - (1+\frac{\xi y+\xi}{\tilde{\sigma}})^{-1/\xi}]*d_{y+u}\right\} \qquad (2)$$

An approximation of the above log-likelihood function can be obtained using the probability density function of the GPD.

$$l(\tilde{\sigma},\xi) = -\sum_{y=0}^{\omega_l-u-1}\left\{[\log(\tilde{\sigma}) + (1+\frac{1}{\xi})\log(1+\frac{\xi y}{\tilde{\sigma}})]*d_{y+u}\right\} \qquad (3)$$

This approximation sometimes produces a faster computational speed and an alternative way if the original one does not converge.

### 3.3 Parameter estimation

To select a threshold, we need to consider a relatively high threshold for accuracy in the approximation as well as enough survivors to reach the threshold in the life table. A threshold $u = 90$ is selected to compute the maximum likelihood estimates (MLEs) of $(\tilde{\sigma},\xi)$ for the survivorship group of each year during 1901–1999. As an illustration, for the male group in the year 1901 (we use superscript for gender and subscript for year to distinguish different groups), the MLEs are:

$$(\hat{\tilde{\sigma}}^m_{1901},\, \hat{\xi}^m_{1901}) = (3.8978,\, -0.2535)$$

The negative shape parameter $\xi$ indicates a Weibull distribution, which has an upper limit at $y^m_{1901} = -\hat{\tilde{\sigma}}^m_{1901} / \hat{\xi}^m_{1901}$. Combined with the threshold, the maximum likelihood estimate of the limiting age $\omega^m_{1901}$ is

$$\hat{\omega}^m_{1901} = u^m_{1901} - \hat{\tilde{\sigma}}^m_{1901} / \hat{\xi}^m_{1901} = 105.38$$

The related variance-covariance matrix of $(\hat{\tilde{\sigma}}^m_{1901},\, \hat{\xi}^m_{1901})$ is $\begin{pmatrix} 0.01991 & -0.002089 \\ -0.002089 & 0.0003396 \end{pmatrix}$. Applying the delta method (see Coles [2] for details), the estimated variance of $\hat{\omega}^m_{1901}$ is 0.55954. Therefore, a 95 percent confidence interval estimation of $\hat{\omega}^m_{1901}$ is (103.91, 106.85).

To conduct a time series analysis of the limiting age, we need to obtain the estimate of the limiting age for each year between 1901 and 1999 for both male and female. However, a constant threshold may not be a good choice. Obviously, for a constant threshold $u = 90$, the probability that one from the 1901 group survives to the age 90 is significantly smaller than one from the 1999 group. Therefore, we need to find ways to have thresholds such that the proportion of survivors who survive to the threshold is consistent throughout time.

### 3.4 Last-*k*-years thresholds and linear regression of the limiting age

There are a few ways to handle the choice of thresholds. A common way is to add time $t$ in a covariate model for the thresholds: $u_t = u_0 + c \cdot t$ where $u_0$ and $c$ are constants. The drawback is that we have to estimate one more parameter $c$ in our estimation. Also, since the threshold almost surely is no longer an integer in this setting, we need to convert $d_x$ to the number of deaths in a fraction of year.

Since our data is from life tables, the largest age observed for each year group is relatively stable throughout time. We propose an alternative solution by using last-*k*-years thresholds $u_t = \omega_{l,t} - k$, where $\omega_{l,t} - 1$ is the largest age observed in year $t$ group. With this type of varying thresholds, the data used in our analysis is always the last $k$ years in the life table. For instance, for the male group in 1901, the last observed age group is at age $\omega_{l,1901} - 1 = 104$. The choice of $u_{1901} = 90$ is equivalent by choosing $k = 15$. After obtaining estimates for groups from each year, a simple linear regression can be conducted on the results of the following:

For Male: $\omega_t^m = -46.5 + 0.0793\, t$ For Female: $\omega_t^f = -93.4 + 0.1039\, t$

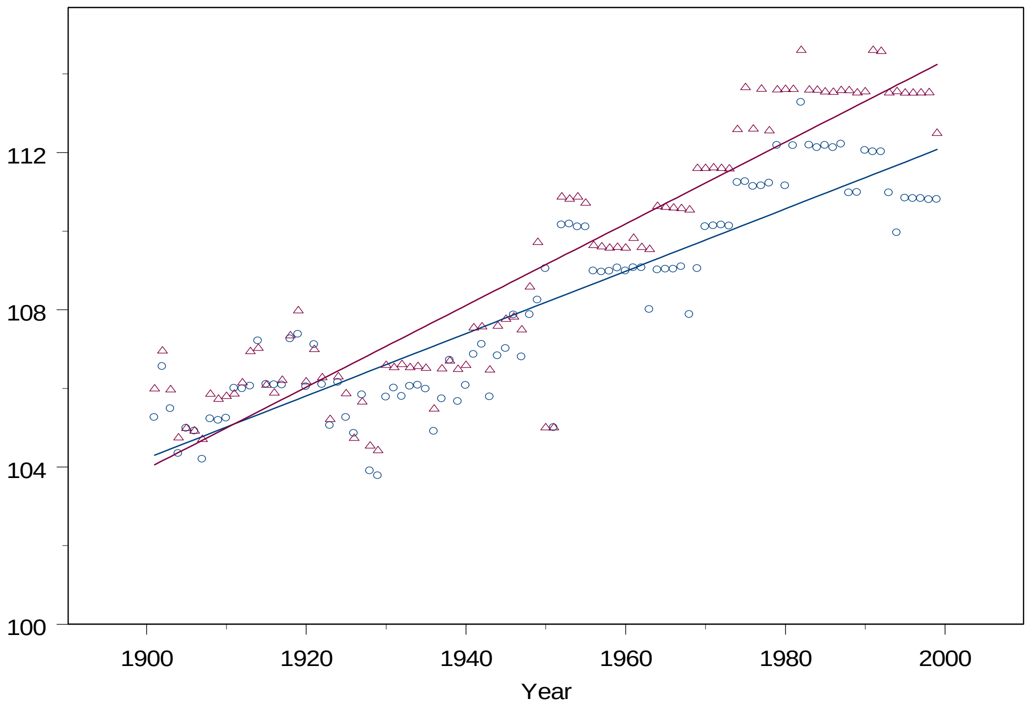

**Figure 1: Simple Linear Regression Plot of Limiting Age vs. Year in GPD Model**

Figure 1 gives the plots of linear regression analysis of the estimated limiting age by time, where blue circles stand for male and red triangles for female. The female group has a greater coefficient of the slope, which indicates that the increase of limiting age for the female group is slightly faster than the increase for the male group. Comparing with the annual increase of 0.18–0.30 in life expectancy, the increase in the limiting age is slower. However, after comparing the result with the life table, we can quickly find that maximum likelihood estimate of the limiting age heavily depends on $\omega_{l,t}$. In addition, it is easy to tell from Figure 1 that the estimates from 1950 to 1999 are clustered at several values.

## 4. Transformed GPD Model

### 4.1 Model and log-likelihood function

For a positive shape parameter $\xi$, the distribution does not have an upper bound, so an estimate of the limiting age is not viable. A solution can be found in Han [4] by using a transformed distribution. Using the hyperbolic transformation with $p=1$,

$$Z = \frac{wY}{w-Y} \tag{4}$$

The distribution function for Y = $(X-u \mid X>u)$ can be obtained (see Han [4] page 52).

$$H(y) = 1 - \left(\frac{u}{\omega-u}\right)^{1/\xi}\left(\frac{\omega}{y+u}-1\right)^{1/\xi} \quad \textit{where}\ 0<y<\omega-u \quad \textit{and}\ \xi>0 \tag{5}$$

With this $H(y)$, the mortality function $\mu(y) = \dfrac{\omega/\xi}{(\omega-y-u)(y+u)}$ has a quadratic form in its reciprocal. The corresponding log-likelihood function is

$$l(\omega,\xi) = \sum_{y=0}^{\omega_l-u-1}\left\{\log\left[H(y+1)-H(y)\right] * d_{y+u}\right\} \tag{6}$$

The approximation using the probability density function can also be found:

$$l(\omega,\xi) = \sum_{y=0}^{\omega_l-u-1}\left\{\left[\log(\omega/\xi) + \log\left(\frac{u}{\omega-u}\right) - 2\log(y+u) + (1/\xi-1)\log\left(\frac{\omega}{y+u}-1\right)\right] * d_{y+u}\right\} \tag{7}$$

### 4.2 Parameter estimation and linear regression of the limiting age

Using equation (7) and thresholds $u_t = \omega_{l,t} - 15$, we can obtain the MLEs of $(\omega_t, \xi_t)$ for each year group t = 1901, 1902, …, 1999. For instance, for the male group in 1901, ( $\hat{\omega}_{1901}^m, \hat{\xi}_{1901}^m$ ) = (104.8, 0.296) with variance-covariance matrix $\begin{pmatrix} 0.3543 & -0.00858 \\ -0.00858 & 0.0002992 \end{pmatrix}$. Similarly, we obtain the simple linear regression for the estimated limiting age for both male and female:

For Male: $\omega_t^m = -47.5 + 0.0797\,t$ For Female: $\omega_t^f = -96.6 + 0.1054\,t$

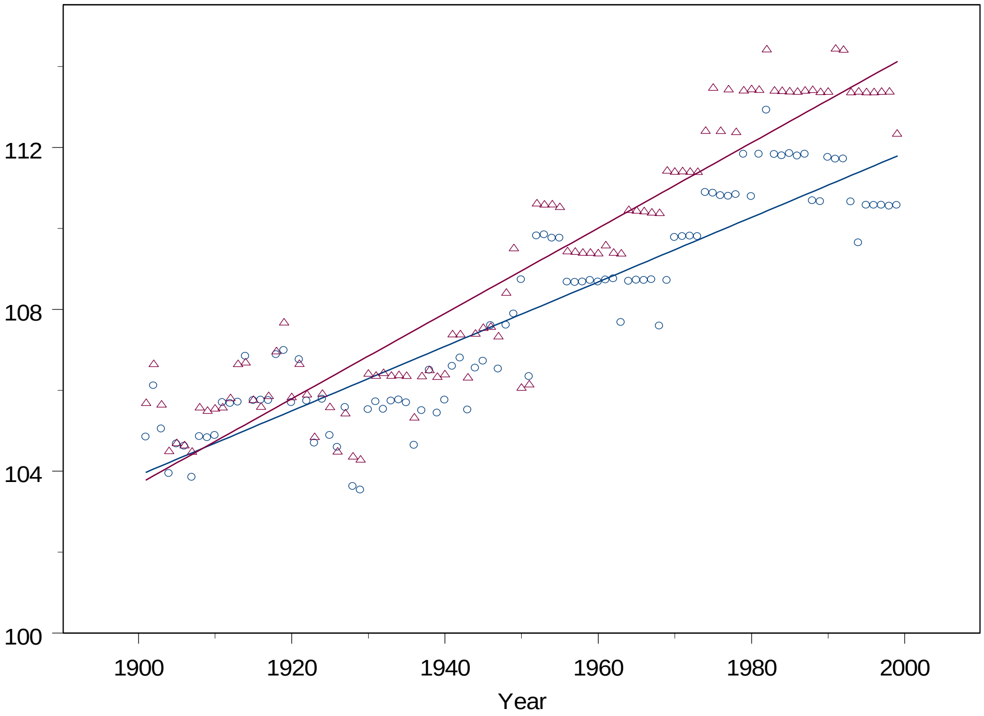

**Figure 2: SLR Plot of Limiting Age vs. Year in Transformed GPD Model**

Figure 2 shows a very similar pattern as in the previous model. Still, the estimated limiting age is slightly larger than the last observed age, and the result is not satisfactory. Next, we consider the case when the shape parameter is approaching 0.

## 5. Transformed Exponential Distribution

### 5.1 Log-likelihood function and model interpretation

When the shape parameter $\xi = 0$ , the generalized Pareto distribution degenerates to an exponential distribution with mean $\tilde{\sigma}$:

$$H(y) = 1 - \exp(-\frac{y}{\tilde{\sigma}}), \qquad y > 0 \tag{8}$$

The distribution is unbounded from the right. Applying the transformation in equation (4), we can obtain the transformed distribution for $Y = (X - u \mid X > u)$:

$$H(y) = 1 - \exp\left(-\frac{\omega}{\tilde{\sigma}}\left[\frac{(y+u)}{w-(y+u)} - \frac{u}{\omega - u}\right]\right) \quad \textit{where } 0 < y < \omega - u \quad \textit{and } \tilde{\sigma} > 0 \tag{9}$$

The mortality function $\mu(y)=\dfrac{\omega^2/\tilde{\sigma}}{(\omega-y-u)^2}$ also has a quadratic form in its reciprocal. The log-likelihood function then can be found from equation (6). An approximation can also be found by using its probability function distribution.

$$l(\omega,\xi)=\sum_{y=0}^{\omega_l-u-1}\left\{\left[2\log(\omega-y-u)-\log(\omega^2/\tilde{\sigma})\right]*\frac{\omega}{\tilde{\sigma}}\left[\frac{(y+u)}{w-(y+u)}-\frac{u}{\omega-u}\right]*d_{y+u}\right\} \qquad (10)$$

With this model structure, the limiting age for every human being is $\omega$. At an age of $x$, the model age can be calculated through the transformation equation (4), which is $\dfrac{\omega x}{\omega-x}$. If the parameter $\tilde{\sigma}$ is chosen to be a constant, then the model implies that the distribution of the model age follows a constant force of mortality assumption with $\mu=1/\tilde{\sigma}$ . After the transformation, the actual age distribution follows an accelerated force of mortality.

**5.2 Threshold selection**

Parameter estimates are obtained by maximizing the log-likelihood in equation (9) for each year group, using the last-15-years thresholds. For instance, for the 1901 male group, the estimates are $(\omega_{1901}^{m},\tilde{\sigma}_{1901}^{m})=(109.3613, 155.7952)$ with variance-covariance matrix $\begin{pmatrix}0.65895 & -10.8549\\ -10.8549 & 197.432\end{pmatrix}$.

So far, all the results we have shown are drawn using the last 15 years of observation in the life table. However, is 15 a good choice, or can we find a better one? One method is to fit the model at a range of thresholds and then check the stability of parameter estimates. As an illustration, we pick 1901, 1950 and 1999 for the selection of $k$. We seek for $k$ such that the parameter estimation is stabilized. For each specific year, we find the MLEs of the limiting age for a range of $k$ from 13 to 20. The procedure finds the best choice of $k$. Results can be seen in Table 1. In Figure 3, we plot the parameter estimation of the limiting age by having $k$ range from 10 to 20. By drawing a horizontal line to help us to choose $k$, we can tell that $k=15$ and $k=16$ are good selections for all 1901, 1950, 1999 data.

| **Year / k** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** | **Selected** |
|---|---|---|---|---|---|---|---|---|---|
| 1901 | 109.16 | 109.29 | 109.36 | 109.40 | 109.41 | 109.42 | 109.44 | 109.51 | 17 |
| 1950 | 112.56 | 112.64 | 112.68 | 112.71 | 112.73 | 112.73 | 112.73 | 112.72 | 18 |
| 1999 | 113.98 | 113.98 | 113.97 | 113.93 | 113.86 | 113.77 | 113.67 | 113.55 | 14 |

**TABLE 1**
**Estimates of the Limiting Age by Varying $k$**

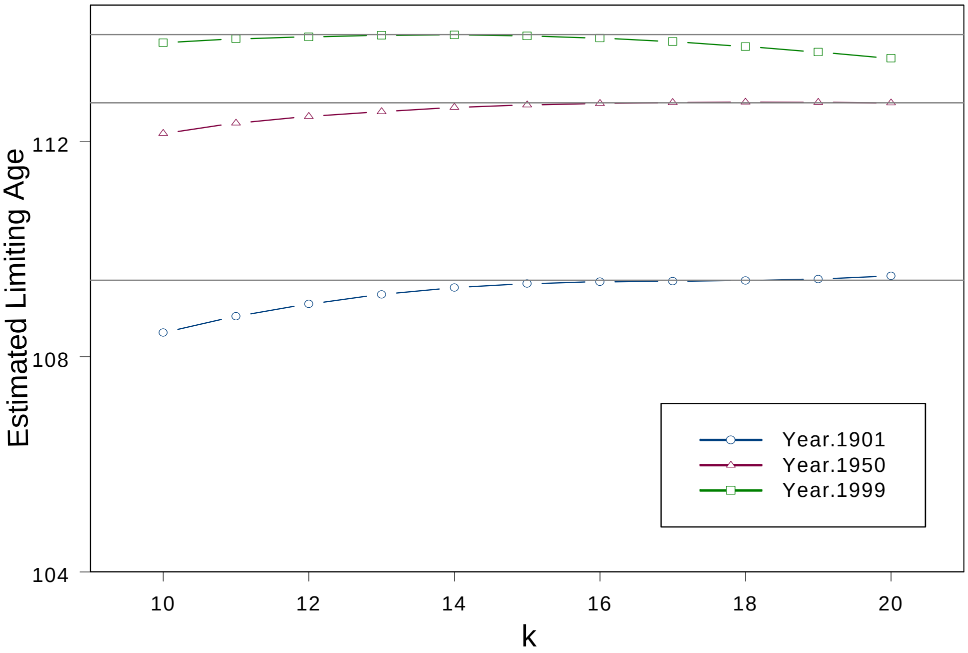

**Figure 3: Estimated Limiting Age by Varying Threshold $u$**

### 5.3 Simple linear regression of the limiting age and the limiting model age

As before, we conduct simple linear regression of those estimates. The result we obtain from this model does not differ greatly from the models we had before. But it is clear that the problem of the heavy dependence of the estimate of the limiting age upon the age of death of the last survivor is much less serious.

For Male: $\omega_t^m = -36.5 + 0.0761\,t$ For Female: $\omega_t^f = -76.5 + 0.097\,t$

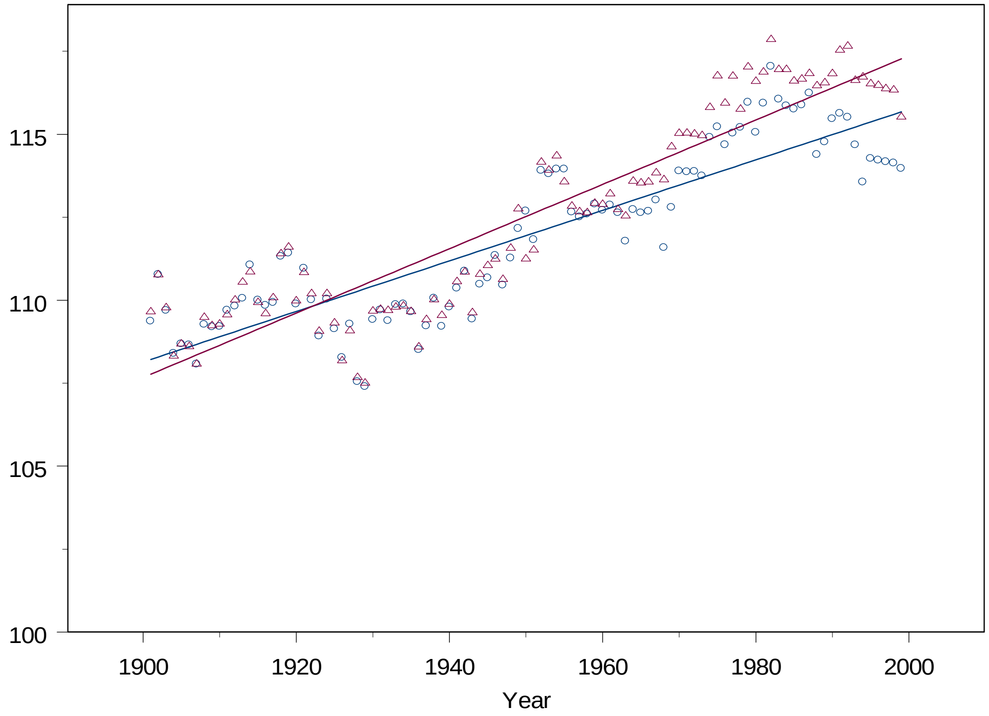

**Figure 4: SLR Plot of Limiting Age vs. Year in Transformed Exponential Model**

Now we take a look at the parameter $\tilde{\sigma}$, which can be explained as the life expectancy of the model age. The result from a simple linear regression follows in Figure 5. It is surprising that there is a significant difference between the two slopes of the linear fits. The average life expectancy in the model age increases 0.157 year for males, and this number is almost tripled for females.

For Male: $\tilde{\sigma}_t^m = -136.7 + 0.1570\, t$ For Female: $\tilde{\sigma}_t^m = -682.7 + 0.4447\, t$

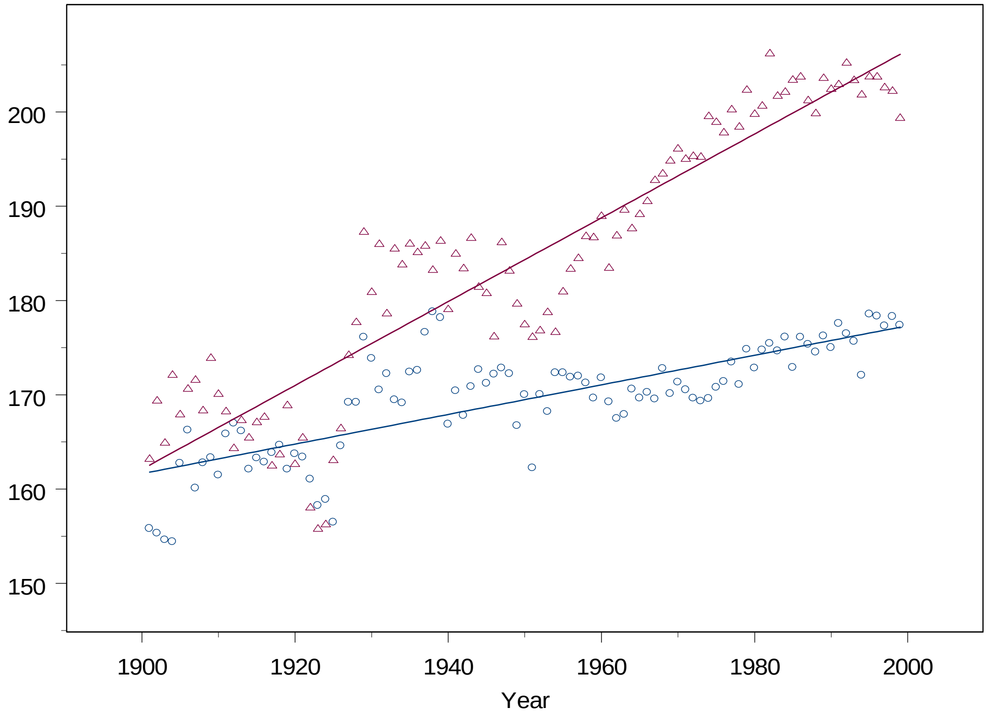

**Figure 5: SLR Plot of Model Age vs. Year in Transformed Exponential Model**

### 5.4 Confidence interval for parameter estimation and model validation

As a simple validation of our model, consider the male group in 1901. Since the estimated limiting age is 109.3613, the model age for 90 is $\frac{109.3613*90}{109.3613-90}=508.36$, and the model age for 91 is $\frac{109.3613*91}{109.3613-91}=542.00$, with an estimated $\tilde{\sigma}_{1901}^{m}=155.7952$. The estimated ${}_{1901}^{\ m}\hat{q}_{90}$ is: ${}_{1901}^{\ m}\hat{q}_{90}=1-\exp[-(542-508.36)/155.7952]=0.194$. This is fairly close to the exact ${}_{1901}^{\ m}q_{90}$, 0.268, in the life table. Since we have the variance-covariance matrix, we can construct a confidence interval for all related functions by applying the delta method. Figure 6 shows an overall comparison between the estimated annual death rates with the observed ones in the life table. This approximation can be improved if we choose $k=16,17,18$.

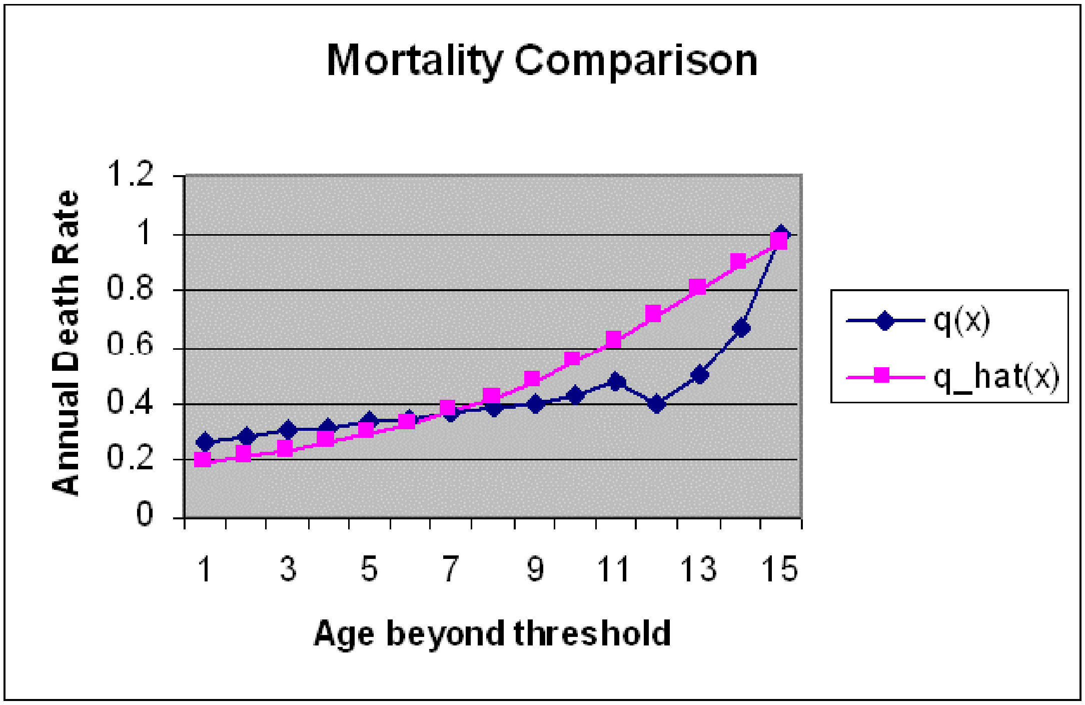

**Figure 6: Annual Death Rates Comparison in Life Table vs. in Model (1901, Male Group)**

## Conclusion and Discussion

All the suggested models indicate the gradual increase in the limiting age for both male and female during the 20th century. We feel the transformed exponential model has the overall best fit. The male group has an annual increase of 0.08 per year in the limiting age versus the female group's 0.10 per year. Yet these numbers are smaller than the annual increase in life expectancy.

In an earlier draft, we proposed that the transformed exponential model might be improved by adding an additional power $p$ to control the transformation. The proposed transformation was:

$$Z = X\left(\frac{w}{w-X}\right)^{p} \tag{11}$$

However, the numerical results indicate that both estimates of $\omega$ and $p$ would approach infinite; therefore, it seems redundant to add this parameter.

Theoretically, a more meaningful analysis can be conducted on the basis of well-recorded data instead of life tables, such as fully detailed Medicare data with birth cohort contributes. In our analysis, we used last-$k$-years thresholds with fixed $k$ to do a time series study from 1901 to 1999. This method is robust in the estimation of the limiting age; the estimated limiting age varies slightly for a reasonable range of the thresholds. However, the estimation of the model age varies significantly for different selection of thresholds. Therefore, we would monitor the estimation of the limiting age to select the optimal $k$ in case our goal is to fit the best model to a specific data set.

## Acknowledgements

I gratefully thank Dr. S.G. Coles for providing the s-plus codes together with the wonderful book. I am also grateful to the Human Life-Table Database (HLD), which provides life tables free of charge to all interested users. Last but not least, I would like to thank T. P. Edwards and the anonymous editor for their valuable comments and suggestions.

## References

[1] Bowers, Gerber, Hickman, Jones, Nesbitt (1997). *Actuarial Mathematics, Second Edition.* Society of Actuaries, Michigan.

[2] Coles, S.G. (2001). *An Introduction to Statistical Modeling of Extreme Values.* Springer-Verlag, London. 74–91.

[3] Embrechs, P., Kluppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin.

[4] Han, Z. (2003). "Actuarial Modeling of Extremal Events Using Transformed Generalized Extreme Value Distribution and Transformed Generalized Pareto Distribution," Ph.D. thesis, The Ohio State University.

[5] The Human Life-Table Database (HLD) at http://www.lifetable.de.