Nazarbayev University

MATH 499 Capstone project

Nonlinear Regression Analysis of the generalized Logistic Model as an Actuarial life contingency model.

Aida Kadenova

Research Supervisor:

Dr. Dongming Wei

Second reader: Dr. Yogi Erlangga

Abstract

The aim of this project is to analyze three different population models such as Gompertz, Logistic and Generalized Logistic based USA population data. Finding the appropriate model is essential in actuarial application. Firstly, the parameters of the two models are estimated using the special function ~nls in the R language program. But, due to some complexities, parameters of the generalized logistic model are evaluated using the new method from Causton`s paper. Secondly, two different comparison methods such as residual plot and AIC are used to analyze what model is appropriate for USA statistical data. Lastly, suitable models are used to estimate the force of mortality.

1 Introduction

The Gompertz model is the most popular sigmoid model, which is used to fit growth or other data. The concept of this model is introduced by Mr. Benjamin Gompertz in 1825. He established the connection between age and rising death rate. Then, insurance industry started to implement the Gompertz model to analyze death risk. However, only the probability density function was introduced by Gompertz. It was Makeham who developed the Gompertz model and presented well-known cumulative form. Thus, it was called Gompertz-Makeham model. It was firstly mentioned in Greenwood's discussions [12].

The second most frequently used model is logistic model, which was introduced by Pierre-Francois Verhulst in 1838 [12], [11]. Using this model, he represented population growth in a limiting environment. Lately, this model is widely applied to describe the behavior of natural growth, socio-technological and economic systems [7].

The logistic curve is symmetrical around the time where the maximum growth rate is achieved. Richards in 1959 added new parameter γ to handle the problem when growth rate is not symmetrical [14]. Richard's model depending on γ converges to Gompertz if $\gamma \to 0$ and Logistic model if $\gamma = 1$ [2].

In this paper, three different population models such as Gompertz, Logistic and Generalized Logistic models are evaluated according USA population. In Section 2, mathematical representations of three classical population growth models are introduced. Then, in Section 3, the unknown parameters of the two models using the R language program are estimated. However, for Generalized Logistic model R program language do not contain automatic package. Thus, the new method from Causton's research to determine four parameters are used. After the estimation of parameters for our three models, in Section 4, models are analyzed with statistical tools such as residual plots and ACI. In section 5, suitable models for identifying survival function and force of mortality are used.

2 Population Growth Models

2.1 Gompertz model

The mathematical equation of the Gompertz model where k is the rate of growth, which affects the slope at an inflection, A is upper asymptote [11]:

$$\frac{dP}{dt} = kP(t)\ln(\frac{A}{P(t)})$$

$$\frac{dP}{P(t)\ln(\frac{A}{P(t)})} = kdt \quad \rightarrow$$

$$\frac{(\frac{-A}{P^2(t)})dP}{\frac{A}{P(t)}\ln(\frac{A}{p(t)})} = -kt + C \quad \rightarrow$$

$$\ln(\ln\left(\frac{A}{P(t)}\right) = -kt + C \quad \rightarrow$$

$$\ln\left(\frac{A}{P(t)}\right) = \exp(-kt + c) \quad \rightarrow$$

$$\frac{A}{P(t)} = \exp(\exp(-kt + c) \quad \rightarrow$$

$$P(t) = Aexp(-\exp(-kt + c)) \quad \rightarrow$$

Parametrize $c = km$

$$P(t) = Aexp(-\exp(-kt + km))$$

where m indicates the time at inflection. This formula is more convenient since the inflection point can be calculated directly [11].

2.2 Logistic model

The mathematical equation of Logistic model where k is the rate of growth and K is upper asymptote [11]:

$$\frac{dP}{dt} = kP(t)\left(1 - \frac{P(t)}{K}\right) \quad \rightarrow$$

$$\int \frac{dP}{P(1 - \frac{P}{K})} = \int kdt \quad \rightarrow$$

In order to evaluate left hand side:

$$\frac{1}{P(1 - \frac{P}{K})} = \frac{K}{P(K - P)} = \frac{1}{P} + \frac{1}{K - P}$$

Thus;

$$\int \frac{dP}{P} + \int \frac{dP}{K - P} = \int kdt \quad \rightarrow$$

$$ln|P| - ln|K - P| = kt + C \quad \rightarrow$$

$$\ln\left(\frac{K-P}{P}\right) = -kt + c \quad \rightarrow$$

$$\frac{K-P}{P} = e^{-kt+c} \quad \rightarrow$$

$$P(t) = \frac{K}{1+e^{-kt+c}}$$

where $c = \ln\left(\frac{K}{P(o)} - 1\right)$ $\quad P(0) -$ population at time $t = 0$

2.3 Generalized logistic model

The mathematical equation of the generalized logistic model, where a is the rate of growth and K is upper asymptote and γ as mentioned earlier parameter which describe asymmetric curves [11], [14]:

$$\frac{dP}{dt} = aP(t)(1-(\frac{P(t)}{K})^{\gamma})$$

The explicit solution of is:

$$P(t) = \frac{P_0 K}{\left(P_0^{\gamma} + \left(K^{\gamma} - P_0^{\gamma}\right)e^{-at}\right)^{1/\gamma}}$$

where P_0 is the population at time $t = 0$ [2].

3 Estimation of parameters

3.1 USA population

Every decade, U.S. Census Bureau conducts a census to estimate the actual population size in the US. The statistical information can be represented in the table below [13]:

Year	Population (mln)	Year	Population (mln)	Year	Population (mln)
1790	3.93	**1860**	31.44	**1930**	123.20
1800	5.31	**1870**	38.56	**1940**	132.16
1810	7.24	**1880**	50.19	**1950**	151.33
1820	9.64	**1890**	62.98	**1960**	179.32
1830	12.87	**1900**	76.21	**1970**	203.21
1840	17.07	**1910**	92.23	**1980**	226.55
1850	23.19	**1920**	106.02	**1990**	248.71
				2000	281.42
				2010	309.05

Based on this information the parameters of models can be found.

3.2 Parameters of Gompertz and Logistic models

R programming language has a specific package in order to identify automatically the unknown parameters of classical models. It automatically determines the initial values and uses ~nls (nonlinear least squares) function in order to estimate the unknown coefficients [5].

3.2.1 Nonlinear Least Squares Regression

Let $y_i = f(x_i,\theta) + e_i$ where $i = 1,2, \ldots .N,\ \theta \in R^m, e$ is an approximation error, $f(x,\theta)$ is general nonlinear function m parameter model.

Suppose $e_i = y_i - f(x_i,\theta)$

To find the parameters $\theta_1, \theta_2, \ldots, \theta_m$ this constraint problem should be solved

$$\min_{\theta}\left\{\sum_{i=1}^{N} e_i^2\right\}$$

It can be solved by Gauss-Newton iterative method. Suppose $\bar{\theta}$ is initial guess solution. Let $f(x_i,\theta)$ linearize by Taylor expansion around $\bar{\theta}$. Then,

$$y_i = f(x_i,\bar{\theta}) + (\frac{\partial f(x_i,\bar{\theta})}{\partial \theta_i})(\theta - \theta_i)$$

$$\begin{bmatrix} y_1 - f(x_1,\bar{\theta}) \\ y_2 - f(x_2,\bar{\theta}) \\ \cdots \\ y_N - f(x_N,\bar{\theta}) \end{bmatrix} = \begin{bmatrix} \frac{\partial f(x_1,\overline{\theta})}{\partial \theta_1} & \cdots & \frac{\partial f(x_1,\overline{\theta})}{\partial \theta_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(x_N,\overline{\theta})}{\partial \theta_1} & \cdots & \frac{\partial f(x_N,\overline{\theta})}{\partial \theta_m} \end{bmatrix} \begin{bmatrix} \Delta\theta_1 \\ \cdots \\ \Delta\theta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \cdots \\ \varepsilon_N \end{bmatrix}$$

This can be written as

$$\boldsymbol{\Delta Y} = \boldsymbol{A}(\overline{\boldsymbol{\theta}})\boldsymbol{\Delta\theta} + \boldsymbol{\varepsilon}$$

$$\min_{\Delta\theta} \varepsilon_1^2 + \ldots + \varepsilon_N^2$$

Can be solved analytically $\theta_{LS} = \left[A(\overline{\theta})^T A(\bar{\theta})\right]^{-1} A(\bar{\theta})^T \Delta Y$

$$\theta_{new} = \bar{\theta} + \theta_{LS}$$

This procedure is repeated

$$\min_{\theta}\left\{\sum_{i=1}^{N} e_i^2\right\}$$

until appropriate θ is estimated [1].

Both of R-codes to estimate parameters are included in Appendix A. The output of this R-code for Gompertz model is

```
Nonlinear regression model
  model: y ~ SSgompertz(x, phi1, phi2, phi3)
   data: parent.frame()
    phi1      phi2      phi3
1.370e+09 5.757e+00 9.406e-01
 residual sum-of-squares: 2.06e+14
Number of iterations to convergence: 0
Achieved convergence tolerance: 2.573e-06
```

Mathematical equation of Gompertz model in R is [5]

$$P(t) = \varphi_1 * e^{-\varphi_2 \varphi_3^t}$$

Our Gompertz model is $\quad P(t) = Aexp(-\exp(-kt + km))$

where $\quad \varphi_3 = e^{-k} \quad \rightarrow \quad k = -l\,n(\varphi_3) \quad \rightarrow \quad k = 0.061$

$$\varphi_2 = e^{km} \quad \rightarrow \quad ln(\varphi_2) = km \quad \rightarrow \quad m = \frac{ln(\varphi_2)}{k} = 28.582$$

The output of this R-code for Logistic model is

```
Nonlinear regression model
  model: y ~ SSlogis(x, Asym, xmid, scal)
   data: parent.frame()
    Asym     xmid     scal
4.852e+08 1.952e+01 4.806e+00
 residual sum-of-squares: 5.235e+14
```

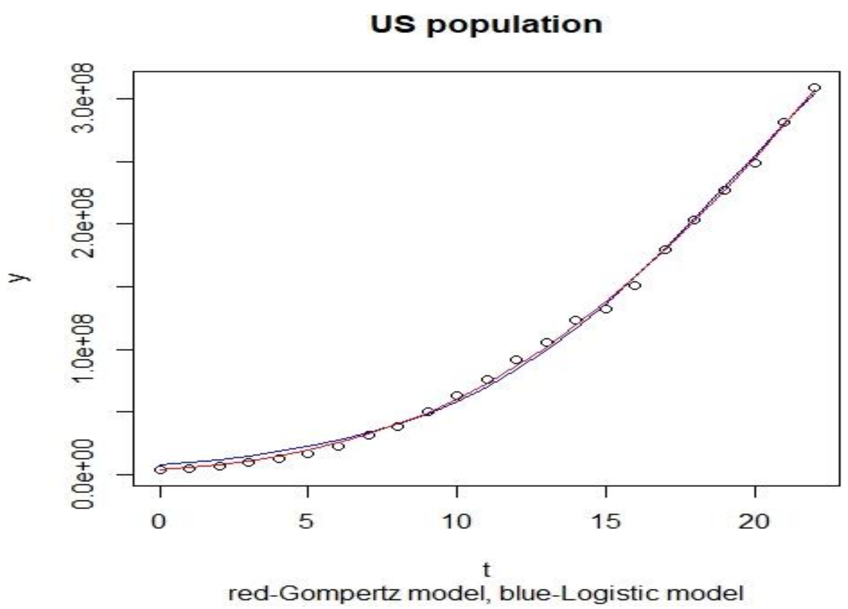
Mathematical equation of Logistic model in R $\qquad P = \frac{Asym}{1+\exp(\frac{x_{mid}-t}{scal})} \qquad$ [5]

Our equation is

$$P(t) = \frac{K}{1 + e^{-kt+c}}$$

Where $K = Asym = 4.852\text{e} + 08$, $\quad k = \frac{1}{\text{scal}} = 0.208$, $\quad \text{c} = \frac{x_{mid}}{\text{scal}} = 4.06$

The graphics of the USA population with Gompertz and Logistic model are represented below:



3.3 Generalized Logistic Model

The self-starting function for the generalized logistic model is not contained in the R programming language. Therefore, it creates some complexity of identifying parameters. To solve this problem new method from Causton`s method is used [3]. According to this method we parameterize

$$P(t) = \frac{P_0 K}{\left(P_0^{\gamma} + (K^{\gamma} - P_0^{\gamma})e^{-at}\right)^{1/\gamma}} \quad \rightarrow \quad V(t) = \frac{K}{\left(1 + \left(\frac{K^{\nu}}{V_0^{\nu}} - 1\right)e^{-at}\right)^{1/\nu}} \quad \text{as}$$

$$b = \frac{K^{\nu}}{V_0^{\nu}} - 1, \quad k = a, \quad A = K, \quad \gamma = n$$

then

$$P(t) = A(1 + be^{-kt})^{-1/n}$$

ODE is

$$\frac{dP}{dt} = \frac{kP}{nA^n}(A^n - P^n) \quad \rightarrow$$

$$\frac{dP}{dt} \cdot \frac{1}{P} = \frac{k}{nA^n}(A^n - P^n)$$

$$R = \frac{k}{n} - \frac{kP^n}{nA^n}$$

can be re-parametrized as $R = \alpha + \beta P^n$

where k $= \alpha n$ $\qquad$ $A = \left(-\frac{k}{\beta n}\right)^{1/n}$

In this case, R can be computed using the Fisher method

$$\bar{R}_i = {(logW_{i+2} - logW_i)}\big/{(t_{i+2} - t_i)}$$

R_i values can be calculated in R program, which represents in Appendix 2. In order to estimate parameters $R = \alpha + \beta P^n$ of this nonlinear model start with $n = -1$ which is the lowest sensible value for Richards' function and represents the monomolecular curve and considered equation as linear regression in Excel [3]. Then increment n by 0,1 until R^2 is highest. From Excel it is estimated that $n = 0.3$, $\alpha = 0.4213$, $\beta = -0.001$.

R-code is

```
> nonlin_mod=nls(R~a+b*(P^n),start=list(a=0,4213,b=-0,001,n=0,3))
```

The output is

```
Formula: R ~ a + b * (W^n)
Parameters:
  Estimate Std. Error t value Pr(>|t|)
a  0.449048   0.123901   3.624  0.00194 **
b -0.002165   0.006651  -0.326  0.74855
n  0.263378   0.141829   1.857  0.07975 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.02384 on 18 degrees of freedom
Number of iterations to convergence: 22
Achieved convergence tolerance: 9.811e-06
```

where n =0,263378 k $= an$=0,11827 $A = \left(-\frac{k}{bn}\right)^{1/n}$=626071402

Then, only b parameter should be estimated

$$log\left|\left(\frac{A}{W_i}\right)^n - 1\right| = \mathrm{logb} - \mathrm{k}t_i \quad (1)$$

In order to identify starting values of b it is used guessing method. It means that we choose one point in the graph of USA population and apply (1) formula to find b and use this result to check the convergence in R-language program. If it does not converge, we use other points. The R-code is

```
> nonlin_mod4=nls(log((626374056.5/y)^0.263372-1)~log(B)
0.1182679*t,start=list(B=2.743136676))
> summary(nonlin_mod4)
```

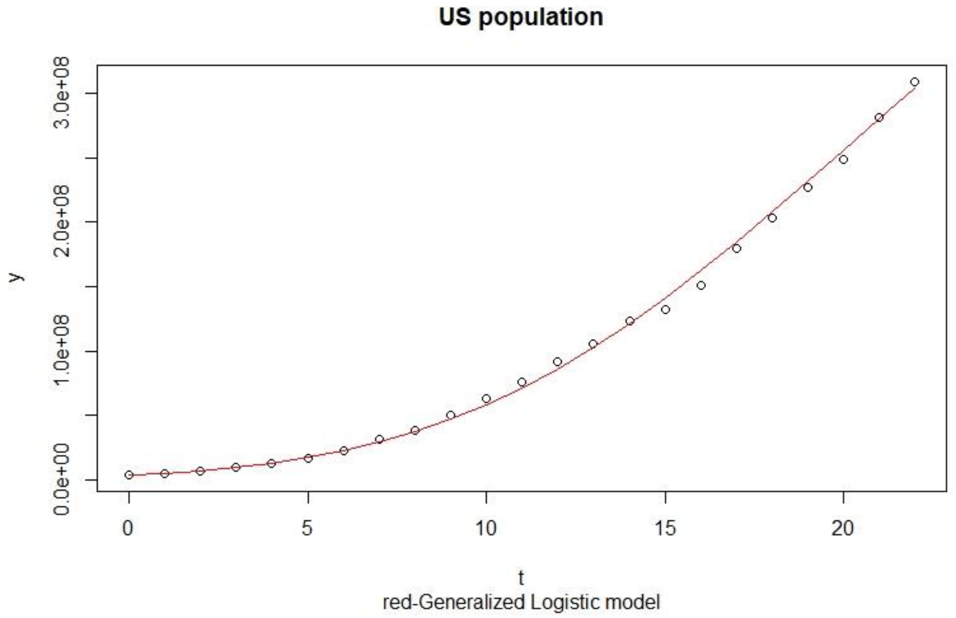
It is finally converge and the output is

```
Formula: log((626374056.5/y)^0.263372 - 1) ~ log(B) - 0.1182679 * t

Parameters:
  Estimate Std. Error t value Pr(>|t|)
B  2.83224    0.01868   151.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03023 on 20 degrees of freedom

Number of iterations to convergence: 2
Achieved convergence tolerance: 4.24e-06
```

Eventually, estimation of parameters leads to this equation:

$$P(t) = 626071402 \cdot (1 + 2{,}83224 \cdot e^{-0{,}11827t})^{-1/0{,}263378}$$

The graphical representation of this model:



4 Comparisons of the models

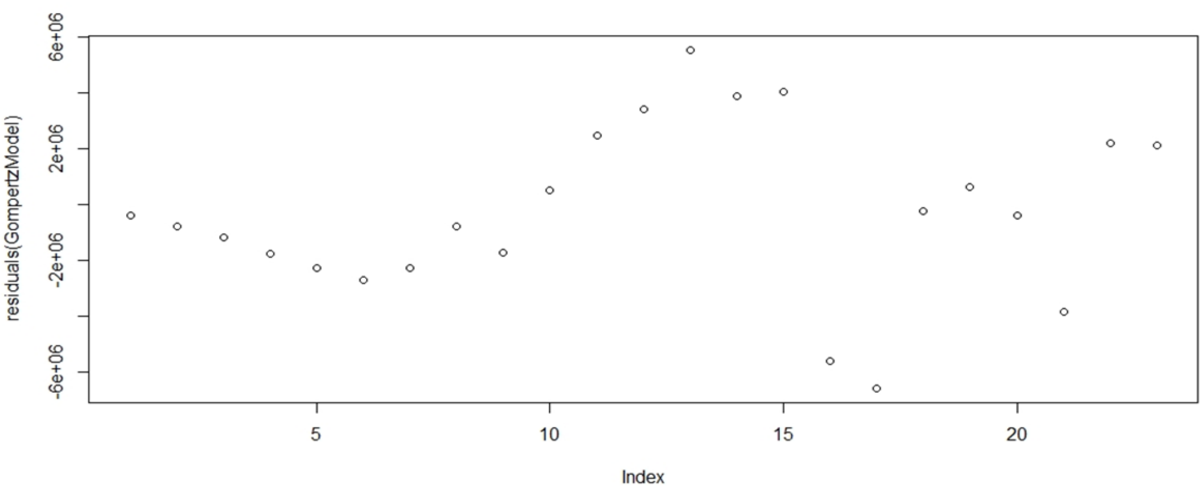
4.1 Residual plots

Residual plot is a powerful tool to evaluate the appropriateness of the model. It is a graph of residuals versus independent variable (time). Residuals are calculated as follows:

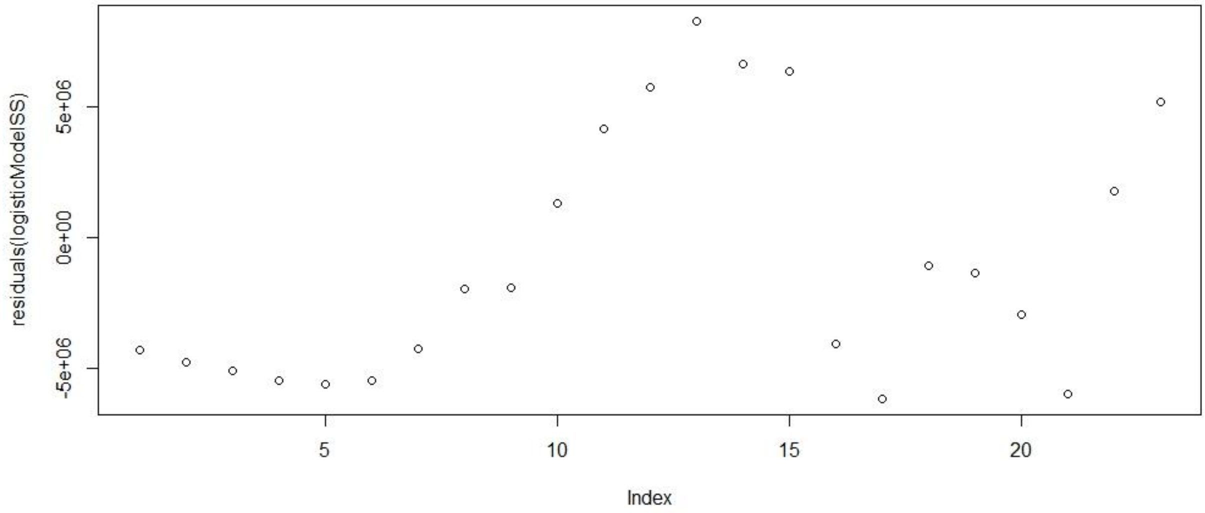
$$Res = \bar{y} - y$$

where $\bar{y}$ is observed value and y is predicted value. If the residual plot of non-linear model reveals some pattern then this model is suitable [9].

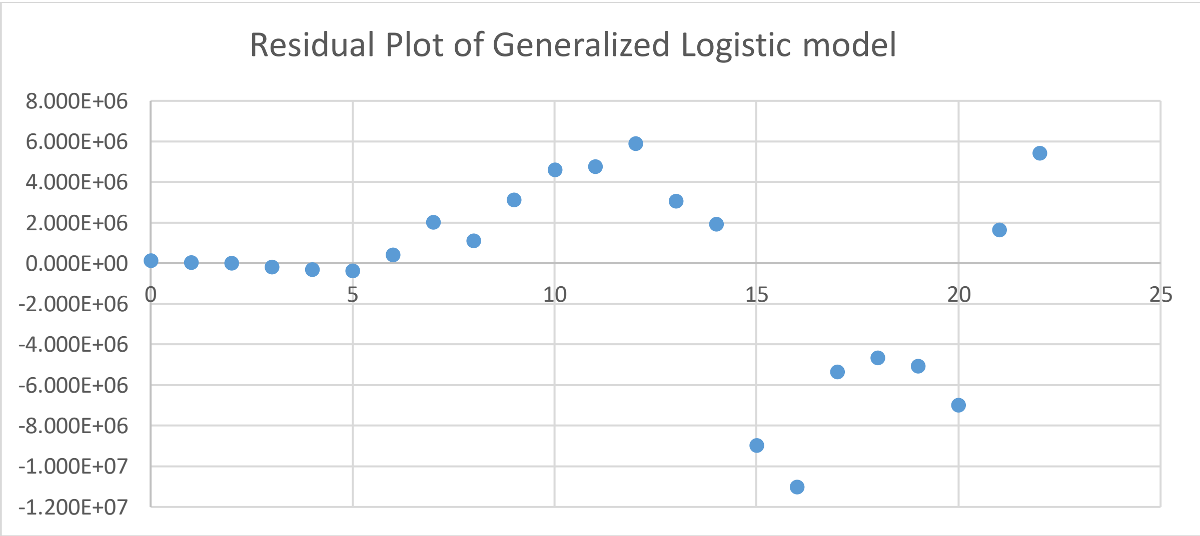
The residual plot of Gompertz Model:



The residual plot of Logistic Model:



The residual plot of Generalized Logistic model:



According to these graphs, it can be concluded that three residual plots display some pattern. It means that these three models are appropriate non-linear models to USA population.

4.2 AIC

AIC stands for "Akaike's information criteria" is a frequently used tool for analyzing models which was developed by Hirotugu Akaike. The more suitable model has lower AIC. AIC chooses the appropriate model from some set. If a set contains unsuitable models, it will choose among them [6]. However, in our case based on residual plots it can be concluded that AIC selects between three proper models.

The value of AIC can be estimated from

$$AIC = N + Nlog(2\pi) + Nlog\left(\frac{SSE}{N}\right) + 2(K+1)$$

where N is the number of recorded measurements, K is the number of parameters [8]. SSE (sum of squared errors) is calculated by this formula:

$$SSE = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

The R-code of AIC calculation is represented in Appendix 3. The output is

```
> glance(GompertzModel)
# A tibble: 1 x 8
   sigma isConv    finTol logLik   AIC   BIC deviance df.residual
   <dbl> <lgl>      <dbl>  <dbl> <dbl> <dbl>    <dbl>       <int>
1 3209645. TRUE   0.00000257  -376.  759.  764.  2.06e14         20
> glance(logisticModelSS)
# A tibble: 1 x 8
   sigma isConv     finTol logLik   AIC   BIC deviance df.residual
   <dbl> <lgl>       <dbl>  <dbl> <dbl> <dbl>    <dbl>       <int>
1 5115982. TRUE   0.000000548  -386.  781.  785.  5.23e14         20
```

The SSE and AIC of Generalized Logistic Model is 4.65733E+14 and 779.971, respectively.

5 Actuarial Application

Since Logistic model is not appropriate in comparison with two other models only Gompertz and Generalized Logistic models can be applied in the calculation of actuarial quantities such as the survival functions and mortality rate (also called hazard function in biostatistics and epidemiology).

5.1 Actuarial Application of Gompertz model

The survival function $S_x(t)$, the probability of (x) survival for at least t years:

$$S_x(t) = {}_tp_x = \frac{l_{x+t}}{l_x}$$

where l_x is the the expected number of survivors at age x.

The force of mortality at age x+t : $\mu_{x+t} = -\frac{d}{dx}\ln({}_tp_x)$ [4]

For the Gompertz model, the force of mortality is obtained by the following equations:

$$\mu_{x+t} = -kexp(-k(x+t)+km)$$

Survival function: $S_x(t) = \exp(-\exp(-k(x+t)+km) + \exp(-k(t+m))$

The parameters k, m of Gompertz model are estimated in Section 3.

5.2 Actuarial Application of Generalized Logistic model

The survival function is

$$S_x(t) = \frac{(1 + be^{-k(x+t)})^{-1/n}}{(1 + be^{-kt})^{-1/n}}$$

The force of mortality

$$\mu_{x+t} = -\frac{bkexp\big(-k(t+x)\big) * (1 + bexp(-k(x+t))^{-\left(\frac{1}{n}\right)-1}}{n * (1 + bexp(-kt))^{-1/n}}$$

The parameters k, b, n of the generalized logistic model are estimated in Section 3.

Conclusion

This project evaluated three different population models: Gompertz, Logistic, and Generalized Logistic according to USA statistical information. From comparing residual plots and AIC values, it is concluded that Gompertz model is the most proper model for USA population. In addition, the same results were obtained in Pflaumer`s paper in 2012. However, he only compared Gompertz, Logistic and Polynomial models [10]. In our research it is found that the Generalized Logistic Model is more appropriate than Logistic Model since have lower SSE and AIC. The future work is expected to compare other population models to find a more appropriate mathematical equation to describe the USA population.

Reference list

[1] Bates, D. M., & Watts, D. G. (2007). *Nonlinear regression analysis and its applications*. New York: John Wiley & Sons.

[2] Benzekry, S., Lamont, C., Beheshti, A., Tracz, A., Ebos, J. M., Hlatky, L., & Hahnfeldt, P. (2014). Classical Mathematical Models for Description and Prediction of Experimental Tumor Growth. *PLoS Computational Biology,10*(8). doi:10.1371/journal.pcbi.1003800

[3] Causton, D. R. (1969). A Computer Program for Fitting the Richards Function. *Biometrics,25*(2), 401. doi:10.2307/2528797

[4] Dickson, D. C., Hardy, M., & Waters, H. R. (2009). *Actuarial mathematics for life contingent risks*. New York: Cambridge University Press.

[5] Fox, J., & Weisberg, S. (2019). *An R companion to applied regression*. Thousand Oaks, CA: SAGE Publications.

[6] Golla, S. S., Adriaanse, S. M., Yaqub, M., Windhorst, A. D., Lammertsma, A. A., Berckel, B. N., & Boellaard, R. (2017). Model selection criteria for dynamic brain PET studies. *EJNMMI Physics,4*(1). doi:10.1186/s40658-017-0197-0

[7] Kucharavy, D., & Guio, R. D. (2015). Application of Logistic Growth Curve. *Procedia Engineering,131*, 280-290. doi:10.1016/j.proeng.2015.12.390

[8] Larget, B. (2003). AIC and BIC [lecture note]. Retrieved from http://www.stat.wisc.edu/courses/st333-larget/aic.pdf

[9] Martin, J., Adana, D. D., & Asuero, A. G. (2017). Fitting Models to Data: Residual Analysis, a Primer. *Uncertainty Quantification and Model Calibration*. doi:10.5772/68049

[10] Pflaumer, P. (2012). Forecasting the U.S. Population with the Gompertz Growth Curve. *Social Statistics Section – JSM,*4967-4981. doi:10.17877/DE290R-4505

[11] Tjørve, E., & Tjørve, K. M. (2010). A unified approach to the Richards-model family for use in growth analyses: Why we need only two model forms. *Journal of Theoretical Biology,267*(3), 417-425. doi:10.1016/j.jtbi.2010.09.008

[12] Tjørve, K. M., & Tjørve, E. (2017). The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the Unified-Richards family. *Plos One,12*(6). doi:10.1371/journal.pone.0178691

[13] U.S. Census Bureau. (n.d.). Decennial Census of Population and Housing by Decade. Retrieved from https://www.census.gov/programs-surveys/decennial-census/decade.2010.html

[14] Yin, X., Goudriaan, J., Lantinga, E., Vos, J., & Spiertz, H. (2002). A Flexible Sigmoid Function of Determinate Growth. *Annals of Botany,91*(3), 361-371. doi:10.1093/aob/mcg029

Appendix A

R-code for estimation of parameters of Gompertz and Logistic model

```
 > population=read.table(file.choose(),header=TRUE,sep="\t")
> Index <- population$Index
> Amount <- population$Population
> t=c(Index)
>  y=c(Amount)
> plot(t,y)
> logisticModelSS <- nls(y~SSlogis(x, Asym, xmid, scal))
>   lines(x,predict(logisticModelSS),col="blue")
> logisticModelSS
> popGompertz <- nls(y ~ SSgompertz(x, phi1, phi2, phi3))
> lines(x,predict(popGompertz),col="red")
> title(main="US population", sub="red-Gompertz model, blue-Logistic model")
> popGompertz
```

A2

```
> population=read.table(file.choose(),header=TRUE,sep="\t")
> Index <- population$Index
>   Amount <- population$Population
> t=c(Index)
>   y=c(Amount)
> R<-0
> for(i in 1:21){R[i]<-(log(population[i + 2, "Population"]) - log(population[i,
+ "Population"]))/(population[i + 2, "Index"] -
+ population[i, "Index"])
+ print(R[i])}
[1] 0.3054909
[1] 0.2981646
[1] 0.2876389
[1] 0.2857007
[1] 0.2944111
[1] 0.3053792
[1] 0.2542472
[1] 0.2338674
[1] 0.2453009
[1] 0.2088384
[1] 0.1907341
[1] 0.1650675
[1] 0.1447618
[1] 0.1101928
[1] 0.1028269
[1] 0.1525793
[1] 0.1473885
[1] 0.1168969
[1] 0.1010238
[1] 0.1084413
[1] 0.1086078
```

A3

```
> population=read.table(file.choose(),header=TRUE,sep="\t")
>  Index <- population$Index
> Amount <- population$Population
>    t=c(Index)
>    y=c(Amount)
>  GompertzModel<- nls(y ~ SSgompertz(x, phi1, phi2, phi3))
> library(broom)
> glance(GompertzModel)
# A tibble: 1 x 8
     sigma isConv     finTol logLik   AIC   BIC deviance df.residual
     <dbl> <lgl>       <dbl>  <dbl> <dbl> <dbl>    <dbl>       <int>
1 3209645. TRUE   0.00000257  -376.  759.  764.  2.06e14          20
> glance(logisticModelSS)
# A tibble: 1 x 8
     sigma isConv      finTol logLik   AIC   BIC deviance df.residual
     <dbl> <lgl>        <dbl>  <dbl> <dbl> <dbl>    <dbl>       <int>
1 5115982. TRUE   0.000000548  -386.  781.  785.  5.23e14          20
```