Chapter 4: A Comparison of Parametric Models for Graduation of Mortality of Districts of Assam

4.1 Introduction

The first step, and perhaps one of the fundamental ways in which statistics plays a part, is the graduation of mortality data. The graduation is a name for a class of techniques which produce smooth estimates for probabilities of dying in the life tables obtained from empirical data. In practice, the graduation methods suggested in the literature can be classified into two fundamental types: parametric and non-parametric, depending on whether they adjust the data to a function or simply achieve smoothness. Within the first type, there have been many attempts to find an appropriate model that represents mortality. Traditionally, a parametric curve, like the ones suggested by DeMoivre [16], Gompertz [23], and Weibull [62], is used to fit annual death rates. The earliest Gompertz [23] model is still one of the most important parametric models for graduation of mortality. The law of Makeham [46], which improved the Gompertz law by adding an extra parameter to this model to take into account the force of accidental death, assumed to be a constant independent of age. There are several models which are known under different names but which are essentially the same, because in all of them the force of mortality is a logistic function of age. These include the models proposed by Perks [49], by Beard [3], by Vaupel et al. [60] and by Le Bras [38].

4.1.1 A Review of Parametric Models to be Compared:

The selection of models is based on an extensive study of the proper literature and examinations with other workers in this area.

Gompertz model: The first attempt to develop a parametric model of mortality was that of Gompertz [23]. Gompertz modelled the aging or senescent component of mortality with two positive parameters. The force of mortality in the Gompertz model is

$$\mu_x = ae^{bx}. \tag{4.1}$$

where a is the scale parameter that varies with level of mortality, and b is the shape parameter that measures the rate of increase in mortality with age.

Makeham model: Makeham [46] modified the Gompertz model by adding a constant term, so that

$$\mu_x = c + ae^{bx}. \tag{4.2}$$

The new parameter c represents mortality resulting from causes, such as accidents or sexually transmitted diseases, unrelated to either maturation or senescence, which is the same for all ages.

Logistic model: The Logistic was first found by Perks [49], who discovered experimentally that the estimations of μ_x in a life table which he was looking at could be fitted by a specific curve, which was in certainty a logistic function (however he didn't describe it all things considered at the time). This model is known under a variety of names. The force of mortality in the Logistic model considered here is:

$$\mu_x = c + \frac{ae^{bx}}{1 + de^{bx}}. \tag{4.3}$$

It is noted that Makeham model $(d = 0)$ is a special case of the Logistic model. When d is small, any theories which may clarify why should follow a Logistic function will also help to explain why the Makeham and Gompertz laws work so well over much of the age range.

Beard model: The three-parameter Beard model obtained by assuming that the parameter $c = 0$ in **(4.3)**,

$$\mu_x = \frac{ae^{bx}}{1 + de^{bx}}. \tag{4.4}$$

Beard identified **(4.4)** as a Logistic function and showed how it could arise in a simple model of a heterogeneous population. If the members of the population are subject to hazards of the Makeham form **(4.2)**, the parameter a varying from individual to individual in such a way that they have a gamma distribution at birth, then the normal estimation of μ_x for the survivors who achieve age x will have the logistic form **(4.3)**. Beard also demonstrated how the Logistic curve could arise from a very simple type of stochastic process which assumed that individuals accumulate "shots" from random firings and is assumed to be dead when the total reaches a given figure. Special assumptions were about initial conditions. Thatcher et al. [56] also mention two ways in which the Logistic model could arise from stochastic processes.

In this chapter, the different parametric graduation models namely Gompertz, Makeham, Logistic, and Beard have been examined for mortality of district in Assam. Here, five high population growth rate districts and five low population growth rate districts of Assam are considered. The high population growth rate districts: Darang, Dhubri, Goalpara, Morigaon and Nagaon and the low population growth rate districts: Golghat, Jorhat, Sibsagar, Dibrugarh, and Tinsukia. Here the mortality pattern is divided by two age groups 20-60 and 60-100. These age groups are chosen, somewhat arbitrarily to separate working years and retirement. In addition, all the mortality models can not apply to the entire age range. In fact, Gompertz [23] suggested that there are 4 distinct periods in the life span between which separate laws of mortality apply: birth to 12 months, 12 months to 20 years 20 years to 60 years and 60 years to 100 years.

4.2 Objectives

The main objective of this chapter is to find a suitable mortality model for people of mortality of districts of Assam. This study proposes a revision of the most commonly used parametric mortality models namely Gompertz, Makeham, Logistic and Beard for describing the mortality pattern of districts of Assam. Ten districts have a sufficiently reliable data to be useful for the specialized purpose of the present analysis. The best fit model has been selected based on a selection criterion.

4.3 Methods and Materials

4.3.1 Data Source

The original data of abridged life tables of Assam are taken from sample registration system (SRS). Then the required complete life tables are constructed by using the methodology used Saikia and Borah [1].

The four mortality models i.e. Gompertz, Makeham, Logistic, and Beard have been considered here. These nonlinear models can be written in the form as

$$y_i = f(x_i, \mathbf{B}) + \varepsilon_i, \tag{4.5}$$

$i = 1,2,\cdots,n$, where y is the response variables, x is the independent variable, **B** is the vector of parameters β_j to be estimated $(\beta_1, \beta_2, \cdots, \beta_p)$, ε_i is a random error term, p is the number of unknown parameters, and n is the number of observations [8]. The estimators of β_j's are found by minimizing the sum of squares residual (SS_{Res}) function

$$(SS_{Res}) = \sum_{i=1}^{n} [y_i - f(x_i, \boldsymbol{B})]^2, \tag{4.6}$$

under the assumption that the ε_i are normal and independent with mean zero and common variance σ^2. Since y_i and x_i are fixed observations, the sum of squares residual is a function of **B**. Least squares estimates of **B** are values which when substituted into equations **(4.6)** will make the (SS_{Res}) a minimum and are found by differentiating equations **(4.6)** with respect to each parameter and setting the result to zero. This provides the p normal equations that must be solved for $\hat{\mathbf{B}}$. These normal equations take the form

$$\sum_{i=1}^{n}\{y_i - f(x_i, \mathbf{B})\}\left[\frac{\partial f(x_i, \mathbf{B})}{\partial \beta_j}\right] = 0 \tag{4.7}$$

for $j = 1,2, \cdots, p$. The parameters of the models are estimated using the Levenverg - Marquardt iteration method based complete life table for 10 districts of Assam both males and females.

The Levenberg-Marquardt method is a standard iterative technique for nonlinear least-squares problems. The Levenberg-Marquardt curve-fitting method is a combination of steepest descent and the Gauss-Newton method. In the gradient descent method, the sum of the squared errors is reduced by updating the parameters the steepest-descent direction. In the Gauss-Newton method, the sum of the squared errors is reduced by assuming the least squares function is locally quadratic, and finding the minimum of the quadratic. When the parameters are far from the correct one, the method acts more like a gradient-descent method; descent method: slow, but guaranteed to converge. When the parameters are close to the correct solution, it becomes a Gauss-Newton method. The LM algorithm is given by

$$(J^T W J + \lambda I)\delta = J^T r$$

where J is the Jacobin matrix of derivatives of the residuals with respect to the parameters, λ is the damping parameter and r is the residual vector. Matlab version 7.11.0 has been used for the estimation of the parameters.

4.3.2 Selection Criteria of Best Fit Model

After fitting the models using different districts of Assam, the best fit model is selected based on the following selection criteria. The two goodness-of-fit measures are considered here. The first measure is the root mean square error (RMSE) and the other is the sum squared error (SSE). Matlab software evaluates these goodness-of-fit statistics for parametric models:

Sum of Squares Due to Error: This statistic measures the total deviation of the response values from the fit to the response values. It is also called the summed square of residuals and is usually labeled as *SSE*.

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where y_i is observed values and $\hat{y}_i$ is the predicted values for $i = 1, 2, \cdots, n$. A value closer to 0 indicates that the model has a smaller random error component and that the fit will be more useful for prediction.

The Root Mean Square Error (RMSE): The RMSE measures to aggregate the residuals into one measure of predictive power. The RMSE of a model prediction with reference to the calculable variable is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}},$$

Where y_i is observed values and $\hat{y}_i$ is the predicted values for $i = 1, 2, \cdots, n$. By comparing the RMSE, an attempt has been made to select the best result. A *RMSE* value closer to 0 indicates a fit that is more useful for prediction.

Coefficient of determination($\boldsymbol{R^2}$): The coefficient of determination (R^2) is also evaluated. R-square is the square of the correlation between the response values and the predicted response values. This statistic measures how successful the fit is in explaining the variation of the data. This value indicates how well data point fits a mortality model. R-square can take on any value between 0 and 1, with a value closer to 1 indicating that a greater proportion of variance is accounted for by the model. Generally it, is not impossible for R^2 to actually attain1, if pure error exist. In practice, sometime negative value of R^2 may occur if the fit is actually worse than just fitting a horizontal line then R-square is negative. In this case, R-square cannot be interpreted as the square of a correlation. Theoretically the value 1 indicates a perfect fit, 0 reveals that the model is not a better than the simple average and negative value indicate a poor model. If the value of R^2 is above 0.9, it is accepted as efficient. The mathematical formulation of the coefficient of determination is,

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2},$$

where $\bar{y}$ is the mean of the response variables.

4.4 Results and Discussion

Four mortality models are fitted to 10 districts of Assam for both male and female. Here we consider five high population growth rate districts and five low population growth rate districts. The high population growth rate districts considered here are Darang, Dhubri, Goalpara, Morigaon and Nagaon and low population growth rate districts are Golaghat, Jorhat, Sibsagar, Dibrugarh, and Tinsukia. In view of our model choice criteria as discussed above, the results are summarised as below.

The estimation of SSE, R^2, RMSE are evaluated for fitting of each model for the age groups 20-60 and 60-100 for both male and female and are presented from **Table 4.1** to **Table 4.4**. **Table 4.1** and **Table 4.2** represent the values of SSE, R^2, RMSE for male and female respectively for the age group 20-60. **Table 4.3** and **Table 4.4** represent the same for the age group 60-100. It is clear from our results that for all cases R^2 value is significant. R-squared is a statistical measure of how close the data are to the fitted regression line. All in all, the higher the R-squared, the better the model fits your information. It is watched that estimations of the SSE and RMSE are minimum for Logistic model when contrasted with the others. It is likewise observed that Makeham show gives preferred outcomes over the others. For the age group 20-60, logistic model is the first preferred model for both high and low population growth rate districts for male. The fit is good in individual districts, the deviations small. On the other hand for female population the three parameter makeham model is a preferable model. Beard model also gives a better result for low population growth rate districts Dibrugarh, Golaghat, Jorhat, Tinsukia, and Sibsagar. On the other hand for the female population, Makeham model is appropriate for high population growth rate districts as well as low population growth rate districts.

It is also noted from our results that for the other age group 60-100, the logistic model is more suitable model for every one of the districts for male population whereas Makeham model is the preferable model for the female population.

Table 4.1: Estimated Values of SSE, $\boldsymbol{R^2}$, RMSE for the age group 20-60 (Male).

Districts	Gompertz			Makeham			Logistic			Beard		
	SSE	R^2	RMSE	SSE	R^2	RMSE	SSE	R^2	RMSE	SSE	R^2	RMSE
Darang	.00027	.9253	.00216	.00027	.9254	.00217	.00025	.9255	.00119	.00027	.9253	.00217
Dhubri	.00031	.9213	.00232	.00031	.9221	.00232	.00029	.9236	.00132	.00031	.9213	.00234
Dibrugarh	.00035	.8849	.00244	.00035	.8868	.00246	.00017	.9446	.00174	.00023	.9249	.00201
Goalpara	.00027	.9254	.00216	.00027	.9255	.00218	.00025	.9256	.00120	.00027	.9254	.00218
Morigaon	.00027	.9210	.00214	.00027	.9210	.00216	.00024	.9233	.00114	.00026	.9216	.00215
Nagaon	.00028	.9123	.00221	.00028	.9124	.00222	.00022	.9296	.00201	.00027	.9157	.00218
Golaghat	.00029	.9069	.00226	.00039	.7412	.00322	.00021	.9335	.00194	.00027	.9137	.00219
Jorhat	.00034	.8882	.00242	.00034	.8882	.00244	.00015	.9513	.00163	.00023	.9231	.00203
Sibsagar	.00031	.9005	.00232	.00038	.8514	.00366	.00017	.9455	.00174	.00026	.9153	.00215
Tinsukia	.00031	.9015	.00231	.00031	.9015	.00233	.00017	.9447	.00176	.00027	.9148	.00216

Table 4.2: Estimated Values of SSE, $\boldsymbol{R^2}$, RMSE for the age group 20-60 (Female).

Districts	Gompertz			Makeham			Logistic			Beard		
	SSE	R^2	RMSE	SSE	R^2	RMSE	SSE	R^2	RMSE	SSE	R^2	RMSE
Darang	.00034	.8349	.00241	.00013	.9855	.00203	.00033	.8847	.00209	.00029	.8533	.00227
Dhubri	.00034	.8289	.00241	.00013	.9799	.00201	.00024	.8521	.00205	.00029	.8504	.00227
Dibrugarh	.00038	.7082	.00256	.00018	.9814	.00220	.00029	.7526	.00224	.00032	.7516	.00238

Goalpara	.00035	.7896	.00245	.00012	.9995	.00201	.00025	.8494	.00211	.00030	.8184	.00229
Morigaon	.00036	.7590	.00249	.00016	.9549	.00210	.00026	.8249	.00216	.00031	.7934	.00232
Nagaon	.00037	.7383	.00251	.00017	.9907	.00213	.00037	.8175	.00219	.00031	.7764	.00234
Golaghat	.00036	.7672	.00248	.00016	.9817	.00211	.00036	.8317	.00214	.00031	.8001	.00231
Jorhat	.00036	.7335	.00250	.00018	.9922	.00212	.00029	.7922	.00225	.00031	.7722	.00233
Sibsagar	.00037	.7383	.00251	.00015	.9077	.00212	.00027	.7079	.00219	.00031	.7764	.00234
Tinsukia	.00037	.7414	.00251	.00017	.9404	.00213	.00037	.8105	.00219	.00031	.7790	.00234

Table 4.3: Estimated Values of SSE, $\boldsymbol{R^2}$, RMSE for the age group 60-100 (Male).

Districts	Gompertz			Makeham			Logistic			Beard		
	SSE	R^2	RMSE	SSE	R^2	RMSE	SSE	R^2	RMSE	SSE	R^2	RMSE
Darang	.00126	.9961	.00358	.00062	.9981	.00253	.00055	.9983	.00239	.00071	.9878	.00271
Dhubri	.00553	.9670	.00752	.00227	.9864	.00484	.00222	.9867	.00381	.00248	.9752	.00505
Dibrugarh	.00775	.9950	.00890	.00746	.9952	.00877	.00225	.9979	.00382	.00325	.9779	.00579
Goalpara	.00141	.9954	.00379	.00069	.9978	.00266	.00052	.9980	.00254	.00081	.9774	.00289
Morigaon	.00105	.9982	.00327	.00084	.9986	.00295	.00030	.9993	.00203	.00043	.9893	.00210
Nagaon	.00238	.9972	.00492	.00022	.9974	.00477	.00102	.9988	.00227	.00103	.9918	.00327
Golaghat	.00337	.9967	.00587	.00319	.9969	.00573	.00148	.9985	.00293	.00149	.9945	.00392
Jorhat	.00745	.9951	.00872	.00716	.9953	.00859	.00314	.9979	.00272	.00314	.9919	.00569
Sibsagar	.00467	.9961	.00690	.00446	.9993	.00678	.00204	.9983	.00361	.00205	.9913	.00459
Tinsukia	.00446	.9962	.00675	.00426	.9983	.00662	.00195	.9983	.00251	.00199	.9953	.00449

Table 4.4: Estimated Values of SSE, $\boldsymbol{R^2}$, RMSE for the age group 60-100 (Female).

Districts	Gompertz			Makeham			Logistic			Beard		
	SSE	R^2	RMSE	SSE	R^2	RMSE	SSE	R^2	RMSE	SSE	R^2	RMSE

Darang	.07156	.9291	.02702	.05076	.9498	.02103	.07086	.8298	.02717	.07133	.9293	.02712
Dhubri	.07055	.9277	.02683	.05975	.9484	.02283	.06984	.9244	.02697	.07033	.9279	.02693
Dibrugarh	.08192	.8361	.03487	.05333	.9384	.03080	.07458	.8121	.04121	.1185	.837	.03495
Goalpara	.07557	.9101	.02777	.05463	.9415	.02176	.07473	.9011	.0279	.07534	.9102	.02787
Morigaon	.08758	.8874	.02989	.05643	.9488	.02187	.08651	.8888	.03002	.08728	.8878	.03001
Nagaon	.09871	.8687	.03174	.05745	.9204	.02917	.09742	.8704	.03186	.09832	.8692	.03184
Golaghat	.08382	.8941	.02925	.05284	.9154	.02122	.08282	.8954	.02937	.08355	.8945	.02935
Jorhat	.09965	.8480	.03372	.05912	.9501	.02410	.09899	.8501	.03383	.1109	.8488	.03381
Sibsagar	.09871	.8687	.03174	.05745	.9704	.02171	.09742	.8704	.03186	.09832	.8692	.03184
Tinsukia	.09689	.8716	.03144	.05567	.9732	.02144	.09564	.8733	.03156	.09652	.8721	.03154

4.5 Conclusion

In this study, four mortality models namely Gompertz, Makeham, Logistic and Beard have been fitted for ten different districts of Assam by using Levenberg Marquardt iteration method. The investigation in this study is somewhat lengthy, but the results can be described concisely. All the models are all far closer to the observed values, and at first sight, it is not easy to choose between them. No single model is always best. For both the age groups, the four parameter logistic model is the most successful of the four in describing the mortality in high and low population growth rate districts for male population. From a theoretical point of view, the logistic model likewise has some imperative points of interest. It is more broad than alternate models and will work in circumstances where they won't. It additionally has some informative support, there are hypotheses regarding why it works and about conditions in which it may come up short. On the other hand, in case of female mortality the Makeham model gives better results than the others. The other models also give satisfactory results. Based on our results it is noted that for both the age groups and for high and low population growth rate districts logistic and Makeham model are preferred first for male and female. As per the aftereffects of our estimation, it can be presumed that the logistic model and Makeham model are more reasonable over the remaining models for describing the mortality pattern of districts of Assam for male and female respectively.

*****