

PREDICTION NUMBER OF DEATHS BY OCCURRENCE IN MALAYSIA: A COMPARISON BETWEEN SIMPLE LINEAR REGRESSION MODEL AND HOLT'S LINEAR TREND MODEL

Mohamad Adam Bujang, Tassha Hilda Adnan, Premaa Supramaniam,
Abd Muneer Abd Hamid, Jamaayah Haniff¹

Abstract

The purpose of this paper is to determine the better model between Simple Linear Regression model (SLR) or Holt's Linear Trend model to predict the number of deaths in Malaysia. The analysis was compared using two different methods, two different approaches and three different time periods. The best prediction techniques were measured using the smallest root mean squared error (RMSE). Findings indicated that RMSE of Holt's Linear Trend showed advantage in one-step-ahead approach for a long period of time and Simple Linear Regression showed slight advantage in one, two and three-step-ahead approach for a short period of time. Hence, the better model for predicting the number of deaths by occurrence in Malaysia depends on the approach used and time period of duration.

Keywords: Simple Linear Regression, Holt's Linear Trend, and Root Mean Squared Error.

Introduction

Mortality prediction has its own significant benefits and the most crucial element would be on human resource management and health care allocation planning. Realising the importance of death prediction, this paper was produced with the main purpose on which method between Simple Linear Regression model and Holt's Linear Trend model could provide a better prediction.

Somehow, the more complicated a model is, the better the prediction. It is because the complicated models accounted for few strong predictors and this will of course incur more time, energy and cost in collecting the information. Therefore, sometimes the simpler model is preferable because it will reduce time, energy and cost and could predict as good as or most likely how the complicated models could predict. Perhaps not for all cases simple model could be a better predictor but for prediction of number of deaths, this paper proved it is worth and still reliable to do so. The mortality records by occurrence from the Department of Statistics, Malaysia (DOSM) for year 1980 to 2008 were used for this study. Mortality by occurrence is defined as death statistics compiled on the basis of where the death actually occurred (Health Statistics Centre, 2004).

¹ All the authors are working in Clinical Epidemiology Unit, Clinical Research Centre, Malaysia. Mohamad Adam Bujang and Premaa Supramaniam are working as Research Officer, Tassha Hilda Adnan and Abd Muneer Abd Hamid are Statistician (Department of Statistics), and Jamaayah Haniff is the Head of Unit for Clinical Epidemiology Unit, Clinical Research Centre, Malaysia.

There are three benefits emphasized in this paper. First, authors would like to prove in certain conditions, simple model could also give good prediction. Second, the authors would like to establish the extent of the population estimates from the DOSM could effectively predict the number of deaths. It can be one of the indicators to check the reliability of the population estimates in predicting the death based on the coefficient of determination, R^2 derived from Simple Linear Regression model. Third, the public can use the model if prediction on the number of deaths is a concern. The statistics on population projections and number of deaths used in this study have been compiled by DOSM since 1960s.

Literature Review

The future of human survival has attracted renewed interest in recent decades. The historic rise in life expectancy shows little sign of slowing, and increased survival is a significant contributor to ageing population. Due to that, mortality projection is an essential input for life expectancy that is used for the projections of the financial development of pension schemes. Government and insurance companies all over the world rely on good mortality projections for efficient administration of their pension commitments.

There are many studies done with regards to mortality forecasting. There were researchers who used sophisticated model such as Keyfitz (1982) assessed various established and rudimentary demographic theories: demographic transition, effects of development, Caldwell's theory concerning education and fertility, urbanisation, income distribution, Malthus theories on population, human capital, the Easterlin effect, opportunity cost, prosperity and fertility and childbearing intentions.

Time series method is one of the common methods for mortality forecasting. This approach is based on the assumption that historical values of variables of interest have been generated by means of statistical model, which also holds for the future. A widely used method is Autoregressive Integrated Moving Average (ARIMA) models. Lazim (2001) stated in his book, "The term ARIMA is in short stands for the combination that comprises of Autoregressive/Integrated/Moving Average Models". Carter (1996) used ARIMA to forecast United States mortality and compare the findings with Structural Time Series models. Results showed marginal differences in fit and forecasts between the two statistical approaches with a slight advantage to structural models.

The Lee-Carter method is regarded as among the best currently available and has been widely applied (e.g., Lee and Tuljapurkar, 1994; Wilmoth, 1996; Lundstrom and Qvist, 2004; Buettner and Zlotnik, 2005). The underlying principle of the Lee-Carter method is the extrapolation of past trends. The method was designed for long-term forecasting based on lengthy time series of historical data.

Methodology

The statistical techniques applied to analyse the data collected from Department of Statistics, Malaysia are briefly described in this section (Table 1). As the purpose of

this paper is to determine the best prediction technique to predict future values of number of deaths in Malaysia, two methods were used; Simple Linear Regression model and Holt's Linear Trend model. These models are good to use when the number of deaths is displaying a trend (growth or decline). The procedure of the analysis was based on past observations in a given time series from 1980 to 2008.

Table 1: Number of deaths (by occurrence) and population in Malaysia, 1980 – 2008

Year	No. of deaths	Population	Year	No. of deaths	Population
1980	73,203	13,879,200	1995	95,103	20,681,800
1981	70,311	14,256,900	1996	95,982	21,222,600
1982	73,254	14,651,100	1997	97,457	21,769,300
1983	76,679	15,048,200	1998	106,166	22,333,500
1984	77,940	15,450,400	1999	112,452	22,909,500
1985	79,015	15,882,700	2000	105,370	23,494,900
1986	77,103	16,329,400	2001	104,609	24,012,900
1987	75,811	16,773,500	2002	105,982	24,526,500
1988	80,057	17,219,100	2003	111,644	25,048,300
1989	81,676	17,662,100	2004	112,700	25,580,900
1990	83,724	18,102,400	2005	113,714	26,127,700
1991	84,221	18,547,200	2006	115,084	26,640,200
1992	86,040	19,067,500	2007	118,167	27,173,600
1993	87,626	19,601,500	2008	123,286	27,728,700
1994	90,079	20,141,700			

Source: Department of Statistics, Malaysia (number of deaths by occurrence) and <http://www.statistics.gov.my/portal/index.php?lang=e> (population estimate)

The data were split into three categories; long, moderate and short period. Each period was then analysed using two approaches; one-step-ahead approach and one, two and three-step-ahead approach. Table 2 summarizes the method used for analysis.

Table 2: The summary of method used in analysis

Duration	One-step-ahead		One, two and three-step-ahead	
	Fitted	To forecast	Fitted	To forecast
Long	1980-2005	2006	1980-2005	2006
	1980-2006	2007		2007
	1980-2007	2008		2008
Moderate	1990-2005	2006	1990-2005	2006
	1990-2006	2007		2007
	1990-2007	2008		2008
Short	2000-2005	2006	2000-2005	2006
	2000-2006	2007		2007
	2000-2007	2008		2008

For one-step-ahead, each period of duration has three different series of data while in the second approach, each duration has only a series of data. As such, methodology was designed to answer whether different approaches in different periods could affect the forecast or prediction. Linear Regression and Time Series Analysis methods were applied to predict or forecast the number of deaths by occurrence in Malaysia and analysis was made using Simple Linear Regression model and Holt's Linear Trend model. Simple Linear Regression model used population estimates from the DOSM as the only predictor for the number of deaths by occurrence.

There is no seasonal pattern observed in the trend of mortality. Abridged Life Tables, Malaysia (2009) published by DOSM confirmed that the trend of deaths was consistent. Thus, these two models were selected because the trend of deaths cases is linear. Holt's Linear Trend was selected for univariate forecasting technique because this method is more stable and flexible as compared to the other univariate forecasting technique.

Subsequently, the error measure of RMSE was performed in order to ensure the accuracy and consistency of the models and to compare the model's forecasting performance. The model that stated the smallest RMSE is thus selected as the most suitable model.

In summary, the steps involved are as follows:

- (i) Model development: Data from different period duration (as in Table 1) were used for model development using Simple Linear Regression and Holt's Linear Trend models. As a result, the model provided the equation model together with the RMSE.
- (ii) Model application: Use model generated from the 'Model Development' to forecast number of deaths in 2006, 2007 and 2008.
- (iii) Model validation: Calculate RMSE based on the difference between actual number of deaths (2006 to 2008) and the predicted/forecasted values from the two models (2006 to 2008). RMSE values were compared between the two models and thus, the better model was determined by the smallest RMSE.

(a) Simple Linear Regression Model (SLR)

Regression analysis is a statistical methodology that utilises the relation between two quantitative variables, so that one variable can be predicted from the other. For this study, the number of deaths was predicted by using the population estimate data obtained from the Department of Statistics, Malaysia. The formula is as follows:

$$y = \hat{\alpha} + \hat{\beta}(x)$$

x denotes the independent variable (population estimate from the Department of Statistics, Malaysia) and y is the dependent variable (number of deaths).

(b) Holt's Linear Trend Model

Holt's method is the basic trend model, \hat{Y}_{t+1} = estimated level t + trend t , combined with exponential smoothing. The following are the equations required to apply this method:

The exponentially smoothed series,

$$S_t = \alpha Y_t + (1 - \alpha)(S_{t-1} + T_{t-1})$$

The trend estimate,

$$T_t = \beta (S_t - S_{t-1}) + (1 - \beta)T_{t-1}$$

Forecasts for l-step-ahead are computed using,

$$F_{t+l} = S_t + T_t(l)$$

Where, α is a smoothing constant ($0 < \alpha < 1$)

β is a smoothing constant for trend estimate ($0 < \beta < 1$)

The basic consideration is the search for the best values of α and β that produced the smallest value of the RMSE.

(c) Measuring Models' Performance by Root Mean Squared Error (RMSE)

Taking the square root of MSE yields the root mean squared error or RMSE. The RMSE is a frequently-used measure of the differences between values predicted by a model and the values of actual observation, and it is a good measure of accuracy. It was written as,

$$RMSE = \sqrt{\frac{\sum_{t=1}^n e_t^2}{n}} \quad \text{for which } e_t = Y_t - \hat{Y}_t$$

Where, Y_t is the actual observed value in time t

\hat{Y}_t is the fitted value in time t generated from the origin

$t = 1, 2, 3, \dots, n$

n is the number of period summed

The RMSE also gives equal weights to all the errors, irrespective of any time period.

Results

Based on Figure 1, the actual trend is approaching linear and therefore, there is a possibility to predict the trend based on these two models, and results indicated that the two models predict the actual trend effectively. Figure 1 showed that Holt's Linear Trend method follows the actual trend compared to SLR, whereby SLR only followed the actual trends based on the straight line.

Chart 1: Actual number of deaths predicted from SLR and forecasted from Holt's Linear Trend, mortality by occurrence from 1980 to 2008

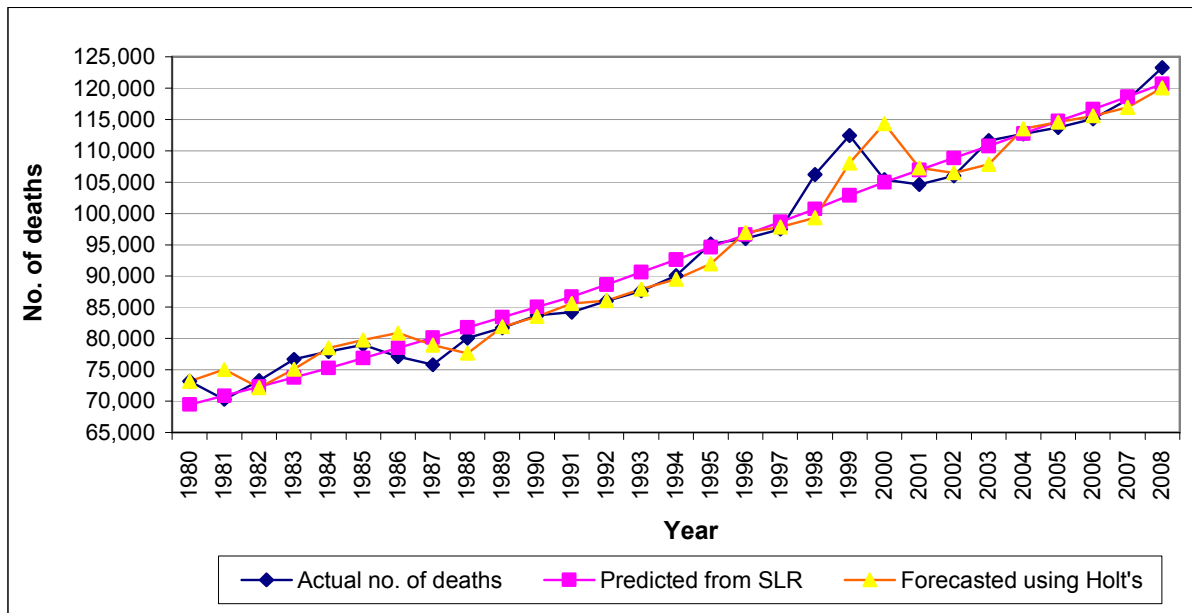


Table 3: Comparison of predicted/forecasted values, RMSE by two different models and period of duration using one-step-ahead approach

Duration		Year	Actual no. of deaths	Predicted no. of deaths: One-step-ahead				Which method is better?
				Simple Linear Regression		Holt's Method		
				Predicted value	RMSE	Fitted value	RMSE	
Long period	1980 - 2005	2006	115,084	116,473	1,934.52	115,529	1,117.55	Holt's
	1980 - 2006	2007	118,167	118,216		119,615		
	1980 - 2007	2008	123,286	120,237		122,081		
Moderate period	1990 - 2005	2006	115,084	117,872	2,103.06	117,764	2,071.22	Holt's
	1990 - 2006	2007	118,167	119,364		117,200		
	1990 - 2007	2008	123,286	121,271		121,106		
Short period	2000 - 2005	2006	115,084	116,178	2,056.16	116,019	2,138.18	SLR
	2000 - 2006	2007	118,167	117,628		117,480		
	2000 - 2007	2008	123,286	119,940		119,769		

Based on one-step-ahead approach, there were slight differences in terms of predicted/forecasted values using both models, whereby RMSE under Holt's Linear Trend showed slight advantage in long and moderate period of duration.

Table 4: Comparison of predicted values, RMSE by two different models and period of duration using one, two and three-step-ahead approach, mortality by occurrence, 1980 - 2008

Duration	Year	Actual no. of deaths	Predicted no. of deaths: One, two and three-step-ahead				Which method is better?	
			Simple Linear Regression		Holt's Method			
			Predicted value	RMSE	Fitted value	RMSE		
Long period	1980 - 2005	2006	115,084	116,473	1,816.79	115,529	2,443.19	SLR
		2007	118,167	118,434		117,344		
		2008	123,286	120,475		119,159		
Moderate period	1990 - 2005	2006	115,084	117,872	2,022.85	117,764	1,974.50	Holt's
		2007	118,167	120,012		119,937		
		2008	123,286	122,238		122,111		
Short period	2000 - 2005	2006	115,084	116,178	1,772.71	116,019	1,910.15	SLR
		2007	118,167	118,256		118,066		
		2008	123,286	120,418		120,114		

Based on one, two and three-step-ahead approach, there were also slight differences in terms of predicted/forecasted values using both models, whereby RMSE under SLR showed slight advantage in long and short period of duration.

Discussion

In short period, SLR showed a slight advantage compared to Holt's Linear Trend in one-step-ahead and also for one, two and three-step-ahead approach, while in moderate period of time, Holt's Linear Trend showed better prediction. As for long period, Holt's Linear Trend model is better in one-step-ahead approach while SLR is better in another approach.

The consistent result was found when predicting the CDR, except in one-step-ahead under moderate time period, the RMSE for both models were the same (result in two decimal points).

There were several points highlighted in this paper. First, both techniques whether regression method (SLR) or forecasting technique (Holt's Linear Trend) are reliable techniques to predict number of deaths by occurrence yearly in Malaysia although there were a slight difference in terms of result. SLR is more appropriate to use in one, two, and three-step-ahead while Holt's Linear Trend is more suitable to use in one-step-ahead approach.

Second, duration plays an important role in the prediction. Referring to one-step-ahead approach, both methods could predict better in longer series of data although Holt's Linear Trend showed more advantage. While based on the second approach, shortened series of data tends to give smaller RMSE and this was somewhat inconsistent with the earlier approach.

Third, this study has also found that approximately 87 to 96 per cent of the variation in the number of deaths can be explained by the population estimates based on SLR model from three different periods. Thus, population estimates from the Department of Statistics, Malaysia is a good and reliable predictor to estimate the number of deaths by occurrence in Malaysia. This is because the number of deaths in Malaysia is almost consistent over the years.

Fourth, the findings indicated that, both models were not reliable for long term forecast. In addition, several studies indicated that forecast accuracy is better for shorter forecast durations, and that it is better for large populations rather than for small populations (Keilman and Pham, 2004). Duration dependence of forecast accuracy is explained by the fact that the more years a forecast covers, the greater is the chance that unforeseen developments will produce unexpected changes in fertility, mortality, or migration (Keilman, 2005). Therefore, the model developed in this study is only relevant to predict one to two years ahead.

Fifth, different approaches also play an important factor in the prediction. One-step-ahead approach is a better approach for Holt's Linear Trend in longer series of data, while different finding for SLR whereby the one, two and three-step-ahead showed smaller RMSE.

This study had proven that there is a possibility to predict death cases in Malaysia using these two techniques. Since the number of deaths can be predicted, so crude death rate (CDR) can also be obtained. The crude death rate for year per thousand populations is calculated using this formula:

$$CDR_t = \left(\frac{a_t}{b_t} \right) \times 1000$$

Where, a_t denotes total number of deaths in year t

b_t denotes total population estimates in year t

Table 5: Comparison of Crude Death Rate (CDR per '000) values by two different models and duration using one-step-ahead approach

Duration		Year	Actual CDR	Crude death rate: One-step-ahead				Which method is better?
				SLR		Holt's Method		
				CDR	RMSE	CDR	RMSE	
Long period	1980 - 2005	2006	4.32	4.37	0.07	4.34	0.04	Holt's
	1980 - 2006	2007	4.35	4.35		4.40		
	1980 - 2007	2008	4.45	4.34		4.40		
Moderate period	1990 - 2005	2006	4.32	4.42	0.08	4.42	0.08	Same
	1990 - 2006	2007	4.35	4.39		4.31		
	1990 - 2007	2008	4.45	4.37		4.37		
Short period	2000 - 2005	2006	4.32	4.36	0.07	4.36	0.08	SLR
	2000 - 2006	2007	4.35	4.33		4.32		
	2000 - 2007	2008	4.45	4.33		4.32		

Based on one-step-ahead approach, there was a slight difference in terms of CDR values using both models. RMSE under Holt's Linear Trend showed a slight advantage in long period, while SLR showed a slight advantage in short period.

Table 6: Comparison of Crude Death Rate (CDR per '000) values by two different models and duration using one, two and three-step-ahead approach

Duration		Year	Actual CDR	Crude death rate: One, two and three-step-ahead				Which method is better?
				SLR		Holt's Method		
				CDR	RMSE	CDR	RMSE	
Long period	1980 - 2005	2006	4.32	4.37	0.07	4.34	0.09	SLR
		2007	4.35	4.36		4.32		
		2008	4.45	4.34		4.30		
Moderate period	1990 - 2005	2006	4.32	4.42	0.08	4.42	0.07	Holt's
		2007	4.35	4.42		4.41		
		2008	4.45	4.41		4.40		
Short period	2000 - 2005	2006	4.32	4.36	0.06	4.36	0.07	SLR
		2007	4.35	4.35		4.34		
		2008	4.45	4.34		4.33		

Based on one, two and three-step-ahead approach, there were also a slight differences in terms of CDR values using both models, whereby RMSE under SLR showed a slight advantage in long and short periods.

Chart 2: Actual number of deaths versus Holt's Linear Trend fitted values and Simple Linear Regression predicted values from long period using one-step-ahead approach

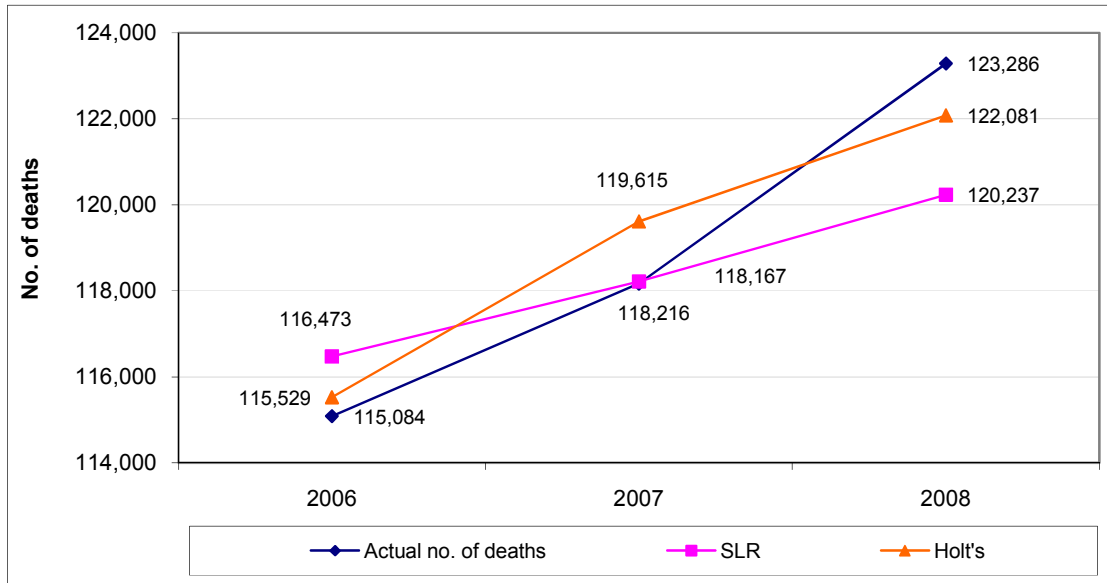


Chart 3: Actual number of deaths versus Holt's Linear Trend fitted values and Simple Linear Regression predicted values from moderate period using one-step-ahead approach

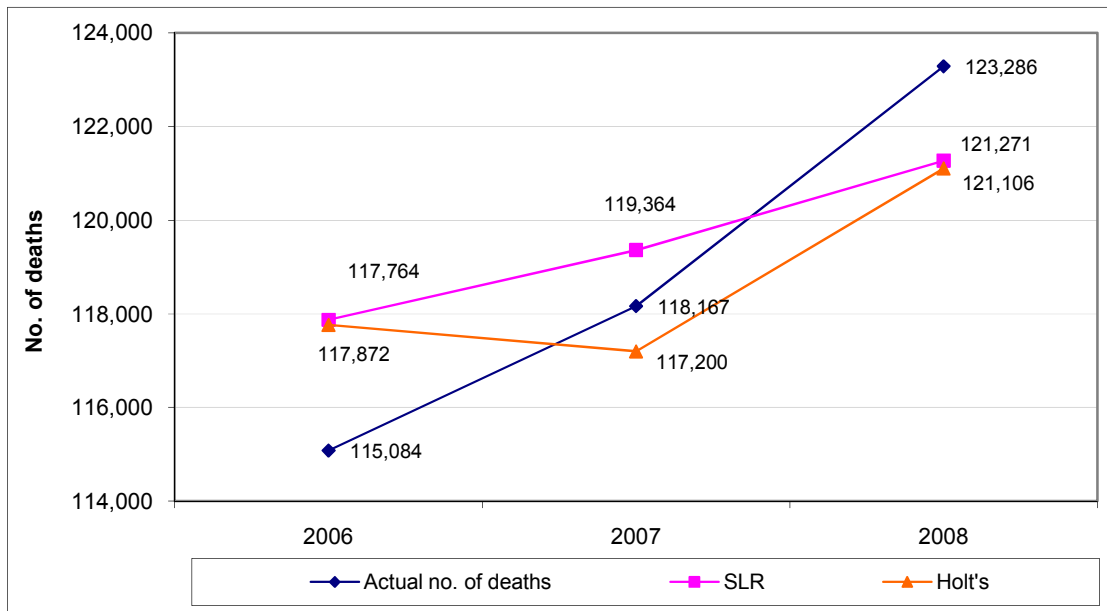


Chart 4: Actual number of deaths versus Holt's Linear Trend fitted values and Simple Linear Regression predicted values from short period using one-step-ahead approach

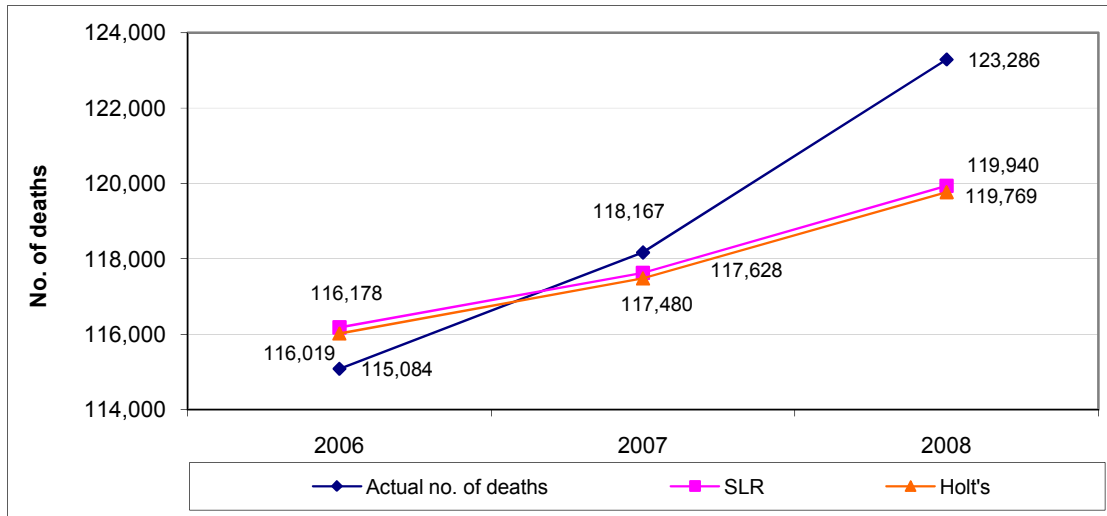


Chart 5: Actual number of deaths versus Holt's Linear Trend fitted values and Simple Linear Regression predicted values from long period using one, two and three-step-ahead approach

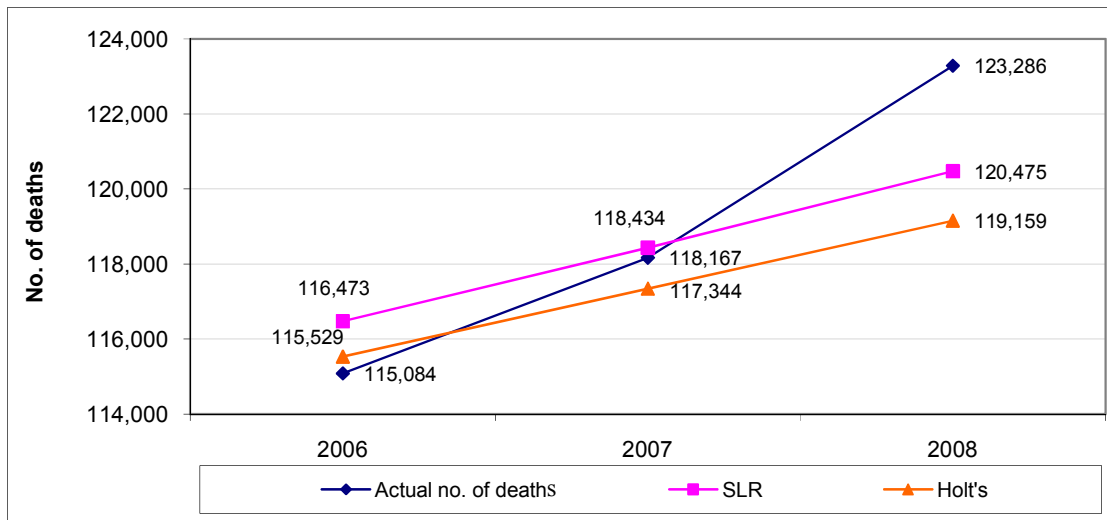


Chart 6: Actual number of deaths versus Holt’s Linear Trend fitted values and Simple Linear Regression predicted values from moderate period using one, two and three-step-ahead approach

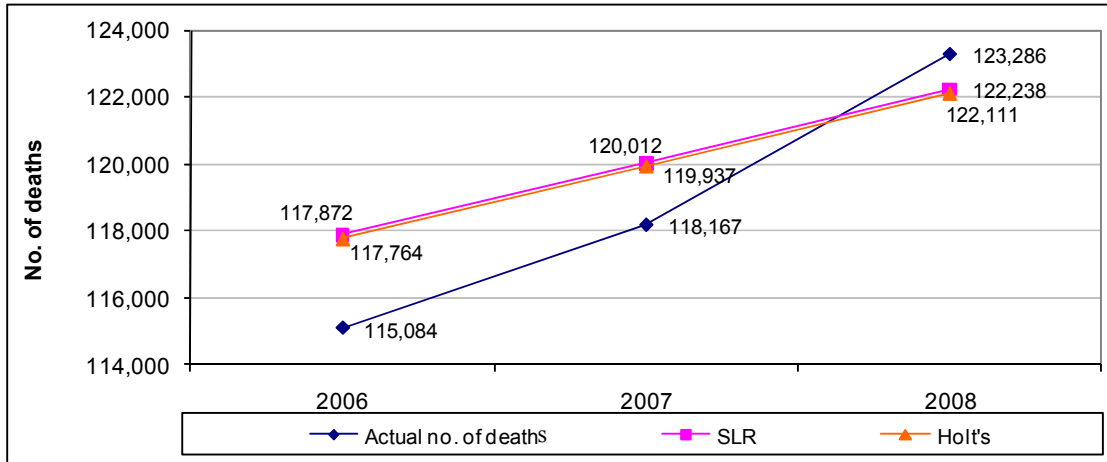
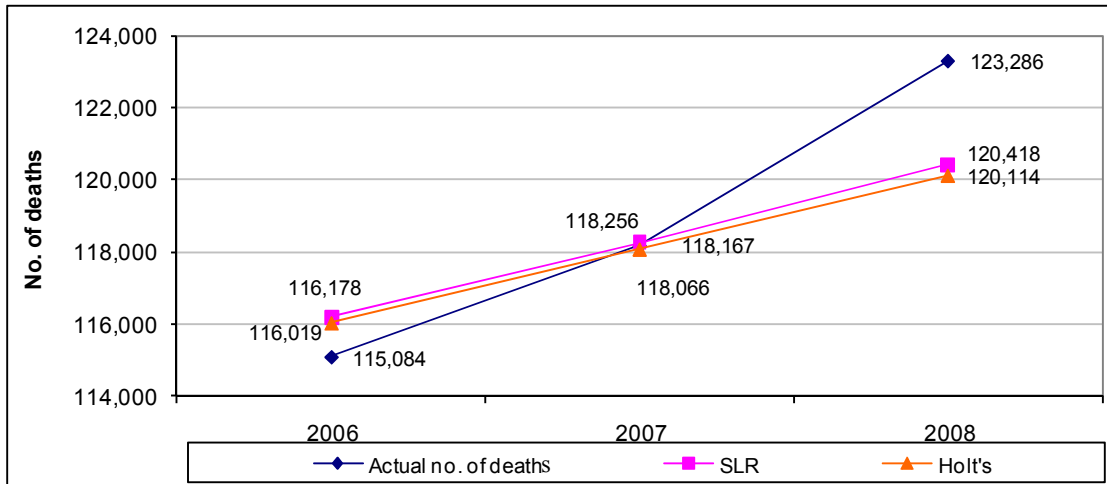


Chart 7: Actual number of deaths versus Holt’s linear Trend fitted values and Simple Linear Regression predicted values from short period using one, two, three-step-ahead approach



Conclusion

Both techniques either SLR or Holt’s Linear Trend can be used to predict number of deaths by occurrence yearly in Malaysia. Few considerations have to be made such as the series of data and the approach to be used, for example whether to apply one-step-ahead or predict few years ahead using the same model.

Acknowledgement

The authors extend their deepest gratitude to the Department of Statistics, Malaysia for providing the data sets for mortality records, 1980 - 2008 and population estimates, 1980 – 2008.

References

- Abridged Life Tables 2006 – 2008. (2009). Department of Statistics, Malaysia.
- Anderson, D. R., Sweeney, D.J. & Williams, T.A. (2005). *Statistics for Business and Economics*. Thomson South-Western.
- Bowerman, B.L., O'Connell, R.T. & Koehler, A.B. (2005). *Forecasting, Time Series, and Regression*. Thomson Brooks/Cole.
- Buettner T, Zlotnik H. (2005). "Prospects for increasing longevity as assessed by the United Nations", *Genus* (1) 213-233.
- Carter. (1996). A Comparison of Box-Jenkins ARIMA and Structural Time Series Models, *The Sociological Quarterly*, 37 (1): 127-144.
- Cohen (1992), A Power Primer, *Psychological bulletin*, American Psychological Association, 112 (1),155-159
- Health Statistics Centre. (2004), *Vital Statistics 2004*. Online search on the June 15th 2010: <http://www.wvdhhr.org/bph/oehp/vital04/METHOD.HTM>
- Lazim, M.A. (2001). *Introductory Business Forecasting: A Practical Approach*, Univision Press Sdn. Bhd.
- Lee R D, Tuljapurkar S. (1994). "Stochastic population forecasts for the United States: Beyond high, medium, and low", *Journal of the American Statistical Association*, 89:1175-1189.
- Lundstrom H, Qvist J, (2004). "Mortality forecasting and trend shifts: An application of the Lee-Carter model to Swedish mortality data". *International Statistical Review*, 72(1): 37-50.
- Keilman, N and D Q Pham (2004). "Time series errors and empirical errors in fertility forecasts in the Nordic countries". *International Statistical Review*; 72 (1), 5-18.
- Keilman. N (2005) "Perspective on Mortality Forecasting II", *Social Insurance Study No.2*, Swedish Social Insurance Agency
- Keyfitz N (1982). "Population change and social policy". Cambridge, Mass: Abt Books.
- Wilmoth J R (1996), "Mortality projections for Japan: A comparison of four methods". In: Caselli G, Lopez A, editors. *Health and mortality among elderly populations*. New York: Oxford University Press: 266-287.