Proc. of Int. Conf. on Current Trends in Eng., Science and Technology, ICCTEST

Trend Analysis to Forecast the Prevalence of Lung Cancer in Bengaluru

Manjula S. Dalabanjan[1], Dr. Pratibha Agrawal[2]
[1] Department of Mathematics, Don Bosco Institute of Technology, (Affiliated to VTU), Bengaluru, India
Email:msd2670@gmail.com
[2]AMC Engineering College (Affiliated to VTU), Bengaluru, India

***Abstract*—The trend analysis is carried out using the variate difference method. The variate difference method is useful to decide the nature of trend (1st degree or 2nd degree or 3rd degree etc). There are several models to represent a given time series. No formula can directly be used to measure the random component of the series at any point of time. To select an appropriate polynomial for fitting of time series we can make use of variate difference method. This method also enables us to estimate the random component in the series and forecast the future Age Adjusted Rates.**

I. INTRODUCTION

Time series data of lung cancer in males and females in Bengaluru is obtained from the National Cancer Registry Programme from the year 1982 to 2009[a].
The values corresponding to 1982, 1983 and 1984 are outliers in the data. If we observe the trend of Age Adjusted Rates (AAR) of cancer over the years, we observe oscillations in the graph. Then we can say that the oscillatory AAR of cancer are due to unknowable risk factors affecting cancer. One such risk factor is increasing tendency of pollution in the air. In metropolitan cities, due to pollution in the air, the lungs of most of the children are found to be weak. Air pollution mainly affects the respiratory system of the body. Some organs associated with the respiratory system are Larynx, hypopharynx, lungs and oral cavity.

Literature Survey

Neelabh and Ramanathan (2011) developed a new approach for order selection in autoregressive moving average models using the focused information criterion. This criterion minimizes the asymptotic mean squared error of the estimator of a parameter of interest. Simulation studies indicate that the suggested criterion is quite effective and comparable to the Akaike information criterion, the corrected Akaike information criterion and the Bayesian information criterion in autoregressive moving average order selection. The use of focused information criterion for simultaneous selection of regression variables and order of the error process in a linear regression model with autoregressive moving average errors is also considered. [1]
In time series analysis, once the data have been transformed to fit a zero - mean autoregressive moving average (ARMA) model, we face the problem of order selection. Three celebrated procedures, namely the Akaike information criterion (AIC), the corrected Akaike information criterion and the Bayesian information criterion have been extensively used for ARMA model-order selection. [1]

Grenze ID: 02.ICCTEST.2017.1.158

T. Jaisankar and M Ravikumar made an effort to develop an ARIMA model for tourist arrival to Tamilnadu and to apply the same in forecasting for the years to come. [2]
The study by Jai Sankar and J. Poorvaraghavan, (2012) targets the forecasting export of liquid bulk in Chennai port by using different forecasting techniques. Export through Chennai is done in four categories Container, Break bulk, Dry bulk and Liquid bulk. Liquid bulked is important for Indian revenue and it has been exported through Chennai port. The data for export of liquid bulk from 1987-88 to 2010-11 has taken from Chennai port and analysed using Autoregressive Integrated Moving Average models. ARIMA (0, 1, 1) was selected as its best fit for the data. The forecast of the model illustrates that the export of Liquid bulk from Chennai port would raise to 20, 29, 154 tonnes in 2014-15. [3]
C. Umasankar et al (2012), developed a linear statistical model with first order autoregressive scheme for the errors has been specified and estimated the parameters of the model by an iterative method of estimation using studentized residuals. Later, this proposed model has been used to obtain the feasible forecasts. [4]
In time series analysis the moving average model is common approach for modeling univariate time series models. A moving average model is conceptually a linear regression of the current value of the series against the white noise error terms of one or more prior values of the series. The errors at each time point are assumed to come from the same distribution with location at zero and constant scale. The distinction in this model is that these errors are propagated to future values of the time series. Since the errors are unobservable, the fitting of the moving average models is more complicated than autoregressive models.
P Balasiddamuni et al proposed two modified estimation procedures for time series linear statistical models involving MA(1) and MA(q) process errors.[5]

Variate difference method

Notations:

Let t denote the time in years. Let Z_t denote the Age adjusted rate in the year t.

If the series can be represented as sum of functional part and random component as given below

$$Z_t = a_0 + a_1 t + a_2 t^2 + \ldots\ldots\ldots\ldots a_{k-1} t^{k-1} + \varepsilon_k$$

Where ε's are independent and identically distributed with the following assumptions that

$$E(\varepsilon_i) = 0$$
$$Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$$
$$Var(\varepsilon_i) = V \text{ (say)}$$

The estimate of V is given by $\hat{V}_k = \dfrac{Var(\Delta^k Z_t)}{\binom{2k}{k}}$

Homogeneity of two successive estimate of V cannot be tested by variance ratio test (F-test) since the consecutive terms are not independent. O. Anderson obtained the standard error of $(V_k - V_{k+1})$ and found for large samples

$$R_k = \frac{V_k - V_{k+1}}{V_k} H_{kN}, \qquad H_{kN} \to N(0,1)$$

Where V_k and V_{k+1} are consecutive estimates of V from the $kth, (k+1)th$ differences of Z_t and H_{kN} is a function of k and N. If $|R_k| > 1.96$ the difference is significant at 5% level of significance otherwise not. [7]

$$V_1 = 0.714915,\ V_2 = 0.605725,\ V_3 = 0.532273,\ V_4 = 0.426517$$

$$R_1 = \frac{V_1 - V_2}{V_1} H_{1,25} = 1.3843 (< 1.96)$$

$$R_2 = \frac{V_2 - V_3}{V_2} H_{2,25} = 1.4539 (< 1.96)$$

$$R_3 = \frac{V_3 - V_4}{V_3} H_{3.25} = 2.57$$

Where $H_{1,25} = 9.065$ $H_{2,25} = 11.99$ $H_{3,25} = 12.944$

$R_1 < 1.96$ implies that there is no significant difference between V_1 and V_2 values. Therefore first degree or second degree model can be treated as good fit for forecasting the future values. Also we can observe the fitted trend in Figure: 1. 1

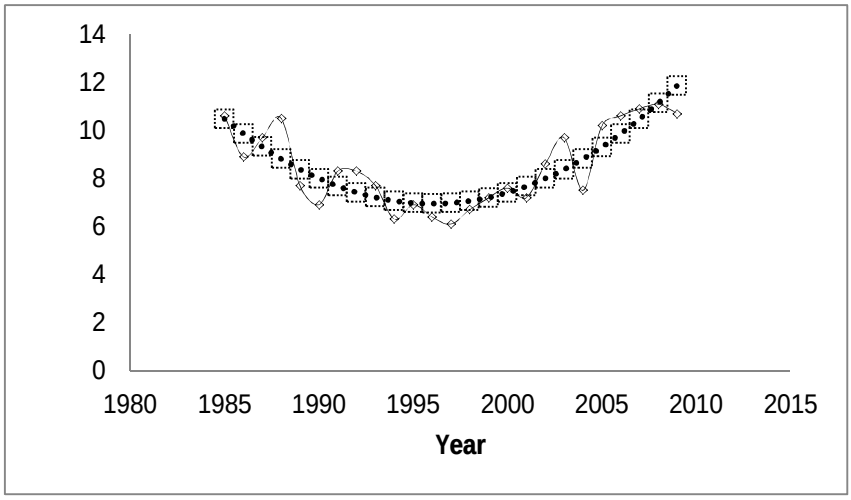

Figure 1. A second degree trend of lung cancer in males in Bengaluru

From the second degree equation,

$$Z_t = a_0 + a_1 t + a_2 t^2, t = 1,2,3 - - - - - - - - - -$$

$$Z_t = 11.1547 - 0.7010t + 0.02919t^2$$

The forecasts of prevalence of lung cancer in the coming years is available in the following table

TABLE I. AGE ADJUSTED RATES OF LUNG CANCER IN MALES IN BENGALURU ARE FORECASTED FOR THE FOLLOWING YEARS

t	26	31	36	41
Year	2010	2015	2020	2025
AAR	12.66	17.47	23.75	31.48

Time series data of lung cancer in females in Bengaluru is obtained from the National Cancer Registry Programme. [a]

In case of females we have $V_1 = 0.293959$, $V_2 = 0.21058$, $V_3 = 0.186932$, $V_4 = 0.17949$

$$R_1 = \frac{V_1 - V_2}{V_1} H_{1.25} = 2.571211$$

$$R_2 = \frac{V_2 - V_3}{V_2} H_{2.25} = 1.358996$$

$R_3 = \frac{V_3 - V_4}{V_3} H_{3.25} = 0.515317$ Since $R_2 < 1.96$ a second degree or third degree model fits well for the data to forecast the future Age Adjusted Rates.

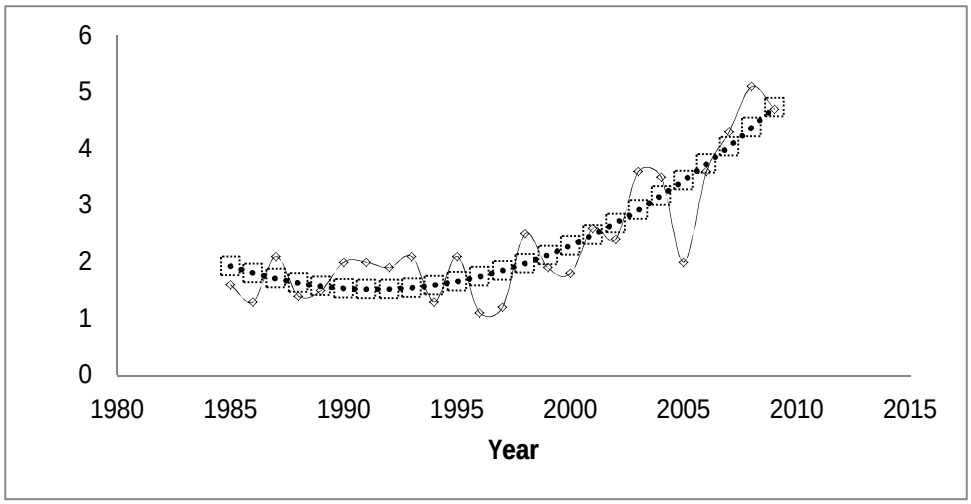

Figure 2. A second degree trend for lung cancer in females in Bengaluru

First we fit a second degree trend to the prevalence of lung cancer in females. The trend equation is

$$Z_t = a_0 + a_1 t + a_2 t^2, t = 1,2,3 - - - - - - - - - -$$

i.e $$Z_t = 3.58389 + 0.14946t + 0.01026t^2, t = 1,2,3 - - - - - - - - - -$$

TABLE II. AGE ADJUSTED RATES OF LUNG CANCER IN FEMALES IN BENGALURU ARE FORECASTED FOR THE FOLLOWING YEARS

t	26	31	36	41
Year	2010	2015	2020	2025
AAR	5.10	8.79	9.96	13.16

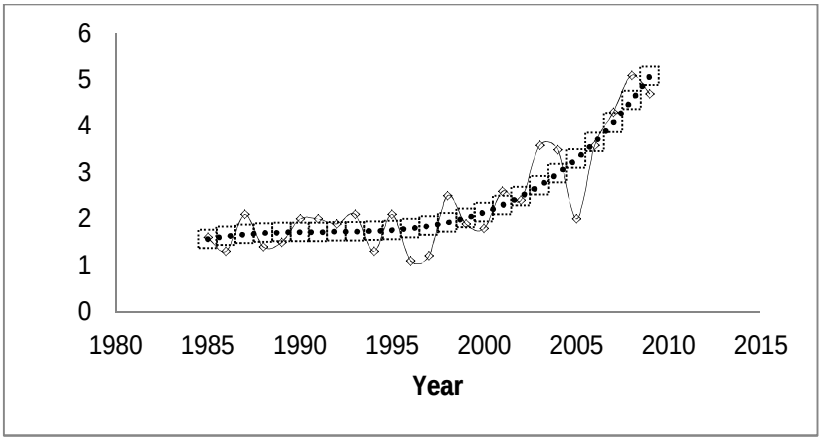

Figure 3. A third degree trend for lung cancer in females in Bengaluru

Secondly we fit a third degree trend to the prevalence of lung cancer in females. The trend equation is

$$Z_t = a_0 + a_1 t + a_2 t^2 + a_3 t^3, t = 1,2,3 - - - - - - - - - -$$

i.e $$Z_t = 1.4855 + 0.09377t - 0.0127t + 0.0005887t^3, t = 1,2,3 - - - - - - - - - -$$

TABLE III. AGE ADJUSTED RATES OF LUNG CANCER IN FEMALES IN BENGALURU ARE FORECASTED FOR THE FOLLOWING YEARS

t	26	31	36	41
Year	2010	2015	2020	2025
AAR	5.64	9.72	15.86	24.54

The forecasted values for the future years give best values when cubic polynomial is fitted for the data.

II. ALTERNATE MEASURES OF FORECASTING ERRORS

The following are the some of the important measures of forecasting errors which help us in selection of an appropriate model for forecasting.

Root Mean Squared Error (RMSE)

This is the statistic whose value is minimized during the parameter estimation process, and it is the statistic that determines the width of the confidence intervals for predictions. The 95% confidence intervals for one-step-ahead forecasts are approximately equal to the point forecast " plus or minus 2 standard errors" ie plus or minus 2 times the RMSE. The RMSE can only be compared between models whose errors are measured in the same units.

The RMSE is given by $$RMSE = \sqrt{\frac{\sum (Z_t - \hat{Z_t})^2}{N}}$$

Mean Absolute Percentage errors (MAPE)

The mean absolute percentage error (MAPE) is also often useful for purposes of reporting, because it is expressed in generic percentage terms which will make some kind of sense even to someone who has no idea what constitutes a "big" error in terms of given units in the data. The MAPE can only be computed with respect to data that are guaranteed to be strictly positive. The MAPE of 10% is considered very good, a MAPE in the range 20% -30% or even higher is quite common.

The MAPE is given by $$MAPE = \frac{100}{N}\sum_{i=1}^{N}\frac{\left|Z_t - \hat{Z}_t\right|}{Z_t}$$

Mean absolute error (MAE)

MAE is another popular error measure that corrects the canceling out effects by averaging the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average.

MAE is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. Where a prediction model is to be fitted using a selected performance measure, in the sense that the least square approach is related to the mean squared error, the equivalent for mean absolute error is least absolute deviations.

The mean absolute error is given by $$\text{MAE}=\frac{1}{N}\sum_{i=1}^{N}\left|Z_t - \hat{Z}_t\right|$$

At the end, we should put more weights on the error measures in the estimation period, most often the RMSE, but sometimes MAE or MAPE, when comparing among models. A model which fails some of the residual tests or reality checks in only a minor way is probably subject to further improvement, where as it is the model which flunks such tests in a major way that cannot be considered as a good model.

TABLE IV. ERROR IN FORECASTING

Model for Lung cancer	RMSE	MAPE	MAE
Males second degree Equation	0.7952	8.0435	0.6696
Lung cancer in Females, second degree Equation	0.510245	22.0209	0.4268
Lung cancer in Females, third degree Equation	0.4804	17.788	0.3949

III. CONCLUSIONS

Since the above models were constructed using variate difference method all the above models are good. The error involved in the estimation of the Z value is not significant.

REFERENCES

[1] Neelabh Rohan and T V Ramanathan, "Order selection in ARMA models using the focused information criterion", *Australian and New Zealand Journal of Statistics*, 2011.217-231

[2] T. Jai Sankar and M Ravikumar, "Time series analysis of tourists arrivals in Tamilnadu*", Proceedings of the Second International Conference on Stochastic Modelling and Simulation* (ICSMS 2012).

[3] T. Jai Sankar and J. Poorvaraaghavan, "Forecasting Export of Liquid Bulk in Chennai Port", *Proceedings of the Second International Conference on Stochastic Modelling and Simulation* (ICSMS 2012).

[4] C. Umasankar, M Vijaybhaskar Reddy, P. Balasiddamuni, K Murali, S. Yadavendra Babu and Sk. Khadar Babu, "An Iterative Statistical First Order Autoregressive Forecasting Model", *Proceedings of the Second International Conference on Stochastic Modelling and Simulation* (ICSMS 2012).

[5] P Balasiddamuni, B. Sarojamma, P Ramakrishna Reddy, S Durga Prasad, J P Naik and Karunakar. "Estimation of Time Series Linear Statistical Model with Moving Average Process Errors",. *Proceedings of the Second International Conference on Stochastic Modelling and Simulation (ICSMS 2012).*

[6] Ling Wu "Stochastic Modeling and Statistical Analysis", *thesis submitted to University of South Florida, copyright 2010.*

[7] P. J. Brockwell and R. A. Davis, "Introduction to Time Series and Forecasting", *Springer,New York,(1996).*