Modelling Breast Cancer Incidences in Sri Lanka using Time Series Analysis

Abstract

In Sri Lankan women, Breast cancer is the most prevalent type of cancer. Sri Lanka has relatively low breast cancer survival rates when compared to more developed countries. Over 3000 new cases of breast cancer are diagnosed each year, according to analysis of Sri Lanka national Cancer Registry. The breast cancer incidence data published in National Cancer Registry, Sri Lanka during 2005 – 2020 was obtained. By understanding and predicting breast cancer rates, healthcare systems and organizations can be better equipped to prevent, detect and treat this disease effectively. The main objective of the study was to predict a suitable model and forecast future breast cancer incidences. In this study, we predicted Quadratic trend model, Exponential trend model and Holt's trend corrected exponential smoothing and did a comparison based on the residuals to find the best model. According to the results, the best prediction model was determined based on accuracy measures Mean Absolute Percentage Error (MAPE),Mean Absolute Deviation (MAD) and Mean Standard Deviation(MSD). By comparing these accuracy measures, Quadratic trend model is selected as the best fitted model for the breast cancer incidences during these 15 years. The forecasted values for years of 2021, 2022 and 2023 are 5267, 5602 and 5950 respectively.

Introduction

Breast cancer is the most common cancer type worldwide. When considering the breast cancer incidences in Sri Lanka, breast cancer incidences are rising year by year. According to the Sri Lanka National Cancer Registry, it was obtained that over 3000 new cases were reported each year.

By using ARIMA, Bootstrap and Bayesian methods, Janbabaee, G., et al aimed to predict the incidence of breast, colorectal, and bladder cancer cases n north of Iran until 2020.According to the ACF and PACF models after differentiation estimated the model parameters and ARIMA(1,1,1) model seems to be the best model. Also this model had the least Akaike Information Criterion (AIC) value.In their conclusion, an increasing trend for breast, bladder and colorectal cancers in northern part of Iran was predicted[1].

Sani Mary, Olanrewaju S.O., and Oguntade E.S. were assessed to determine the trend movement and forecast the trend of reported cases of breast cancer for both rural and urban women in University Teaching Hospital Gwagwalada Abuja. In this research, logistic re-

gression and ARMA were used for analysis and ARMA(23,12) was the best model by using information criterion, AIC, Schwarz Criterion (SIC) values[2].

J Med Discov, E-Discovery & Xie, Liming investigated to analyze the data of cancer incidence rates(CIR) and its residuals in the United States for 1975 – 2014.By using Autoregressive integrated moving average (ARIMA) model and Seasonal Autoregressive Integrated Moving Average (SARIMA) models were applied to this data set and found a fitted model and made reasonably forecast the trends for 2015 to 2020.ARIMA(1,1,2)(1,1,1)[4] was confirmed as the best model for the CIR[3].

In this paper, we investigated the reported breast cancer incidences in Sri Lanka from 2005 to 2020 and predicted a suitable model, and forecasted future breast cancer incidences using Quadratic trend model, Exponential trend model and Holt's trend corrected exponential smoothing methods. By comparing accuracy measures, we selected the Quadratic trend model is the suitable model for this data set.

Materials & Methods

The breast cancer incidences were collected from the National Cancer Registry, Sri Lanka rom 2005 to 2020.The data was analyzed using Minitab and R software version 4.2.1 .

Holt's Trend corrected Exponential Smoothing

Holt's trend corrected exponential smoothing methods are used, when the data series has a trend and no seasonal pattern. This method is based on two smoothing equations - one for the level, one for the trend.

$L_t = \alpha\ y_t + (1-\alpha)\ (L_{t-1} - T_{t-1}) \quad 0 \leq \alpha \leq 1$ (Level equation)

$T_t = \beta\,(L_t - L_{t-1}) + (1-\beta)\,T_{t-1}$ (Trend estimating equation)

$F_{t+m} = \ L_t + mT_t$ (Forecast equation)

Where smoothing parameters α, β and γ ; F_{t+m} is the forecast for m periods ahead at time t.

Quadratic Trends

In a quadratic trend, the values of a time series tend to rise or fall at a rate that is not constant. That means it changes over time. As a result, the trend is expressed as,

$$TR_t = \beta_0 + \beta_1 t + \beta_2 t^2$$

Exponential Trends

An exponential trendline is a curved line that is most useful when data values rise or fall at increasingly higher rates. An important case is ,it cannot be created an exponential trendline if there are any zero or negative values. The exponential trend is expressed as,

$$Y_t = \alpha e^{\beta t}$$

Results and Discussion

According to time series plot we determined that there was an increasing trend and there was no any seasonal pattern. So, we applied Holt trend corrected exponential smoothing(Double exponential smoothing) to model the dataset. Then forecast values for three years and evaluate the forecast accuracy and we have got 8.797153 as Mean Absolute Percentage Error (MAPE).Analyzing residuals using Shapiro test and Box Ljung- Box test, it was given that the residuals of forecasted values are normal and independent each other.

To fit Autoregressive integrated moving average (ARIMA) model, we checked the stationarity of the series using ADF (Augmented Dickey Fuller) test and as the result of ADF test , the series is not stationary. After differencing two times , ADF test is given that the series is stationary. Then we applied fit a suitable ARIMA model for the dataset. But it didn't give any suitable model for this data set.

Then we model using Exponential trend model and Quadratic trend model and compare accuracy measures such as MAPE, MAD, MSD.

The results are shown in following table:

Accuracy Measure	Holt trend corrected exponential smoothing model	Quadratic Trend Model	Exponential Trend Model
MAPE	8.1	5.4	5.6
MAD	244.8	173.0	177.9
MAPE	94483.8	53905.5	54941.6

Table 1: Accuracy Measures of models.

Conclusion

By comparing Holt trend corrected exponential smoothing model, Quadratic trend model and exponential trend model, the Quadratic trend model is the best model for modelling breast cancer incidences in Sri Lanka.

The Quadratic trend model equation is: $Y_t = 1729 - 0.4441 + 89.0t + 7.01t^2$

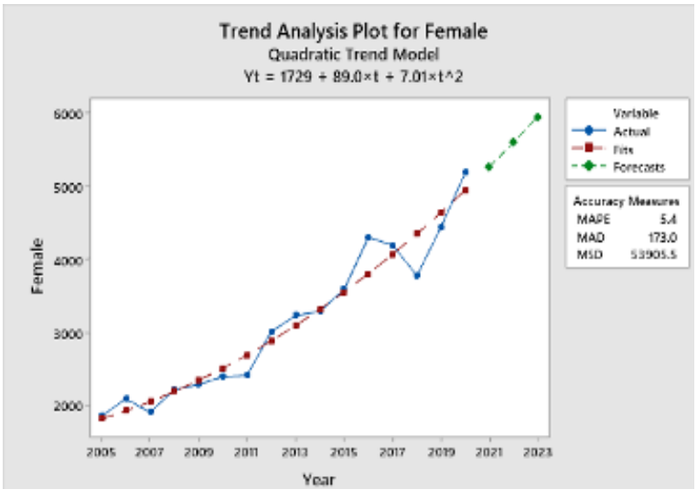

Figure 1: Trend analysis plot of Quadratic trend model

The forecasted values for years of 2021, 2022 and 2023 are 5267, 5602 and 5950 respectively.

References

[1] G. Janbabaee, A. Nadi-Ghara, M. Afshari, S. R. Moghadam, M. Y. Ashrafi, M. Aarabi, A. Hedayatizadeh-Omran, R. Alizadeh-Navaei, M. E. Jouybari, and M. Moosazadeh. Forecasting the incidence of breast, colorectal and bladder cancers in north of iran using time series models; comparing bayesian, arima and bootstrap approaches. *Asian Pacific Journal of Environment and Cancer*, 4(1):3–7, 2021.

[2] S. MARY, O. SO, and O. ES. Analysis of reported cases of breast cancer in abuja, nigeria: a time series approach. *IRE J*, 3, 2020.

[3] L. Xie. Time series analysis and prediction on cancer incidence rates. *Journal of Medical Discovery*, 2(3):1–10, 2017.