Abstract—Breast cancer incidences is steadily increasing in Fiji and accurate forecasting can have major implications in controlling this deadly illness. The best forecasting model is the one that does not underestimates or overestimates the true number of breast cancer cases and gives minimal prediction errors. This paper proposes Linear Regression model for forecasting breast cancer cases using the reported number of cases in Fijian population from the year 1995 to 2016. The proposed model is also compared with the Naïve Forecast Method as the benchmark. The performances of the two models were analyzed based on measures such as the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The proposed model was then further validated using the diagnostic measures such as Goodness-of-fit ($R^2$), Tracking Signal (TS) and Bias. The results showed that the proposed Linear Regression model outperformed the Naïve Forecast Method. It also satisfies the validity diagnostic measures and is a better tool in forecasting Fiji’s yearly breast cancer cases.

Keywords— Modeling, Breast Cancer Cases, Linear Regression Model, Time Series Model, Naïve Method, Error Measures, Validity.

I. INTRODUCTION

Cancer is one of the most life threatening non-communicable diseases. It has become a global epidemic and affects people of all ages. Although there are many different kinds of cancer but all cancers develop as a result of abnormal growth of cells which gets out of control.

There is steadily increase in cancer cases over the past several years in Fiji where cases of breast, cervical and other reproductive tract cancers are mostly increasing in women while cancers such as prostrate, liver, rectum and lung cancer are becoming more common in men (Ministry of Health & Medical Services, 2015). In Fiji, breast cancer is the leading cause of death in all cancer illnesses followed by cervical cancer on the second place (Tuicakau, 2014).

Breast cancer develops from the growth and division of cells in the breast tissue without normal control. This uncontrolled growth of cells forms a lump called a tumor which is either benign (not cancerous) or malignant (cancerous). The first sign of breast cancer is “a breast lump or an abnormal mammogram” (Ministry of Health & Medical Services, 2015). Other signs also include swelling in the armpit, pain in the breast, nipple dimpling, unusual discharge of fluid from the nipple etc. Some of risk factors of breast cancer are genetic and related to family history while others are based on personal factors such as reproductive history and medical history (Tyrer et al., 2004). Youlden et al. (2014) reported that the consumption of certain food items and indulgence in certain behaviors are associated with an increased risk of breast cancer, some of which include eating unhealthy foods, drinking alcohol, smoking and obesity.

The thorough literature exploration shows that a very few descriptive statistical analysis of breast cancer cases has been carried out for Fijian population. One such study was done by Singh et al. (2012) in which it was reported that 19 cases of male breast cancer were recorded over the past 11 years from January 2000 to December 2010 in Fiji with an average of 2 new cases of male breast cancer per annum.

It was also noted through literature review that no studies has been done so far on developing the forecasting model for Fiji’s breast cancer cases. Identifying a model to give reliable forecast is essential as accurate forecasting can have major implications for relevant authorities and stakeholders to take necessary actions in controlling this illness. Initiating more awareness programs around the country, encouraging early detection and organizing treatment procedures could be some of the immediate actions.

Many of the literature has been cited and studied on forecasting models. Moller et al. (2002) carried out a study in Nordic countries on the prediction of cancer incidences up to the year 2020. According to Moller et al. (2002), the prediction of the cancer incidences were calculated by first forecasting the incidence rates based on the observed rates up to 1997 followed by multiplying these rates by the population forecasts for the future periods. The incidence rates based on the observed rates up to 1997 was predicted using the power model (Engeland et al., 1993 cited in Moller et al., 2002) which was derived from the Osmond (1995, cited in Moller et al., 2002), age-period-cohort (APC) model. According to Moller et al. (2002), this method was used for predicting the cancer incidences for all cancer types expect for breast and prostate cancer. For breast cancer prediction, the information on mammographic screening was also included in the model.

A very similar work has been carried out by Erbas et al. (2007) who forecasted the age-specific breast cancer mortality using functional data models. According to this paper, age-period-cohort (APC) methods to estimate mortality and incidence of cancers such as breast cancer, prostate cancer and cervical cancer has been commonly used approach. These models as stated by Erbas et al. (2007) are based on regressions with the outcome defined as mortality or incident cases, and modelled using a Poisson error distribution and a log link function for the mean. These models are then used to make linear projections of mortality and incidence. However, Erbas et al. (2007) critiqued the APC models for estimating future cancer rates by highlighting some common problems associated with it which includes “issues with non-
identifiability of the parameters, strong parametric assumptions, and sensitivity of the projections to most recent changes in cohort effects” (Erbas et al., 2007). Erbas et al. (2007) hence proposed an alternative approach for forecasting cancer mortality and incidence rate which was developed for demographic forecasting (Hyndman and Ullah, 2007 cited in Erbas et al., 2007). The approach proposed by Erbas et al. (2007) uses functional data analysis techniques, and assumes the age-specific mortality curves as the “units of analysis rather than discrete observations” (Erbas et al., 2007).

Moreover, other works on forecasting models were also studied. Vanualailai et al. (2003) studied dynamics of the HIV-positive population size in Fiji using the reported numbers of HIV-positive individuals in Fiji between and including the years 1989 and 2002. From the data description which showed the “S” shaped growth pattern or nonlinear pattern, the Logistic and Gompertz curve was used to predict the short-term behaviour of the observed HIV-positive population size at the initial stages of growth (Vanualailai et al., 2003).

Furthermore, Qiokata and Khan (2015) proposed an Artificial Neural Network (ANN) model for forecasting emigration for Fiji population components during the years from 1986 to 2012. The ANN model that was developed was further compared to Naïve Method as the benchmark method for modeling emigration of Fiji’s population. The paper reported that ANN model outperformed the Naïve Method using quantitative evaluation. Kimata et al. (2015) in his paper also proposed ANN model for forecasting exchange rates of Solomon Islands dollar (SBD) against Euro (EUR) using daily exchange rate data during the period of January 5, 1998 to June 30, 2014. The proposed model was compared with single exponential smoothing; double exponential smoothing with trend; and Holt-Winter multiplicative and additive seasonal and multiple linear regression models. The empirical results confirmed that the proposed ANN model is an efficient tool for predicting SBD against EUR more accurately when compared to regression and time series models.

In this paper, we provide descriptive statistical analysis of breast cancer cases and also explain the cases with respect to gender, ethnicity and age for Fijian population. We also propose a Linear Regression model to forecast yearly breast cancer cases and compare this with the Naïve Forecast method as the benchmark using various error measures such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The proposed Linear Regression model is further validated using the diagnostic measures such as Goodness-of-fit ($R^2$), Tracking Signal (TS) and Bias.

II. DATA ACQUISITION AND ANALYSIS

This study used breast cancer cases reported in Fiji from the year 1995 to 2016. The data was obtained from the Health Information Unit, Ministry of Health & Medical Services, Fiji. The number of breast cancer cases for each year was further classified with respect to gender, ethnicity and age. Fig. 1 shows the time series plot for 22 years of breast cancer cases that depicts that there is approximate growth pattern in the breast cancer cases in Fijian population. There is slow growth pattern in the 1995-2009 interval followed by a rapid growth pattern in the 2010-2016 interval, in which the highest breast cancer case was registered in the year 2016. The rapid growth pattern may be a serious concern that needs to be studied so that government and relevant stakeholders undertake major initiative of creating awareness on the seriousness of breast cancer around the country.

Fig. 2 shows the histogram and a series of statistics for breast cancer cases. A total of 2828 breast cancer cases were reported with annual mean of 129 cases and a standard deviation of 83 cases. The histogram also shows a positively skewed distribution (Skewness = 1.32) indicating that most cases fall towards the lower tail of the distribution while very few extreme number of breast cancer cases has been recorded.

Fig. 3 shows the time series plot of breast cancer cases by gender. The number of cases for females’ shows significant growth after the year 2010 compared to males showing slow growth. The total number of breast cancer cases reported for females was 2749 out of 2828 reported cases with the annual mean of 125 cases (and the standard deviation of 82 cases) whereas the annual mean number of cases for males was 4 (and the standard deviation of 2 cases). This strongly shows that breast cancer is mostly common in females than in males.
Itaukei is significantly higher compared to Fijian of Indian descent (FID) and Fijian of Other descent (FOD) which shows similar growth pattern. There is a significant increase in the breast cancer cases in Itaukei from the year 2012. The total number of cases reported for Itaukei was 1665 out of 2828 reported cases with the annual mean of 76 cases (and the standard deviation of 57 cases). The annual average number of cases for FID was 47 (and the standard deviation of 24 cases) whereas for FOD, the annual mean number of cases was 6 (and the standard deviation of 5 cases). Conclusively, it can be stated that the breast cancer cases is more commonly reported for Itaukei females compared to FID and FOD.

Fig. 4 shows that breast cancer cases by ethnicity. The figure shows that breast cancer cases in Itaukei is significantly higher compared to in Fijian of Indian descent (FID) and Fijian of Other descent (FOD) which shows similar growth pattern. There is a significant increase in the breast cancer cases in Itaukei from the year 2012. The total number of cases reported for Itaukei was 1665 out of 2828 reported cases with the annual mean of 76 cases (and the standard deviation of 57 cases). The annual mean number of cases for FID was 47 (and the standard deviation of 24 cases) whereas for FOD, the annual mean number of cases was 6 (and the standard deviation of 5 cases). Conclusively, it can be stated that the breast cancer cases is more commonly reported for Itaukei females compared to FID and FOD.

Fig. 5 shows the breast cancer cases in different age group. The figure depicts that breast cancer is mostly reported for people above 25 years of age. This means that people in this age be more cautious of the signs and symptoms of breast cancer. They should also consult medical specialist and consider getting tested for breast cancer.

Fig. 4. Breast cancer cases by ethnicity

Fig. 5. Breast cancer cases in different age group

III. MODEL DEVELOPMENT

Time series models are one of the simplest and inexpensive type of forecasting models. These models only requires observed values of the number of breast cancer cases and other explanatory variables to be able to forecast future cases. Time series models can either be univariate which requires only the past observed values of the number of breast cancer cases or multivariate model which requires not only past observed values of the number of breast cancer cases, but past or present observed values of other explanatory variables.

Some of the time series models include: moving average (MA) model, autoregressive model (AR), autoregressive moving average (ARMA) model, autoregressive integrated moving average (ARIMA) model and linear regression based models for instance Multiple Linear Regression (MLR) model.

In this study, no other explanatory variables except the past 22 yearly observed values of the number of breast cancer cases were available thus only univariate time-series models were investigated. These models were developed using the training dataset that had all 22 observations.

We first considered ARIMA model which requires the series to be stationary. The Augmented Dickey-Fuller (ADF) test was used to check for the presence of unit root. The test indicated that there is a unit root present in the series (with p-value of 0.9948), hence, the series is non-stationary. A further ADF test confirmed that the series is stationary after the first difference, however, no suitable model could be achieved. Hence, a Linear Regression model is considered, where the independent variables are the time-lags of number of breast cancer cases and the dependent variable is the number of breast cancer cases.

A. Linear Regression Model

The linear regressive model with p time-lagged period, LR (p), is defined by

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \ldots + \beta_p Y_{t-p} + \epsilon_t$$

where, $t = 1, 2, \ldots, n$; $Y_t$ is the number of breast cancer cases with time $t$; $\beta_0, \beta_1, \ldots, \beta_p$ are the regression coefficients; $Y_{t-1}$ is the number of breast cancer cases with time-lag 1; $Y_{t-2}$ is the number of breast cancer cases with time-lag 2 and so on; $\epsilon_t$ is the error term and $n$ is the number of observations.

The selection of the number of time-lags that fits a best Linear Regression model to forecast the annual breast cancer cases was done using the following criteria:

a) **AIC and SIC Model Selection Criteria:** In LR model, determining the optimal lag is very crucial as it is sensitive to lag length (p). It is imperative that model have the “accurate” number of lags included as too many lags will cause a loss in degrees of freedom and too few lags will cause the model to be imprecise. The number of lags also provides information on how much previous time interval (years) affects the forecast model. The normal likelihood is maximized by choosing $p$ to minimize $SSE_p/n$ which is the estimated error in the covariance in the sample $n$. Akaike Information Criteria (AIC) is the most popular information criterion used to determine the preferred model. AIC modifies the likelihood $\ln(SSE_p/n)$ by adding a penalty on each additional lag and is given by

$$AIC_p = \ln \left( \frac{SSE_p}{n} \right) + \frac{2p}{n}$$

where $SSE_p$ is the sum squared error, $SSE_p/n$ is the maximum likelihood, $r = p+1$ is the number of parameters in the model,
p is the lag length and n is the number of observation. Schwarz Information Criterion (SIC) is another model selection criteria which suggests that p values are too large by adding larger penalty on the parameters r. The SIC in this study is defined by

$$SIC_p = \ln \frac{SSE_p}{n} + \frac{r \ln n}{n}$$  

(3)

The preferred model is one with the minimum value of AIC and SIC when compared against its corresponding candidate models.

b) Mean Square Error ($S_p^2$) and Adjusted R-square ($R_{adj}^2$) Selection Criteria: The Adjusted R-square ($R_{adj}^2$) is a modified version of the coefficient of determination $R^2$. The $R_{adj}^2$ gives the percentage of variation explained by only the independent variables that actually affect the dependent variable. It is more preferable compared to $R^2$ as it penalizes for adding independent variables that do not fit the model. The $R_{adj}^2$ in this study is defined by

$$R_{adj}^2 = 1 - \frac{S_p^2}{S_y^2}$$  

(4)

where $S_p^2 = SSE_f/n - p - 1$ is the mean square error of the residuals, $S_y^2 = SST/n - 1$ is the variance of the dependent variable and SST is the total sum square. The smaller value of $S_p^2$ and higher value of $R_{adj}^2$ identifies the best time-lag model.

B. Lag Selection

The optimal time-lag for LR model was determined by also scrutinizing the influential observations using the Cook’s distance ($D_i$) measure for different time-lag. A large value of $D_i$ usually $D_i > 1$ implies that the ith point is influential and must be considered for deletion. Fig. 6 shows the plot of Cook’s Distance against ith observations for lag 1. The plot suggests that 18th observation (with $D_i > 1$) be considered for deletion.

Appendix 1 and 2 shows the plot of Cook’s Distance against ith observations for lag 2 and lag 3 respectively. The plots suggests that 17th observation (with $D_i > 1$) be considered for deletion for lag 2 while 16th observation (with $D_i > 1$) be considered for deletion for lag 3. Table I shows the model selection criteria, $AIC_p$, $SIC_p$, $S_p^2$ and $R_{adj}^2$ with corresponding probability ($F$-statistic) of the regression model (including and excluding the influential observation) for three lags.

**TABLE I. MODEL SELECTION CRITERIA OF LR MODEL**

<table>
<thead>
<tr>
<th>Model with lags (p)</th>
<th>$S_p^2$</th>
<th>$R_{adj}^2$</th>
<th>$AIC_p$</th>
<th>$SIC_p$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lag 1*</td>
<td>1723.33</td>
<td>0.7373</td>
<td>10.43</td>
<td>10.53</td>
<td>3.8×10^-7</td>
</tr>
<tr>
<td>Lag 2*</td>
<td>1591.90</td>
<td>0.7516</td>
<td>10.46</td>
<td>10.61</td>
<td>3.0×10^-8</td>
</tr>
<tr>
<td>Lag 3*</td>
<td>1581.22</td>
<td>0.7488</td>
<td>10.57</td>
<td>10.77</td>
<td>2.4×10^-7</td>
</tr>
<tr>
<td>Lag 1p</td>
<td>867.30</td>
<td>0.8717</td>
<td>9.75</td>
<td>9.85</td>
<td>1.1×10^-3</td>
</tr>
<tr>
<td>Lag 2p</td>
<td>863.33</td>
<td>0.8693</td>
<td>9.86</td>
<td>10.01</td>
<td>3.3×10^-4</td>
</tr>
<tr>
<td>Lag 3p</td>
<td>827.22</td>
<td>0.8723</td>
<td>9.94</td>
<td>10.14</td>
<td>4.3×10^-7</td>
</tr>
</tbody>
</table>

* Including the influential observation

b. Excluding the influential observation

From Table I, it can be noted that the Linear Regression model with lag 1 (excluding the influential observation), LR (1), is the most preferred model as it is the most significant model at the 5% level of significance, although Linear Regression model with lag 3 (excluding the influential observation) have a slightly higher value of $R_{adj}^2$. Moreover, the lowest values of $AIC_p$ and $SIC_p$ strongly indicate that lag 1 (excluding the influential observation) is the optimal lag. Eviews results of the LR (1) model is shown in Table II. It is imperative to note that the model is significant ($F= 130.1088$) with p-value = 0.

**TABLE II. EVIEW RESULTS FOR LR (1) MODEL**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>-13.31366</td>
<td>14.22756</td>
<td>-0.935765</td>
<td>0.3618</td>
</tr>
<tr>
<td>LAG1</td>
<td>1.289062</td>
<td>0.113011</td>
<td>11.40652</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

R-squared 0.878468 Mean dependent var 129.4500
Adjusted R-squared 0.871716 S.D. dependent var 84.47701
S.E. of regression 0.256941 Akaike info criterion 9.751968
Sum squared resid 16478.68 Schwarz criterion 9.851541
Log likelihood -95.51968 Hannan-Quinn criteria 9.771405
F-statistic 130.1088 Durbin-Watson stat 1.450558
Prob(F-statistic) 0.000000
C. Model Adequacy

The three principal assumptions that validate the use of Linear Regression models for purposes of forecasting are: linearity relationships between dependent and independent variable(s), normality of error distribution (with error term, ε, having zero mean) and homoscedasticity of the errors (the error term, ε, having constant variance, \(\sigma^2\)). If any of these assumptions are violated then the forecasts or scientific interpretations obtained by the model may be inefficient or seriously biased. Each of these assumptions is tested for LR (1) model to validate its robustness.

a) Linearity: The relationship between dependent and independent variables for LR can only be accurately estimated if the relationship between them is linear in nature as the analysis of non-linear relationship between the variables will underestimate the relationship. The scatter plot of the dependent variable, the number of breast cancer cases, against the independent variable, lag 1 of number of breast cancer cases, as shown in Fig. 7 was used to check for linearity. The plot indicates linearity between the dependent variable and independent variable.

![Fig. 7. Scatter plot of breast cancer cases against lag 1](image)

b) Normality: The two normality tests were performed to investigate the normality of error distribution of LR (1) model. Firstly, the Shapiro-Wilk test shown in Table III with the normal Q‒Q plot of the residuals shown in Fig. 8. These were obtained using SPSS software. Secondly, Jarque–Bera statistics for normality with its histogram in Fig. 9 which was obtained using Eviews software.

<table>
<thead>
<tr>
<th>Table III. Shapiro-Wilk Test for Normality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistic</td>
</tr>
<tr>
<td>Standardized Residual</td>
</tr>
</tbody>
</table>

When examining a normal Q-Q plot, the data is said to be normally distributed, if the data points are close to the diagonal line. If the data points stray from the line in an obvious non-linear fashion, the data are not normally distributed. Fig. 8 shows a normally distributed Q-Q plot. Table III shows the results from well-known test of normality, namely the Shapiro-Wilk Test. The test statistics value indicates that the distribution is normal (p-value > 0.05).

Fig. 9 shows the histogram and summary statistics of residual of LR (1) model. The histogram also shows that the distribution is approximately normally distributed with mean of zero and standard deviation of 29.45. The Jarque-Bera test probability as shown in Fig. 9 also endorses that the errors are normally distributed (p-value = 0.81137).

<table>
<thead>
<tr>
<th>Table III. Shapiro-Wilk Test for Normality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistic</td>
</tr>
<tr>
<td>Standardized Residual</td>
</tr>
</tbody>
</table>

When examining a normal Q-Q plot, the data is said to be normally distributed, if the data points are close to the diagonal line. If the data points stray from the line in an obvious non-linear fashion, the data are not normally distributed. Fig. 8 shows a normally distributed Q-Q plot. Table III shows the results from well-known test of normality, namely the Shapiro-Wilk Test. The test statistics value indicates that the distribution is normal (p-value > 0.05).

Fig. 8. Normal Q-Q plot of residuals of LR (1) model

Fig. 9. Histogram and summary statistics of residual of LR (1) model

c) Homoscedasticity: Homoscedasticity means that the variance of errors is constant. The homoscedasticity in this study was tested using the White heteroscedasticity test in Eviews software and its results are shown in Appendix 3. The probability (F-statistic) is 0.5467 (p-value > 0.05), therefore the test fails to reject the null hypothesis concluding that the variance of the error is constant.

The three assumptions of the proposed Linear Regression model were tested for LR (1) model and all three were satisfied and thus validates its robustness. This model can be used to forecast the annual breast cancer cases. The forecasted equation for LR (1) model estimated using the least square method as estimated by Eviews shown in Table II is given as

\[
\hat{Y}_t = -13.3137 + 1.2891Y_{t-1}
\]  

(5)

The Equation (5) was used to forecast the number of breast cancer cases for the training period and the results of the predicted number of cases and residuals along with the actual number of breast cancer cases are presented in Fig. 10.
Fig. 10. Actual and predicted number of breast cancer cases and residual plot of LR (1) model

IV. MODEL COMPARISON

In this section, a comparison study of the proposed LR (1) model is made against the Naïve Forecast Method using the same training dataset.

A. Naïve Forecast Method

Naïve forecasting is an estimating technique in which the last period’s actual values are used as this period’s forecast, without any adjustments or attempting to establish any casual factors. The simplest Naïve Forecast Method was used as a comparator, where the last actual datum is used to forecast the next period, and has an equation

\[ F_{t+1} = Y_t \]  

where \( F_{t+1} \) is the forecasted value for the next period and \( Y_t \) is the actual number of breast cancer cases at time \( t \).

B. Model Evaluation

The performance of the forecasting models, LR (1) and Naïve Forecast Method is compared using the following error measures:

a) Mean Absolute Error (MAE):

\[ MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{Y}_i - Y_i| \]  

b) Mean Absolute Percentage Error (MAPE):

\[ MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\hat{Y}_i - Y_i}{Y_i} \right| \times 100 \]  

c) Root Mean Square Error (RMSE):

\[ RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2} \]  

The following diagnostics measures are also used to test the validity of the LR (1) model:

d) Goodness of fit \( (R^2) \):

\[ R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \]  

e) Tracking Signal (TS):

\[ TS = \frac{1}{n} \sum_{i=1}^{n} |Y_i - \hat{Y}_i| \]  

f) Bias:

\[ Bias = \sum_{i=1}^{n} (Y_i - \hat{Y}_i) \]  

where \( \hat{Y}_i \) is the forecasted value and \( \bar{Y} \) is the mean of the actual number of breast cancer cases with \( t = 1, 2, 3, ..., n \).

The model is considered superior if it gives the lowest error for all error measures (MAPE, RMSE and MAE). The forecast bias in (12) gives the direction of the error. The positive value of the bias implies that the forecasting model is underestimating while negative values implies that the forecasting model is overestimating. By adding the error terms if there is no bias, the positive and negative error terms will cancel each other out and the mean error term will be zero resulting in the value of TS in (11) to be zero or close to zero. In such a case, the forecasting model does not result in bias indicating that there is no underestimation or overestimation.

The error and validity measures of Naïve Forecast method and LR (1) model were calculated using Microsoft Excel 2013. Table IV shows the error measurement statistics of LR (1) model and Naïve Forecast method. The result shows that the optimum model, LR (1) outperforms the Naïve Forecast Method since LR (1) model produces small error values against the Naïve Forecast Method in the training dataset.

<table>
<thead>
<tr>
<th>TABLE IV. ERROR MEASUREMENT STATISTICS BETWEEN LR (1) MODEL AND NAÏVE FORECAST METHOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>Naïve</td>
</tr>
<tr>
<td>LR (1)</td>
</tr>
</tbody>
</table>

Table V shows the validity measurement statistics of LR (1) model. The table depicts that \( R^2 \) value of 0.88 implies that 88% of the total variation is explained by the model. While TS and Bias values are very close to 0 indicating that the forecasting model, LR (1) does not overestimate and underestimate the true annual breast cancer cases. This validates that LR (1) model is a better tool in forecasting Fiji’s annual breast cancer cases.
TABLE V. VALIDITY MEASUREMENT STATISTICS OF LR (1) MODEL

<table>
<thead>
<tr>
<th>Model</th>
<th>$R^2$</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR (1)</td>
<td>0.88</td>
<td>-2.46 × 10^{-11}</td>
</tr>
</tbody>
</table>

V. CONCLUSION AND FUTURE WORK

Breast cancer is steadily increasing in Fijian population with more cases reported for females compared to males. The results also showed that breast cancer cases in Itaukei is significantly higher compared to in Fijian of Indian descent (FID) and Fijian of Other descent (FOD). The graph of breast cancer cases in different age group showed that breast cancer is mostly reported for people above 25 years of age.

Furthermore, it is proposed that the model, i.e. LR (1) be used to accurately make forecasts of the total annual breast cancer cases in Fiji. This proposition is made after concluding that the model gives the least error values when compared to the Naïve Forecast method in the training data set. The validity measurement statistics also validated that LR (1) model does not overestimate and underestimate the true annual breast cancer cases and also explains 88% of the total variation.

In future research, we can use other modeling techniques to forecast the total number of breast cancer cases. And more specifically model the cases for females as the results in this study showed that the breast cancer is significantly higher is females than in males. We can also carry out comparison study of various breast cancer forecasting models.

ACKNOWLEDGMENT

The authors would like to thank Health Information Unit, Ministry of Health & Medical Services, Fiji for providing the breast cancer cases data for Fiji.

REFERENCES

[2] Arzu, A. Forecasting wind speed of Suva (Fiji) and Abaiang (Kiribati) using artificial neural network. unpublished.

APPENDICES

Appendix 1: The plot of Cook’s Distance against 1th observations for lag 2.

Appendix 2: The plot of Cook’s Distance against nth observations for lag 3.
Appendix 3: White heteroscedasticity test for LR (1) Model

Heteroskedasticity Test: White

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-statistic</td>
<td>0.377442</td>
<td>0.5467</td>
</tr>
<tr>
<td>Obs*R-squared</td>
<td>0.410766</td>
<td>0.5216</td>
</tr>
<tr>
<td>Scaled explained SS</td>
<td>0.253048</td>
<td>0.6149</td>
</tr>
</tbody>
</table>

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Sample: 2 21
Included observations: 20

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>704.5329</td>
<td>306.5306</td>
<td>2.298410</td>
<td>0.0337</td>
</tr>
<tr>
<td>LAG1^2</td>
<td>0.007533</td>
<td>0.012262</td>
<td>0.614363</td>
<td>0.5467</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-squared</td>
<td>0.020538</td>
<td>Mean dependent var</td>
<td>823.9341</td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>-0.033876</td>
<td>S.D. dependent var</td>
<td>1042.575</td>
</tr>
<tr>
<td>S.E. of regression</td>
<td>1060.087</td>
<td>Akaike info criterion</td>
<td>16.86473</td>
</tr>
<tr>
<td>Sum squared resid</td>
<td>20228108</td>
<td>Schwarz criterion</td>
<td>16.96430</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>-166.6473</td>
<td>Hannan-Quinn criter.</td>
<td>16.88417</td>
</tr>
<tr>
<td>F-statistic</td>
<td>0.377442</td>
<td>Durbin-Watson stat</td>
<td>1.838027</td>
</tr>
<tr>
<td>Prob(F-statistic)</td>
<td>0.546663</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>