**TIME SERIES ANALYSIS ON THE STUDY OF BREAST CANCER IN ILORIN METROPOLIS**

**N.R. Zakariyau Olufemi-Ojo Olukemi B. and Okechukwu Ijeoma Eberechukwu**

Department of Statistics, Federal Polytechnic, Oko, Anambra State.

E-mail: [rufaistatistics@gmail.com](mailto:rufaistatistics@gmail.com)

**Abstract**

*Breast cancer is the commonest cancer in women and characterized by regional variations which is prevalently common in low and middle income countries including Nigeria. It is now the most common disease and the second leading cause of death among women. This study assessed the reported cases of breast cancer for women in University of Ilorin Teaching Hospital, Kwara State. The data used in this research is a secondary data which was sourced from the Record and Information Department of University of Ilorin Teaching Hospital, Kwara State. The main objectives is to determine the appropriate time series model for the number of breast cancer patients in Ilorin* **between 2007 -2016** *and to forecast for years to come. The statistical techniques used for the analysis is Autoregressive Moving Average (ARMA). The result of the analysis revealed that ARMA (1,0) is the best model based on criterion considered. The model obtained is Xt = 5.85838+0.366856Xt-1 which was used to forecast for the year2017-2020. It is recommended that there should be an awareness and enlightenment programme for the populace on how to prevent breast cancer, the cause, effect and remedy for it.*

**Introduction**

Cancer is one of the fatal diseases that involve abnormal growth of cells that increases compoundedly. It is a malignant tumor which snicks the surrounding tissues without any wall or boarder through the roots and spread to other parts of human body. If the cancer spread, small tads of cancer cell chuck off the original tumor and move to the other parts of the body. The spread can either be direct, through blood or through lymphatic system. Cancers can touch various organs, and each type of cancer has its unique characteristics. Cancer of various types are known in relation to the location of cancers namely cervical cancer, lung cancer, gynecological cancer, skin cancer, brain cancer, breast cancer etc. the specific type of cancer that is more common to an area or community than the other cancer is breast cancer.

Breast cancer is common form of cancer affecting all women of different ages. Breast cancer affects the breast tissue and lobules. The classification of breast cancer is resulted from its origination, if breast cancer is originated from milk ducts then its known as ductal carcinoma while cancer cells found in lobules makes the cancer termed as “lobular carcinoma”. The screening of breast cancer is primary step which filter out the symptoms that can be used to diagnose the patients actual pathological condition. Breast cancer is most frequent cause of death in older women but at same time it is important to note that younger women who does not go under cancer screening process remains in danger circle of breast cancer.

According to the survey conducted in 2017, in US alone there are 252,710 cases of breast cancer. So the number of breast cancer all around the world will be a value very huge. In US the death rate of women due to breast cancer is higher compared to other cancer categories. After skin cancer breast cancer is most commonly found among women. In India it is found that breast cancer is now commonly found in younger age groups also. It is the most common kind of cancer in Indian urban areas and second common in rural areas. All these facts points hand towards the importance of curing breast cancer at an early stage.

This disease is highly spread among the population most especially older women and can kill people within a short period. It is essential to clearly accurately map out strategies towards their reduction. This study will make an appropriate contribution if its finding is implemented.

The aim of this research is to determine the appropriate time series model for the number of breast cancer patients in Ilorin **between 2007 -2016.** The significance of the study will be immense importance to the following groups: The researcher, other researchers, the government and the public at large. The significance of the study is to serve as reference for more finding on how to tackle the breast cancer disease and for subsequent researchers who intent to carry out studies related to this topic.

Health is not a concern of health sector alone, but also of other socio-economic sectors and disciplines such as Information, Agriculture, Water Supply, Sanitation and Education.

According to Mohammed (2013), he said “Breast cancer usually begins in the cell of the lobules which are the milk producing gland or the duct. The passage that drains the milk from the lobules to the nipples”. He said that women whose mother or sister has breast cancer have a high risk of developing the disease themselves. It has been discovered that breast cancer susceptibility genes are from one of the parents. The most common of the genes is the BRCA (gene mutated in breast and ovarian cancer) i.e breast cancer gene. The gene account for about 10% of all breast cancer cases in families that have these genes, the risk of breast cancer can be very high. However, it is important to realize that 85-90% of breast cancers are not from their families.

Dangana *et al.,* (2015) Epidemiology of Breast Cancer among Male in University Of Abuja Teaching Hospital Gwagwalada examined the trends in the prevalence rate of breast cancer in men among tissues submitted to histopathology laboratory university of Abuja Teaching Hospital. A total of 544 data collected consisting of men between the age 17-86years with the mean aged group of 56years and was analyzed using Epi-Info version 6.1. It was found that the prevalence of breast cancer among men was 4(2.6%), fibroadenoma197 (36.8%), fibrocystic disease 120(22.4%), granulomatouse mastitis 14(2.6%) lactating adenoma 11(2.1%), sclerosingadenosis 8(1.5%), the highest prevalence rate was found between the age group of 39-48years (50%) followed by 39-48years (25%) and 79-88years (25%) respectively.

**Research Methodology**

Here it is tacitly assumed that information about the past is available in the form of numerical data. Ideally, at least 50 observations are necessary for performing time series analysis/modeling, as propounded by Box and Jenkins who were pioneers in time series modeling.

# **Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)**

Autocorrelation function (ACF) and partial Autocorrelation function (PACF) of a stationary series are collected over time. The procedure usually serve as a prelude to selecting series of models that are subjected to statistical test to confirmation as the best and optimal model for a particular case. The methodology emphasized the use of some vital statistical approach such as determination of the model order, parameter estimation as well as model testing and forecast behavior of ACF and PACF provide a suggestion of where the model can build from and corresponding model order. While use of ACF and PACF to determine model order is because both give sight information about the behavior of the time series. It is paramount and necessary requirement if a given series is cyclical, irregular, seasonal secular or process any of the two attributes must be stationary (Box and Jenkins 1970). When the series is stationary, the order of the model that is the coefficients number can be determined.

# 

# **Autoregressive (AR) Model**

Autoregressive model are base on the idea that the current value of the series, *x*t can be explained as a function of p pass value …. 

where p determines the number of steps into the pass needed to forecast the current value.

An autoregressive model of order p, abbreviated as AR (p) can be written as;

 … (1)

where  represent the time series data at point t,  are the parameters of the AR model and = 0 unless otherwise stated.

The letter p denotes the order of autoregressive model, defining how many previous values the current value is related to. The model is called autoregressive because the series is regressed on to past values of itself.

 … (2)

The autoregressive operator  is defined to be

 =  …(3)

More concise the AR model can be expressed as;

 …(4)

The mean of AR (2) model can be compute as follows (bear in mind that the model is stationary)

 Due to stationary condition

Mean = 

and variance is defined as

Variance = 

**Moving Average (MA) Model**

A moving average model an alternative to the autoregressive representation in which the

on the left-hand side of the equation are assumed to be combined linearly, the moving average model of order q, abbreviated as MA (q) assumes the white noise () on the right –hand side of the defining equation are combined linearly to form the observed data.

A moving average model of order q, abbreviated as MA (q) can be written as;

 … (5)

Where are the moving average (MA) parameters in the model and  are the white noise error terms  and  is a Gaussian white noise series, with mean zero and variance ****unless otherwise stated. The Moving Average model in lag operation is of form

 … (6)

The Moving Average operator is of form;

 … (7)

# Autoregressive Moving Average (ARMA) models

This is a process of maxing Autoregressive and Moving Average model to form an Autoregressive Moving Average (ARMA) models for stationary time series. A general ARMA (p,q) model can be written as follows; 

# Autoregressive Integrated Moving Average (ARIMA) models

Autoregressive Integrated Moving Average (ARIMA) models are an extension to the class of ARMA models by adding the possibility to integrate a non-stationary process to a stationary. ARIMA (p,d,q) models are univariate time series models that consist of an autoregressive parameter (p), an order of integration (d) and moving average (q).

A process () is said to be an autoregressive integrated moving average (ARIMA) process if is the non- seasonal differencing operator of order d. to produce non-seasonal stationary of the dth difference. Usually d *= 1 or 2*.

# 

# **Information Criterion**

The Akaike information criterion is a measure of the relative goodness of fit of a statistical model. Akaike (1969, 73, 74) suggests measuring the goodness of fit for some particular model by balancing the error of the fit against the number of parameters in the model. It provides the measure of information lost when a given model is used to describe reality. It can be said to describe the tradeoff between bias and variance in model construction.

Hannan-Quinn information criterion (AIC) is a criterion for model selection. It is an alternative to Akaike information criterion (AIC).

Bayesian information criterion (BIC) or Schwarz criterion is another criterion for model selection among a finite set of models. BIC give a model with smaller order than AIC or AICc. The formula for the AIC, BIC, AIC and AICc are:-









where k is the number of parameters in the statistical model.

L is the maximized value of the likelihood function for the estimated model. AICc is a modification of AIC by Hurvich and Tsai (1989) while Burnham and Anderson (1989) insist on using AICc regardless of sample size since it converges to AIC as n is large and n = the sample size.

# 

# **Augmented Dickey-Fuller (ADF) Test for Stationarity**

In statistics and econometrics, an augmented Dickey–Fuller test (ADF) is a test for a unit root in a time series sample. It is an augmented version of the Dickey-Fuller test for a larger and more complicated set of time series models. If the null hypothesis is rejected it implies that the series is stationary. Otherwise, the unit root exists and is simply means that the series is non-stationary.

# **KPSS Test for Stationarity**

The integration properties of a series may also be investigated by testing the null hypothesis that the series is stationary against a unit root. Kwiatkowski et al (1992) have derived a test for this pair of hypotheses. The null hypothesis is rejected when the test statistic is greater than appropriate significance level.

# **Data Analysis and Discussion of Results**

To identify the model of any time series data, one most make a guess as to the data generation process. In doing this, one most begin by plotting the time series.

**Table 3.1: Data on monthly number of breast cancer patients**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Months/ Years | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
| 2007 | 6 | 7 | 4 | 5 | 5 | 6 | 8 | 9 | 9 | 7 | 4 | 2 |
| 2008 | 3 | 4 | 6 | 3 | 7 | 2 | 3 | 8 | 7 | 5 | 8 | 6 |
| 2009 | 2 | 5 | 3 | 7 | 1 | 2 | 6 | 12 | 7 | 8 | 7 | 4 |
| 2010 | 5 | 8 | 5 | 6 | 4 | 7 | 6 | 11 | 10 | 4 | 5 | 2 |
| 2011 | 8 | 9 | 4 | 6 | 7 | 12 | 10 | 15 | 13 | 10 | 7 | 6 |
| 2012 | 4 | 5 | 9 | 1 | 3 | 13 | 7 | 4 | 6 | 9 | 11 | 9 |
| 2013 | 1 | 4 | 3 | 5 | 6 | 8 | 4 | 8 | 2 | 5 | 3 | 7 |
| 2014 | 2 | 5 | 4 | 7 | 5 | 3 | 1 | 1 | 8 | 5 | 7 | 13 |
| 2015 | 7 | 3 | 5 | 11 | 6 | 5 | 8 | 7 | 7 | 11 | 12 | 7 |
| 2016 | 3 | 2 | 3 | 5 | 1 | 2 | 4 | 6 | 4 | 3 | 2 | 4 |

# Time Series Plot

****

**Figure 1: Time series plot of breast cancer data**

A critical visual inspection of the plot of breast cancer data shows that there is high and low frequency which is a normal pattern. Also the observed time plot showed that the series has constant mean and variance with all evidence of stationarity.

We proceed to examine the correlogram i.e. Autocorrelation function (ACF) and the Partial autocorrelation function (PACF) of the series.

**Autocorrelation Function and Partial Autocorrelation Function (ACF AND PACF)**

****

**Figure 2: The ACF and PACF of breast cancer data**

The autocorrelation function ACF and PACF clearly shows that the data is stationary since most of the spikes are within the control limit.

# **Unit Root Test**

**Result of ADF Test**

H0: The series contains unit root

H1: The series is generated by stationary process

Augmented Dickey-Fuller (GLS) test for Breast Cancer

Model: (1-L)y = b0 + (a-1)\*y(-1) + ... + e

1st-order autocorrelation coefficient for e: 0.030

Lagged Differences: F(19, 80) = 0.798 [0.7025]

Estimated Value of (a - 1): -0.83459

Test Statistic: Tau = -2.61324

Asymptotic p-value 0.008695

**Decision:** From the above result, we can observe that the asymptotic p-values is less than 0.05 level of significance, we therefore reject the null hypothesis (Ho) and conclude that the series is generated by stationary process. (i.e the series is stationary)

**Table 1: Result of KPSS test**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Critical values | 1% | 5% | 10% | Test Statistic |
|  | 0.735 | 0.465 | 0.349 | 0.145307 |

H0: The series is stationary

H1: The series is not stationary

**Decision:** By observing the table 1 above, we can notice that the test statistics for the KPSS test is less than all asymptotic critical values at all level of significance, and therefore we fail to reject the null hypothesis and conclude that the series is stationary.

# Model Identification

Since we achieved Stationarity from the data therefore the order of the Integration is 0, and we proposed (p and q) to be 1, 2 and 3 in which 7 ARIMA models are generated and we select the best among them. By considering Akaike Criterion (AIC), Schwarz Criterion (SC) and Hannan-Quinn Criterion (HQ).

**Table 2: Results of ARMA Model Identification for breast cancer data in Ilorin**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **A i c** | **S Ic** | **H qC** |
| ARMA (0,1) | 594.6581 | 603.0205 | 598.0541 |
| ARMA (0,2) | 595.3295 | 606.4794 | 599.8575 |
| **ARMA (1,0)** | **593.6322** | **601.9947** | **597.0282** |
| ARMA (1,1) | 595.3874 | 606.5373 | 599.9154 |
| ARMA (1,2) | 595.7547 | 609.6922 | 601.4148 |
| ARMA (1,3) | 599.2549 | 615.9799 | 606.0470 |
| ARMA (2,1) | 597.3873 | 611.3248 | 603.0474 |

From the table 2 above, we found that **Arma (1,0)** model has the least values of all the criterion considered, therefore we will make use of these model for further analysis.

# Parameter Estimates of the Model

ARMA model using observations from 2007:01-2016:12 (T = 120)

Estimated using Kalman filter (exact ML)

Dependent variable: Breast Cancer

Standard errors based on Hessian

**Table 3: Results of ARMA (1, 0) Model Estimation for the Breast Cancer Data in Ilorin**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | Coefficient | std. error | Z | p-value |
| Const | 5.85838 | 0.401504 | 14.59 | 3.20e-048 \*\*\* |
| phi\_1 | 0.366856 | 0.0845139 | 4.341 | 1.42e-05 \*\*\* |

Key (1% \*\*\*) (5 %\*\*) (10 %\*)

The constant term and phi\_1 are found to be significant at all level of significance (1%, 5% and 10%). The AR (1) model obtained is *Xt = 5.85838+0.366856Xt-1*

**Figure 3: The acf and pacf of ARMA (1, 0) Model**

As we observed figure 3 indicated that the residual ACF and PACF of ARIMA (1, 0) model lie within the 90% confidence interval. The model has passed the standard test criteria of being white noise.

**Figure 4: The Residual Time Plot of arma (1, 0) Model**

By observing the above plot we obtained that the plot has a wave-like pattern; showing that the series has constant mean and variance which make the series to be stationary.

**Figure 5:** Result of the Forecast and Actual Graphical Representation of Arma (1, 0) Model

**Conclusion**

The use of autoregressive moving average (ARMA) modeling strategy is now among the most popular way of analysis of time series data. We employed this method to model the breast cancer patients from 2007 to 2016. The integration order d (I(d)) was found to be zero i.e. (I(0)). ARMA (1,0) model was found to be the best model based on the criterion considered and it is to described and forecast the breast cancer patients in Ilorin. The model obtained from the analysis is *Xt = 5.85838+0.366856Xt-1.* The model was found to be adequate and best describe the monthly data of breast cancer patients in Ilorin.

**Recommendations**

Based on the result obtained from the above analysis, it is recommended that there should be an awareness and enlightenment programme for the populace on how to prevent breast cancer, the cause, effect and remedy for it. Also, women should endeavor to go for check up at least on a monthly basis.

**References**

Dangana, A., Ishaku, H., Victoria, I., Christy, C. F., Egenti, B. N. (2015). Epidemiology of Breast Cancer among Male in University of Abuja Teaching Hospital Gwagwalada. *Nigerian Biomedical Science Journal*.

García-Martínez, J., Pelechano, V., Pérez-Ortín, J. E. (2011). Genomic-Wide Methods to Evaluate Transcription Rates in Yeast. In: Becskei, A. (eds) Yeast Genetic Networks. *Methods in Molecular Biology*, 734, 25–44. Retrieved from https://doi.org/10.1007/978-1-61779-086-7\_2.

Henry, A. A. (2013) A to Z Community Health and Social Medicine. Published by 3:00am communication Nigeria.

Hirok, K. A., Soliman, S., Hablas, A., Seifeldin, A. I., Ramadan, M., Banerjee, M., Hardord, B. J., Chamberlain, M. R. and Merajver, D. S. (2013). “Trends in Breast Cancer Incidence Rates by Age and Stage at Diagnosis in Gharbiah, Egypt, over 10 Years (1999-2008). *Journal of Cancer Epidemiology*, 916394. Retrieved from http://dx.doi.org/10.1155/2013/916394.

Luo, X., Chen, X., Xu, L., Myneni, R. and Zhu, Z. (2013). Assessing Performance of NDVI and NDVI3g in Monitoring Leaf Unfolding Dates of the Deciduous Broadleaf Forest in Northern China.  *Remote Sensing*, *5*(2), 845–861. Retrieved from http://dx.doi.org/10.3390/rs5020845.

Mohammed, A. M., Moles, R. J. and Chen, T. F. (2015). Medication-related burden and patients’ lived experience with medicine: a systematic review and metasynthesis of qualitative studies. *BMJ Open*, 6, e010035. Retrieved from http://doi:10.1136/bmjopen-2015-010035.

[Politis](https://pubmed.ncbi.nlm.nih.gov/?term=Politis%20M%5BAuthor%5D), M.,  [Higuera](https://pubmed.ncbi.nlm.nih.gov/?term=Higuera%20G%5BAuthor%5D), G., [Chang](https://pubmed.ncbi.nlm.nih.gov/?term=Chang%20LR%5BAuthor%5D), R. L., [Gomez](https://pubmed.ncbi.nlm.nih.gov/?term=Gomez%20B%5BAuthor%5D), B., [Bares](https://pubmed.ncbi.nlm.nih.gov/?term=Bares%20J%5BAuthor%5D), J. and [Motta](https://pubmed.ncbi.nlm.nih.gov/?term=Motta%20J%5BAuthor%5D), J. (2015). “Trend Analysis of Cancer Mortality and Incidence in Panama, Using Joinpoint Regression Analysis”. *Medicine (Baltimore)*, 94(24), p e970.

San, M., Olanrewaju, S. O. and Oguntade, E. S. (2020): Analysis of Reported Cases of Breast Cancer in Abuja, Nigeria: A Time Series Approach. *IRE Journals*, 3(9), 2456-8880.

Sissoko, D., Laouenan, C., Folkesson, E., M’Lebing, A. B., Beavogui, A. H. and Baize, S. (2016). Experimental Treatment with Favipiravir for Ebola Virus Disease (the JIKI Trial): A Historically Controlled, Single-Arm Proof-of-Concept Trial in Guinea. *PLoS Med*, 13(3), e1001967. http://doi:10.1371/journal.pmed.1001967

Sowunmi, A., Alabi, A., Fatiregun, O., Olatunji, T., Okoro, U. S. and Durosinmi-Etti, A. F. (2018). Trend of cancer incidence in an oncology center in Nigeria. *West Africa Journal of Radiology*, 25, 52–56.