Stochastic Modeling and Applications
Vol. 25 No. 2 (July-December, 2021)
ISSN: 0972-3641

Received: 25th July 2021 Revised: 19th October 2021 Selected: 24th November 2021

FORECASTING ANALYSIS FOR LUNG CANCER INCIDENCE AND MORTALITY DATA IN TAMILNADU USING DYNAMIC REGRESSION MODEL

S.POYYAMOZHI[1], A. KACHI MOHIDEEN[2].

Abstract

In this paper, the model and forecast lung cancer incidence in Taminadu using dynamic regression models with finite and infinite lags. These models involve autoregressive models (ARs), distributed lag models (DLMs), polynomial distributed lag models (PDLs) and autoregressive polynomial distributed lag models (ARPDLs) and outline the implementation and forecasting issues with DLM and PDL models respectively and additional to present one-step ahead forecast for the various models. Finally, the lung cancer data in Tamilnadu to evaluate the robustness of the results, we explore ARPDL models and present our summaries of best models and their forecasts

Key words: forecast, lung cancer, autoregressive models, dynamic regression models

1. Introduction

Cancer is a basic term for large group of diseases characterised by the increase of abnormal cells past their ordinary boundaries which could then occupy adjoining elements of the body and spread to other organs. Other common phrases used are malignant tumors and neoplasms. Cancer can have an effect on nearly any a part of the body and has many anatomic and molecular subtypes that each requires specific management techniques. Cancer is the second main reason of dying globally and is estimated to account for 9.6 million deaths in 2018. Lung, prostate, colorectal, belly and liver cancer are the most commonplace forms of most cancers in guys, whilst breast, colorectal, lung, cervix and thyroid most cancers are the maximum commonplace among women.

Worldwide, tobacco use is the single best preventable threat aspect for cancer mortality and kills about 6 million human beings every year, from most cancers and other sicknesses. Tobacco smoke has more than 7000 chemicals, at least 250 are known to be dangerous and greater than 50 are regarded to motive cancer. Tobacco smoking reasons many forms of most cancers, such as cancers of the lung, oesophagus, larynx (voice field), mouth, throat, kidney, bladder, pancreas, stomach and cervix. Second-hand smoke, additionally known as environmental tobacco smoke, has been verified to motive lung most cancers in non-smoking adults. Smokeless tobacco reasons oral, oesophageal and pancreatic most cancers. Nearly 80% of the 1 billion people who smoke within the world live in low- and middle-income countries.

2. Model Description

2.1. Linear Model of First-order Autoregressive AR(1)

In simple terms, autocorrelation means that current values depend on past values. A common starting point of analysis is a simple model of positive first-order AR(1) autocorrelated error-process associated with a regression equation that can be represented by the following two equations

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \quad \cdots (1)$$

and $X_t = S_t \times P_t$

Where Y_t is the incidence at time t (number of cases in month t), X_t is the smoking population in 10,000, S_t is the smoking prevalence and P_t is the population size,

and $\varepsilon_t = \rho\varepsilon_{t-1} + \vartheta_t \quad \cdots (2)$

The consequences when ordinary least squares (OLS) is used to estimate β_0 and β_1 Because (OLS) reported Standard Errors are misleading and the OLS estimators are inconsistent if autocorrelation exists, we need to investigate if the errors are auto correlated (Barretto and Howland, 2006).

2.2. Generalized Least Squares

A transformation of and can be such that the resulting linear model has an independent error structure and transform the model so that we get rid of the errors that the errors that are independent and normally distributed. This transformation is defined as followed by substituting the error-forecasting equation is obtain a model in which the error term is a pure, independently and identically distributed error, ϑ_t

$$Y_t - \rho Y_{t-1} = \beta_0(1-\rho) + \beta_1 X_t + \rho\beta_1 X_{t-1} + \vartheta_t \quad \cdots (3)$$

If defined new variable is $Y_t^* = Y_t - \rho Y_{t-1}$ and $X_t^* = X_t - \rho X_{t-1}$, $then\ we\ have$

$$Y_t^* = \beta_0(1-\rho) + \beta_1 X_t^* + \vartheta_t \quad \cdots (4)$$

Equation (4) is known as the transformed model with a well-behaved error term. OLS may be applied to this equation. However is unknown, so it must be estimated from the regression of residuals on lagged residuals. We then use it to transform the original data to obtain the original parameter values of the model. Running OLS on the transformed model is called generalized least squares (GLS). It generates the right SEs as it is the best linear unbiased estimator. Notice as well that when =0, the transformation reduces to the familiar OLS model. Several procedures have been developed, but the most popular ones are Cochrane-Orcutt iterative procedure, Prais-Winsten (1954) in order to transform the variables in the first observation we need to apply the following formula to the first observation as follows

$Y_1^* = \sqrt{(1-\rho^2)}Y_1$ and $X_1^* = \sqrt{(1-\rho^2)}X_1$

2.3. Distributed Lag Models (DLMs)

Distributed lag models are useful because they allow a dependent variable to depend on past values of an explanatory variable at various lags. When the population is increasing, this means that the age distribution will increase over time. Algebraically, we can demonstrate this lag effect by saying that a change in a policy variable X_t has an effect on the dependent or response variables $Y_t, Y_{t-1}, \cdots$ If we turn this around slightly, then we can say that is affected by the values of $X_t, X_{t-1},$ or

$$Y_t = f\left(X_t, X_{t-1}, X_{t-2,} X_{t-3,} X_{t-4,}, \cdots X_{t-k}\right) \quad \cdots (5)$$

This distributed lag model is finite as the duration of the effects is a finite period of time, namely k periods. This model is said to be dynamic because it describes the reaction over time. In order to convert Equation (5) into a distributed lag model we need a functional form with an error term and then make assumptions about the properties of the error term.

2.4. Finite Distributed Lag Models

To model the finite distributed lag, the functional form is assumed to be linear, so that the finite lag model, with an additive error term, is

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + \cdots + \beta_1 X_{t-k} + \varepsilon_t$$
$$= \alpha + \sum_{i=0}^{k} \beta_i X_{t-i} + \varepsilon_t \quad \cdots (6)$$

Where, assume that $E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma^2, and\ \ Cov(\varepsilon_t, \varepsilon_s) = 0$

In this model the parameter α is the intercept and β_i is the distributed lag weight to reflect the fact that it measures the effect of changes in past values of X on the expected current value of Y, all other things being equal in equ (6). There are many consequences of collinearity. Firstly, the estimates of least squares are imprecise, meaning that a wide interval estimates will be detected. Secondly, high levels of correlation among the regressors imply multicollinearity, which leads to unreliable and inconsistent coefficient estimates with large variances and standard errors.

If we assume that the residuals are also uncorrelated with all future values of X this is called strict exogeneity $| E(\varepsilon_t / X_{t+k+s} \cdots X_t \cdots X_{t+k+s}) = 0$ and there may be estimation techniques other than OLS that can be used to estimate dynamic causal effects (Hill et al, 2000). From empirical studies using real data, it has been shown that short-term forecasts are more reliable than long-term forecasts because the forecast relies more on immediate past observations than long-term observations. The evaluate to finding the relationship between the dependent variables and one current value of the independent variables to get the short and long-run multiplier of lung cancer incidence in Tamilnadu.

Koyck (1954) proposed a geometrically declining scheme for the β s. Therefore, rather than estimate the model with a large number of lags we can transform the data into a more parsimonious form by using the Koyck Transformation procedure.

Begin with a model of Y as a function of X and k lags of X is,

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + \cdots + \beta_1 X_{t-k} + \varepsilon_t$$

The distributed lagged model (DLM) the effect of variable X_t diminishes as the lag gets larger by an amount λ each period. This is reflected in the size of coefficients such that $\beta_1 = \beta_0 \lambda^i$ and $0 < \lambda < 1$,

Where, λ is a fraction, so the larger the value of λ the slower the speed of adjustment Substituting into the DLM in Equation we get,

$$Y_t = \alpha + \beta_0 X_t + \beta_1 \lambda X_{t-1} + \beta_2 \lambda^2 X_{t-2} + \beta_3 \lambda^3 X_{t-3} + \cdots + \beta_1 \lambda^k X_{t-k} + \varepsilon_t \quad \cdots (7)$$
$$= \alpha + \beta_0 (X_t + \lambda X_{t-1} + \lambda^2 X_{t-2} + \lambda^3 X_{t-3} + \cdots + \lambda^k X_{t-k} + \varepsilon_t \quad \cdots (8)$$

If (8) is true at time t it is also true at time t-1, so if we lag Equation (8) one time period,

$$Y_t = \alpha + \beta_0 (+\lambda X_{t-1} + \lambda^2 X_{t-2} + \lambda^3 X_{t-3} + \cdots + \lambda^k X_{t-k} + \varepsilon_{t-1} \quad \cdots (9)$$

Multiplying Equation (9) by λ gives

$$\lambda Y_t = \lambda\alpha + \beta_0 (+\lambda X_{t-1} + \lambda^2 X_{t-2} + \lambda^3 X_{t-3} + \cdots + \lambda^{k+1} X_{t-k+1}) + \lambda\varepsilon_{t-1} \quad \cdots (10)$$

Subtracting Equation (10) from Equation (9), we obtain

$$Y_t - \lambda Y_t = \{\alpha + \beta_0 (+\lambda X_{t-1} + \lambda^2 X_{t-2} + \lambda^3 X_{t-3} + \cdots + \lambda^k X_{t-k} + \varepsilon_{t-1}\} - \{\lambda\alpha + \beta_0 (+\lambda X_{t-1} + \lambda^2 X_{t-2} + \lambda^3 X_{t-3} + \cdots + \lambda^{k+1} X_{t-k+1}) + \lambda\varepsilon_{t-1}\}$$

Simplifying (all lags cancel out) gives

$$Y_t - \lambda Y_{t-1} = (1-\lambda)\alpha + \beta_0 X_t + \varepsilon_t - \lambda\varepsilon_{t-1}$$

Hence,

$$Y_t = (1-\lambda)\alpha + \beta_0 X_t + \lambda X_{t-1} + (\varepsilon_t - \lambda\varepsilon_{t-1}) \quad \cdots (11)$$

Using Equation (11), regress on and to generate estimates of and use these estimates to compute the coefficients at each lag as well as the original intercept This transformation is known as the Koyck transformation. Forecasting with the AR (1) Models are,

Given data $(Y_1, Y_2, Y_3, Y_4, \cdots Y_\tau)$the one period ahead optimal forecast is,

$$\begin{aligned}\hat{Y}_{\tau+1,\tau} &= E\left(Y_{\tau+1}/_{\Omega_\tau}\right)\\ &= \alpha_0 + E\left(\alpha_1 Y_\tau/_{\Omega_\tau}\right) + E\left(Y_{\tau+1}/_{\Omega_\tau}\right)\\ &= \alpha_0 + \alpha_1 Y_\tau\end{aligned}$$

In compute $\hat{Y}_{\tau+1,\tau} = \hat{\alpha}_0 + \hat{\alpha}_1 Y_\tau$ using the estimates. The one-step ahead optimal forecast error of AR (1) is $\hat{Y}_{\tau+1} - \hat{Y}_{\tau+1,\tau} = \epsilon_{\tau+1}$,

The forecast error variance is,

$$Var(\hat{Y}_{\tau+1} - \hat{Y}_{\tau+1,\tau}) = Var(\epsilon_{\tau+1}) = \sigma^2$$

Construct the prediction interval (PI) using the following equation

$$\hat{Y}_{\tau+1} \pm Z_{\alpha/2}\sqrt{\hat{Y}_{\tau+1}\widehat{Var}}(\hat{Y}_{\tau+1})$$

Therefore, the 95% PI is computed in

$\widehat{Y}_{\tau+1} \pm 1.96\sqrt{\hat{Y}_{\tau+1}\widehat{Var}}(\hat{Y}_{\tau+1})$

And the estimated AR(1) model is

$$Y_t = 14.788 + 0.5462Y_{t-1} + \varepsilon_t$$

The AR (1) with T observations has the mean μ = 32.584, α_1= 0.5462, Y_τ= 49, σ^2= 61.2. The AR (1) process is $Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \varepsilon_t$

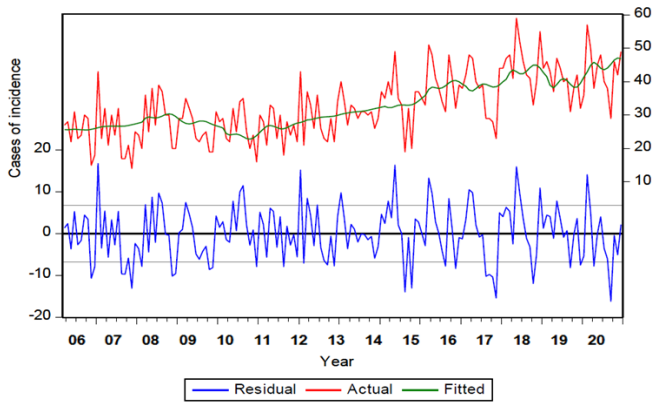
Where, $\alpha_0 = \mu(1-\alpha_1)$ so that $\alpha_0 = \mu(1-\alpha_1$= 32.584 (1-0.5462) = 14.788. The one period ahead forecast is 14.788+ 0.5462× 49 = 41 cases of lung cancer. Thus, the one-step ahead forecast is a fixed amount $\alpha_0 + \alpha_1 Y_{t-1}$ plus the stochastic term ε_t . The fixed amount has a variance of zero, so the variance of the one-step ahead forecast is $\hat{\sigma}^2 = 61.2$. The plots for one-step ahead forecasts and the residuals are shown in Figure 5.4 and Figure 5.5respectively.

Forecasting with the Linear Regression Model with Lagged Covariate

3. Results and Discussion

Forecasting of the best ARPDL model of the total cases of lung cancer on total smoking population. The one-step-ahead forecast is 50 with 95% of adjusted R-squared for the estimated relation is 55.4 and the overall F-test value is 14.68 with p-value 0.00. Thus, the long run effect suggests that there will be on average 50 cases of lung cancer per month for the next 24 months. The estimated yearly lung cancer cases in 2020 and 2021 are 606 and 581 respectively. In a main aim of regressing the total cases of lung cancer on smoking population separately for males and females is that we want to identify the effect of changes in past values of smoking population separately for males and females on the current expected value of total cases of lung cancer. Particularly, we want to see where the effect of smoking is greater among males or females. In addition, we aim to minimize the error as much as possible since there are available data on smoking population for males and females separately in order to obtain reliable residual plots figure-1 for the best PDL model of lung cancer cases per month from 2006 to 2020 in Tamilnadu

Fig-1:Fitted and residual plots for the best PDLmodel of lung cancer cases per month from 2006 to 2020 in Tamilnadu



The estimated model the short and long-run multipliers are the same i.e. = 0.116 (see Table- 1). Therefore, 1% increase in smoking population suggests an approximately immediate and permanent 12% (43) individual cases increase in lung cancer per month. The value of the Durbin-Watson test indicates that there seems to be first order autocorrelation in the data so standard errors are wrongly estimated, but coefficients are unbiased.

Table -1: Regression for result in total cases of Lung Cancer against smoking population

gression equation is

Y_t = - 0.03 + 0.116 X_t

tor Coef SE Coef T P
ant -0.034 2.849 -0.01 0.990
0.11623 0.01007 11.55 0.000

1722 R-Sq = 41.2% R-Sq(adj) = 40.9%

is of Variance

DF SS MS F P
sion 1 6753.9 6753.9 133.33 0.000
al Error 190 9624.4 50.7
191 16378.3
n-Watson statistic = 1.55538

In a Apply for the Koyck transformation to the total cases of Lung Cancer against smoking population, the results are shown in Table-2. From the estimated equation we can find the coefficient parameters as follows:

Table -2: Results of Koyck transformation estimated Coefficient parameters.

gression equation is

Y_t = - 1.18 + 0.0994 X_t + 0.176 Yt-1

ases used, 1 cases contain missing values

tor	Coef	SE Coef	T	P
ant	-1.182	2.682	-0.44	0.660
	0.09939	0.01221	8.14	0.000
	0.17621	0.06786	2.60	0.010

)7419 R-Sq = 46.9% R-Sq(adj) = 46.3%

is of Variance

	DF	SS	MS	F	P
sion	2	7392.0	3696.0	82.97	0.000
al Error	188	8374.4	44.5		
	190	15766.4			

n-Watson statistic = 2.01369

Using the estimates $\hat{Y}_{\tau+1,\tau} = \hat{\beta}_0 + \hat{\beta}_1 X_\tau$ and the estimated regression equation is as follows

$$Y_t = -1.448 + 0.1211 X_{t-1} + \varepsilon_t$$

Therefore the one-step ahead forecast when X_τ= 377.540 is

$$Y_{\tau+1} = \beta_0 + \beta_1 X_\tau$$

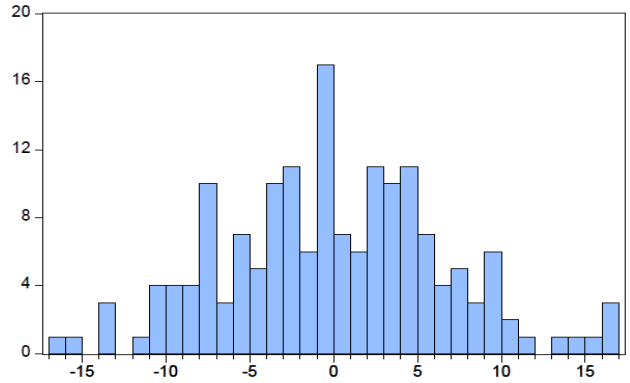
$$Y_{\tau+1} = -1.448 + 0.1211 \times 337.540 = 44 \ cases\ of\ lung\ cancer.$$

Forecast Error in Prediction Interval The 95% PI is computed as follows:

$$\hat{Y}_{\tau+1} \pm 1.96 \sqrt{\widehat{Var}(\hat{Y}_{\tau+1})}$$

$$44 \pm 1.96 \sqrt{45.42}$$

Fig-1: Fitted and residual plots for Ordinary Least Square model of Lung cancer cases per month from 2006 to 2020.



	ıals
e	103 2020M12
vations	
	03
n	391
ıum	45
ıum	439
ev.	73
ess	44
is	45
-Bera	25
ıility	30

Table-3: Results of restricted least squared PDL model

Varible	Coe fficient	t –Statistic	p-value
Z_{0t}	3.93	0.29	0.86
Z_{1t}	-0.32	-1.88	0.04
Z_{2t}	-0.12	-0.39	0.64
Z_{3t}	0.03	1.36	0.28
Z_{4t}	0.02	0.24	0.84
Z_{5t}	-0.00	-0.67	0.55
Z_{6t}	-0.02	-0.26	0.74
Z_{7t}	0.01	0.29	0.70
Z_{8t}	0.01	0.20	0.68
Z_{9t}	-0.02	-0.09	0.83
Z_{10t}	0.68	1.26	0.24
Z_{11t}	-0.38	-1.32	0.29
Z_{12t}	0.01	0.04	0.77
Z_{13t}	0.12	1.42	0.34
Z_{14t}	-0.00	-1.03	0.38
Z_{15t}	-0.00	-1.52	0.02
Z_{16t}	0.02	1.60	0.21
Z_{17t}	0.01	1.51	0.32
Z_{18t}	-0.00	-1.80	0.17

red,	354	rz criterion	83
dependent variance.	08	kelihood -	99
ed R-squared.	02	n-Quinn criteria.	72
ependent variance.	90	tic	69
regression.	54	n-Watson stat	35
information criterion.	92	ility(F-statistic)	00
quared residual.	12		

The created variables from Z_{0t} to Z_{8t} refer to the lag of $x_1 t - i$ where as the variables from Z_{9t} to Z_{17t} refer to the lag of $x_0 t - i$. What the polynomial approximation has done is to reduce the number of parameters that have to be estimated restricted equation.

Table- 4: Summary of Polynomial Models result.

s	$\bar{R}^2$		σ^2	ep head and 5%CI	st value
				2.8, 45.7)	
				8.1, 42.0)	
L				2.6, 35.5)	
- II					
				9.4, 33.5)	
				1.0, 54.5)	
L				5.6, 52.6)	

The fitted model is shown in Figure-2 together with residual diagnostic plots. This is followed by the distribution of the series in the histogram with a complement of standard descriptive statistics displayed along with the histogram. The p-value of the Jarque-Bera test is not less than for 5% significance level and hence we do not reject the null hypothesis that the model is normally distributed level.

4. Conclusion

Finally a fit of new ARPDL model in data used are monthly incidence cases of lung cancer and smoking population in Tamilnadu by gender from 2006-2020.The empirical results suggest that Lung Cancer cases are strongly affected by smoking, and most of the cases are among males. However the value of the sum of t-ratios of the best model ARPDL suggest that the smoking effect is greater for females than for males. The one-step-ahead forecasts for each different Forecasting AR(1) model. The one-step-ahead forecast is 41 with 95% PI. The mean square error is 69.5. The one-step-ahead forecast is 44 with 95% adjusted R-squared for the estimated relation is 45.3 and the mean square error for linear regression model with lagged covariate and AR(1) errors. Estimated relation is 36.4 and the mean square error distributed lagged variable model (DLM). However, when it comes to time series forecasting, because of the inherent serial correlation and potential non stationarity of the data, its application is not straightforward and often omitted by practitioners in favour of an out-of-sample evaluation. Hence we generated our forecasts accordingly using the seasonal ARIMA model.

It is important to mention that cross correlation methods in the time domain and impulse response functions in frequency domain which are generated through cross spectral analysis are other potential methods that can be used for modelling bivariate time series. Consideration of these approaches may lead to models that can be derived more efficiently than using lagged regression models with their many parameters. However, due to time constraints, we have not considered these approaches.

References

1. Bray, F. and Moller, B. (2006) Predicting the future burden of cancer. *Nat Rev Cancer, 6, 63-74.*
2. Clements, M. S., Armstrong, B. K. and Moolgavkar, S. H. (2005) Lung cancer rate predictions using generalized additive models. *Biostatistics, 6, 576-589.*
3. Devesa, S. S., Silverman, D. T., Young, J. L., Pollack, E. S., Brown, C. C., Horm, J. E. (1987) Cancer incidence and mortality trends among whites in the United States, 1947-84. *J Natl Cancer Inst, 79, 701-770.*
4. Doll, R., Payne, P. and Waterhouse, J. (1966) Cancer Incidence in Five Continents.Geneva, UICC. *Berlin: Springer, Volume 1.*
5. Hurvich, C.M., & Tsai, C.L (1989) Regression and Time Series Model Selection in Small Sample. *Biometrika, 76, 297-307.*
6. McCullagh, P. & Nelder, J.A. (1983) Generalized linear models. *New York: Chapman and Hall.*
7. E Sakthivel and S Anitha (2021) Survival Analysis for Diagnosing Tuberculosis Patients, *Stochastic Modeling and Applications,* Volume 25, Issue No. 1, Pages 35-43, MUK Publications
8. Pankratz, A. (1991) Forecasting with dynamic regression models. *New York: John Wiley and Sons.*
9. Wei, W. W. S. (1990) Time Series Analysis: Univariate and Multivariate Methods. *California: Addison-Wesley Publishing Company.*
10. West, M., & Harrison, P.J. (1989) Bayesian Forecasting and Dynamic Models. *New York: Springer-Verlag. (2nd ed., 1997).*
11. World Health Organization, Media Center, Cancer, (2014), accessed date 4-11-2014, available at http://www.who.int/mediacentre/factsheets/fs297/en/
12. World Health Organization, Media Center, Cancer, (2015), accessed date 24-02-2015, available at < http://www.who.int/mediacentre/factsheets/fs297/en/>

S.POYYAMOZHI : Assistant Professor and Head, Department of Statistics,
Government Arts College (Autonomous), Kumbakonam. Affiliated to BharathidasanUniversity, Tiruchirappalli, Tamilnadu, India

A. KACHI MOHIDEEN: Assistant Professor, Department of Statistics,
Periyar EVR College (Autonomous), Trichy ,Affiliated to BharathidasanUniversity, Tiruchirappalli , Tamilnadu, India