ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

REGRESSION MODELS OF DEATHS OCCURRED BY TUBERCULOSIS (TB) DISEASE IN THE SUDAN

Dr. Mohamed Hassan Mahmoud Farg*

* Associate Professor, (Statistics), Faculty of Economics and Political Sciences-Omdurman Islamic University (Sudan) – Shaqra University- KSA

ABSTRACT

This paper aims at studying the regression models that represent the relationship between the deaths that occurred by TB disease as dependent variable and the number of cases of TB. The study was conducted in the Sudan, it includes cross-sectional data covered all the fifteenth states of the Sudan. The data which was used in the research covered the period 2010-2011. The Linear Regression (L.R.) was used to analyze the data. The important result was, there is a significant relationship between the deaths that occurred by TB and the number of cases of TB. The model: Log Y = -0.886 + 0.843 L

KEYWORDS: Tuberculosis (TB), Regression, Model(s), Sudan

INTRODUCTION

This paper deals with fitting models that should be used to estimate the number of deaths that may occur by TB disease in the Sudan.

The problem of the study is that, the researcher does not find the previous study used fitted models to estimate the number of deaths that occur with TB disease in the Sudan; therefore he conducted this study.

The main objective of this paper is to fit a mathematical model used to estimate the number of deaths that occurs with TB disease.

The importance of this paper is that, it will determine the best model that passes all the tests of the significance and the assumptions.

Tuberculosis, MTB, or TB (short for tubercle bacillus), in the past also called phthisis, phthisis pulmonalis, or consumption, is a widespread, and in many cases fatal, infectious disease caused by various strains of mycobacteria, usually Mycobacterium tuberculosis (Kumar, et al; 2007). Tuberculosis typically attacks the lungs, but can also affect other parts of the body. It is spread through the air when people who have an active TB infection, cough, sneeze, or otherwise transmit respiratory fluids through the air (Konstantinos, 2010). Most infections do not have symptoms, known as latent tuberculosis. About one in ten latent infections eventually progresses to active disease which, if left untreated, kills more than 50% of those so infected.

There are many studies that have conducted to study TB disease. Sudan has a large and growing private health sector. No survey was done in Sudan to show the extent of the use of private health care services by the population. Also, precise data on tuberculosis (TB) diagnosis and treatment in the private sector are not available, (Eltilib; et al, 2010).

In fact, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

is a function of the independent variables called the regression function. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function which can be described by a probability distribution.

Really regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which, among the independent variables relate to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However, this can lead to illusions or false relationships, so caution is advisable; (Armstrong; 2012), for example, correlation does not imply causation.

Many techniques for carrying out regression analysis have been developed. Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. A nonparametric regression refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite-dimensional.

The execution of regression analysis methods in practice depends upon the form of the data generating process, and how it relates to the regression approach being used. Since the true form of the data-generating process is generally not known, regression analysis often depends to some extent on making assumptions about this process. These assumptions are sometimes testable if a sufficient quantity of data is available. Regression models for prediction are often useful even when the assumptions are moderately violated, although they may not perform optimally. However, in many applications, especially with small effects or questions of causality based on observational data, regression methods can give misleading results (Freedman; 2005) and (Cook; 1982).

Usually regression models involve the following variables:

The unknown parameters, denoted as β, which may represent a scalar or a vector.

The independent variables, X.

The dependent variable, Y.

In various fields of application, different terminologies are used in place of dependent and independent variables.

A regression model relates Y to a function of X and β is $Y= f(X, \beta)$.

The approximation is usually formalized as $E(Y \mid X) = f(X, \beta)$. To carry out regression analysis, the form of the function f must be specified. Sometimes the form of this function is based on knowledge about the relationship between Y and X that does not rely on the data. If no such knowledge is available, a flexible or convenient form for f is chosen.

Assume now that the vector of unknown parameters β is of length k. In order to perform a regression analysis the user must provide information about the dependent variable Y:

If N data points of the form (X, Y) are observed, the most common situation is $N > k$ data points are observed. In this case, there is enough information in the data to estimate a unique value for β that best fits the data in some sense, and the regression model when applied to the data can be viewed as an overdetermined system in β.

In the last case, the regression analysis provides the tools for:

Finding a solution for unknown parameters β that will, for example, minimize the distance between the measured and predicted values of the dependent variable Y (also known as the method of least squares). Under certain statistical assumptions, the regression analysis uses the surplus of information to provide statistical information about the unknown parameters β and predicted values of the dependent variable Y.

Necessary number of independent measurements:

Consider a regression model which has two unknown parameters, β_0, β_1. If we have 10 pairs of (X,Y), the best one can do is to estimate the average value and the standard deviation of the dependent variable Y. Similarly, measuring at two different values of X would give enough data for a regression with the two unknowns, but not for three or more unknowns.

If the experimenter had performed measurements at three different values of the independent variable vector X, then regression analysis would provide a unique set of estimates for the three unknown parameters in β.

In the case of general linear regression, the above statement is equivalent to the requirement that the matrix $(X^T.X)$ is invertible.

Statistical assumptions:

When the number of measurements, N, is larger than the number of unknown parameters, k, and the measurement errors εi are normally distributed, then the excess of information contained in $(N - k)$ measurements is used to make

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

statistical predictions about the unknown parameters. This excess of information is referred to as the degrees of freedom of the regression.

Classical assumptions for regression analysis include:

- The sample is representative of the population for the inference prediction.
- The error is a random variable with a mean of zero conditional on the explanatory variables.
- The independent variables are measured with no error. (Note: If this is not so, modeling may be done instead using errors-in-variables model techniques).
- The predictors are linearly independent, i.e. it is not possible to express any predictor as a linear combination of the others.
- The errors are uncorrelated, that is, the variance–covariance matrix of the errors is diagonal and each non-zero element is the variance of the error.
- The variance of the error is constant across observations (homoscedasticity). If not, weighted least squares or other methods might instead be used.

These are sufficient conditions for the least-squares estimator to possess desirable properties; in particular, these assumptions imply that the parameter estimates will be unbiased, consistent, and efficient in the class of linear unbiased estimators. It is important to note that actual data rarely satisfies the assumptions. That is, the method is used even though the assumptions are not true. Variation from the assumptions can sometimes be used as a measure of how far the model is from being useful. Many of these assumptions may be relaxed in more advanced treatments. Reports of statistical analyses usually include analyses of tests on the sample data and methodology for the fit and usefulness of the model.

Assumptions include the geometrical support of the variables, Cressie (1996).Independent and dependent variables often refer to values measured at point locations. There may be spatial trends and spatial autocorrelation in the variables that violate statistical assumptions of regression. Geographic weighted regression is one technique to deal with such data, Fotheringham; 2002. Also, variables may include values aggregated by areas. With aggregated data the modifiable areal unit problem can cause extreme variation in regression parameters, Fotheringham; 1991. When analyzing data aggregated by political boundaries, postal codes or census areas results may be very distinct with a different choice of units.

Linear regression:

In linear regression, the model specification is that the dependent variable, $\mathbf{y_i}$ is a linear combination of the parameters (but need not be linear in the independent variables). For example, in simple linear regression for modeling "n" data points there is one independent variable: $\mathbf{x_i}$, and two parameters, $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$:

straight line:

$$\mathbf{y_i = \beta_0 + \beta_1 x_i + E_i} \quad \text{.....(1)}$$

$$\mathbf{i = 1,2,3\ldots,n}$$

In multiple linear regression, there are several independent variables or functions of independent variables.

Adding a term in x_i^2 to the preceding regression gives a parabola as in equation 2:

$$\mathbf{y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i} \quad \text{.....(2)}$$

$$\mathbf{i = 1,2,3\ldots,n}$$

This is still linear regression; although the expression on the right hand side is quadratic in the independent variable $\mathbf{x_i}$, it is linear in the parameters $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$.

In both cases, $\mathbf{E_i}$ is an error term and the subscript "i" indexes a particular observation.

Given a random sample from the population, we estimate the population parameters and obtain the sample linear regression model 3:

$$\hat{\mathbf{y}}_\mathbf{i} = \hat{\boldsymbol{\beta}}_\mathbf{0} + \hat{\boldsymbol{\beta}}_\mathbf{1}\mathbf{x_i} \quad \text{.....(3)}$$

The residual, , $e_i = y_i - \hat{y}_i$, is the difference between the value of the dependent variable predicted by the model, $\mathbf{y_i}$, and the true value of the dependent variable, $\mathbf{y_i}$. One method of estimation is ordinary least squares. This method obtains parameter estimates that minimize the sum of squared residuals, SSE, Kutner, et al; 2004 and Ravishankar and Dey; 2002. Similarly with respect to model or equation 2, we estimate the population parameters and obtain the sample linear regression model 4:

$$\hat{\mathbf{y}}_\mathbf{i} = \hat{\boldsymbol{\beta}}_\mathbf{0} + \hat{\boldsymbol{\beta}}_\mathbf{1}\mathbf{x_i} \quad \text{.....(4)}$$

The residual, , $e_i = y_i - \hat{y}_i$, is the difference between the value of the dependent variable predicted by the model, $\mathbf{y_i}$, and the true value of the dependent variable, $\mathbf{y_i}$. One method of estimation is ordinary least squares.

Also, sometimes denoted RSS:

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

$$SSE = \sum_{i=1}^{n} e_i^2.$$

Minimization of this function results in a set of normal equations, a set of simultaneous linear equations in the parameters, which are solved to yield the parameter estimates, β_0 and β_1 .

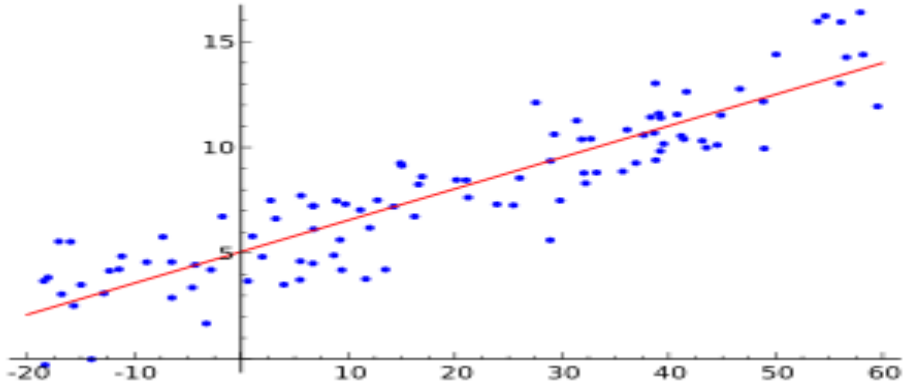

Illustration of linear regression on a data set.

In the case of simple regression, the formulas for the least squares estimates are

$$\widehat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \text{ and } \hat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

where $\bar{x}$ is the mean (average) of the x values and $\bar{y}$ is the mean of the y values.

Under the assumption that the population error term has a constant variance, the estimate of that variance is given by:

$$\hat{\sigma}_\varepsilon^2 = \frac{SSE}{n-2}.$$

This is called the mean square error (MSE) of the regression. The denominator is the sample size reduced by the number of model parameters estimated from the same data, (n-p) for p regressors or (n-p-1) if an intercept is used.[22] In this case, p=1 so the denominator is n-2.

The standard errors of the parameter estimates are given by

$$\hat{\sigma}_{\beta_0} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}, \quad \text{and} \quad \hat{\sigma}_{\beta_1} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{\sum(x_i - \bar{x})^2}}.$$

Under the further assumption that the population error term is normally distributed, the researcher can use these estimated standard errors to create confidence intervals and conduct hypothesis tests about the population parameters.

In the more general multiple regression model, there are p independent variables can be written as follows:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad \text{.....(5)}$$

where xij is the ith observation on the jth independent variable, and where the first independent variable takes the value 1 for all i (so β_1 is the regression intercept).

The least squares parameter estimates are obtained from p normal equations. The residual can be written as

$$\varepsilon_i = y_i - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}. \quad \text{.....(6)}$$

The normal equations are

$$\sum_{i=1}^{n} \sum_{k=1}^{p} X_{ij} X_{ik} \hat{\beta}_k = \sum_{i=1}^{n} X_{ij} y_i, \ j = 1, \ldots, p. \quad \text{.....(7)}$$

In matrix notation, the normal equations are written as

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

$$(\mathbf{X}^\top \mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{Y},$$ .....(8)

where the ij element of X is xij, the i element of the column vector Y is yi, and the j element of $\hat{\beta}$ is $\hat{\beta}_j$. Thus X is n×p, Y is n×1, and $\hat{\beta}$ is p×1. The solution is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}.$$ .....(9)

Once a regression model has been constructed, it may be important to confirm the goodness of fit of the model and the statistical significance of the estimated parameters. Commonly used checks of goodness of fit includes the R-squared, analyses of the pattern of residuals and hypothesis testing. Statistical significance can be checked with an F-test of the overall fit, followed by t-tests of individual parameters.

Interpretations of these diagnostic tests rest heavily on the model assumptions. Although examination of the residuals can be used to invalidate a model, the results of a t-test or F-test are sometimes more difficult to interpret if the model's assumptions are violated. For example, if the error term does not have a normal distribution, in small samples the estimated parameters will not follow normal distributions and complicate inference. With relatively large samples, however, a central limit theorem can be invoked such that hypothesis testing may proceed using asymptotic approximations, Steel, et al; 1960 and Chiang; 2003.

.

MATERIAL AND METHODS

A cross-sectional survey was carried out of reports of Ministry of Health (MOH) at the Sudan. Data of two years 2101 and 2011 are used in the data analysis. The data covered all the fifteen states of the Sudan.

The important reasons for using the data for the years 2010 and 2011 are:

- The data that found in the studied states is typical of that in the united MOH.
- Over the years 2010 and 2011, Sudan, separated into two countries.
- Over the years 2010 and 2011, States of Darfor and Kordofan separated into new states..

The models were used in this paper are given below:

Firstly: For the data of year 2010, equations 1 and 2 are used in fitting regression models of death of TB disease in cases of TB.

where: Y is the number of deaths that occurred by TB disease.
X is the number of cases of TB disease.
β_0 is the number of deaths when TB disease is zero.
β_1 is the rate of change of deaths when TB disease cases, changes by one unit.
E_i is the error term ~NID(0, σ^2).

Secondly: For the data of year 2011, equations 1 and 2 are used in fitting regression models of death from TB disease in cases of TB.

Thirdly: For the mean of combined data from the years 2010 and 2011:

In addition to the equations 1 and 2 that mentioned above, equations 10, 11 and 12 are used

$Y = \beta_0 + \beta_1 (1/X) + E_i$ .....(10)

$\text{Log } Y = \log \beta_0 + \beta_1 \log X + E_i$ .....(11)

$\text{Log } Y = \beta_0 + \beta_1 X + E_i$ .....(12)

where: Y is the mean of the number of deaths that occurred by TB disease.
X is the mean of the number of cases of TB disease.
A is the number of deaths when TB disease is zero.

By using an ordinary least square method (OLSM), the estimated value of β_0 and β_1 can be obtained as " $\beta_0^{\wedge}$" and " $\beta_1^{\wedge}$" respectively. The analysis of variance, test of autocorrelation, t-test of the models coefficients were used.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Data in appendix 1 was analyzed by SPSS program. We have descriptive statistics is shown in table 1, includes, Mean, Std. Error of Mean, Median, Mode, Std. Deviation, Variance, Skewness, Std. Error of Skewness, Std. Error of Skewness, Kurtosis, Std. Error of Kurtosis, Range, Minimum, Maximum, Sum and Percentiles

Table (1): Descriptive Data of number of deaths by TB and number of cases of TB in 2010-2011

		No. of Death By TB 2010	No. of TB Incidence 2010	No. of Death By TB 2011	No. of TB Incidence 2011	Combined No. of Death By TB 2010 and 2011	Combined No. of TB Incidence 2010 and 2011	Log Mean Combined No. of Death By TB 2010 and 2011	Log Mean Combined No. of TB Incidence 2010 and 2011
N	Valid	15	15	15	15	15	15	15	15
	Missing	0	0	0	0	0	0	0	0
Mean		36.6000	621.3333	32.9333	550.2667	69.5333	1171.6000	0.990328	2.22581
Std. Error of Mean		23.15575	397.38123	25.41306	359.99778	48.51989	757.31257	0.14842	0.15864
Median		10.0000	137.0000	7.0000	97.0000	16.0000	211.0000	0.90309	2.02325
Mode		5.00[a]	19.00[a]	.00[a]	82.00	4.00	36.00[a]	0.30103	1.25527[a]
Std. Deviation		89.68182	1539.05090	98.42435	1394.26542	187.91672	2933.05895	0.57483	0.61440
Variance		8042.829	2368677.667	9687.352	1943976.067	35312.695	8602834.829	0.330	0.377
Skewness		3.762	3.747	3.846	3.737	3.819	3.744	1.399	0.940
Std. Error of Skewness		.580	.580	.580	.580	.580	.580	0.580	0.580
Kurtosis		14.372	14.281	14.852	14.228	14.695	14.262	3.229	1.728
Std. Error of Kurtosis		1.121	1.121	1.121	1.121	1.121	1.121	1.121	1.121
Range		356.00	6111.00	388.00	5520.00	742.00	11631.00	2.27068	2.51066
Minimum		2.00	19.00	.00	17.00	4.00	36.00	0.30103	1.25527
Maximum		358.00	6130.00	388.00	5537.00	746.00	11667.00	2.57171	3.76593
Sum		549.00	9320.00	494.00	8254.00	1043.00	17574.00	14.85493	33.38711
Percentiles	25	5.0000	76.0000	1.0000	50.0000	8.0000	153.0000	0.60206	1.88366
	50	10.0000	137.0000	7.0000	97.0000	16.0000	211.0000	0.90309	2.023253
	75	27.0000	368.0000	12.0000	327.0000	33.0000	695.0000	1.21748	2.540955
a. Multiple modes exist. The smallest value is shown									

Table 2 shows models summary consists of R, R^2, Adjusted R^2 and Durbin –Watson Statistics for the data of Number of Deaths by TB (Y) and Number of Cases of TB(X) in 2010 and 2011.

Table (2): Models Summaryb of Number of Deaths from TB (Y) and Number of Cases of TB(X) in 2010-2011, Including R, R^2, Adjusted R^2 and Durbin –Watson Statistics

YEAR	Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin -Watson
2010	$Y = \beta_0^{\wedge} + \beta_1^{\wedge} X$	.991[a]	.983	.981	12.23441	1.998
	$Y = \beta_0^{\wedge} + \beta_1^{\wedge} X + \beta_2^{\wedge} x^2$	.994[a]	.987	.985	10.92196	1.953
2011	$Y = \beta_0^{\wedge} + \beta_1^{\wedge} X$	.994[a]	.987	.986	11.48727	2.398
	$Y = \beta_0^{\wedge} + \beta_1^{\wedge} X + \beta_2^{\wedge} x^2$	.999[a]	.998	.997	5.26575	1.413
Mean Combined 2010 And 2011	$Y = \beta_0^{\wedge} + \beta_1^{\wedge} X$	.994[a]	.987	.986	22.07588	2.156
	$Y = \beta_0^{\wedge} + \beta_1^{\wedge} X + \beta_2^{\wedge} x^2$	.997[a]	.994	.993	15.15060	1.780
	$Y = \beta_0^{\wedge} + \beta_1^{\wedge} (1/X)$	.271[a]	.073	.002	187.71766	2.182
	$\text{Log } Y = \log \beta_0^{\wedge} + \beta_1^{\wedge} \log X$	.901[a]	.812	.797	0.25895	1.885
	$\text{Log } Y = \beta_0^{\wedge} + \beta_1^{\wedge} X$	.811[a]	.657	.631	.80400	1.921
	a. Predictors: ($\beta_0^{\wedge}$ = Constant), X = TB.INCIDENCE.					
	b. Dependent Variable: Y = TB.DEATH.					

Table 3 shows analysis of variance of the studied. The table consists year to which the data belongs, Models, Components of Sum of Squares, Sum of Squares, Degrees of Freedom, Mean Squares, Calculated Fs and P-values.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Table (3): Analysis of Variance (ANOVAa) of the Fitted Models of number of deaths from TB (Y) and number of cases of TB(X) in 2010-2011

YEAR	Model	Sum of Squ Component	Sum of Squares	df	Mean Square	F	Sig.
2010	$Y = \beta_0^{\wedge} + \beta_1^{\wedge} X$	Regression	110653.749	1	110653.749	739.265	0. 000[b]
		Residual	1945.851	13	149.681		
		Total	112599.600	14			
	$Y = \beta_0^{\wedge} + \beta_1^{\wedge} X + \beta_2^{\wedge} x^2$	Regression	111168.129	2	55584.064	465.960	.000[b]
		Residual	1431.471	12	119.289		
		Total	112599.600	14			
2011	$Y = \beta_0^{\wedge} + \beta_1^{\wedge} X$	Regression	133907.487	1	133907.487	1014.778	.000
		Residual	1715.447	13	131.957		
		Total	135622.933	14			
	$Y = \beta_0^{\wedge} + \beta_1^{\wedge} X + \beta_2^{\wedge} x^2$	Regression	135290.195	2	67645.098	2439.581	.000
		Residual	332.738	12	27.728		
		Total	135622.933	14			
Mean Combined 2010 And 2011	$Y = \beta_0^{\wedge} + \beta_1^{\wedge} X$	Regression	488042.254	1	488042.254	1001.432	.000
		Residual	6335.479	13	487.345		
		Total	494377.733	14			
	$Y = \beta_0^{\wedge} + \beta_1^{\wedge} X + \beta_2^{\wedge} x^2$	Regression	491623.244	2	245811.622	1070.884	.000[b]
		Residual	2754.489	12	229.541		
		Total	494377.733	14			
	$Y=\beta_0^{\wedge}+ \beta_1^{\wedge} (1/ X)$	Regression	36284.782	1	36284.782	1.030	.329[b]
		Residual	458092.951	13	35237.919		
		Total	494377.733	14			
	$Log\ Y = log\ \beta_0^{\wedge} + \beta_1^{\wedge} log\ X$	Regression	3.754	1	3.754	55.988	.000[b]
		Residual	0.872	13	0.067		
		Total	4.626	14			
	$Log\ Y = \beta_0^{\wedge} + \beta_1^{\wedge} X$	Regression	16.124	1	16.124	24.943	.000[b]
		Residual	8.403	13	.646		
		Total	24.527	14			

a. Predictors: (a = Constant), . Independent Variable: X = TB.INCIDENCE.

b. Dependent Variable: Y = TB.DEATH.

Table 4 shows Estimating and Testing the Significance of the Coefficients of the Fitted Models of number of deaths by TB (Y) and number of cases of TB(X) in 2010-2011. The table contains nine columns, include year to which the data belongs, Dependent Varable, **Independent Variable,** Unstandardized Coefficients**,** Standardized Coefficients, Std. Error, t-statistic, P-values and 95.0% Confidence Interval for B.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Table (4): Estimating and Testing the Significance of the Coefficients of the Fitted Models of number of deaths by TB (Y) and number of cases of TB(X) in 2010-2011

YEAR	Dependent Varable	**Independent Variable**	Unstandardized **Coefficients**[a] B	Standardized Coefficients Std. Error	Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	Upper Bound
2010	Y	Constant	$\beta_0^{\wedge}$ =0.709	3.424		0.207	0.839	-6.688-	8.105
		X	$\beta_1^{\wedge}$ =0.058	0.002	0.991	27.189	0.000	0.053	0.062
	Y	Constant	$\beta_0^{\wedge}$ =7.217	4.378		1.649	0.125	-2.321-	16.755
		X	$\beta_1^{\wedge}$ =0.026	0.015	0.447	1.693	0.116	-0.007-	0.060
		X^2	$\beta_2^{\wedge}$ =5.082E-6	0.000	0.548	2.077	0.060	0.000	0.000
2011	Y	Constant	$\beta_0^{\wedge}$ = -5.665-	3.204		-1.768-	0.100	-12.586-	1.257
		X	$\beta_1^{\wedge}$ =0.07	0.002	0.994	31.856	0.000	0.065	0.075
	Y	Constant	$\beta_0^{\wedge}$ =3.854	1.994		1.933	0.077	-0.489-	8.198
		X	$\beta_1^{\wedge}$ =0.015	0.008	0.216	1.941	0.076	-0.002-	0.032
		X^2	$\beta_2^{\wedge}$ =9.778E-6	0.000	0.785	7.062	0.000	0.000	0.000
Mean Combined 2010 And 2011	Y	Constant	$\beta_0^{\wedge}$ =-5.047-	6.168		-.818-	0.428	-18.372-	8.278
		X	$\beta_1^{\wedge}$=0.064	0.002	0.994	31.645	0.000	0.059	0.068
	Y	Constant	$\beta_0^{\wedge}$ =11.313	5.922		1.910	0.080	-1.591-	24.217
		X	$\beta_1^{\wedge}$=0.02	0.011	0.319	1.857	0.088	-0.004-	0.044
		X^2	$\beta_2^{\wedge}$ =3.642E-6	0.000	0.679	3.950	0.002	0.000	0.000
	Y	Constant	$\beta_0^{\wedge}$ =111.182	63.512		1.751	0.104	-26.027-	248.391
		(1/X)	$\beta_1^{\wedge}$=-7023.73-	6921.669	-0.271-	-1.015-	0.329	-21977.088-	7929.624
	Log Y	Constant	Log$\beta_0^{\wedge}$=-.886-	0.259		-3.413-	.005	-1.446-	-0.325-
		Log X	$\beta_1^{\wedge}$ =0.843	0.113	.901	7.483	.000	0.600	1.086
	Log Y	Constant	$\beta_0^{\wedge}$ =2.545	0.225		11.328	.000	2.059	3.030
		X	$\beta_1^{\wedge}$ =0.000	0.000	0.811	4.994	.000	.000	.001
a. Predictors: (a = Constant), X = TB.INCIDENCE.									
b. Dependent Variable: Y = TB.DEATH.									
c. The Cubic model could not be fitted due to near-collinearity among model terms.									

Figure 1, illustrate the scatter (dots) of the transformed data (Log X, Log Y) and the straight line of the best fitted model: Log Y = -0.886 + 0.843 Log X

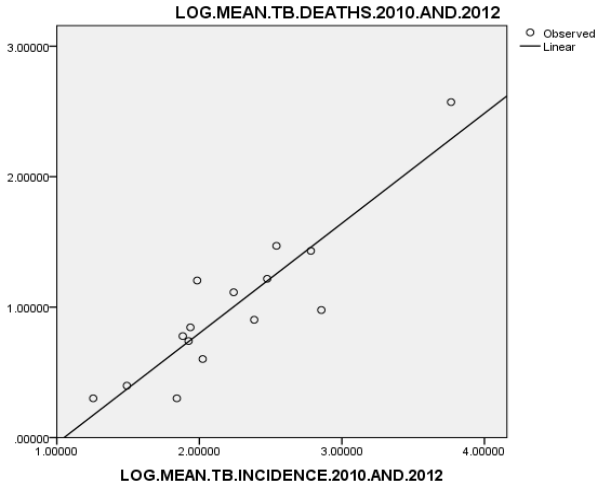

Figure (1): Scatter of the transformed data (Log X, Log Y) and the straight line Log Y = -0.886 + 0.843 Log X

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Figure 2, shows that there is no correlation between the residuals and the dependent variable Log Y.

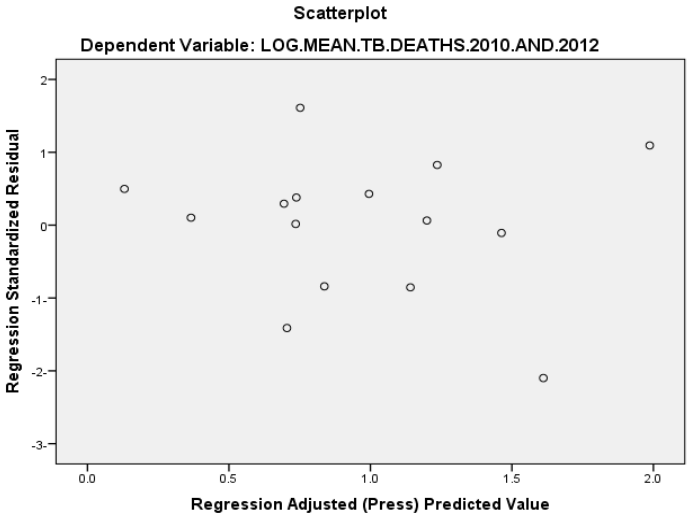

Figure (2): Scatter of residuals against the dependent variable Log Y.

RESULTS

According to the skewness and kurtosis coefficients, which are shown in table 1, the best model is the logarithmic model: Log Y = Log a + b Log X, because its data is nearly normally distributed. From the summary that was shown in table 2, the model $Y = a + b X + Cx^2$, has DW = 1.413, that is far from 2.0, so there is positive autocorrelation in this model (This is with respect to the data of the year 2011). From the same table 2, for the other models DW statistics are near to equate 2.0, therefore there are no autocorrelations. Also from table 2, with respect to the model Y = a + b (1/ X), the adjusted R^2 = 0.002 which is approach to zero, therefore this model cannot be able to estimate the number of deaths that may occur by TB. The other models have no problems at adjusted R^2 values. According to the values of the adjusted R^2 that are shown in table 2, 63.1% to 99.7% of the changes that occurred in the number of the deaths return to TB disease. From table 3 all the fitted models have P- values = 0.000 with F calculated values greater than 24 accept the model Y = a + b (1/ X), which has a p-value = 0.329 with F = 1.03, so this model is insignificant. From table 4, most of the coefficients of the fitted models are insignificant, accept of the models Y = a + b X, Log Y = Log a + b Log X and Log Y = a + b X are significant. Finally, from the results mentioned above, the model: Log Y = -0.886 + 0.843 Log X is the best model for fitting the studied combined data.

DISCUSSIONS

In consequence of the above mentioned results, the following points discussed:
To conduct similar studies in each state.
To conduct similar studies by using time series data to fit models.
To take the advantages of this study in the planning and improvement of Health status.

ACKNOWLEDGEMENT

Foundation item: Ministry of Health in Sudan. Author is grateful to MOH at Sudan for data support to carry out this work

REFERENCES

1. Chiang, C.L, (2003) Statistical methods of analysis, World Scientific. ISBN 981-238-310-7 - page 274 section 9.7.4 "interpolation v's extrapolation"
2. Cook, R. Dennis; Sanford Weisberg Criticism and Influence Analysis in Regression,Sociological Methodology, Vol. 13. (1982), pp. 313–361

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

3. Cressie, N. (1996) Change of Support and the Modiable Areal Unit Problem. Geographical Systems 3:159–180.
4. Eltilib et al; 2010. Management of TB in the private sector in Khartoum, Sudan: quality and impact on TB control, Sudan JMS Vol. 5, no. 1, Mar 2010.
5. Fotheringham, A. Stewart; Brunsdon, Chris; Charlton, Martin (2002). Geographically weighted regression: the analysis of spatially varying relationships (Reprint ed.). Chichester, England: John Wiley. ISBN 978-0-471-49616-8.
6. Fotheringham, AS; Wong, DWS (1 January 1991). "The modifiable areal unit problem in multivariate statistical analysis". Environment and Planning A 23 (7): 1025–1044. doi:10.1068/a231025.
7. Freedman, David A., Statistical Models: Theory and Practice, Cambridge University Press (2005)
8. Konstantinos A ; 2010. "Testing for tuberculosis". Australian Prescriber 33 (1): 12–18.
9. Kumar V, Abbas AK, Fausto N, Mitchell RN (2007). Robbins Basic Pathology (8th ed.). Saunders Elsevier. pp. 516–522. ISBN 978-1-4160-2973-1.
10. Kutner, M. H., C. J. Nachtsheim, and J. Neter (2004), "Applied Linear Regression Models", 4th ed., McGraw-Hill/Irwin, Boston (p. 25)
11. Ravishankar, N. and Dey ,D. K. (2002), "A First Course in Linear Model Theory", Chapman and Hall/CRC, Boca Raton (p. 101)
12. Steel, R.G.D, and Torrie, J. H., Principles and Procedures of Statistics with Special Reference to the Biological Sciences., McGraw Hill, 1960, page 288.

APPENDIX:

Appendix (1): TB cases and deaths in the states the Sudan in 2010 and 2011

	No. of Death 2010	No. of Incidence 2010	No. of Death 2011	No. of Incidence 2011	Combined No. of Death 2010 and 2011	Combined No. of Incidence 2010 and 2011
NORTH	3	19	1	17	2	18
RIVER.NILE	2	129	6	82	4	105.5
RED.SEA	9	247	7	238	8	242.5
GADAREF	37	368	22	327	29.5	347.5
KASSALA	21	318	12	280	16.5	299
KHARTOUM	358	6130	388	5537	373	5833.5
GAZERA	37	611	17	600	27	605.5
SINNAR	5	76	9	97	7	86.5
WHITE.NILE	15	171	11	177	13	174
BLUE.NILE	10	137	1	31	5.5	84
N.KORDOFAN	10	775	9	660	9.5	717.5
S.KORDOFAN	4	89	00	50	2	69.5
W.DARFUR	27	135	5	58	16	96.5
S.DARFUR	5	44	00	18	2.5	31
N.DARFUR	6	71	6	82	6	76.5
Source: Health Annual Statistical Report (2010-2011), Ministry of health, Sudan.						