American Academic Scientific Research Journal for Engineering, Technology, and Sciences

ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

https://asrjetsjournal.org/index.php/American_Scientific_Journal/index

Mathematical Analysis of Regression Model Epidemiology

Prof. Dr. Sumit Kumar Banerjee[a*], Mr. Boaz Andrews[b]

[a,b]Papua New Guinea University of Technology, Department of Mathematics & Computer Science, PMB,
Morobe Province, Lae 411, Papua New Guinea
[a]Email: sumit.banerjee@pnguot.ac.pg
[b]Email: boaz.andrews@pnguot.ac.pg

Abstract

Statistical modeling techniques, specifically regression line analysis have become important analytical tools and are contributing immensely to the field of epidemiology. However, many users do not understand their effective use and applications. Underlying epidemiological concepts and not the statistics should govern or justify the proper use and application of any modeling exercise. Main utility of the regression line analysis lies in its ability to provide a general but practical conceptual framework for casual problems, explaining and evaluating the role of biases, confounders and effect modifiers. Successful modeling of complex data is a part science, part statistics and part experience, but the major part is logic or common sense. Findings of this research article focuses on the contributions of regression analysis towards the pedagogical study of epidemiological models by enhancing the research process and serving as an effective tool for communicating findings to public health managers and policymakers and fostering interdisciplinary collaboration.

Keywords: Epidemiology; Regression Line; Logic; Public health; Policymakers.

1.0 Introduction

Use of statistical models has been dominating the analysis strategies in epidemiological research. Goal of analysis is epidemiological studies should be to quantify information in an objective and defensible manner, testing out the role of chance and bias, primary analysis of epidemiological studies is crude- that is without accounting for the role of other factors, whereas secondary analysis is refined and is multi-factorial in nature, wherein the main interest lies in studying role of confounders and/or effect modifiers [1, 2]. A confounder is a covariate that is associated with both the outcome variable and a risk factor or predictor.

Received: 3/5/2023
Published: 4/25/2023

----

* Corresponding author.

An effect modifier is a covariate that interacts with a risk factor and produces different effects for different levels of the covariate [3]. Fitting a series of univariate models will rarely provide an adequate analysis of the data in any study since the independent variables (predictors or risk factors) are usually associated with each other and may have different distributions with the levels of the outcome variable [4,5].

Thus one generally considers a multivariate analysis for a more comprehensive modeling of the data [6]. The term 'multivariable' is preferable to 'multivariate' in epidemiological research which usually focuses on assessing the relationship between a single dependent (outcome) variable and multiple independent (explanatory) variables [7]. Statisticians generally use the term 'multivariable analysis' to describe a method in which several dependent variables can be considered simultaneously. Researchers in biomedical and health sciences who are not statisticians however, use this term to describe any statistical technique involving several variables, even if only one dependent variable is considered at a time. Regression analysis is one type of multivariable technique commonly used in epidemiological research to assess relationships amongst a set of variables. A well-known application of multivariate regression model to Framingham Heart study data by Trutt and colleagues (1967) demonstrated full power and broad applicability of these techniques to the field of biology and medicine [8, 9, 10]. Researchers think that analysis done in an epidemiological study is incomplete without the use of some kind of model [11]. However, there is no dearth of examples of inappropriate uses of these techniques. For example, some researchers have used linear regression models in a situation where non-linear regression models should have been used or vice-versa (Gail and colleagues 1984) [12, 13]. The assumptions required in modeling are often not met. Nonetheless, statistical modeling and other statistical methods have become important analytical tools in the field of epidemiology [14, 15]. Here we want to emphasize that the underlying epidemiological concepts along with the appropriate statistical methodology should be used in any model building process.

1.1 Role of regression model in epidemiology

If certain statistical assumptions are plausible, then epidemiological models can be used to make inferences from a study population to a larger target population. Statistical modeling in epidemiology provides understanding of the underlying mechanisms that influence the spread of disease and in the process, it suggests control strategies. A statistician may choose to start with a simple but incomplete model to obtain quantitative information, while an epidemiologist may think that the model is too simplistic, sketchy and omits important aspects of cause and effect associations, or links in disease occurrence, or disease progression and transmission. In another contrasting situation, the statistician may suggest a very complex model, while epidemiologists may find it too quantitative, difficult to apply and analyze. The main utility of the statistical model lies in its ability to provide a general but practical conceptual framework for causal problems. For example, a health manager might be interested in planning strategies for response to and management of a disease outbreak in a community. The data available to him are in the form of surveillance reports. But he finds that reported data are grossly inaccurate and incomplete. In order to describe or predict the course of future disease outbreak, formulation of a statistical model may be the only way available with him to work out and compare the effects of different management strategies.

The mathematical modeling of epidemics has been the objective of a vast number of studies over the past century. Given the importance of epidemics for life on Earth in general, it is not in the least astonishing that the desire to understand their mechanism has led to formulation of models which make possible the simulation of events for which laboratory experiments cannot be conducted easily. The reason we have chosen the SIR model in our research is that there is not enough evidence that the patient might not be immune to the disease. Prominent among the mathematical models of epidemics, the great historical importance is the susceptible-infected-removed (SIR) model initially proposed by Kermack and Mckendrick in 1991. The model has been defined with three groups of healthy people who are susceptible (S), infected individuals (I), removed individuals either by them being recovered and immunized or by their death ®. Since the number of susceptible, infected and recovered people may fluctuate over the time, the SIR model is dynamic. Flowing from susceptible to infected and then recovered could be showed in the below figure.

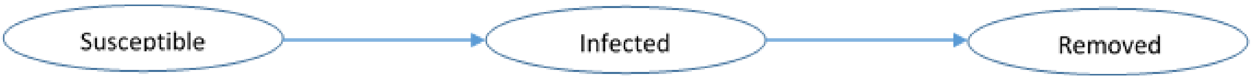

Figure 1

In this model, the infection rate β which is the probability of transmitting disease between a susceptible and an infectious individual. γ is the recovery rate. N is defined as population and is equal to $N = S + I + R$. We can write the SIR model as the following differential equation:

$$\left.\begin{aligned} \frac{dS}{dt} &= -\frac{\beta SI}{N} \\ \frac{dI}{dt} &= \frac{\beta SI}{N} - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned}\right\} - - - - - - - - - - - - - - -(1)$$

To model the relationship between the response and the explanatory variable we may use linear regression. Simple linear regression is a model with a single regressor 'x' that corresponds to a response variable 'y'. Simple linear regression can be formulated from the above differential equation model (1) as follows:

$$Y = \beta_0 + \beta_1 X + E - - - - - - - - - - -(2)$$

There is a basic difference between a mathematical model and a statistical model. The equation $Y = \beta_0 + \beta_1 X$ represents a mathematical model. In a mathematical model, there is no random error term in the equation, whereas the regression line model includes the random error term. The simplest example of a regression model is a "straight line model" which can be described by the following equations:

$$\left.\begin{aligned}\mu_{y/x} &= \beta_0 + \beta_1 X \quad -----(3)\\ Y &= \beta_0 + \beta_1 X + E -----(4)\end{aligned}\right\}$$

Where β_0 is an intercept, β_1 is the slope of the straight line (for a given population) and E is the random error component. The above two equations (3) and (4) describe a statistical model to express the relationship between a dependent variable (Y) and an independent variable (X) when both X and Y are continuous variables. These statistical models can be used to predict a mean value of Y for a given value of X (i.e. $\mu_{y/x}$) in the population. A line that fits data well ought to have small errors or deviations between what is observed and what is predicted by the fitted model. The principal objective of many epidemiologic studies is to evaluate association between exposure to a single risk factor or a cause and the occurrence of a specific disease or a disorder. In this type of investigation, it is essential to isolate the effect of interest from the effects of other risk factors or covariates. This can be handled by using a multiple regression model which elucidates the relationship between a single outcome variable and a set of independent predictors.

The effect of other factors that explain or produce confounding can either be controlled in the design stage or analysis stage of the research exercise. The distortion due to confounder if large can not only lead to underestimation or over-estimation of the effect but can altogether change the apparent direction of an observed effect; e.g. a truly beneficial effect might be considered as a harmful one or vice-versa. This concept of confounding, which is a central one in epidemiology can be effectively addressed; either on ad-hoc basis i.e. at the planning stage of the study or on post-hoc basis i.e. at the analysis stage of the research exercise.

Randomization (only for experimental designs), restriction and matching (for all study designs) are the logical approaches available to control the confounders at the study design stage. Two common approaches to the analytical control of the covariates and effect modifiers are 'stratification' and 'statistical modeling'.

2.0 Dominance of modeling over stratification

Stratification was the predominant approach to the analysis of epidemiologic studies in the past because of its simplicity, clarity and minimal statistical assumptions.

Prime analytic concern that motivates the epidemiologists to use stratification is in the evaluation and control of confounding. Though simple, this approach has several inherent limitations: (i) Stratification upon even a modest number of covariates may result in the allocation of only a few subjects to individual strata. (ii) Estimation of the main association of interest becomes imprecise and unreliable. (iii) It may be difficult to interpret the resulting pattern of effect estimates across several strata and (iv) Stratified analysis is based on a non-parametric approach. Therefore analysis of stratified data though simpler but is often difficult to apply in epidemiologic research.

On the contrary analysis of multivariate data though looks much more complex, but is easier to analyze and apply in actual practice.

The use of statistical modeling in epidemiologic research can overcome many of the limitations of other analytic methods. In contrast to stratification, statistical modeling can easily accommodate several covariates besides having the ability to smoothen out or dampen variation attributable to unimportant factors, capacity to incorporate continuous independent variables and ability to ascertain interactions between two or more variables.

3.0 Models with confounding and interaction

Modeling helps the researcher in explaining and evaluating the role of confounders as well as effect modifiers. Determining if a covariate is a confounder or effect modifier involves several issues. In practice, the confounder status of a covariate is ascertained by comparing the estimates of effect measure for a risk factor from two models- one containing the covariate and the other not containing the covariate. Any biologically important change in the estimated effect measure for the hypothesized risk factor as indicated by such a comparison would dictate the decision to include or not to include that covariate in the statistical model, regardless of the statistical significance of the estimates of the effect measure. If the covariate is an effect modifier then an interaction term is added to the model only when it is found to be statistically significant and biologically meaningful. Some sensitivity analysis is to be done to assess presence or absence of interactions. Firstly, if there is no interaction, the estimates of effect measures between a covariate and outcome variable will remain same within each level of the risk factor. Visual effect of interaction or no interaction can be easily accessed via graphical presentations (see Fig. 1 and Fig. 2). For example, a plot of estimated effect measures would show parallel lines if no interaction exists. Generally a model with interaction would include second or higher order terms involving two or more variables. If a covariate is an effect modifier, then knowing its status as a confounder will be of secondary importance to the epidemiologist.

4.0 Modeling process

The entire process of modeling epidemiologic data becomes transparent, coherent and more meaningful when it is conducted through a logical sequence of decisions. Complete analysis of epidemiologic data may involve examination of relationships between an outcome variable and a set of covariates with more than one candidate model. Thus through analysis of a set of epidemiologic data should not be limited only to a single model form. In the statistical modeling of disease transmission, there is a trade-off between simple models, which omit most details and are designed only to highlight general qualitative behavior; and detailed models usually designed for specific situations including short-term quantitative predictions. Detailed models are generally difficult or impossible to solve analytically and hence their usefulness for theoretical purposes is limited, although their strategic value may be high in actual practice. Successful modeling of complex data is part science, part statistical methods, part experience but the major part is logic or common sense.

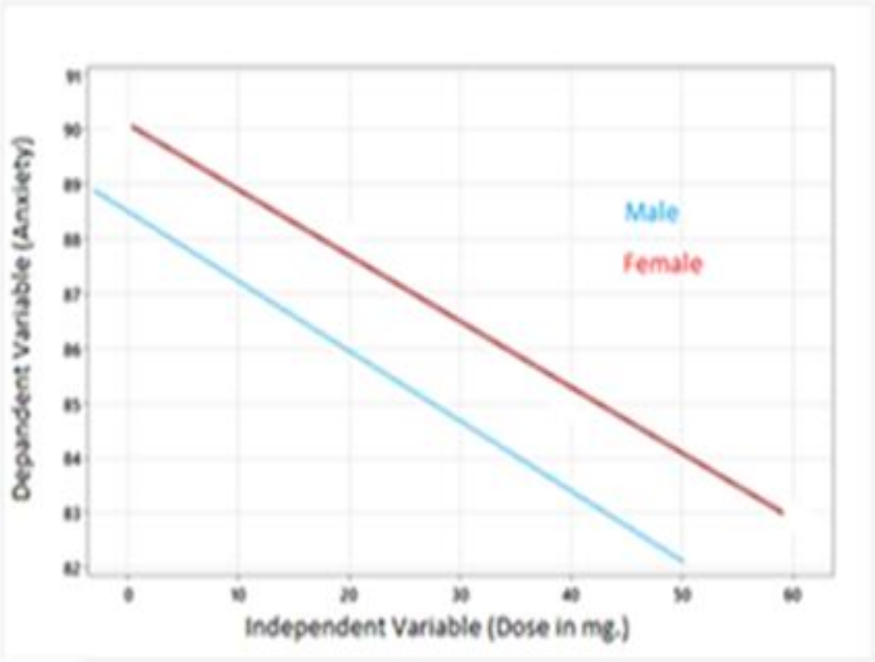

Figure 2: No interaction between independent variable and gender *Data have been collected from a survey made on male and female staff members of MCS Dept. of Unitech, PNG, based on doses of medicine they have consumed versus their corresponding anxiety level.

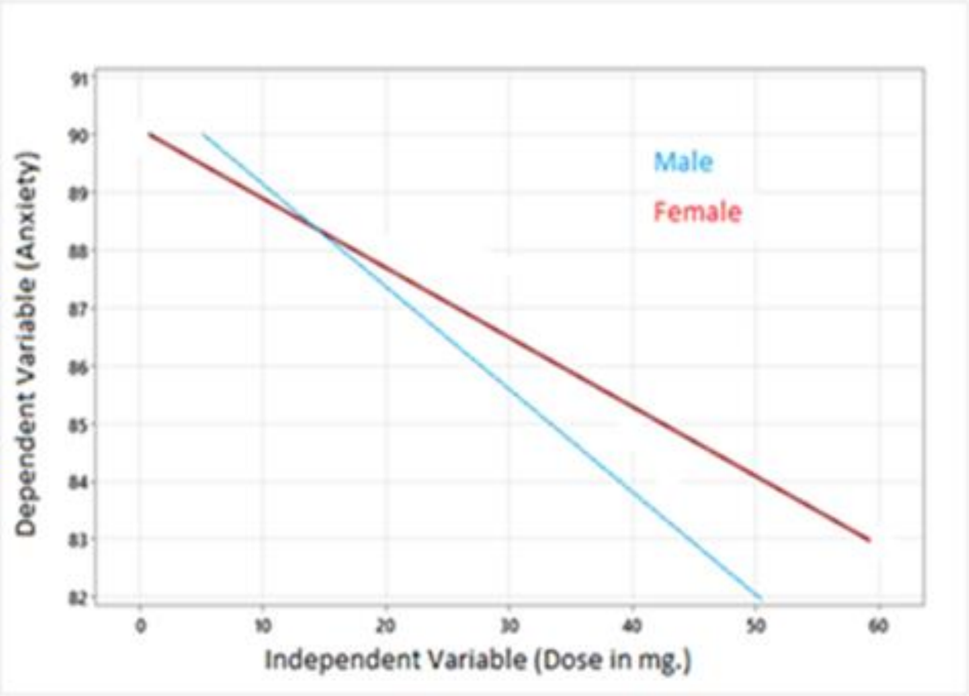

Figure 3: interaction between independent variable and gender **Data have been collected from a survey

made on male and female staff members of MCS Dept. of Unitech, PNG, based on doses of medicine(changed) they have consumed versus their corresponding anxiety level.

5.0 Regression models

Regression analysis is one type of modeling technique used in epidemiologic research to assess relationships amongst a set of variables. Objective is to predict response or outcome from a set of explanatory variables. Sometimes interest lies in simply describing the structure of variation in a set of variables for which there are no obvious outcome or predictor variables.

A variety of regression models are available including linear and non-linear forms, namely multiple linear regression, multiple logistic regression, Log-linear models, Generalized Linear models and Generalized estimating equations. Some of these models follow a parametric approach and some others the non-parametric approach for statistical inference. Some models use basic designs while others might use hybrid designs. Some models deal only with a dichotomous outcome variable, some others might deal with polychromos-nominal or ordinal outcome variables and some others might even deal with continuous outcome variables. Of these, Logistic Regression is the most widely utilized parametric approach currently in epidemiologic research, particularly when the outcome variable is dichotomous. The theory and appropriate use of these models require some familiarity with the regression techniques.

6.0 Model-building strategies

Choice of model, variable selection, design parameterization, model fitting, adequacy and goodness of model fit, step-wise inclusion or exclusion of predictors, model diagnostics assessment of reliability and validity, testing hypothesis about model parameters etc. are some of the important issues in modeling exercise.

If epidemiologists wish to obtain the best model within the scientific context and objective of the study, then he/she needs to first select all the relevant hypothesized variables in the model. Next step will be to conduct univariate analyses independently for each selected variable. Then fit a full model with only those variables with p-value < 0.25 in the univariate analyses plus some others with known biological importance for multivariate analysis. A detailed discussion on the choice of an alpha level for the model containing more variables is available in a text book on “Applied Logistic Regression” by Hosmer and Lameshow (1989, John Wiley & Sons). Bendel and Afifi (1977) and Constanza and Afifi (1979) studied the choice of p_E (‘E’ stands for entry) value for stepwise linear regression, discriminant analysis and logistic regression and they have shown that fixing p_E value = 0.05 is too stringent, often excluding important variables from the model. The authors have recommended choosing p_E in the range of 0.15-0.20. Sometimes the goal of the analysis may be broader and models containing more variables are sought to provide a more complete picture of possible models. In those cases use of p_E =0.25 might be a reasonable choice. After fitting a multivariate model, the importance of each variable entered in the model should be verified. Variables that do not contribute to the model as per criterion of the "Best fit" should be eliminated and a new reduced model be fitted. Once we get the model that we feel contains essential variables we should consider the need to include interaction terms among the variables. After assessing the significance of the interaction terms, we decide to retain or exclude them from the final model on statistical, practical and biological perspectives.

7.0 Choice of Model

Choice of a specific model depends on type of the outcome variable (dichotomous, polytomous or continuous) adequacy of sample size, the sampling process, constraints of statistical procedures and validity of certain underlying assumptions apart from the context and objectives of research.

8.0 Additive versus multiplicative models

In some situations an 'additive model' is appropriate whereas in some other circumstances a 'multiplicative model' is preferred. For example, when several etiologic agents act interchangeably at a single step of a multistage pathway an additive model is implied. In contrast when etiologic agents act at different steps, a multiplicative model may be obtained. Usually for effect measures expressed on interval scale additive models are used, whereas multiplicative models are used for effect measures expressed on ratio scale.

9.0 Association versus prediction models in epidemiology

There are different types of models in epidemiology requiring different assumptions and mathematical techniques. Basically an epidemiologist needs to choose between two types of models that are commonly used: an 'Association model" and a "prediction model'.

An 'Association model' is how most epidemiologists usually think of regression: they identify an exposure variable of interest (say X) and look at its relationship with the outcome variable (say Y). Other independent variable (say Z_1, Z_2, Z_3..........) are simply moderators of the effect of X or not. If yes, the interaction term and the main effect are included in the model. If not, we ignore the interaction term and further evaluate if the independent variable is a confounder or not. Accordingly we decide whether the main effect is to be included or not to be included in the final or reduced model.

In the 'prediction model' we want a model that fits best to the given data and explains the outcome variable (Y). All possible independent variables (say X_1, X_2, X_3..........) interactions (say $X_1 * X_2, X_1 * X_3, X_2 * X_3$..........) transformations (logarithmic, exponential, quadratic… etc.) can be tried with these models. The models are then evaluated for goodness of fit, R-square or predictive ability using different validation techniques called regression diagnostics.

10.0 Regression Diagnostics

Once we have applied a model we need to assess how well it fits the data or how close the model predicted values are to the corresponding observed values. Deviations of predicted values from observed values should be essentially normal. The test statistics that assess model fit in this manner are known as "Goodness-of-fit statistics'. We expect that if the model fits the data adequately, the test statistic should be non-significant. If this statistics is larger than a tolerable value then we might be fitting an over simplified model. Hence we need to identify some other factors which can explain the observed variation in the data in a better manner. While goodness-of-fit statistics tell us how well a particular model fits the data but they tell us a little about the lack of fit or why and where a particular model fails to fit the data. Some deviance statistics have been devised as a function of observed values and their model predicted values. By taking a look at the individual components of these statistics one can gain insight into a model's lack of fit. Graphical display of residual errors (difference between predicted and observed values scaled in units of standard deviation of observed values) can tell us if there are unusually large errors (possibly indicative of outliers) or systematic pattern of variation (possibly indicative of a poor model choice). Regression diagnostics primarily focus on methods for analyzing residuals, assessing the influence of outliers and assessing problems of collinearity (i.e. inter-relationship between two independent variables or predictors).

Researchers should understand that the selected mathematical models need not always be able to explain all the variation in the observed data. Statistical modeling techniques using maximum likelihood estimation (MLE) or weighted least squares (WLS) are often employed to describe variation in terms of a parsimonious model. In some situations the choice of the model is governed by the statistical considerations, in some others it is governed by the epidemiologic understanding or the public health impact point and in yet another it is based upon the hypothesized underlying biologic process.

11.0 Conclusion

Successful modeling of complex data is part science, part mathematics, part statistical methods, part experience but the major part is logic or common sense. Fundamental concepts of epidemiology do not depend on empirical results, hence considerable importance should be placed on effective modeling strategies in epidemiologic research. It should be understood clearly that the goal of any model building technique is to find the best fitting and most parsimonious, yet biologically reasonable model that will largely describe relationship between an outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables teasing out the role of chance and bias. Model may help explain observed variation in the data but some part of the relationship may remain unexplained although the best techniques and the most relevant variables have been used by the researcher for modeling purposes.

Statistical modeling has the potential to make significant contributions to the field of epidemiology by enhancing the research process, helping the researchers in filling the gaps and explaining variations in the observed phenomenon, serving as an effective tool for communicating findings to public health managers and policymakers and fostering interdisciplinary collaboration. With availability of some good statistical and epidemiological software packages the task of fitting the multivariate models and evaluating the role of confounders and effect modifiers have become much easier. Mathematical models are now becoming accessible to a wider audience of clinical and epidemiological researchers and therefore their application is likely to be seen more often in epidemiology, public health and clinical research and practice.

References

[1]. Fred Brauer, Pauline van den Driessche and J Wu, “Mathematical Epidemiology”, Springer Book Archives, vol. 2, pp. 79-91, 2008.

[2]. Mikayla Chubb and Kathryn H Jacobsen, “Mathematical modeling and the epidemiological research process”, Eur. J Epidemiol, vol. 25, pp. 13-19, 2010.

[3]. Mitchell H Gail, H S Wieand and Steven Piantadosi, “Biased estimates of treatment effects in randomized experiments with non-linear regressions and omitted covariates”, Biometrika, vol. 71, pp. 431-444, 2012.

[4]. Walter W Holland, Roger Detels and George Knox, “Methods of public Health”, Oxford University press, vol. 2, pp. 291-311, 2009.

[5]. David W Hosmer and Stanley Lemeshow, “Applied Logistic Regression”, NY:John Wiley & Sons, vol. 2, pp. 67-85, 2011.

[6]. Stephen B Hulley and Steven R Cummings, “Designing Clinical Research”, Williams and Wilkins, vol. 7, pp. 131-145, 2017.

[7]. Sharon Kleinbaum, Leo Kupper and Thomas Muller, “Applied Regression Analysis and Multivariate Methods”, CA: Brooks/Cole Publishing company, vol. 3, pp. 89-97, 2018.

[8]. Kenneth J Rothman and Sander Greenland, “Modern Epidemiology”, PA: Lippincott-Raven Publishers, vol. 2, pp. 237-252, 2018.

[9]. Herbert Hethcote and Dimitri Breda, “Mathematics of epidemiological stabalities”, SIAM review, vol. 8, pp. 217-235, 2017.

[10]. Elyas Hayeti and Dimitri Breda, “An SIR epidemic model with inconsistent latency infectious period”, Mathematical Biosciences and Engineering, vol. 3, pp. 654-676, 2017.

[11]. Andrea Margheri and Peter Helderson, “Nonlinear seasonally forced epidemiological models”, Journal of Mathematical Biology, vol. 52, pp. 833-862, 2016.

[12]. Herbert Hethcote and Lawrence D Wang, “Mathematics of inconsistent incidence rate of infectious diseases”, SIAM review, vol. 10, pp. 233-287, 2017.

[13]. Nicolas Bacaer, Herbert Hethcote and Dimitri Breda, “Wavering in seasonally forced epidemiological models”, Journal of Mathematical Biology, vol. 89, pp. 118-148, 2018.

[14]. Carlos Rebelo, Andrea Margheri and Nicolas Bacaer, “Persistence in seasonally forced epidemiological models”, Journal of Mathematical Biology, vol.64, pp. 933-949, 2012.

[15]. Elizabeth Beretta and Dimitri Breda, “An SEIR epidemic model with constant latency time and infectious period”, Mathematical Biosciences and Engineering, vol. 8, pp. 931-952, 2018.