www.cafetinnova.org

Indexed in
Scopus Compendex and Geobase Elsevier, Chemical
Abstract Services-USA, Geo-Ref Information Services-USA,
List B of Scientific Journals, Poland,
Directory of Research Journals

ISSN 0974-5904, Volume 06, No. 04 (01)

International Journal of Earth Sciences and Engineering

August 2013, P.P.751-756

Model to Identify the Affect of Water Quality Parameters on Incidence of Hepatitis

H. A. P. P. MADHUBANI AND M. ROSHINI SOORIYARACHCHI
Department of Statistics, University of Colombo, Sri Lanka
Email: *peshala_uoc@yahoo.com, roshini@stat.cmb.lk*

Abstract: Viral Hepatitis is one of the main water born diseases and occurs primarily in third world countries. Hepatitis has been endemic in the country for many years. Several recent outbreaks were reported in last few years in Sri Lanka. The dangerous aspect of the disease is, more than 90% of the infected were children. The organisms are transferred through contaminated food, water and feces of infected people. Through the literature review, it was found that there were no broad research studies on viral hepatitis in Sri Lanka. Contaminated water is one of the major causes of water born disease. It is vital to identify the main water quality parameters which affect Hepatitis. The purpose of the research was to identify the factors affecting the incidence of Hepatitis. As the number of hepatitis cases are ordinal scale and the water quality parameters are continuous the Spearman's correlation coefficients were calculated to explore the primary association of hepatitis cases and water quality parameters. Incidence of hepatitis was modeled as a function of explanatory variables in this research. Since data are collected on the same units (districts) across the successive structure and the responses are correlated within the cluster (district) and time (month), Generalizing Estimating Equation (GEE) methodology was used. The Negative binomial distribution was chosen for the response which is appropriate in cases such as when there are an excess of zeros. Significant variables for the model were selected by considering the p-value associated with the Wald test statistic, while following forward selection procedure. The parameter estimates of the model produces the contribution of each variable to the log of expected number of hepatitis cases recorded in each district of the country throughout the year.

Keywords: *Incidence, ordinal scale, Spearman's correlation coefficient, Successive structure, Correlation, Forward selection*

1. Introduction:

Viral Hepatitis is one of the main water born diseases in Sri Lanka. It mainly affects the liver of a human. Viral infection of the liver makes the liver swell up and stops working well. The liver is an important organ and performs several vital functions. The most common causes of viral hepatitis are the five unrelated viruses Hepatitis A, Hepatitis B, Hepatitis C, Hepatitis D, and Hepatitis E. Prevalence of hepatitis disease can be caused by several factors including viruses, drugs, chemicals and alcohol. A patient exposed to the virus takes about a month to show symptoms and during that period the patient might unknowingly spread the disease. Most affected are poor people who are living in shanties. Hepatitis has shown a significant rise especially in the flood affected areas.[14] Lack of surveillance system of communicable diseases would result to spread the disease. Therefore it's vital to identify the factors that affect the survival of hepatitis patients in Sri Lanka. The objective of this paper is to identify the affect of water quality parameters on incidence of Hepatitis.

2. The Data:

The corresponding data for incidence of Hepatitis was gathered from the Epidemiological unit (EU) from 2005 to 2008 as per the availability of computerized data. The data consisted of all the patients reporting to a private or government hospital and collected by the EU. On the whole, there were 38 variables in the main dataset. Those are, Particulars of the patient, Date of onset, date of admission, outcome (cured/died) and date of discharge, clinical & laboratory details, information of disease transmission and immunization.

Due to the inconvenience of obtaining the water quality parameters throughout the above period with respect to each district, the scope was limited to the year 2007 since the number of patients had significantly increased in 2007. Related data on water quality was collected from the Industrial Technological unit (WEB). The research was based on limited water quality parameters for instance pH level, Turbidity (NTU), Hardness (CaCo3mg/l), Fecal coli form /100 ml and Conductivity ((µC/cm). These secondary data were measured at different places throughout the year in each district. The time effect was incorporated by considering effect of the month on the response. The brief description of each are comprised below.

The term "pH" is positive Hydrogen. The balance of positive hydrogen ions (H+) and negative hydroxide ions (OH-) in water determines how acidic or basic the water is. The pH scale ranges from 0 (strongly acidic) to 14 (strongly basic). Turbidity (NTU), it has no health effects. However, turbidity can interfere with disinfection and provide a medium for microbial growth. Turbidity may indicate the presence of disease causing organisms. These organisms include bacteria, viruses, and parasites. Hardness (CaCo3mg/l), Hard water is water that contains cations with a charge of +2, especially Ca2+ and Mg2+. These ions do not pose any health threat, but they can engage in reactions that leave insoluble mineral deposits. These deposits can make hard water unsuitable for many uses. Fecal coli form /100 ml, it's a bacterium whose presence indicates that the water may be contaminated with human or animal wastes. Microbes in these wastes can cause short-term effects. And finally, conductivity ((µC/cm), it's the ability or power to conduct or transmit heat and electricity.

3. Theory and Methodology:

A. Fitting models to clustered correlated data:

1) GLM and Generalized Estimating Equations:

Generalized Estimating Equations (GEE) are methods of parameter estimation for correlated data. When data are collected on the same units across the successive points in time, these repeated observations are correlated over time. If the correlation is not taken into account then the standard errors of the parameter estimates will not be valid and hypothesis testing results will be non replicable. GEE was introduced by Liang & Zeger(1986) and McCullah and Nelder (1989) as a method of estimation of regression model parameters when dealing with correlated data. GEE methodology is a common choice when the outcome measure of interest is discrete (e.g. binary or count data, possibly from a binomial, poison or negative binomial distribution) rather than continuous. To define Generalized Linear models using the GEE methodology, the followings need to be defined; The distribution of the response variable must be a member of the exponential family, The link function, The explanatory variables and the covariance structure of the repeated measurements.

2) Correlation of responses:

When dealing with data which are correlated within a cluster, the correlation within responses must be accounted for. Otherwise incorrect inferences about the model coefficients can be made. (Because of incorrect estimation of the variances) Ordinary Least Squares (OLS) regression models have been adopted for analysis of correlated responses when the dependent variable is normally distributed.

B. Selecting the distribution of the response variable and the link function:

1) Distribution of the response variable:

GEE's permit specification of distributions from the exponential family of distributions. In fitting a GEE (or any GLM), the user should make every reasonable effort to correctly specify the distribution for the response variable so that the variance can be efficiently calculated as a function of the mean and the regression coefficients can be properly interpreted. Generally, some prior knowledge of the distribution of the response variable must be used. In cases in which the responses are counts, first a Poisson distribution should be selected and then the extent of dispersion in the outcome predictor should be examined. When the variances derived from the data are higher or lower than those assumed in the model, the data may be over or under dispersed. When analysing counted data, a negative binomial distribution should be specified in cases in which the dispersion is high (Bliss, C.J. and Owen, A.R.G. (1958). The probability density function of a non-negative, integer valued random variable Y having a Negative Binomial distribution with parameters µ and α is denoted by Y~NB(µ;α) and is typically given as;

$$\mathrm{pr}(Y = y) = \left(\frac{\mu}{\mu + k}\right)^{y}\left(1 + \frac{\mu}{k}\right)^{-k}\frac{\tau(y + k)}{y!\,\tau(k)}$$

for k>0, µ>0. Ancombe, F.J. (1950). Here µ=E(Y) and k=1/ α.
Also, var(Y) = µ+ 2µ/k = µ+ 2αµ. This distribution has numerous applications in the biological sciences. A more useful parameterization involves α=1/k, giving

$$\mathrm{pr}(Y = y) = \frac{\tau(y + \alpha^{1})}{y!\,\tau(\alpha^{-1})}\left(\frac{\alpha\mu}{\alpha\mu + 1}\right)^{y}(1 + \alpha\mu)^{\frac{-1}{\alpha}}$$

by Bliss and Owen (1958). Here µ is the mean and α is the dispersion parameter.

Let Yij be the event of interest (e.g. hepatitis cases) in the jth unit (e.g. month within year) of the ith cluster (e.g. district) then Yij ~ NB (μij ;α). Here the associated probability of an event in the jth unit in the ith cluster is πij which is related to the cluster level covariates xij, can be expressed as follows.

$$\pi_{ij} = \mathrm{pr}(Y_{ij} = y_{ij}) = \tau_i \mu_{ij}$$

Where τi>0 is the random effect for cluster i; and log (μij) = βT xij

2) ***Link function:***

There are several choices available for link function. In cases in which count data are being modeled (with Poisson or Negative Binomial), the most appropriate link function is modeling the logarithm of the mean. Model coefficients represent the expected change in the log of the mean of the dependent variable for each change in a covariate. These coefficients that result from GEE models for log links need to be exponentiated before they are meaningful. The most appropriate link function for the negative binomial distribution is the log link.

C. Selecting the correlation structure within response variable:

A most important step involves specification of the form of correlation of responses. It is the working correlation matrix that allows GEEs to estimate models that account for the correlation of the responses. Method of GEEs has several options to specify this form and this specification will differ based on the nature of the data. For data that are correlated within cluster over time, an AutoRegressive (AR) correlation structure is appropriate. (Ballinger, G.A. (2004).

D. Theory of GEE's:

Let Y_{ij}, j = 1, ... , n_i, i = 1, ... ,k represent the jth measurement on the ith cluster. There are ni measurements on cluster i and $\Sigma\, n_i$ total measurements. Correlated data are modelled using the same link function and linear predictor setup (systematic component) as the independence case. The random component is described by the same variance functions as in the independence case, but the covariance structure of the correlated measurements must also be modeled. Let the vector of measurements on the ith cluster be Y_i = [Y_{i1}, ... , Y_{ini}]' with corresponding vector of means μ_i =[μ_{i1},……., μ_{ini}]^T and let V_i be the covariance matrix of Y_i. Let the vector of independent, or explanatory variables for the jth measurement on the ith cluster be X_{ij} = [x_{ij1}, ... , x_{ijp}]'.

The Generalized Estimating Equation of Liang and Zeger (1986) for estimating the p×1 vector of regression parameters β is an extension of the independence estimating equation to correlated data and is given by

$$S(\beta) = \sum_{i=1}^{k} \frac{\partial \mu'_i}{\partial \beta} V_i^{-1}\left[Y_i - \mu_i(\beta)\right] = 0$$

Since $g\left(\mu_{ij}\right) = X'_{ij}\beta$ where g is the link function, the p×ni matrix of partial derivatives of the mean with respect to the regression parameters for the ith subject is given by

$$\frac{\partial \mu'_i}{\partial \beta} = \begin{bmatrix} \frac{x_{i11}}{g'(\mu_{i1})} & \cdots\cdots & \frac{x_{inip}}{g'(\mu_{ini})} \\ \cdot & \cdots\cdots & \cdot \\ \frac{x_{i1p}}{g'(\mu_{i1})} & \cdots\cdots & \frac{x_{inip}}{g'(\mu_{ini})} \end{bmatrix}$$

E. Model Selection:

The Wald test statistic Balinger,G.A. (2004) can be calculated by dividing the estimate of the parameter by its standard error. This has a standard normal distribution for large samples. It can be used to test the significance of individual parameters.

F. Goodness of fit of the model:

Residuals from GEE regression models should be checked for the presence of outliers that may seriously affect the results. (Have and Chinchilli, 1998) Assessment of fit of the negative binomial model may proceed with analyses of Pearson residuals and evaluation of the generalized Pearson statistic,

$$\sum_{i=1}^{n} \frac{\left[Y_i - E\left(\hat{Y}_i\right)\right]^2}{\mathrm{var}\left(\hat{Y}_i\right)}$$

Where $E\left(\hat{Y}_i\right)$ and $\mathrm{var}\left(\hat{Y}_i\right)$ are the estimated expectation and variance of Yi under a given model. The associated degrees of freedom are n-p, where p is the number of estimated parameters in the model.

G. Interpretation of the selected model:

Let the selected model be

$$\log\left(\mu_i\right) = \beta^T x_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$$

Suppose the value of the variable x1 is increased by 1 unit when all the other variables are held constant. Let the expected response before and after this increment be μ1 and μ2 respectively.

$$\log\left(\mu_2\right) = \beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2 + \cdots + \beta_q x_q \text{---- (01)}$$

$$\log\left(\mu_1\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q \text{------ (02)}$$

(01) - (02) => $\log\left(\mu_2\right) - \log\left(\mu_1\right) = \beta_1$

$$\frac{\mu_2}{\mu_1} = e^{\beta_1}$$

This implies that when x1 increases by 1 unit while all other variables are held constant the ratio of expected response will change by exp (β1). The remaining parameters could also be interpreted in this manner.

4. Analysis and Results:

In this Univariate analysis, one explanatory variable was considered at a time and its association with the response variable was investigated. As the number of hepatitis cases are ordinal scale and the water quality parameters are continuous the Spearman's correlation coefficients were calculated to explore the primary association of hepatitis cases and water quality parameters.

Table I: *Correlation Coefficients and associated p values*

Variable1	Variable 2	coefficient	P value
cases	time	-0.054	0.4432
cases	ph	0.230	0.0009*
cases	conductivity	-0.155	0.0273*
cases	turbidity	0.061	0.386
cases	hardness	-0.027	0.7032
cases	Fecal coli form	0.033	0.6426

**significance at 20% level*

The correlation coefficients (ρ) implied that the incidence of hepatitis has positive relation with Ph (ρ=0.17 with p-value=0.016) and Fecal coli form (ρ=0.1 with p-value=0.15), while Conductivity (ρ = -0.216 with p-value=0.002) was having a negative relationship (20% significance level).

According to table I, few coefficients are highly significant at 20% level. Hence, they influence the response and must be considered in the model fitting procedure. All six variables should be considered in the model fitting procedure, since the other variables may have some associations with each other.

Initially to model the Y_{ij} as a function of the explanatory variables, a generalized linear model (GLM) was fitted using Generalized Estimating Equations (GEE) methodology, with a poison distribution for the responses. The GEE methodology was used as data are collected on the same units (districts) across successive points in time (month within year). The poison distribution was used as the response corresponds to counts. The characteristic link function for poison distribution i.e. the log link was used. Autoregressive correlation structure was used since the responses are correlated within cluster (district) over time (month). Significant variables for the model were selected by using forward selection procedure and most appropriate model was chosen. The chosen poison model was highly over-dispersed and thus this required another distribution for the response to be pursued. When the poison model is highly over dispersed and there are a large number of zero counts, the negative binomial is an alternative for modelling the data.

Next, a generalized linear model was refitted under GEE methodology for the data, with negative binomial distribution for the response, log link (the usual link function for negative binomial distribution) and autoregressive correlation structure. The next step involves selecting the most appropriate negative binomial model. PROC GENMOD procedure in SAS & forward selection method are used to fit the model.

A. *Selecting the most appropriate model using Wald test:*

The Wald test was used to select the most significant main effect. Initially, the p value was looked at and variable giving the smallest p-value corresponding to the most significant variable was selected. In instances where the p-value has the same magnitude the value of the Wald statistic (Z) was used. The largest absolute value of Z corresponds to the most significant variable. Fecal coli form, conductivity and the water are the significant main effects from the model and mentioned below.

Log(μij) = β0 + β1 (fecal coli form) + β2 (time) + β3 (conductivity)

The other main effects should be further investigated to determine if these remaining covariates could improve the model when it is already adjusted for combined effect of selected variables. Since, none of variables are significant at 5% level the procedure was stopped and the final model was selected as;

Log(μij) = β0 + β1 (fecal coli form) + β2 (time) + β3 (conductivity)

B. *Adding interaction terms to the model:*

The other interactions should be further investigated to determine if these remaining covariates could improve the model when it is already adjusted for combined effect of selected variables. Since all remaining interaction terms are insignificant, the process has been concluded with the previous best model.

C. *Diagnostics of the fitted model:*

Assessment of fit of the negative binomial model was made by analysing the Generalized Pearson residuals and using the

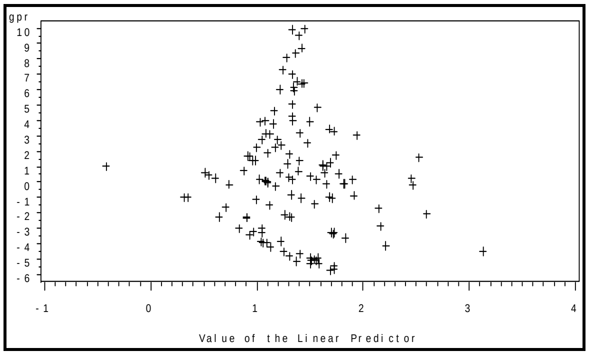

Fig1: *Residual plot*

Fig. 1 gives a plot of the generalized Pearson residuals versus fitted linear predictor. The plot shows no specific pattern and residuals are distributed symmetrically about zero apart from very few observations which are lying in right and left tails of the distribution.

D. Parameter estimates and Interpretation of the model:

The best model selected can be represented as,

$$\log(\mu ij) = \beta 0 + \beta 1(\text{fecal coli form}) + \beta 2(\text{time}) + \beta 3(\text{conductivity})$$

After adequacy of a selected model is established, the parameters estimates of model should be interpreted. Table II gives the parameter estimates and standard errors together with the corresponding Wald statistic and p-value.

Table II: *Parameter estimates of the selected Negative Binomial model*

Parameter	**Estimate**	**Standard Error**	**Z value**	**P value**
Intercept	1.8560	0.3289	5.64	<.0001
fecal coli form	0.0006	0.0001	4.73	<.0001
conductivity	-0.0012	0.0003	-4.62	<.0001
time	-0.0580	0.0184	-3.15	0.0017

The parameter estimates give the contribution of each variable to the log of the expected number of hepatitis cases recorded in each district of the country throughout the year. All variables are significant at 5% significant level. (Including the intercept)

1) ***Effect of fecal coli form on the response:***

The coefficient of fecal coli form is positive indicating that increase of the amount of fecal coli form by 1 unit of the corresponding month will lead to increasing of hepatitis patients. Suppose the fecal coli form of a particular month increases by 1 unit and the expected number of hepatitis cases before and after this increment are $\mu 1i$ and $\mu 2i$ respectively.

$\log(\mu 1i) = \beta 0 + \beta 1$ (fecal coli form) $+ \beta 2$(time) $+ \beta 3$ (conductivity) -------------model (i)

$\log(\mu 2i) = \beta 0 + \beta 1$(fecal coli form $+ 1) + \beta 2$ (time) $+ \beta 3$ (conductivity) ------------model (ii)

(ii)-(i) = > $\log(\mu 2i/ \mu 1i) = \beta 1 = 0.0006$
=> $\mu 2i/ \mu 1i = \exp(\beta 1) = \exp(0.0006)$
⇨ 1.0006
$\mu 2i = 1.0006\ \mu 1i$

This procedure can be applied for each effect. The findings are summarized in the table III.

Table III: *calculations on the selected effects*

Effect	**$\log(\mu_{2i}/\mu_{1i})$**	**μ_{2i}/ μ_{1i}**	**Relationship**
fecal coli form	0.0006	1.0006	$\mu 2i = 1.0006\mu 1i$
conductivity	-0.0012	0.9988	$\mu 2i = 0.9988\mu 1i$
time	-0.0580	0.9436	$\mu 2i = 0.9436\mu 1i$

The parameter estimates of the model produces the contribution of each variable to the log of expected number of hepatitis cases recorded in each district of the country throughout the year. This indicates that the expected number of hepatitis cases of a particular month increases (slightly) by a ratio of 1, as a result of 1 unit increment in the fecal coli form of the water of that month. This result implies that expected number of hepatitis cases of a particular month decreases by a ratio of 0.9988, as a result of 1 unit increment in the conductivity of the water of that month. This clearly depicts that expected number of hepatitis cases of a particular month decreases by a ratio of 0.9436, as a result of 1 unit increment in time.

5. Conclusion:

Fecal coli form and Conductivity of the water are the main factors which affect the incidence of hepatitis. One unit increment in the fecal coli form of the water in a particular month will increase the expected number of hepatitis cases of that month. And also, expected number of hepatitis cases of a particular month decreases as a result of one unit increment in the conductivity of the water.

6. Acknowledgment:

First and foremost I gratefully acknowledge Prof. Mrs. M. R. Sooriyarachchi for this research. Her supervision, guiding me throughout, encouragement given and for her advices helped me to perfect this research. Her alternative comments vigilant implications and supervision assisted me to succeed this. I would like to pen a special word of thanks to staff of Statistical unit in

EDU who provided me the datasets and relevant guide lines without which this study could have been impractical.

7. References:

[1] Anscombe, F.J. (1950). Sampling theory of the negative binomial and logarithmic series distributions. Biometrika. 37:358-382

[2] Ariyananda, T. (2001). Quality of Collected Rainwater from Sri Lanka. Colombo 5: Lanka Rainwater Harvesting Forum.

[3] Ballinger, G.A. (2004). Using generalized estimating equations for longitudinal data analysis. Organizational Research methods, vol. 7, No. 2, pp127-150

[4] Bliss, C.J. and Owen, A.R.G. (1958). Negative binomial distribution and a common k. Biometrika. 45:37-58

[5] C.L.Abayasekara, C. (2007). Short commiunication water quality of Maha oya stream in Peradeniya 134-137.

[6] C.Shanthi De Silva, N. Impact and Intensive Vegetable Cultivation on Agro-Well Water Quality in Malsiripura Region of Kurunagala District. Open University of Sri Lanka.

[7] Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73, 13-22

[8] McCullah, P. and Nelder, J.A. (1989). Generalized linear models (2nd Editon). London: Chapman and Hall

[9] R.M.P Rajakaruna, K. N. Quality of Shallow Ground Water in an Intensively Cultivated Hilly Cantena in a Up country Intermediate Zone of Sri Lanka. Peradeniya: Faculty of Agriculture, University of Peradeniya.

[10] S.A.Amarasinghe, C. S. Water quality assessment of agro wells in vavuniya district for the use of agricultural and domestic purposes.

[11] S.A.M.S.Dissanayake, S. Microbial Quality Assuarence of Drinking Water Supplies through Survellance. Colombo: Environmental division, national Building Research Organization.

[12] Ten-Have, T.R. and Chinchilli, V.M. (1998) Two-stage negative binomial and over dispersed poison models for clustered developmental toxicity data with random cluster size. Journal of Agricultural, Biological and Environmental Statistics, vol. 3, pp 75-98

[13] U.K.Piyadasa, K. W. (2005). Hydrochemical distribution and characteristics of groundwater in Weligama area in southern Sri Lanka.
www.medicinenet.com/viral_hepatitis/article