

Volume 10, Number 1 (2014)

# ESTEEM

Academic Journal UiTM (Pulau Pinang)



MANAGING EDITOR  
CHIEF EDITOR

Dr. Maryam Farooqui  
Dr. Chang Siu Hua

p-ISSN 1675-7939

e-ISSN 2289-4934



UNIVERSITI  
TEKNOLOGI  
MARA

## EDITORIAL BOARD

---

### ESTEEM ACADEMIC JOURNAL

VOLUME 10, NUMBER 1, JUNE 2014

Universiti Teknologi MARA (Pulau Pinang)

## ENGINEERING, SCIENCE & TECHNOLOGY

### ADVISORS

Tan Sri Prof. Ir. Dr. Sahol Hamid Bin Abu Bakar, FASc

Assoc. Prof. Dr. Ngah Ramzi Hamzah

### PANEL OF REVIEWERS

Prof. Dr. Nor Sabirin Binti Mohamed (*University Malaya*)

Assoc. Prof. Ir. Dr Hj. Ramli Nazir (*Universiti Teknologi Malaysia*)

Dr. Ahmad Safuan bin A Rashid (*Univeristi Teknologi Malaysia*)

Dr. Ahmad Zia ul-Saufie Mohamad Japeri (*Universiti Teknologi MARA (Pulau Pinang)*)

Dr. Aryati Bakri (*Universiti Teknologi Malaysia*)

Dr. Farid Ezanee Bin Mohamed Ghazali (*Universiti Sains Malaysia*)

Dr. Fatimah De'nan (*Universiti Sains Malaysia*)

Dr. Johan Mohamad Sharif (*Universiti Teknologi Malaysia*)

Dr. Leo Choe Peng (*Universiti Sains Malaysia*)

Dr. Mahibub Mahamadsa Kanakal (*Universiti Teknologi MARA (Pulau Pinang)*)

Dr. Mohd Mahadzir Mohammad@Mahmood (*Universiti Teknologi MARA (Pulau Pinang)*)

Dr. Ng Kok Shien (*Universiti Teknologi MARA (Pulau Pinang)*)

Dr. Noor Akma Ibrahim (*Universiti Putra Malaysia*)

Dr. Siti Meriam Zahari (*Universiti Teknologi MARA*)

Dr. Suriati Paiman (*Universiti Putra Malaysia*)

Dr. Vincent Lee Chieng Chen (*Curtin University, Sarawak*)

### CHIEF EDITOR

Dr. Chang Siu Hua

### MANAGING EDITOR

Dr. Maryam Farooqui

### TECHNICAL EDITORS

Dr. Ong Jiunn Chit

Sundara Rajan Mahalingam

## **LANGUAGE EDITORS**

Emily Jothee Mathai (*Universiti Teknologi MARA (Pulau Pinang)*)

Liaw Shun Chone (*Universiti Teknologi MARA (Pulau Pinang)*)

Rasaya A/L Marimuthu (*Universiti Teknologi MARA (Pulau Pinang)*)

Zeehan Shanaz Ibrahim (*Universiti Sains Malaysia*)

## FOREWORD

---

















Welcome to the 10<sup>th</sup> volume and 1<sup>st</sup> issue of the ESTEEM Academic Journal (EAJ), an online peer-refereed academic journal of engineering, science and technology. Since the beginning of this year, a number of articles have been sent to us; some of which still being under review in their first or second phase, and the first eight of them are being published now, others following in the subsequent issue. Article submissions came from different UiTM branch campuses across the country and the manuscripts covered a wide range of engineering, science and technology topics, all of them being interesting and innovative.

First and foremost, we would like to extend our sincere appreciation and utmost gratitude to Associate Professor Dr. Ngah Ramzi Hamzah, Rector of UiTM (Pulau Pinang), Dr. Mohd Mahadzir Mohammad@Mahmood, Deputy Rector of Academic Affairs and Dr. Mohd Subri Tahir, Deputy Rector of Research, Industry, Community & Alumni Network for their generous support towards the successful publication of this issue. Not to be forgotten also are the constructive and invaluable comments given by the eminent panels of external reviewers and language editors who have worked assiduously towards ensuring that all the articles published in this issue are of the highest quality. In addition, we would like to thank the authors who have submitted articles to EAJ, trusting Editor and Editorial Board and thus endorsing a new initiative and an innovative academic organ and, in doing so, encouraging many more authors to submit their manuscripts as well, knowing that they and their work will be in good hands and that their findings will be published on a short-term basis. Last but not least, a special acknowledgement is dedicated to those members of the Editorial Board who have contributed to the making of this issue and whose work has increased the quality of articles even more. Although there will always be cases in which manuscripts will be rejected, our work so far has shown that the board members' motivation has been, and will be, to make publications possible rather than to block them. By means of intensive communication with authors, academic quality is and will be guaranteed and promising research findings are and will be conveyed to the academia in a functional manner.

Dr. Chang Siu Hua  
Chief Editor  
ESTEEM Academic Journal  
Vol. 10, No. 1 (2014)  
(Engineering, Science & Technology)

## CONTENTS OF JOURNAL

---

1. International Market Entry Mode Choices by Malaysian Construction Firms Using Multinomial Regression Model 1  
Che Maznah Mat Isa, Hamidah Mohd Saman, Siti Rashidah Mohd Nasir and Christopher Nigel Preece  
 
2. Analysis of Thin Walled Tube Al 3003 H12 Under Quasi-Static Axial Crush Mode Using Finite Element Method 22  
Mansol M. M., Ismail N. I., Hisyam Basri M., Tang S. H. and Anuar M. K.  
 
3. Discharge Equation of Contracted Rectangular Falat-Crested Slit Weirs 31  
Rosley Jaafar and Ishak Abdul Azid  
 
4. The Procedure of Poisson Regression Model Using Lower Respiratory Illness In Infants 40  
Zuraira Libasin, Suryaefiza Karjanto and Shamsunarnie Mohamed Zukri  
 
5. Parametric Study On the Settlement Improvement Factor of Stone Column Groups 55  
Ng Kok Shien and Tan Siew Ann  
 
6. ICT Integration In Classrooms: The Educators' Perspective Based On Their School and Home ICT Use 66  
Johan@Eddy Luanan, Ahmad Danial Mohd Ghazali and Jasmine Jain  
 
7. A Comparative Study of the Structural Analysis Between the Integral and the Simply Supported Bridge 75  
Mohd Ashaari Masrom  
 
8. Effect of Cationic and Anionic Dye Adsorption by Base-Modified Papaya Seed (Fixed-Bed System) 89  
Norhaslinda Nasuha, Nurulhuda Amri and Hawaiah Imam Maarof  
 

# THE PROCEDURE OF POISSON REGRESSION MODEL USING LOWER RESPIRATORY ILLNESS IN INFANTS

Zuraira Libasin<sup>1</sup>, Suryaefiza Karjanto<sup>2</sup> and Shamsunarnie Mohamed Zukri<sup>3</sup>

<sup>1,2</sup>Department of Mathematical and Computer Sciences, Universiti Teknologi MARA (Pulau Pinang), Malaysia.

<sup>3</sup>Faculty of Mathematical and Computer Sciences, Universiti Teknologi MARA Cawangan Kelantan, Malaysia.

<sup>1</sup>zuraira946@ppinang.uitm.edu.my; <sup>2</sup>suryaefiza016@ppinang.uitm.edu.my;  
<sup>3</sup>shamsunarnie077@kelantan.uitm.edu.my

## ABSTRACT

*This study considers an analysis using a Poisson regression model where the response outcome is a count, with large outcomes being rare events. Estimates of the parameters are obtained by using the maximum likelihood estimates. Inferences about the regression parameters are based on Wald test and likelihood ratio test. In the model building process, the stepwise selection method were used to determine important predictor variables, diagnostic tools were used in detecting multicollinearity, non-constant variance, outliers, and also analysis of residual were used to measure the goodness fit of the model. Applications of these methods are illustrated by employing a study from LaVange, Keyes, Koch, and Margolis (1994) where a case study of lower respiratory illness data in infants which took repeated observations of infants over one year. Six explanatory variables involve the number of weeks during that year for which the child is considered to be at risk, crowded conditions occur in the household, family's socioeconomic status, race, passive smoking, and age group. We found that the explanatory variables which contribute significantly are passive smoking and crowding. Social economic status and race do not appear to be influential, and neither does age group. The value of  $R^2$  is 0.0562 which indicate that about 5.62% from the total variation can be explained by the Poisson regression model. This number does not give a better result since the variance is non-constant. It simply means the existence of overdispersion.*

**Keywords:** Poisson distribution; Poisson Regression Model; Nonlinear Model; Model Building.

## 1. INTRODUCTION

Poisson regression became popularized as an analysis method in the 1970s and 1980s (research done by Frome, Kutner, and Beauchamp (1973), and Charnes, Frome, and Yu (1976)). Later in the 1990s, Poisson's regression modeling technique was widely used in a homicide incidence study (Cyrus & Guohua, 1999), a study of injuries incurred by electrical

utility workers (Loomis, Dufort, Kleckner, & Savitz, 1999), and an evaluation of the risk of endometrial cancer as related to occupational physical activity (Moradi et al., 1998). Currently, the applications of Poisson regression models are widely used in various fields such as biomedicine (Waltoft, 2009), accident analysis and prevention (El-Basyouny & Sayed, 2009), insurance (Morata, 2009), biostatistics and epidemiology (Shults Sun, Tu, & Amsterdam, 2005), environmental sciences (Agarwal, Gelfand, & Citron-Pousty, 2002), criminology (Osgood, 2000) and agriculture (Hall, 2000).

Poisson regression model is one of the nonlinear regression models where the response outcomes are discrete; therefore all the theories including the model development, model building, diagnostics and inferences that have been used in the analysis of Poisson regression model are carried out in a similar fashion as for the nonlinear regression models (Neter, Kutner, Nachtsheim, & Wasserman, 2003). Estimation of the parameters of a nonlinear regression model is usually carried out by the method of least squares or the method of maximum likelihood (Spiegelman & Hertzmark, 2005; Linda & Julio, 2001). Unlike in linear regression, it is usually not possible to find analytical expressions for the least squares and maximum likelihood estimators for nonlinear model regression models. Inferences about the regression parameters in nonlinear regression are usually based on large-sample theory. This theory states that the least squares or maximum likelihood estimators for nonlinear regression models with normal error terms, when the sample size is large, are approximately evenly distributed and almost unbiased, and have almost minimum variance (Neter et al., 2003). In Poisson regression model, a large-sample test of a single regression parameter can be constructed by using Wald test (Yang, Hardin, & Addy, 2009). For several regression parameters, a large-sample test can be constructed by using likelihood ratio test (Hardin, Yang, Addy, & Vuong, 2007).

The model building process of the regression model considers the selection of variables, diagnostic tools and remedial measures. The automatic selection procedures that have been used are stepwise method and all-possible-regressions method (Kleinbaum, Lawrence, Kupper, Muller, & AzharNizam, 1998). However, stepwise selection procedures are frequently occupied in Poisson regression model (Neter et al., 2003). The model building process for nonlinear regression models often differs somewhat from the linear regression models. The reason is that the functional form of many nonlinear models is less suitable for adding or deleting predictor variables and interaction effects in the direct fashion that is feasible for linear regression models. The use of diagnostic tools to examine the appropriateness of a fitted model plays an important role in the process of building a nonlinear regression model. Plots of residuals can be helpful in diagnosing departures from the assumed model. Two goodness fit tests that can be determined are the Pearson chi-square (Neter et al., 2003) and the deviance (Wang, Kalwani, & Akcura, 2007). If unequal error variances are present, weighted least squares can be used in fitting the nonlinear regression model (Bender & Heinemann, 1995). Alternatively, transformations of the response variable that may stabilize the variance of the error terms and also permit use of a regression model can be investigated (Neter et al., 2003). Multicollinearity can be verified by using the collinear quantity and condition index that can be obtained from the eigen system (Lazaridis, 2007). Plots of deviance residuals can help to identify the outliers indicated in the model (Shrestha, 2007).

The Poisson regression model is one of the nonlinear regression models where the response outcomes are discrete (Frome et al., 1973). Therefore all the theories including the model development, model building, diagnostics and inferences that have been used in the analysis of Poisson regression model are carried out in a similar fashion as for the nonlinear regression models so that the accurate explanation will be obtained (Neter et al., 2003). However, most researchers take a simple way to model the Poisson regression without taking into account the assumptions corresponding to the model before doing any further analysis. This problem also happens when modelling the linear regression model (Karjanto, Abdul Razak, & Mahlan, 2007). Failure to fulfil the assumptions will cause an inaccurate model which leads to an insignificant study. The objectives of the study are: (1) To identify the existence of departures in Poisson regression model through the appropriate diagnostic tools, (2) To fix the departures by using the appropriate goodness measurement.

## 2. MATERIAL AND METHOD

### 2.1 Data

Data of lower respiratory illness (LRI) in infants which took repeated observations of infants over one year was used as a case study to model the Poisson regression model. This data was taken from an article of the National Blood, Heart and Lung Institute, United States namely "Application of Sample Survey Methods for Modeling Ratios to Incidence Densities (LaVange et al., 1994). About 284 children participated in the study and the outcome interest was the total number of times or counts of lower respiratory infection recorded for a year. The variable **COUNT** is the total number of infections that year.

Six explanatory variables were evaluated including the variable **RISK** that is the number of weeks during that year for which the child is considered to be at risk (when a lower respiratory infection is ongoing, the child is not considered to be at risk for a new one), **CROWDING** is an indicator variable for whether crowded conditions occur in the household, **SES** is an indicator variable for whether the family's socioeconomic status was considered low (0), medium (1), or high (2). The variable **RACE** is an indicator for whether the child was white (1) or not (0), and the variable **PASSIVE** is an indicator for whether the child was exposed to cigarette smoking. Finally the **AGEGROUP** variable takes the value 1, 2, and 3 for below four months, four to six months, and more than six months respectively.

### 2.2 Methodology

Poisson distribution is suitable for the data where the response outcome is a count ( $Y_i = 0, 1, 2, \dots$ ) with a large number of occurrences being a rare events (Heinzl & Mittlböck, 2003). The probability of Poisson distribution can be stated as follows:-

$$f(y) = \frac{\mu^y e^{-\mu}}{y!} \quad Y = 0, 1, 2, \dots \quad (1)$$

where  $f(y)$  denotes the probability of  $Y$  outcomes and  $Y! = Y(Y-1)\dots 3.2.1$ . The mean and variance of Poisson distribution are  $E\{Y\} = \mu$  and  $\sigma^2\{Y\} = \mu$ , respectively.

Poisson regression model is one of the nonlinear regression model (Neter et al., 2003). Hence it is generally stated as



$$\underbrace{Y_i}_{E(Y_i)} = f(X_i, \beta) + \varepsilon_i \quad i = 1, 2, \dots, n \quad (2)$$

Where  $Y_i$  is the mean response for the  $i$ th case,  $f(X_i, \beta)$  is the response function that is also known as  $E(Y_i)$ ; the mean response of  $Y_i$  and  $\varepsilon_i$  is the error terms. Therefore, from equation (2), the Poisson regression model can best be written as

$$\underbrace{Y_i}_{E(Y_i)} = \mu_i + \varepsilon_i \quad i = 1, 2, \dots, n \quad (3)$$

Where  $\mu_i$  denotes the mean response and  $\varepsilon_i = Y_i - \mu_i$  is the error terms with  $E(\varepsilon_i) = 0$ . In Poisson regression model, two properties that should be considered are (a) Since  $Y_i \sim \text{Poisson}(\mu_i)$ , therefore the error  $\varepsilon_i = Y_i - \mu_i$  follows approximately a Poisson distribution with  $E(\varepsilon_i) = 0$ , (b) Errors of the variance are constant since  $\text{Var}(Y_i) = \mu_i$ , therefore  $\text{Var}(\varepsilon_i) = \text{Var}(Y_i - \mu_i) = \mu_i$ . The mean response for the  $i$ th case that is  $E(Y_i)$  can be written as  $\mu_i$  in which it is the function of the set of predictor variables  $X_1, \dots, X_{p-1}$ . In Poisson regression model, the notation  $\mu(X_i, \beta)$  is denoted as the function that relates the mean response ( $\mu_i$ ) to  $X_i$  (the values of the predictor variables for case  $i$ ) and  $\beta$  (the values of the regression coefficients). Some commonly used functions for Poisson regression are:

$$\mu_i = \mu(X_i, \beta) = X_i' \beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{p-1} X_{p-1} \quad (4)$$

$$\mu_i = \mu(X_i, \beta) = e^{X_i' \beta} = (e^{\beta_0})(e^{\beta_1 X_1})(e^{\beta_2 X_2}) \dots (e^{\beta_{p-1} X_{p-1}}) \quad (5)$$

$$\mu_i = \mu(X_i, \beta) = \log(X_i' \beta) = \log(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{p-1} X_{p-1}) \quad (6)$$

Estimation of the parameters of a nonlinear regression model is usually carried out by the method of least squares or the method of maximum likelihood (Spiegelman & Hertzmark, 2005; Linda & Julio, 2001). The probability distribution of response variable  $Y_i$  is

$$F_i(Y) = \frac{[\mu(X_i, \beta)]^{Y_i} \exp[-\mu(X_i, \beta)]}{Y_i!} \quad (7)$$

From equation (7), the likelihood function is as follows:

$$L(\beta) = \prod_{i=1}^n f_i(Y_i) = \frac{\prod_{i=1}^n [\mu(X_i, \beta)]^{Y_i} \exp[-\mu(X_i, \beta)]}{\prod_{i=1}^n Y_i!} = \frac{\{ \prod_{i=1}^n [\mu(X_i, \beta)]^{Y_i} \} \exp[-\sum_{i=1}^n \mu(X_i, \beta)]}{\prod_{i=1}^n Y_i!} \quad (8)$$

By taking logarithm function into equation (8), therefore the log-likelihood function is

$$\ln L(\beta) = \sum_{i=1}^n Y_i \ln[\mu(X_i, \beta)] - \sum_{i=1}^n \mu(X_i, \beta) - \sum_{i=1}^n \ln Y_i \quad (9)$$

The maximum likelihood estimates  $b_0, b_1, \dots, b_{p-1}$  can be obtained by differentiating equation (9) with respect to the vector of regression coefficients  $\beta$  then equals to zero.

The test concerning a single regression parameter  $\beta_k$  is commonly referred to as the Wald test (Yang et al., 2009). It is based on standard normal variable  $z$  which is based on large-sample test. On the other hand, the test concerning several regression parameters  $\beta_k$  is called the

likelihood ratio test (Hardin et al., 2007) where it is based on a comparison between full model and reduced model.

In building the Poisson regression model, the significant variables should be entered into the model. The criterion for adding or deleting the variables is based on the test of regression parameter. The method of stepwise selection procedures are frequently used in Poisson regression model building (Kleinbaum et al., 1998). It includes regression models in which the selection of predictive variables is carried out by an automatic procedure. The selection criteria for Poisson models is by taking the largest log-likelihood calculation as the best since Poisson regression does not have a measure equivalent to R-squared (Neter et al., 2003). The log-likelihood is evaluated with a chi-squared test to determine the relative significance for the “*p* to enter” and “*p* to stay” where *p* denotes the *p*-value.

The appropriateness of the fitted Poisson regression model needs to be examined before it is accepted for use. Therefore some diagnostics testing are applied to check the adequacy of a Poisson regression model. These include goodness of fit test which are Pearson Chi-Square goodness of fit (Neter et al., 2003): The test assumes that the  $Y_{ij}$  observations are independent and the sample size is rationally large. The test can detect major departures from a Poisson response function. However it is not responsive to small departures from a Poisson response function, and deviance goodness of fit (Neter et al., 2003 and Wang et al., 2007): Another test of fitting the Poisson response function is based on the model deviance (Wang et al., 2007). It can be defined as follows:

$$DEV(X_0, X_1, \dots, X_{p-1}) = -2 \left[ \sum_i^n Y_i \ln \left( \frac{\hat{\mu}_i}{Y_i} \right) + \sum_i^n (Y_i - \hat{\mu}_i) \right] \quad (10)$$

Where  $\hat{\mu}_i$  is the fitted value for the *i*th case. If the Poisson response function is the exact response function and the sample size *n* is large, therefore the deviance will follow approximately a chi-square distribution with  $n - p$  degrees of freedom. Large values of the deviance show that the fitted Poisson model is inexact. Multicollinearity (Lazaridis, 2007): The issue of multicollinearity arises when there is a high degree of correlation (either positive or negative) between two or more predictor variables. If this happens, the model might be not adequate to represent the data set. Therefore the diagnostic tool is considered for identifying multicollinearity which is suitable in modeling Poisson regression by using the collinear quantity and condition index that can be obtained from the eigensystem (Lazaridis, 2007). The detection of multicollinearity is usually based on condition number (CN) or condition index of the data matrix where the matrix consists of the predictor variables. Three (3) conditions to determine the result of multicollinearity are (Neter et al., 2003):

1. If the collinear quantity is less than 100 ( $\kappa < 100$ ), therefore no serious multicollinearity exists.
2. If the collinear quantity is between 100 and 1000 ( $100 < \kappa < 1000$ ), therefore moderate multicollinearity exists.
3. If the collinear quantity is greater than 1000 ( $\kappa > 1000$ ), therefore serious multicollinearity exists.

The condition index should be considered to verify the evidence firmly. By looking at the value from the condition index, if the condition index for every variables is greater than 30 ( $\kappa_j > 30$ ), it indicates serious multicollinearity exists. Outliers (Neter et al., 2003; Shrestha, 2007): Outliers are extreme observations which may lead departures to Poisson regression model (Neter et al., 2003). Once the outliers are encountered, the first suspicion is that the observations resulted from a mistake or other extraneous effect. Therefore residual plots against predicted value as well as box plots, stem and leaf plots and dot plots are useful to identify the outlier. Moreover, it's helps to examine the adequacy of the linear part of the Poisson regression model (Neter et al., 2003). Non-constant error variance (Neter et al., 2003): To model the Poisson regression where the data is counted, the variance must be constant (Neter et al., 2003). If not, there exist a departure namely over-dispersion or under-dispersion. This problem normally happens in data calculation because it is difficult to have a constant variance. Figure 1 contains the prototype of residual plots. The y-axis represents the residual whereas the x-axis represents the predicted value. Figure 1(a) indicates the constant error term variance whereas figure 1(b)–(d) indicates the non-constant error term variance (Neter et al., 2003).

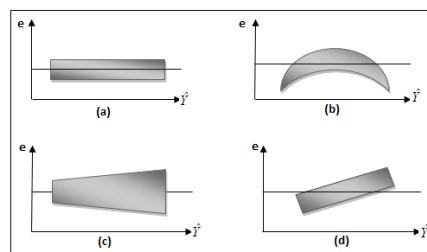


Figure 1: Prototype Residual Plots.

This study provides a flow chart to summarize the procedure in building the Poisson regression model. Figure 2 indicates the flow chart on how to develop the best Poisson regression model.

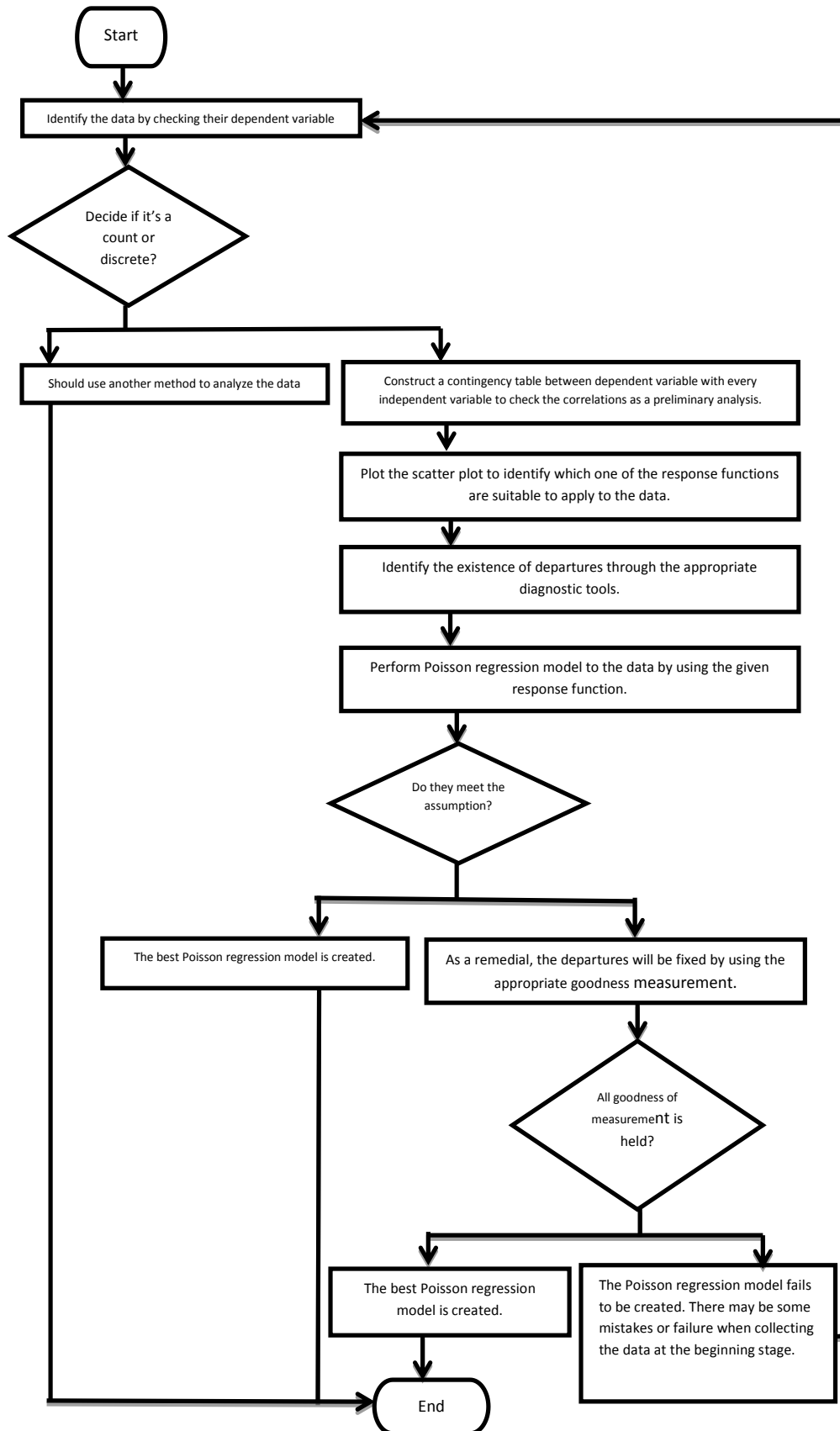


Figure 2: The flow chart of Poisson regression model.

### 3. RESULTS AND DISCUSSION

Statistical relationships between variables rely on notions of correlation as well as regression (Neter et al., 2003). These two concepts aim to describe the ways in which variables relate to one another. Therefore, as a preliminary analysis, data of lower respiratory illness in infants are presented into correlation matrix (Table 1) to identify the relationship among the variables. By referring to Table 1, it indicates that there are positive correlation between COUNT\*PASSIVE and COUNT\*CROWDING the significant level at 0.01. For COUNT\*SES, it indicates a significant positive correlation at 0.05.

To identify which functions are to be used into this data, scatter plot for every variables were displayed. Figure 2 shows the scatter plots. From the scatter plots, the pattern looks like an exponential function. It shows that the exponential function is a suitable function to be applied into this data.

Table 1: Correlation matrix.

		COUNT	RISK	PASSIVE	CROWDING	SES	AGEGROUP	RACE
COUNT	Pearson Correlation	1	-.059	.165**	.198**	.150*	-.034	.019
	Sig. (2-tailed)	.	.326	.005	.001	.012	.563	.751
	N	284	284	284	284	284	284	284
RISK	Pearson Correlation	-.059	1	-.187**	-.139*	-.278**	-.193**	-.257**
	Sig. (2-tailed)	.326	.	.002	.019	.000	.001	.000
	N	284	284	284	284	284	284	284
PASSIVE	Pearson Correlation	.165**	-.187**	1	.190**	.350**	-.064	.182**
	Sig. (2-tailed)	.005	.002	.	.001	.000	.284	.002
	N	284	284	284	284	284	284	284
CROWDING	Pearson Correlation	.198**	-.139*	.190**	1	.385**	-.009	.241**
	Sig. (2-tailed)	.001	.019	.001	.	.000	.887	.000
	N	284	284	284	284	284	284	284
SES	Pearson Correlation	.150*	-.278**	.350**	.385**	1	.070	.261**
	Sig. (2-tailed)	.012	.000	.000	.000	.	.242	.000
	N	284	284	284	284	284	284	284
AGEGROUP	Pearson Correlation	-.034	-.193**	-.064	-.009	.070	1	.064
	Sig. (2-tailed)	.563	.001	.284	.887	.242	.	.284
	N	284	284	284	284	284	284	284
RACE	Pearson Correlation	.019	-.257**	.182**	.241**	.261**	.064	1
	Sig. (2-tailed)	.751	.000	.002	.000	.000	.284	.
	N	284	284	284	284	284	284	284

\*\* . Correlation is significant at the 0.01 level (2-tailed).  
 \* . Correlation is significant at the 0.05 level (2-tailed).

Regression analysis attempts to determine the best "fit" between two or more variables. Thus, Poisson regression model is fitted using SAS software to predict the values of a dependent variable namely COUNT. Based on the pattern in Figure 3, the general Poisson regression model is

$$\hat{\mu}_i = \exp \left( \beta_0 + \beta_1 PASSIVE + \beta_2 SES + \beta_3 CROWDING + \beta_4 RISK + \beta_5 RACE + \beta_6 AGEGROUP \right) \quad (11)$$

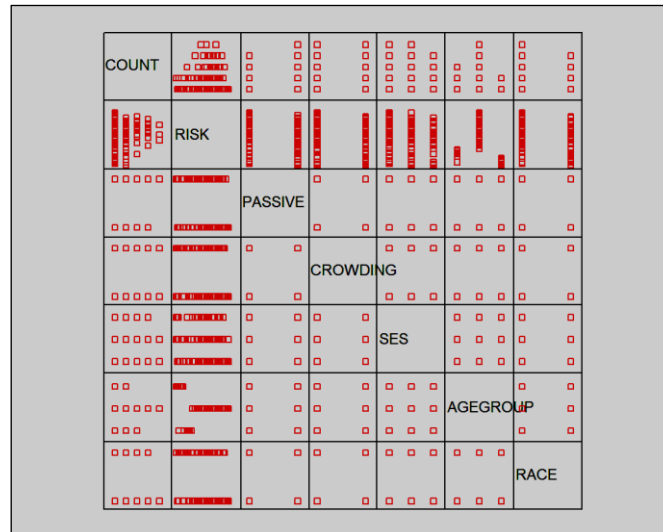


Figure 3: Scatter plots.

By taking the natural log to the above equation, the Poisson regression model will become a linear model where it can be written as follows:

$$\log_e(\hat{\mu}) = \log_e e^{\beta_0 + \beta_1 PASSIVE + \beta_2 SES + \beta_3 CROWDING + \beta_4 RISK + \beta_5 RACE + \beta_6 AGEGROUP} \quad (12)$$

$$\log_e(\hat{\mu}) = \beta_0 + \beta_1 PASSIVE + \beta_2 SES + \beta_3 CROWDING + \beta_4 RISK + \beta_5 RACE + \beta_6 AGEGROUP \quad (13)$$

In the LRI data, the model building began with the six independent variables that were considered fundamental as the key explanatory variables. They are RISK, PASSIVE, CROWDING, SES, AGEGROUP and RACE. By applying the stepwise method, researchers recommended the **PASSIVE** and **CROWDING** variables in the model since there is significance which is shown by the stepwise method. Table 3.4.5 shows the analysis of parameter estimates after considering the stepwise method. It indicates that all the variables are strongly significant since p-values are less than 0.05. Hence, the Poisson regression model will be

$$\log_e(COUNT) = 0.36282 + 0.29039PASSIVE + 0.38339CROWDING \quad (14)$$

According to Table 2, the value of collinear quantity is  $\kappa = 9$  which is less than 100. It indicates that no serious multicollinearity exists. Besides, the condition index for every variable is less than 30 which indicates that no serious multicollinearity exists.

Table 2: Collinearity diagnostics of LRI data.

Variable	Eigenvalue	Condition Index	Proportion of Variation		
			Intercept	Passive	Crowding
Intercept	2.31813	1.00000	0.05921	0.06536	0.07145
Passive	0.42432	2.33733	0.01945	0.39591	0.76848
Crowding	0.25755	3.00010	0.92134	0.53873	0.16006

To detect the existence of departures, the goodness of fit test was applied to the Poisson regression model (Neter et al., 2003). The tests are Pearson Chi-Square and Deviance goodness of fit tests. Both tests used the following hypothesis testing;

$$H_0: E\{Y\} = \log(X'\beta) \quad H_a: E\{Y\} \neq \log(X'\beta) \quad (15)$$

From the SAS program, the result shows that  $X^2 = 520.8575$  and  $DEV(X_0, X_1, X_2, X_3) = 416.7219$ . The critical value for both tests is  $\chi^2(0.95, 281) > \chi^2(0.95, 100) = 124.3$ . Since  $X^2 = 520.8575 < 124.3$  and  $DEV(X_0, X_1, X_2, X_3) = 416.7219 < 124.3$ , this result leads to accept the null hypothesis. It indicates that this Poisson regression model is an appropriate model. The value of  $R^2$  is 0.0562 which indicates that about 5.62% from the total variation can be explained by the Poisson regression model. This number does not give a better result because the variance is non constant. Figure 4 shows the plot of residual versus predicted value.

It indicates that there is a strong statistical evidence that count variation is dependent on the values of the predictor variables. A transformation of count might be useful in reducing non-constant variance.

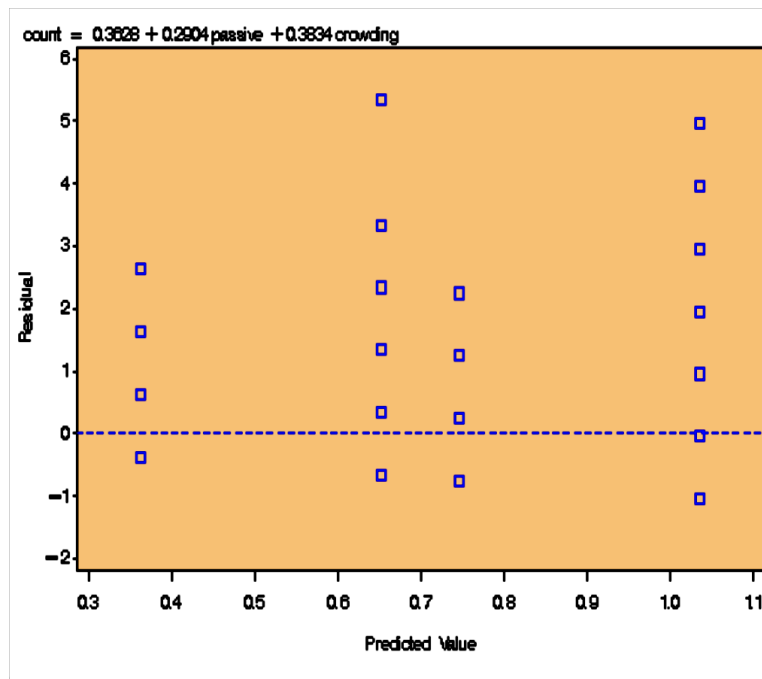


Figure 4: Plot of Residual versus Predicted Value.

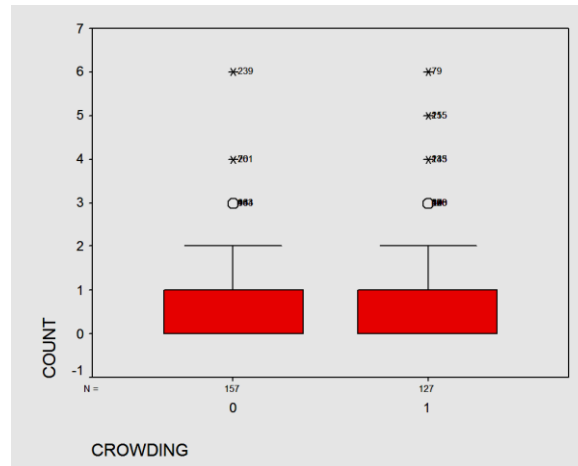


Figure 5(a): Count Vs Crowding.

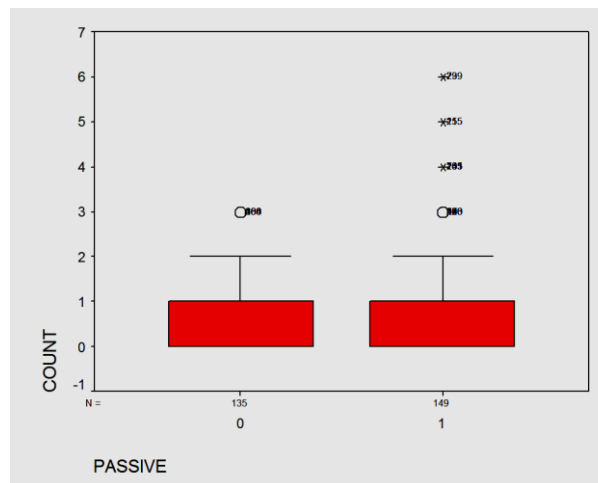


Figure 5(b): Count Vs Passive.

Figure 5(a) and 5(b) contain a box plot for variable CROWDING and PASSIVE. It shows that there exist and outliers. From the SAS program, there are six observations that are encountered as an outlier. Table 3 indicates the list of the outlying observation. After deleting the six observations, no observations are qualified as outliers.

Table 3: List of the outlying observations.

Potential Outlier Observations: Prob < 0.05							
Omit?	Obs	count	crowding	passive	Residual	Studentized Residual	Pr >  t
*	76	4	0	1	.	.	.
*	79	6	1	1	.	.	.
*	115	5	1	1	.	.	.
*	201	4	0	1	.	.	.
*	239	6	0	1	.	.	.
*	255	5	1	1	.	.	.



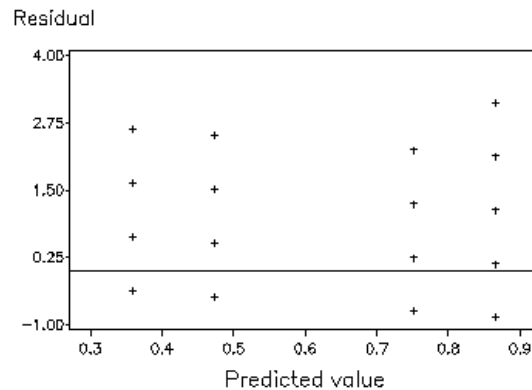


Figure 6: Plot of Residual versus Predicted Value (After removing the outliers).

Figure 6 contains the plot of residual versus predicted value after removing the outliers. There is strong statistical evidence that count variation is dependent on the values of the predictor variables although six observations that are qualified as outliers are deleted. A transformation of count might be useful in reducing non-constant variance.

Although the goodness of fit test shows that the model is appropriate, we cannot conclude that this model is the best model because there is strong statistical evidence that count variation is dependent on the values of the predictor variables although six observations that are qualified as outliers are deleted. A transformation of count might be useful in reducing non-constant variance (Neter et al., 2003). Moreover the value of  $R^2$  is 0.0562 which indicate that about 5.62% from the total variation can be explained by the Poisson regression model. This number does not give a better result since the variance is non-constant. This means there exist an overdispersion. According to Maura, Charles, and Gary (2000), to manage an overdispersion, the researcher should assume a more flexible distribution such as the negative binomial since this data is a count data. Overdispersion happens when the observed variance is larger than the nominal variance for a particular distribution (Dean, 1998). The Poisson regression model assumes that the mean and variance of the dependent variable is equal but in practice the data may display overdispersion or extra-Poisson variation, i.e a situation where the variance exceeds the mean (Ismail & Jemain, 2005). Since overdispersion is able to have a major impact on inference, further analysis needsto be done by using Generalized Estimating Equations (GEE) based approach (Maura et al., 2000; Morel & Neerchal, 2008) instead of using another approach such as scaling factor method (McCullagh, 1989).

#### 4. CONCLUSION AND RECOMMENDATION

This study used the total number of lower respiratory infection data among infants recorded for a year. Specifically, this study isto find the regression relationship between response variable that is the total number of times or counts of lower respiratory infection recorded for the year (**COUNT**) and six explanatory variables such as the number of weeks during that year for which the child is considered to be at risk (**RISK**), the crowded conditions (**CROWDING**), the socioeconomic status (**SES**), the race of the child (**RACE**), smoking exposure (**PASSIVE**), and the age of child (**AGEGROUP**). Since the response variable is a count, therefore Poisson regression model is suitable for this data. The Pearson test of correlation matrix shows that there are positive correlation between **COUNT\*PASSIVE** and **COUNT\*CROWDING** at 0.01 level of significant. For **COUNT\*SES**, it indicates positive

correlation at 0.05 level of significant. The model building process of the Poisson regression model is stepwise method. This method shows that variables which are recommended are **PASSIVE** and **CROWDING** since both are significant. Fortunately, no serious multicollinearity exist in the model.

## REFERENCES

- Agarwal, D. K., Gelfand, A., & Citron-Pousty, S. (2002). Zero-inflated model with application to spatial count data. *Environmental and Ecological Statistics*, 9, 341-355.
- Bender, R. & Heinemann, L. (1995). Fitting nonlinear regression models with correlated errors to individual pharmacodynamic data using using SAS software. *Journal of Pharmacokinetics and Pharmacodynamics*, 23, 87-100.
- Charnes, A., Frome, E. L., & Yu, P. L. (1976). The equivalence of generalized least squares and maximum likelihood estimation in the exponential family. *Journal of the American Statistical Association*, 71, 169-172.
- Cyrus, S. & Guohua, L. (1999). Homicide mortality in the United States, 1935-1994: Age, period, and cohort effects. *American Journal of Epidemiology*, 11, 1213-1222.
- Dean, C. B. (1998). *Overdispersion*, in *Encyclopedia of Biostatistics* (4th vol.) (P. Armitage & T. Colton, Eds.) (pp. 3226-3232). New York: John Wiley & Sons, Inc.
- El-Basyouny, K. & Sayed, T. (2009). Collision prediction models using multivariate Poisson-lognormal regression. *Accident Analysis and Prevention*, 41, 820-828.
- Frome, E. L., Kutner, M. H., & Beauchamp, J. J. (1973). Regression analysis of Poisson distributed data. *Journal of the American Statistical Association*, 68, 935-940.
- Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics*, 56, 1030-1039.
- Hardin, J. W., Yang, Z., Addy, C. L., & Vuong, Q. H. (2007). Testing approaches for overdispersion in Poisson regression versus the generalized Poisson model. *Journal of Biometrical*, 49, 565-584.
- Heinzl, H. & Mittlböck, M. (2003). Pseudo R-squared measures for Poisson regression models with over- or underdispersion. *Computational Statistics & Data Analysis*, 44, 253-271.
- Ismail, N. & Jemain, A. A. (2005). Generalized Poisson regression: An alternative for risk classification. *Jurnal Teknologi*, 43, 39-54.
- Karjanto, S., Abdul Razak, N. A., & Mahlan, S. B. (2007). *Prosedur Pembinaan Model Linear*. Institut Penyelidikan, Pembangunan dan Pengkomersilan, Universiti Teknologi MARA, Shah Alam.
- Kleinbaum, D. G., Lawrence, L., Kupper, L. L., Muller, K. E., & AzharNizam, A. (1998). *Applied Regression Analysis and Multivariate Methods* (3rd ed.). Duxbury Press.

- LaVange, L. M., Keyes, L. L., Koch, G. G., & Margolis, P. E. (1994). Application of sample survey methods for modelling ratios to incidence densities. *Statistics in Medicine*, 13, 343-355.
- Lazaridis, A. (2007). A note regarding the condition number: The case of spurious and latent multicollinearity. *Journal of Quality and Quantity*, 41, 123-135.
- Linda, L. H. & Julio, M. S. (2001). Generalized least squares methods for bivariate Poisson regression. *Communications in Statistics – Theory and Methods*, 30, 263-277.
- Loomis, D., Dufort, V., Kleckner, R. C., & Savitz, D. A. (1999). Fatal occupational injuries among electric power company workers. *American Journal of Industrial Medicine*, 35, 302-309.
- Maura, E. S., Charles, S. D., & Gary, G. K. (2000). *Categorical Data Analysis Using the SAS System* (2nd ed.). Cary, NC: SAS Institute Inc.
- McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.
- Morata, L. B. (2009). A priori ratemaking using bivariate Poisson regression model. *Insurance: Mathematics and Economics*, 44, 135-141.
- Moradi, T., Nyrèn, O., Bergström, R., Gridley, G., Linet, M., Wolk, A., Dosemeci, M., & Adami, H. (1998). Risk for endometrial cancer in relation to occupational physical activity: A nationwide cohort study in Sweden. *International Journal of Cancer*, 76, 665-670.
- Morel, J. G. & Neerchal, N. K. (2008). Ratio estimation via Poisson regression and Generalized Estimating Equations. *Statistics and Probability Letters*, 78, 2188-2193.
- Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (2003). *Applied Linear Regression Models*. Irwin: McGraw-Hill.
- Osgood, D. W. (2000). Poisson-based regression analysis of aggregate crime rates. *Journal of Quantitative Criminology*, 16, 21-43.
- Shrestha, S. L. (2007). *Time series modelling of respiratory hospital admissions and geometrically weighted distributed lag effects from ambient particulate air pollution within Kathmandu valley, Nepal*. Project report, Kathmandu, Nepal.
- Shults, J., Sun, W., Tu, X., & Amsterdam, J. (2005). *On the violation of bounds for the correlation in generalized estimating equation analyses of binary data from longitudinal trials*. Technical Report 200501, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine.
- Spiegelman, D. & Hertzmark, E. (2005). Easy SAS calculations for risk or prevalence ratios and differences. *American Journal of Epidemiology*, 162, 199-200.

- Waltoft, B. L. (2009). A SAS-macro for estimation of the cumulative incidence using Poisson regression. *Computer Methods and Programs in Biomedicine*, 93, 140-147.
- Wang, H. M., Kalwani, M. U., & Akcura, T. (2007). A Bayesian multivariate Poisson regression model of cross-category store brand purchasing behavior. *Journal of Retailing and Consumer Services*, 14, 369-382.
- Yang, Z., Hardin, J. W., & Addy, C. L. (2009). A score test for overdispersion in Poisson regression based on the generalized Poisson-2 model. *Journal of Statistical Planning and Inferenc*, 139, 1514-1521.