Malaysian Journal of Public Health Medicine 2010, Vol. 10 (2): 6-15

ORIGINAL ARTICLE

HANDLING OVERDISPERSION IN MORTALITY DATA IN TIME-SERIES EPIDEMIOLOGIC RESERCH USING SAS SOFTWARE

Wan Rozita WM [1], Rasimah A [2], Mazrura S [3], Lim KH[1], Thana S[1]

[1] *Institute for Medical Research, Jalan Pahang 50588, Kuala Lumpur*
[2] *Universiti Teknologi MARA, Shah Alam, Selangor*
[3] Faculty of Allied Health Sciences, Universiti Kebangsaan Malaysia, Jalan Raja Muda Abdul Aziz, 50300 Kuala Lumpur

ABSTRACT

Analysis of count event data such as mortality cases, were often modelled using Poisson regression model. Maximum likelihood procedures were used by using SAS software to estimate the model parameters of a Poisson regression model. However, the Negative Binomial distribution has been widely suggested as the alternative to the Poisson when there is proof of overdispersion phenomenon. We modelled the mortality cases as the dependent variable using Poisson and Negative Binomial regression and compare both of the models. The procedures were done in SAS by using the function PROC GENMOD. The results showed that the mortality data in Poisson regression exhibit large ratio values between deviance to degree of freedom which indicate model misspecification or overdispersion. This large ratio was found to be reduced in Negative Binomial regression. The Normal probability plot of Pearson residual confirmed that the Negative Binomial regression is a better model than Poisson regression in modelling the mortality data. *The objective of this study is to compare the goodness of fit of Poisson regression model and Negative Binomial regression model in the application of air pollution epidemiologic time series study by using SAS software.*

Key words: *count data, Poisson regression, PROC GENMOD, SAS*

INTRODUCTION

The relationships between air pollution and mortality are most often studied using time-series studies. These time-series studies analyze daily observations of the number of deaths with the daily pollution levels. Regression techniques are used to estimate a coefficient that represents the relationship between exposure to pollution and the health outcome. The usual regression method models the logarithm of the outcome to estimate the relative risk or proportional change in the outcome per increment of ambient pollutant concentration. The most widely and traditionally used regression model for mortality data in environmental epidemiology is the Poisson regression model[1]. A common practical problem with Poisson regression is the variance of the observed counts is greater than the mean, which is also called as overdispersion. Inappropriate usage of Poisson may underestimate the standard errors and overestimate the significance of the regression parameters, thus will give misleading inferences about the regression parameters[2]. An alternative approach to modelling overdispersion is to start from standard Poisson regression and add a random effect factor to represent unobserved heterogeneity. This is the characteristic of a negative binomial distribution (NB). The NB has been proved could handle some situations where the Poisson model is poor fit. Although there are new

methods in statistical modelling done in the area of environmental epidemiology such as generalized linear mixed-effects model which uses the penalized splines as the smoothing methods[3] or generalized additive models that uses natural cubic splines as the smoothing methods[4,5,6] generalized linear model with Poisson and NB regression[7] are still been used widely in Asian countries in the area of environmental epidemiology.

OBJECTIVE OF THE STUDY

The objective of the study is to compare the goodness of fit of Poisson regression model and Negative Binomial regression model in the application of air pollution epidemiologic time series study by using SAS software.

METHODOLOGY

Poisson Regression

Time series modelling of count data using Poisson regression model has been the primary statistical approach in the environmental epidemiology to assess the risks of air pollution studies. Counts of independent and random occurrences across time have typically been modelled as a Poisson process[8] that is daily mortalities are assumed to follow a Poisson distribution. If risk of mortality is influenced by seasonal changes, such as weather, air pollution, or holiday indicator, then the Poisson process will be non stationary. This means that the underlying expected mean mortality count will change over time depending on these variables. Poisson regression modelling provides a formal way to evaluate possible associations between daily mortality counts and daily concentrations of air pollution while controlling for possible confounders such as weather or holiday indicator.

The Poisson model takes the form of:

$$P(y_i) = \frac{\left(\exp(-\mu_t)(\mu_t)^{y_t}\right)}{y_i!}$$

A quadratic term to the variance representing overdispersion has been added in the negative binomial model and takes the form as below:

$$P(y_i) = \frac{\Gamma\left(v_i + \frac{1}{K}\right)}{y_i!\Gamma\left(\frac{1}{K}\right)}\left[\frac{(K\mu_i)}{(1+K\mu_i)}\right]^{y_i}\left(\frac{1}{1+K\mu_i}\right)^{1/K}$$

K is the overdispersion parameter.

The main assumption under the Poisson model is that expected value of the random variable Y_i (mortality counts) for subject i is equal to its variance:

$\mu = E(Y_i) = Var(Y_i)$

If the Variance in a Poisson model is larger than the mean, the model is known as overdispersed model and this phenomenon is known as overdispersion.

The value of the mean must be greater than zero. Therefore the mean can also be written in a more generalized linear form given by:

$\mu_t = \exp\left(\beta_0 + \sum_{j=1}^{n} x_{ij}\beta_j\right)$; μ_i is the number of mortality to be expected.

$X_{i1}, X_{i2,} X_{i3}, \ldots X_{in}$ are the values of the covariates during that time period, and the $\beta_0, \beta_1, \beta_2, \ldots \beta_n$, are the coefficients to be estimated by the modelling.

In a negative binomial model, it allows for extra-poisson variation due to other variables not included in the model. If the K in the negative binomial model equals to zero, the negative binomial reduces to the Poisson model. The larger the value of K, the more variability there is in the data over and above that associated with the mean μ_i.

Negative Binomial Regression (NB)

Mortality data which is an example of count data, often exhibit larger variance than would be expected from the Poisson assumption[9]. There are a number of strategies for accommodating overdispersion. One of the approaches is to retain the use of Poisson error distributions but allow the estimation of a

value of dispersion parameter from data rather than defining it to be unity for these distributions. The estimate is usually the residual deviance divided by its degrees of freedom. Parameter estimates remain the same but the parameter standard errors are increased by multiplying them by the square root of the estimated dispersion parameter. This model is called as Quasi-poisson model. Greenwood et al suggested another approach to the problem was a model in which μ was random variable with a gamma distribution leading to a negative binomial distribution (NB) for the count data. NB regression handles dispersion issues by modelling the dispersion parameter of the response variable. The relationship between variance and mean for NB distribution has the form of:

Var (Y_i)= $\mu + k\mu^2$

Residuals for Generalized Linear Models

Poisson and NB regressions are both categorized as Generalized Linear Models (GLMs). It is very important when fitting GLMs to look at suitable residuals to assess assumptions. The popular measures of the adequacy of the model fit are deviance residuals and Pearson Chi-Square (χ^2) residuals. If the statistical model is correct than both quantities are asymptotically distributed as χ^2 statistics with n-p degrees of freedom (df), where n is the size sample and p is the number of fitted parameters including the intercept. Due to that, if the regression model is adequate, the expected value of both the deviance and the Pearson Chi-Square is equal or close to n-p (the scaled deviance close to 1 or the scaled Pearson Chi-Squared which is χ^2/df is close to 1). Otherwise, the validity of the model will be questioned.

The two residuals useful in assessing fitted GLMs are deviance residuals and Pearson residuals.

The deviance residuals are defined as

$$r_i^D = \text{sign}\left(y_i - \hat{\mu}_i\right)\sqrt{d_i}$$

Where di is the contribution of the ith subject to the deviance, with total deviance given by

$$D = \sum\left(r_i^D\right)^2$$

The Pearson residuals are defined as the contribution of the ith subject to the Pearson x^2 statistic,

$$r_i^P = \frac{\left(y_i - \hat{\mu}_i\right)}{\sqrt{V\left(\hat{\mu}_i\right)}}$$

So that the $x^2 = \sum\left(r_i^P\right)^2$

Both the Pearson and deviance statistics can be used for detecting observations not well fitted by the model. The deviance residuals are more commonly used because their distribution tends to be closer to normal than that of the Pearson residuals.

ANALYSIS AND RESULTS

Below is the illustration on the application of Poisson regression model and NB model in epidemiologic time-series studies. We consider using an example from a case study done in Klang Valley, Malaysia. The data consist of daily measurements of air pollution level which is carbon monoxide (*cokl*), meteorological variables which are mean temperature of the day (*mean24s*) and mean relative humidity (*meanrels*), holiday indicator (*holiday*), day of the week (*daywk*) and natural mortality counts from various causes (*nontrkl*). Although the techniques that we describe are particularly useful for analyzing air pollution and health data; they are certainly applicable in other areas.

Table 1. Summary of variables used in the analysis of the data for Klang Valley.

No	Variables	Description
1	cokl	Daily average of Carbon monoxide levels in ppm. This is the current value of single day exposure
2	meanrels	Daily relative humidity
3	Mean24s	Daily average temperature
4	daywk	Day of the week indicator variables
5	holiday	Complete environmental time series data from 1st January 2000 to 31st Dec 2002
6	nontrkl	Daily natural mortality counts

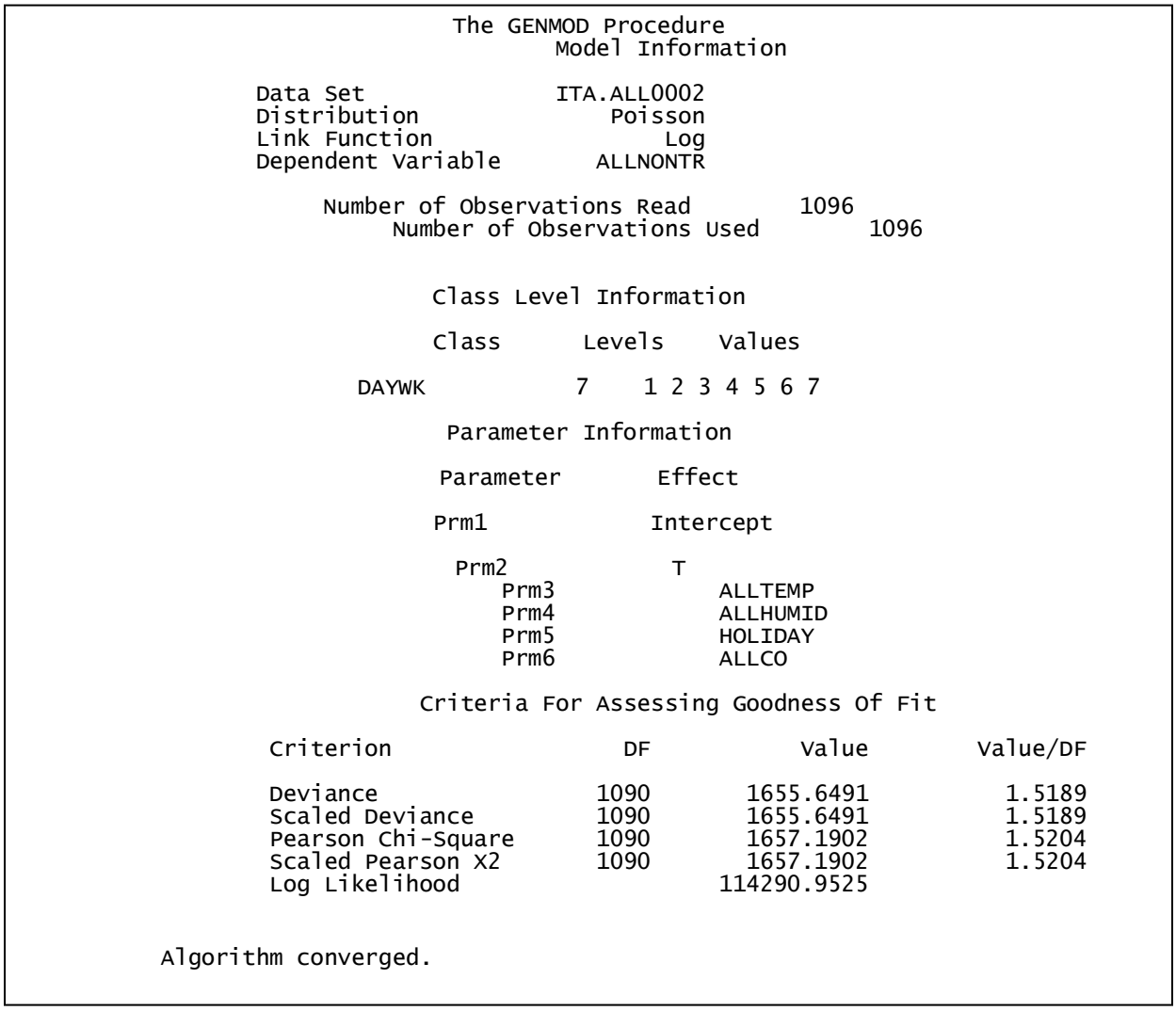
The PROC GENMOD of SAS can fit wide range of generalized linear models. The following SAS statements use PROC GENMOD to fit the Poisson regression

$\log(\mu i) = \log ti + \beta_0 + \beta_1 (\text{alltemp}) + \beta_2 (\text{allhumid}) + \beta_3 (\text{holiday}) + \beta_4 (\text{allco})$

to the Airpoll data with temperature, humidity, holiday indicator and carbon monoxide level as the explanatory variables:

Figure 1 which shows the goodness of fit of the model, indicate that the value of deviance bigger than the n-p or the degree of freedom, suggests that the model is overdispersed. Likewise, the Pearson Chi-square statistic with the value of 1.5204 which is bigger than 1, also indicated that overdispersion existed in the Poisson regression model. The variance (60.412) for mortality data was also found to be larger than the mean (39.105). Therefore, the assumption under the Poisson regression model was found to be violated.

```
proc genmod data=Airpoll.all0002;
class daywk;
 model allnontr=t alltemp allhumid holiday
allco/
       dist=poisson
                link=log;
estimate 'beta CO'   allco   1  / exp;
output out=pgmout reschi=rs;
  run;
```



The GENMOD Procedure
Model Information

Data Set	ITA.ALL0002
Distribution	Poisson
Link Function	Log
Dependent Variable	ALLNONTR

Number of Observations Read 1096
Number of Observations Used 1096

Class Level Information

Class	Levels	Values
DAYWK	7	1 2 3 4 5 6 7

Parameter Information

Parameter	Effect
Prm1	Intercept
Prm2	T
Prm3	ALLTEMP
Prm4	ALLHUMID
Prm5	HOLIDAY
Prm6	ALLCO

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	1090	1655.6491	1.5189
Scaled Deviance	1090	1655.6491	1.5189
Pearson Chi-Square	1090	1657.1902	1.5204
Scaled Pearson X2	1090	1657.1902	1.5204
Log Likelihood		114290.9525	

Algorithm converged.

Figure 1. Output from Poisson regression

The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr >ChiSq
Intercept	1	2.2715	0.3203	1.6438	2.8992	50.30	<.0001
T	1	0.0000	0.0000	-0.0000	0.0000	0.32	0.5712
ALLTEMP	1	0.0383	0.0079	0.0229	0.0537	23.77	<.0001
ALLHUMID	1	0.0037	0.0015	0.0008	0.0066	6.17	0.0130
HOLIDAY	1	-0.0133	0.0233	-0.0589	0.0323	0.33	0.5671
ALLCO	1	0.0206	0.0130	-0.0049	0.0461	2.51	0.1130
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Contrast Estimate Results

Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
beta coall	0.0206	0.0130	0.05	-0.0049	0.0461	2.51	0.1130
Exp(beta coall)	1.0208	0.0133	0.05	0.9951	1.0472		

Figure 2. Continuation from Figure 1

From Figure 1, under the 'Analysis of Parameter Estimates', we can clearly see that temperature and humidity were found to be significant towards mortality but the trend, carbon monoxide and holiday indicator were not significant.

To run NB regression, by specifying option DIST=NB in the model statement. The Deviance has an approximately chi-square distribution with n-p degrees of freedom, where n is the number of observations, p is the number of predictor variables (including intercept), and the expected value of a chi-square random variable is equal to the degrees of freedom. If our model fits the data well, the ratio of the Deviance to DF which is the Value/DF should be about one. Large ratio values may indicate model misspecification or an over-dispersed response variable.

In Figure 3, it was reported that the dispersion parameter (k) was 0.0130 and the scaled deviance was 1.0110, which was about one. This shows that the problem of overdispersion has been settled when we applied the data using negative binomial model. Temperature and humidity were still found to be significant in the model.

Class Level Information

Class	Levels	Values
DAYWK	7	1 2 3 4 5 6 7

Parameter Information

Parameter	Effect
Prm1	Intercept
Prm2	T
Prm3	ALLTEMP
Prm4	ALLHUMID
Prm5	HOLIDAY
Prm6	ALLCO

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	1090	1101.9752	1.0110
Scaled Deviance	1090	1101.9752	1.0110
Pearson Chi-Square	1090	1100.6101	1.0097
Scaled Pearson X2	1090	1100.6101	1.0097
Log Likelihood		114344.0120	

Algorithm converged.

The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	2.2807	0.3921	1.5122	3.0493	33.83	<.0001
T	1	0.0000	0.0000	-0.0000	0.0000	0.24	0.6260
ALLTEMP	1	0.0381	0.0096	0.0192	0.0569	15.65	<.0001
ALLHUMID	1	0.0037	0.0018	0.0001	0.0073	4.03	0.0446
HOLIDAY	1	-0.0134	0.0285	-0.0693	0.0425	0.22	0.6376
ALLCO	1	0.0204	0.0159	-0.0108	0.0516	1.64	0.1998
Dispersion	1	0.0130	0.0016	0.0097	0.0162		

NOTE: The negative binomial dispersion parameter was estimated by maximum likelihood.

Figure 3. Output from Negative Binomial Regression

Contrast Estimate Results

Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
beta coall	0.0204	0.0159	0.05	-0.0108	0.0516	1.64	0.1998
Exp(beta coall)	1.0206	0.0162	0.05	0.9893	1.0530		

Figure 4. Continuation from Figure 3

The PROC UNIVARIATE was done to plot the probability plot to assess the GLMs using the Pearson (CHI) residuals.

```
proc univariate data=pgmout noprint;

var rs;

probplot rs / normal;

run;
```

The probability plots were shown in Figure 5 and Figure 6 The plot associated with Poisson regression shows a bigger residuals, ranging from -6 to 6, while the plot associated with Negative Binomial regression slightly shows a smaller residuals ranging from -4 to 4 Therefore, the plot associated with negative binomial errors model appears to be satisfactory[10].

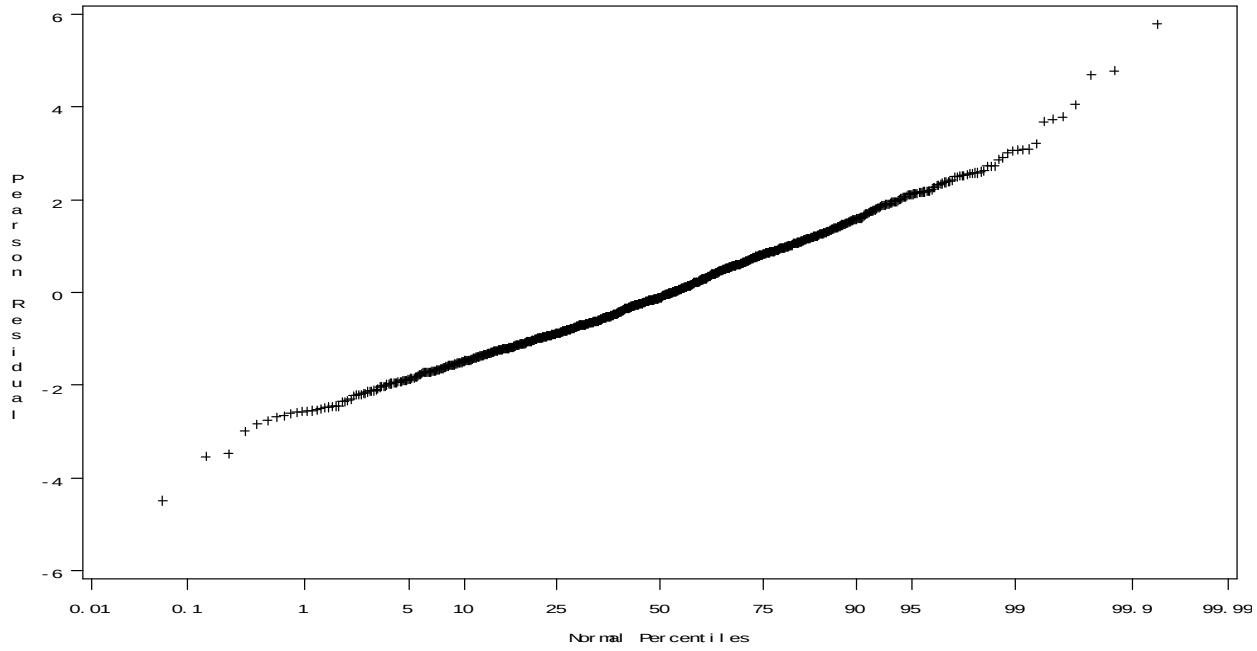


Figure 5. Normal probability plot of Pearson residuals from the Poisson regression model

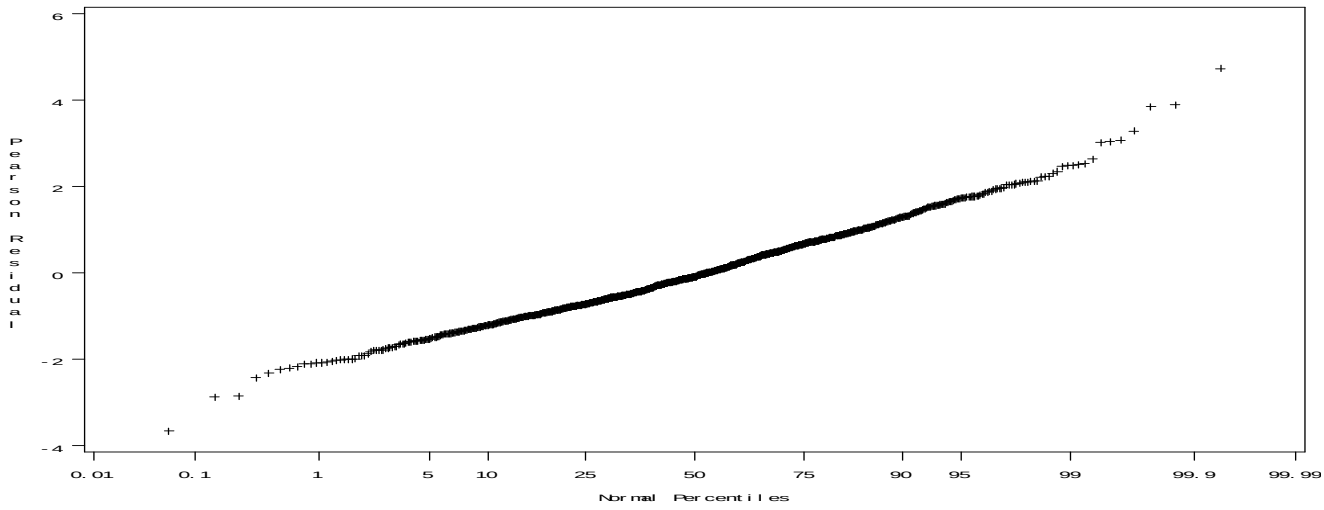


Figure 6. Normal probability plot of Pearson residuals from the Negative Binomial regression model

The probability plots were shown in Figure 4.3 and Figure 4.4. The plot associated with Poisson regression shows bigger residuals, ranging from -6 to 6, while the plot associated with NB regression slightly shows a smaller residuals ranging from -4 to 4 Therefore, the plot associated with NB errors model appears to be satisfactory[10].

CONCLUSION

Ignoring overdispersion in the analysis would lead to underestimation of standard errors and consequently will effects the level of significance in hypothesis testing. Therefore by using the inappropriate model for count data can change a statistical inference[11]. The overdispersion must be accounted for by the analysis method suitable for the data. In this air pollution and health study, we suggest the NB regression model provides a better account of the probability distribution of the data than the Poisson regression model.

ACKNOWLEDGEMENT:

This study has been supported by a MOH research grant (Non- CAM 07-0140). We would like to thank the Director General of Health Malaysia for his permission to publish this paper. We wish to gratefully acknowledge our appreciation to Department of Environment, Statistic Department and Meteorological Department of Malaysia for providing the data and to the Director of Institute for Medical Research for her permission to conduct this study. We also wish to thank Dr Amal Nasir for his administrative support for this study.

REFERENCES:

1. McCullagh, P., & Nelder, J.A. *Generalized Linear Models (2nd ed.)* 1989; London: Chapman & Hall.

2. Ismail, N., & Jemain, A.A. Handling overdispersion with negative binomial and generalized poisson regression models. Casualty Actuarial Society Forum Arlington, Virginia. 2007.

3. Chuang, K.J., Chan, C.C., Lee, C.T., & Tang, C.S. The effect of urban air pollution on inflammation, oxidative stress, coagulation and autonomic dysfunction in young adults. *Am J Respir Crit Care Med* 2007; **176**: 370-376.

4. Morgan, G., Corbett, S., Wlodarczyk, J., Lewis, P. Air pollution and daily mortality

in Sydney, Australia, 1989 through 1993. *American Journal of Public Health* , 1998; **88**(5), 759-764.

5. Michelozzi, P., Forastiere, F., Fusco, D., Perucci, CA., Ostro, B., Ancona, C., *et al*. Air pollution and daily mortality in Rome, Italy. *Occup Environ Med* 1998; **55**(9), 605-610.

6. Boldo, E., Medina, S., LeTertre, A., Hurley, F., Mucke, H.G., Ballester, F. et al. Apheis: Health impact assessment of long-term exposure to PM2.5 in 23 European cities. *European Journal of Epidemiology* 2006; **21**: 449-458.

7. Wong, C.M., Ou, C.Q., Chan, K.P., Chau YK, Thach TQ, Yang L., *et al*. The effects of air pollution on mortality in socially deprived urban areas in Hong Kong, China . *Environmental Health Perspectives,* 2008; **116**(9), 1189-1194.

8. DeGroot, M.H. *Probability and Statistics*. Massachusetts :Addison-Wesley.1986.

9. Greenwood, M., & Yule, G.U. An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents. *Journal of the Royal Statistical Society,* 1920; **83**, 255-279.

10. Geoff, D. & Brian, S.E. *Statistical Analysis of Medical Data Using SAS*, 2005; London: Chapman and Hall.

11. Alex, P. *Analysis of count data using the SAS system*. Paper presented at the SUGI Conference, Long Beach, California, 2001.