Iraqi Journal of Statistical Science (15) 2009
p.p. [1-16]

The Prediction of the Maternal and Fetal Blood Lead Level via Generalized Linear Model

Zakaria Y. AL-Jammal*

Abstract

Generalized linear models (GLMs) are generalization of the linear regression models, which allow fitting regression models to response variable that is non normal and follows a general exponential family. The aim of this study is to encourage and initiate the application of GLMs to predict the maternal and fetal blood lead level. The inverse Gaussian distribution with inverse quadratic link function is considered. Four main effects were significant in the prediction of the maternal blood lead level (pica, smoking of mother, dairy products intake of mother, calcium intake of mother), while in the prediction of the fetal blood lead level two main effects showed significance (dairy products intake of mother and hemoglobin of mother).

Keywords: Generalized linear models, Exponential family, Inverse Gaussian distribution, Link functions

* * Asst. Lecturer / Statistics and Informatics Dept./ Computer Science and Mathematics College / Mosul University / Mosul / IRAQ.

Received:24 /12 /2008 _____Accepted:29 /3 / 2008

التنبؤ بمستوى الرصاص في دم الأم والجنين باستخدام النموذج الخطي المعمم

الملخص

تعتبر النماذج الخطية المعممة تعميما لنماذج الانحدار الخطية عندما لا يتبع متغير الاستجابة التوزيع الطبيعي ويعود إلى العائلة الاسية. الهدف من هذا البحث هو استخدام النماذج المعممة للتنبؤ بمستوى الرصاص في دم الأم والجنين. تم اعتبار توزيع كاوس المعكوس هو التوزيع الملائم لمتغير الاستجابة ثم تم استخدام الدالة التربيعية العكسية كدالة ربط. أربعة متغيرات أساسية أظهرت تأثيرات معنوية عند التنبؤ بمستوى الرصاص في دم الأم في حين عند التنبؤ بمستوى الرصاص في دم الجنين تبين بأن هناك متغيرين فقط اظهرا تأثيرا معنويا.

1-Introduction

Generalized linear models (GLMs), as the name implies, are generalizations of the classical linear regression model. The classical linear model assumes that the mean of the response variable $\boldsymbol{y}$ is a linear function of a set of predictor variables (Hardin & Hilbe, 2007), and that the response variable is continuous and normally distributed with constant variance. As a matter of fact, in many applications, the response variable is categorical or consists of counts or is continuous but non normal, so the ordinary least square method can't be applied to find the regression models (De Jong & Heller, 2008). Generalized linear models were introduced by Nelder and Wedderburn in 1972 to address those limitations. GLMs are a family of models developed for regression models with non normal response variable. In the GLMs the mean of the response variable is

modeled as a monotonic nonlinear transformation of a linear function of the predictor variables.The inverse of the transformation $\boldsymbol{g}$ is known as the link function.

Many applications had been done using GLM. (Vidoni, 2003), (Jiao and Chen, 2004), (Zhukovskaya, 2007).

An example of non normal continuous distribution that has many applications is the inverse Gaussian distribution. It is skewed, takes on only positive values, and its variance is a function of its mean. It is used to model a wide variety of response variables that can take on only positive values, such as income, insurance, survival time,…etc. Models with inverse Gaussian distributed response variables can be models within a GLM framework.

This paper focused on the application of the GLM to predict the maternal and fetal blood lead level, in which the inverse Gaussian distribution with inverse quadratic link function is considered. This article has the following structure. The second section contains the description of the exponential family. The elaboration of the GLMs is presented in the third section. The used distribution for analyzing and predicting maternal and fetal blood lead level are considered in the fourth section. In the fifth and sixth sections the application and its results and the conclusion were given respectively.

2- Exponential Family of Distributions

An important concept underlying GLM is the exponential family of distributions. Members of the exponential family of distributions all have probability density functions for a response y that can be expressed in the form

$$f(y,\theta,\phi)=c(y,\phi)\,Exp\left\{\frac{y\theta-b(\theta)}{a(\phi)}\right\} \qquad \text{.....(1)}$$

where $a(\cdot),b(\cdot)$, and $c(\cdot)$ are specific functions. The parameter θ is a natural location parameter, and ϕ is often called a dispersion parameter. The binomial, Poisson, normal, gamma, and inverse Gaussian distributions are members of this family. (Myers et al., 2002). Here are some properties of the exponential family:

$$E\left(\frac{\partial \ln L(\theta)}{\partial\theta}\right)=0 \qquad \text{.....(2)}$$

$$\mu=E(y)=b'(\theta)=\frac{\partial b}{\partial\mu}\bullet\frac{\partial\mu}{\partial\theta} \qquad \text{.....(3)}$$

$$\text{var}(y)=b''(\theta)a(\phi),b''(\theta)=\frac{\partial^2 b}{\partial^2\mu}\left(\frac{\partial\mu}{\partial\theta}\right)^2+\frac{\partial b}{\partial\mu}\bullet\frac{\partial^2\mu}{\partial\theta^2} \qquad \text{.....(4)}$$

3- Generalized Linear Models

The theory and use of GLMs were introduced by Nelder and Wedderburn (1972). They were developed to allow us to fit regression models for univariate response data not normally distributed. The idea of GLMs is defined in terms of a set of

independent random variables $y_1, y_2, ..., y_n$ each with a distribution from the (1).

There are three components specify a GLM.

1- The random component consists of a response variable y with independent observations ($y_1, y_2, ..., y_n$) from a distribution in the canonical exponential family.

2- The systematic component relates a vector ($\eta_1, \eta_2, ..., \eta_n$) to explanatory variables through a linear model. Let x_i denote the value of predictor k, then

$$\eta = X'\beta = \hat{\beta}_\circ + \sum_{i=1}^{k} \hat{\beta}_i x_i \qquad \text{.....(5)}$$

This linear combination of explanatory variables is called the linear predictor.

3- The link function component connects the random and systematic component. Let $\mu_i = E(y_i), i = 1,2,...,n$, the model links μ_i to η_i, so the link function is

$$\eta_i = g(\mu_i), \quad i = 1,2,....,n \qquad \text{.....(6)}$$

where g is a monotonic differentiable function. The term link is derived from the fact that the function is the link between the mean and the linear predictor (Myers et al., 2002) .The expected response is

$$E(y_i) = g^{-1}(\eta_i) = g^{-1}(x_i'\beta), i = 1,2,....,n \qquad \text{.....(7)}$$

One way of assessing the adequacy of a model is to compare it with a more general model with the maximum number of

parameters that can be estimated. This is called a saturated model, which is a generalized linear with the same distribution and same link function as the models of interest. We define a measure of the fit of the model to the data as twice the difference between the log likelihoods of the model of interest and the saturated models. Since this difference is a measure of the deviation of the model of interest from a perfectly fitting model, this measure is called the deviance. The deviance, D, is given by

$$D = \sum_{i=1}^{n} 2\left[y\{\theta(y_i) - \theta(\mu_i)\} - b\{\theta(y_i)\} + b\{\theta(\mu_i)\}\right] \quad \ldots\ldots(8)$$

In fitting a particular model, we seek the values of the parameters that minimize the deviance. A good rule of thumb is that the lack of fit be good when deviance/ (n-p) is less than 1.0 (Myers et al., 2002).

The maximum likelihood estimates of the parameter β in the linear predictor can be obtained by using iterative weighted least squares (McCullagh & Nelder, 1989).

4- Inverse Gaussian Distribution

The inverse Gaussian distribution is a positively skewed continuous distribution having two parameters μ and σ^2. Several alternative parameterization appear in the literature. In our paper we use the following p.d.f.

$$f(y;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\prod y^3}} Exp\left\{-\frac{(y-\mu)^2}{2\mu^2\sigma^2 y}\right\}, \quad \sigma^2, \mu, y > 0 \quad (9)$$

The mean and variance are $E(y)=\mu$, $var(y)=\sigma^2\mu^3$ where σ^2 is the dispersion parameter (De Jong and Heller, 2008).

From equation (1), the exponential form is

$$f(y,\theta,\phi)=c(y,\phi)\,Exp\left\{\frac{y/\mu^2-2/\mu}{2\sigma^2}\right\} \qquad \ldots\ldots\ldots\ldots(10)$$

where $c(y,\phi)=-\frac{1}{2\sigma^2}\left\{1/y+\sigma^2\,ln(2\prod y^3\sigma^2)\right\}$,

$\theta=1/2\mu^2$,

$a(\phi)=-\sigma^2$, and $b(\theta)=1/\mu$.

The log likelihood function of (10) may be derived as:

$$L=\sum_{i=1}^{n}\left\{\frac{y_i/2\mu_i^2-1/\mu_i}{-\sigma^2}+\frac{1}{2\sigma^2 y_i}-\frac{ln(2\prod y_i^3\sigma^2)}{2}\right\} \ldots\ldots.(11)$$

The link function is

$$\eta_i=\theta_i$$
$$=1/\mu^2 \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots..(12)$$

The sign and coefficient value are typically dropped from (12). (Hardin and Hilbe, 2007).

In GLMs the mean is related to explanatory variables. Thus the mean varies with the explanatory variables. As the mean varies, so does the variance, through $v(\mu)$. So, the variance function, $v(\mu)$, is

$$v(\mu)=\frac{var(y)}{a(\phi)} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(13)$$

Now, the $v(\mu)$ of the inverse Gaussian distribution is

$$v(\mu) = -\mu^3 \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(14)$$

Finally, the deviance function, D, is calculated from the saturated model and the model log-likelihood formulas

$$D = 2\sigma^2 \sum_{i=1}^{n} \left\{ \frac{y_i / 2y_i^2 - 1/y_i}{-\sigma^2} - \frac{y_i / 2\mu_i^2 - 1/\mu_i}{-\sigma^2} \right\}$$

$$D = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{y_i \mu_i^2} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(15)$$

5- Application

Great attention has been directed to study maternal and fetal blood lead levels since pregnant women and young children are the most sensitive populations to the lead exposure from various sources.(AL-Mola, 2007).

The data was taken from AL-Mola (2007), which are representing 350 pregnant women. The obtained data were taken directly from mothers themselves through questionnaire form. In this study we have two separated response variables, one for the maternal blood lead level (MBLL) and the other for the fetal blood lead level (FBLL). Many predictor variables are taken for both response variables.

5-1 Prediction of the Maternal Blood Lead Level

High levels of lead in pregnant women arise from various affected variables. These explanatory variables are:

x_1 (residence, 1 for urban and 0 for rural), x_2 (Pica, 1 for No and 2 for yes), x_3 (Physical activity), x_4 (Chronic disease, 1 for No and 2 for Yes), x_5 (Smoking of mother), x_6 (Smoking of father), x_7 (Diary products intake of mother), and x_8 (Calcium intake of mother).

The GLM equation is

$$\hat{y}_{MBLL} = \hat{\beta}_\circ + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_6 + \hat{\beta}_7 x_7 + \hat{\beta}_8 x_8$$

Figure (1) shows that the response variable y_{MBLL} has a distribution with a heavy right tail, and thus an inverse Gaussian distribution be appropriate. (The value of χ^2=6.893, and $\chi^2(0.05, 8)$ =15.507)

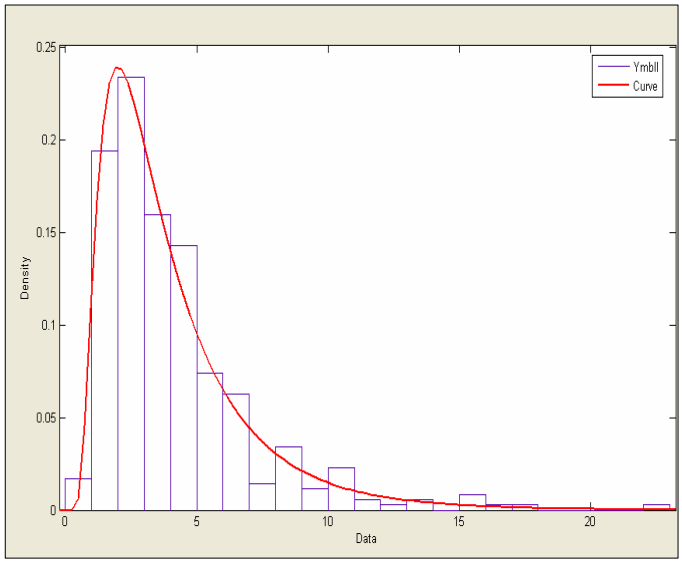

Figure (1): The histogram of the MBLL variable

Using the function ***glm*** in STATA 10 program, the obtained results are shown in table (1).

Table (1): The GLM results using inverse Gaussian distribution.

No. of Iteration=6 Scale parameter=0.1111

Optimization : ML Residual df = 341 Deviance = 32.656998

No. of Observation=350 AIC = 5.523189

Log likelihood = -957.55802

Coef.	Coef. value	Std.Err.	t	P>$\|t\|$	95% Conf. Int.	
Const.	0.08289	0.0608	1.36	0.173	-0.036276	0.202
x_1	-0.00172	0.0068	-0.25	0.8	-0.015	0.0466
x_2	0.03079	0.00808	3.81	0.00	-0.0289	0.007
x_3	-0.01095	0.00916	-1.19	0.232	-0.0166	0.058
x_4	0.02068	0.019	1.08	0.278	-0.00012	0.01715
x_5	0.008515	0.0044	1.93	0.05	-0.00641	0.0147
x_6	0.004158	0.0053	0.77	0.441	0.00562	0.02449
x_7	0.01505	0.00481	3.13	0.002	-0.1116	-0.0515
x_8	-0.08158	0.0153	-5.32	0.000	-0.0362	0.202

The predicted equation is

$$\hat{y}_{MBLL} = 0.08289 + 0.03079x_2 + 0.008515x_3 + 0.015x_7 - 0.08158x_8$$

From Deviance = 32.656998/(Residual df = 341) the lack of fit for this equation is good since it equal to 0.0957 < 1. The normal probability plot of the residuals and the scatter plot between the

deviance residual and the fitted value are shown in figure (2) and (3).

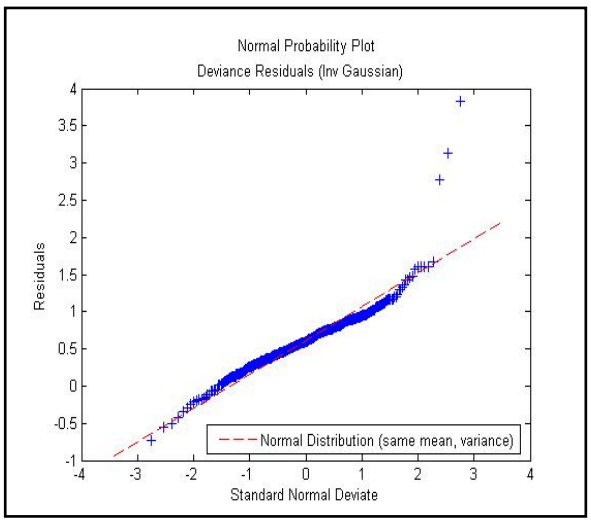

Figure (2): Normal probability plot of the residuals

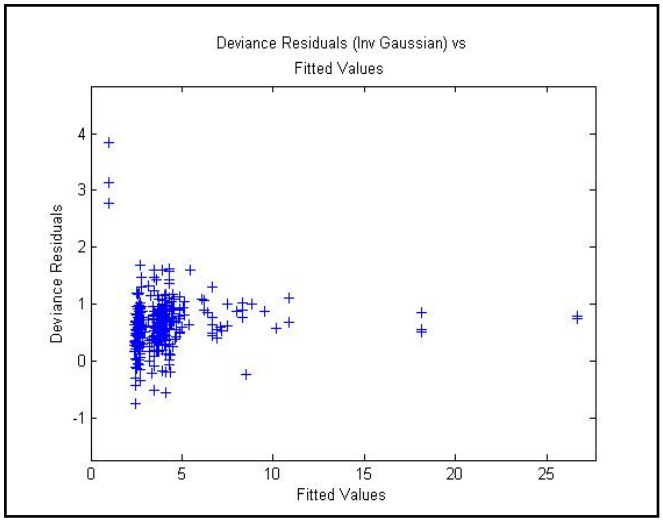

Figure (3): Scatter between deviance and fitted value

5-2 Prediction of the Fetal Blood Lead Level

Maternal blood is one of the important sources of the lead exposure for fetus and infant. There is no apparent maternal -fetal barrier to lead, therefore fetal blood lead level (FBLL) are nearly equal to MBLL.(AL-Mola,2007). The explanatory variables are: x_1 (smoking of mother), x_2 (dairy products intake of mother), x_3 (blood pressure of mother), and x_4 (hemoglobin of mother).

The GLM equation is

$$\hat{y}_{MBLL} = \hat{\beta}_{\circ} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$$

Figure (4) shows the histogram of the response variable y_{FBLL}, thus an inverse Gaussian distribution be appropriate (The value of χ^2=14.9, and χ^2 (0.05, 8) =15.507). Using the function ***glm*** in STATA 10 program, the obtained results are shown in table (2).

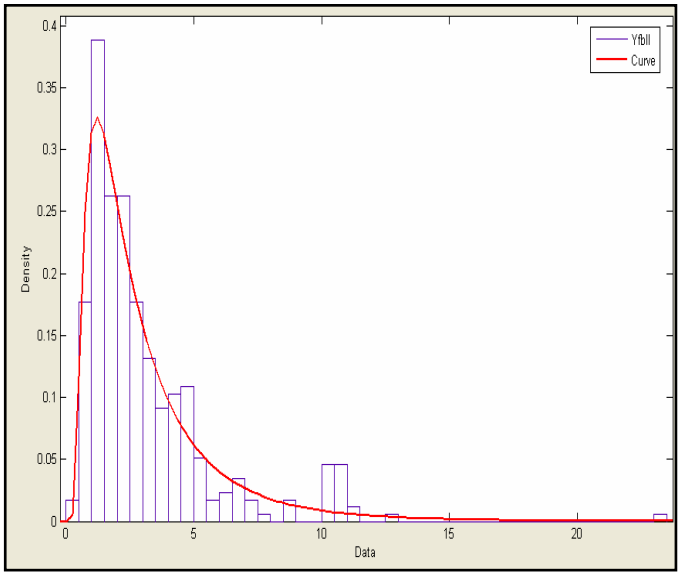

Figure (4): The histogram of the MBLL variable

Table (2): The GLM results using inverse Gaussian distribution.

No. of Iteration=4 Scale parameter=0.24708

Optimization : ML Residual df = 341 Deviance = 74.37703

No. of Observation=350 AIC = 4.56926

Log likelihood = -794.7019908

Coef.	Coef. value	Std.Err.	t	P>$\|t\|$	95% Conf. Int.	
Const.	-0.32515	0.09569	-3.40	0.001	-0.5127	-0.1376
x_1	0.03269	0.0104	1.66	0.098	0-0.00601	0.0623
x_2	0.041919	0.02707	4.03	0.000	0.02153	0.0623
x_3	-0.04032	0.006781	-1.49	0.136	-0.0933	0.01273
x_4	0.028646	0.09569	4.22	0.000	0.015355	0.0419366

The predicted equation is

$$\hat{y}_{MBLL} = -0.325 + 0.0419x_2 + 0.0286x_4$$

From Deviance = 74.37703/(Residual df = 341) the lack of fit for this equation is good since it equal to 0.281< 1. The normal probability of the residuals and the scatter plot between the deviance residual and the fitted value are shown in figure (5) and (6) respectively.

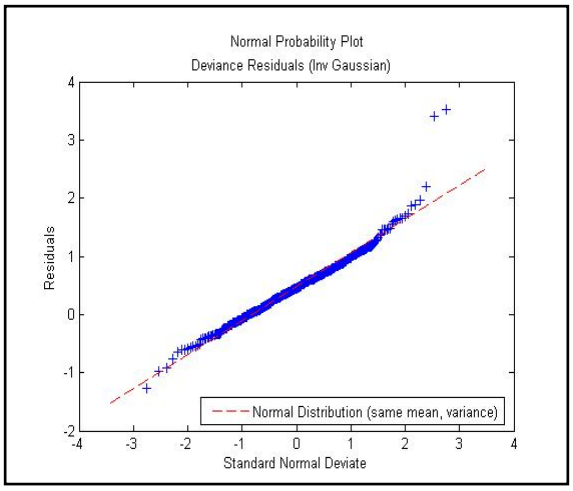

Figure (5): Normal probability plot of the residuals

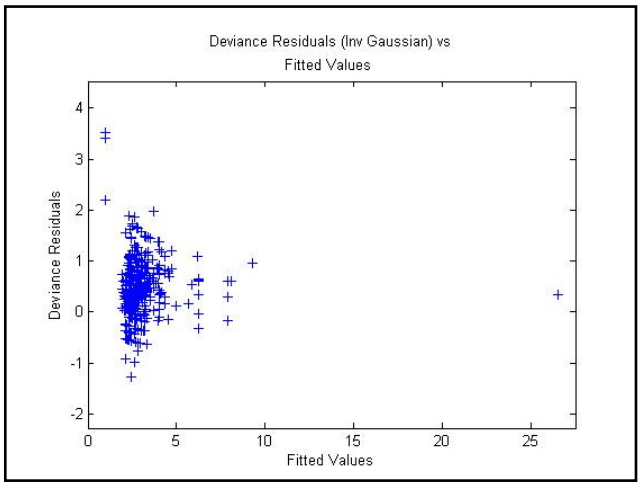

Figure (6): Scatter between deviance and fitted value

6- Conclusion

The generalized linear regression models for predicting MBLL and FBLL assuming the inverse Gaussian distribution as the response distribution are considered. From table (1), four explanatory variables (pica, smoking of mother, dairy products intake of mother, calcium intake of mother) have shown significant effects, while from table (2), dairy products intake of mother and hemoglobin of mother show main effects. The normal probability plot for the residuals for both response variables are represented on figure (2) and (5) which show that the residuals have normal distribution. The scatter plot between deviance residuals and fitted values for both MBLL and FBLL are shown in figure (3) and (6), which points out that the variance is not constant.

7- References

1- Al-Mola, Z. W., 2007, "Maternal and Umbilical Cord Blood Lead Levels and Pregnancy Outcomes Hospital Based Enquiry", M.Sc thesis, College of Medicine, Mosul University.

2- De Jong, P. & Heller, G. Z., 2008, "Generalized Linear Models for Insurance Data", Cambridge University Press.

3- Hardin, J. W. & Hilbe, J., 2007,"Generalized Linear Models and Extensions"2nd edition, Stata Press.

4- Jiao, Y. & Chen, Y., 2004, "An application of Generalized Linear Models in Production Model and Sequential Population Analysis", Fisheries Research, No.70, pp. 367-376.

5- McCullagh, P. & Nelder, J.A., 1989,"Generalized Linear Models", 2nd edition, Chapman and Hall Inc., London.

6- Myers, R. H., Montgomery, D. C., and Vining, G. G., 2002, " Generalized Linear Models with Applications in Engineering and the Sciences", John Wiley & Sons, Inc. New York.

7- Nelder, J. A. & Wedderburn, R. W. M., 1972, " Generalized Linear Models", Journal of Royal Statistics, Soc.A, 135, pp.370-384.

8- Vidoni, P., 2003, "Prediction and Calibration in Generalized Linear Models", The Annals of Institute of Statistical Mathematics, Vol.55, No.1, pp.169-185.

9- Zhukovskaya, C., 2007, "Use of the Generalized Linear Model in Forecasting the Air Passengers' Conveyances from EU Countries", Journal of Computer Modeling and New Technologies, Vol.11, No.1, pp.62-72.