International Journal of Advanced Statistics and Probability, 5 (2) (2017) 65-69

International Journal of Advanced Statistics and Probability

Website: www.sciencepubco.com/index.php/IJASP
doi: 10.14419/ijasp.v5i1.7615
Research paper

Prediction of blood lead level in maternal and fetal using generalized linear model

Zakariya Yahya Algamal [1]*, Intisar Ibrahim Allyas [2]

[1] *Department of Statistics and Informatics, University of Mosul, Mosul, Iraq*
[2] *College of Administration and Economics, Nawroz University, Kurdistan region, Iraq*
**Corresponding author E-mail: zakariya.algamal@gmail.com*

Abstract

Over the past decades, with advanced data collection techniques, a different type of data continues to appear in various biological, sciences, medical, social, and economical studies. Statistical modeling is essential in many scientific research areas because it explains the relationship between the response variable of interest and a number of explanatory variables. 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.*

1. Introduction

Generalized linear models (GLMs), as the name implies, are a generalizations of the classical linear regression model. The classical linear model assumes that the mean of the response variable y is a linear function of a set of predictor variables [1-7], 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 [8-15]. Generalized linear models were introduced by Nelder and Wedderburn in 1972 [16] 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 g is known as the link function.

Many applications used GLM [15, 17-23]. 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 conclusions 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

$$f(y,\theta,\varphi)=c(y,\varphi)Exp\left\{\frac{y\theta-b(\theta)}{a(\varphi)}\right\} \quad (1)$$

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 [16], [24-30]. Here some properties of the exponential family:

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

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

$$\text{var}(y)=b''(\theta)a(\varphi),b''(\theta)=\frac{\partial^2 b}{\partial^2\mu}(\frac{\partial\mu}{\partial\theta})^2+\frac{\partial b}{\partial\mu}\bullet\frac{\partial^2\mu}{\partial\theta^2} \quad (4)$$

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

$$\eta = X'\beta = \hat{\beta}_o + \sum_{i=1}^{k} \hat{\beta}_i x_i \tag{5}$$

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

1) 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

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

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

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

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

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

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) less than 1.0 [6].

The maximum likelihood estimates of the parameter β in the linear predictor can be obtained by using iterative weighted least squares [5].

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.

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

The mean and variance are $E(y) = \mu$, $\text{var}(y) = \sigma^2\mu^3$ where σ^2 is the dispersion parameter [2].

From equation (1), the exponential form is

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

Where

$c(y, \varphi) = -\frac{1}{2\sigma^2}\{1/y + \sigma^2 \ln(2\prod y^3\sigma^2)\}$, $\theta = 1/2\mu^2$, $a(\varphi) = -\sigma^2$, $b(\theta) = 1/\mu$.

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

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

The link function is

$$\begin{aligned} \eta_i &= \theta_i \\ &= 1/\mu^2 \end{aligned} \tag{12}$$

The sign and coefficient value are typically dropped from (12) [3].

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

$$v(\mu) = \frac{\text{var}(y)}{a(\varphi)} \tag{13}$$

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

$$v(\mu) = -\mu^3 \tag{14}$$

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

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

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

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 [1].

The data was taken from [1], 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 effected 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

$$\hat{y}_{MBLL} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_6 + \hat{\beta}_7 x_7 + \hat{\beta}_8 x_8$$

Fig. 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 χ^2 (0.05, 8) =15.507)

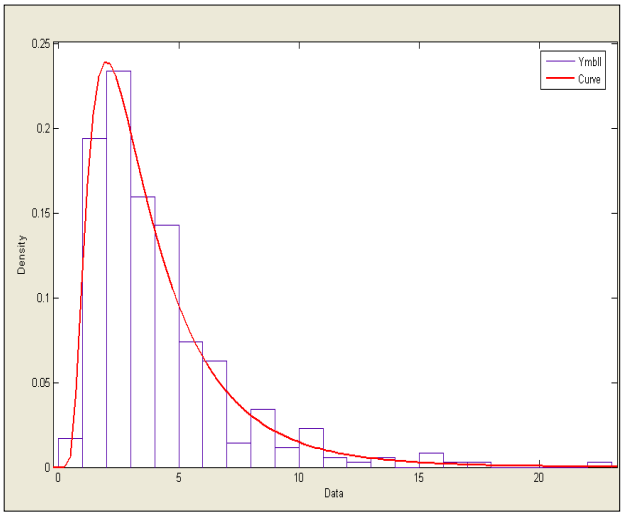

Fig. 1: The Histogram of the MBLL Variable.

Using the function *glm* in STATA 10 program, the obtained results showed 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 Figs. 2 and 3.

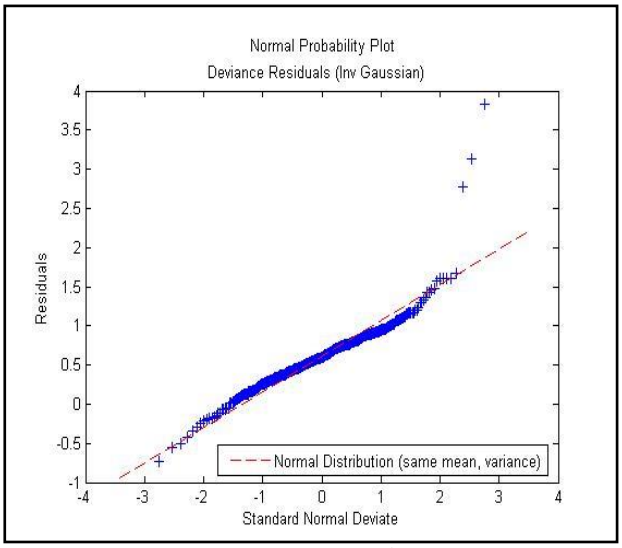

Fig. 2: Normal Probability Plot of the Residuals.

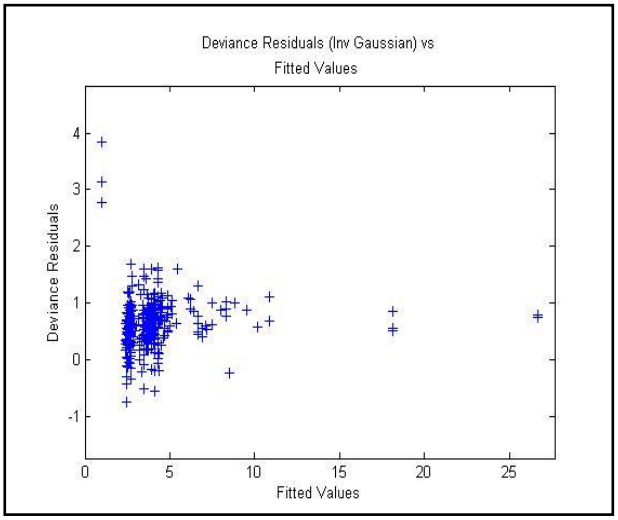

Fig. 3: Scatter between Deviance and Fitted Value.

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. 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

$$\hat{y}_{MBLL} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$$

Fig. 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 showed in Table 2.

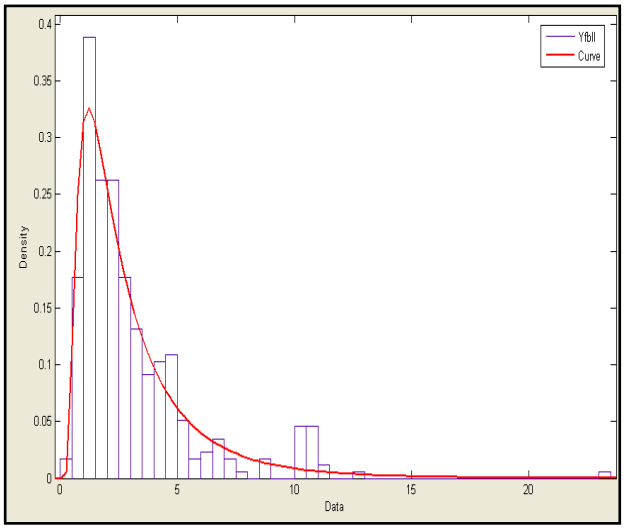

Fig. 4: The Histogram of the MBLL Variable.

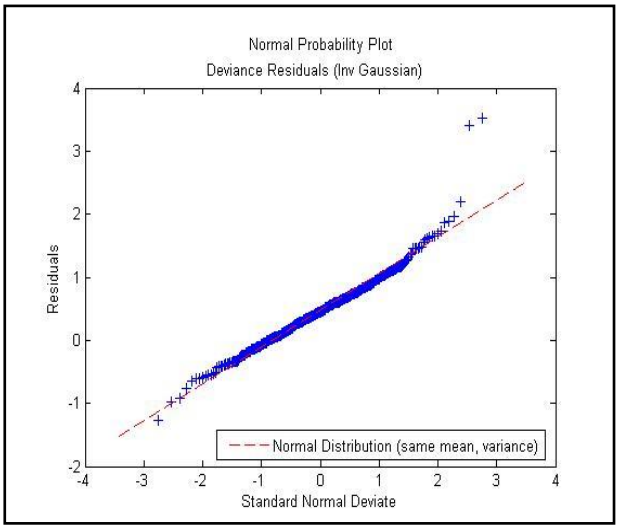

Fig. 5: Normal Probability Plot of the Residuals.

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	0.041936

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 Figs. 5 and 6, respectively.

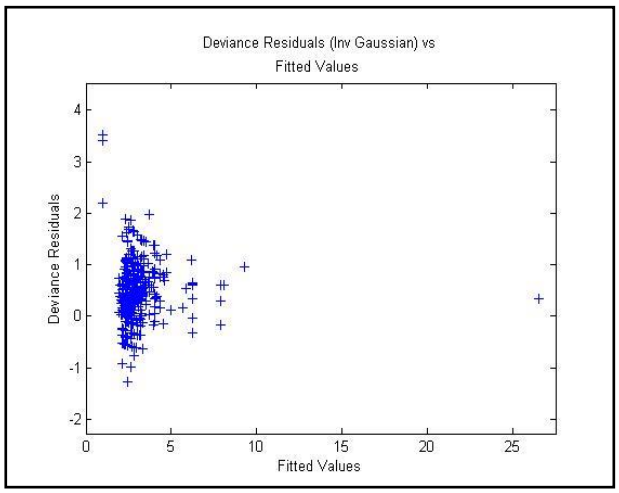

Fig. 6: Scatter between Deviance and Fitted Value.

6. Conclusion

The generalized linear regression models for the 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) shown significant effects, while from Table 2, dairy products intake of mother and hemoglobin of mother, show main effects. The normal probability plots for the residuals for both response variables are represented on Figs. 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 Figs. 3 and 6, which point out that the variance is not constant.

References

[1] A.M. Al-Fakih, Z.Y. Algamal, M.H. Lee, H.H. Abdallah, H. Maarof, M. Aziz, Quantitative structure-activity relationship model for prediction study of corrosion inhibition efficiency using two-stage sparse multiple linear regression, Journal of Chemometrics 30(7) (2016) 361-368. https://doi.org/10.1002/cem.2800.

[2] A.M. Al-Fakih, M. Aziz, H.H. Abdallah, Z.Y. Algamal, M.H. Lee, H. Maarof, High dimensional QSAR study of mild steel corrosion inhibition in acidic medium by furan derivatives, International Journal of Electrochemical Science 10 (2015) 3568-3583.

[3] Z.Y. Algamal, Exponentiated exponential distribution as a failure time distribution, IRAQI Journal of Statistical science 14 (2008) 63-75.

[4] Z.Y. Algamal, Paired Bootstrapping procedure in Gamma Regression Model using R, Journal of Basrah Researches 37(4) (2011) 201-211.
[5] Z.Y. Algamal, Diagnostic in Poisson regression models, Electronic Journal of Applied Statistical Analysis 5(2) (2012) 178-186.
[6] Z.Y. Algamal, Using maximum likelihood ratio test to discriminate between the inverse Gaussian and gamma distributions, International Journal of Statistical Distributions 1(1) (2017) 27-32.
[7] Z.W. Al-Mola, Maternal and Umbilical Cord Blood Lead Levels and Pregnancy Outcomes Hospital Based Enquiry, M.Sc thesis, College of Medicine, Mosul University (2007).
[8] Z.Y. Algamal, H.T.M. Ali, An efficient gene selection method for high-dimensional microarray data based on sparse logistic regression, Electronic Journal of Applied Statistical Analysis 10(1) (2017) 242-256.
[9] Z.Y. Algamal, H.T.M. Ali, Bootstrapping pseudo - R2 measures for binary response variable model, Biomedical Statistics and Informatics 2(3) (2017) 107-110.
[10] Z.Y. Algamal, M.H. Lee, Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification, Expert Systems with Applications 42(23) (2015) 9326-9332. https://doi.org/10.1016/j.eswa.2015.08.016.
[11] Z.Y. Algamal, M.H. Lee, Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification, Computers in Biology and Medicine 67 (2015) 136-45. https://doi.org/10.1016/j.compbiomed.2015.10.008.
[12] Z.Y. Algamal, M.H. Lee, Penalized Poisson regression model using adaptive modified elastic net penalty, Electronic Journal of Applied Statistical Analysis 8(2) (2015) 236-245.
[13] Z.Y. Algamal, M.H. Lee, High dimensional logistic regression model using adjusted elastic net penalty, Pakistan Journal of Statistics and Operation Research 11(4) (2015) 667-676. https://doi.org/10.18187/pjsor.v11i4.990.
[14] Z.Y. Algamal, M.H. Lee, Adjusted adaptive lasso in high-dimensional Poisson regression model, Modern Applied Science 9(4) (2015) 170-176. https://doi.org/10.5539/mas.v9n4p170.
[15] Z.Y. Algamal, M.H. Lee, Applying penalized binary logistic regression with correlation based elastic net for variables selection, Journal of Modern Applied Statistical Methods 14(1) (2015) 168-179.
[16] J.A. Nelder, R.W.M. Wedderburn, Generalized Linear Models, Journal of Royal Statistics 135 (1972) 370-384. https://doi.org/10.2307/2344614.
[17] Z.Y. Algamal, M.H. Lee, A new adaptive L1-norm for optimal descriptor selection of high-dimensional QSAR classification model for anti-hepatitis C virus activity of thiourea derivatives, SAR and QSAR in Environmental Research 28(1) (2017) 75-90. https://doi.org/10.1080/1062936X.2017.1278618.
[18] Z.Y. Algamal, M.H. Lee, A.M. Al-Fakih, High-dimensional quantitative structure-activity relationship modeling of influenza neuraminidase a/PR/8/34 (H1N1) inhibitors based on a two-stage adaptive penalized rank regression, Journal of Chemometrics 30(2) (2016) 50-57. https://doi.org/10.1002/cem.2766.
[19] Z.Y. Algamal, M.H. Lee, A.M. Al-Fakih, M. Aziz, High-dimensional QSAR prediction of anticancer potency of imidazo[4,5-b]pyridine derivatives using adjusted adaptive LASSO, Journal of Chemometrics 29(10) (2015) 547-556. https://doi.org/10.1002/cem.2741.
[20] Z.Y. Algamal, M.H. Lee, A.M. Al-Fakih, M. Aziz, High-dimensional QSAR modelling using penalized linear regression model with L1/2-norm, SAR and QSAR in Environmental Research 27(9) (2016) 703-19. https://doi.org/10.1080/1062936X.2016.1228696.
[21] Z.Y. Algamal, M.H. Lee, A.M. Al-Fakih, M. Aziz, High-dimensional QSAR classification model for anti-hepatitis C virus activity of thiourea derivatives based on the sparse logistic regression model with a bridge penalty, Journal of Chemometrics (2017) e2889. https://doi.org/10.1002/cem.2889.
[22] Z.Y. Algamal, M.K. Qasim, H.T.M. Ali, A QSAR classification model for neuraminidase inhibitors of influenza A viruses (H1N1) based on weighted penalized support vector machine, SAR and QSAR in Environmental Research (2017) 1-12. https://doi.org/10.1080/1062936X.2017.1326402.
[23] P. De Jong, G.Z. Heller, Generalized Linear Models for Insurance Data, Cambridge University Press, UK, 2008. https://doi.org/10.1017/CBO9780511755408.
[24] J.W. Hardin, J. Hilbe, Generalized Linear Models and Extensions, 2nd ed., Stata Press, USA, 2007.
[25] Y. Jiao, Y. Chen, An application of Generalized Linear Models in Production Model and Sequential Population Analysis, Fisheries Research 70 (2004) 367-376. https://doi.org/10.1016/j.fishres.2004.08.027.
[26] M.A. Kahya, W. Al-Hayani, Z.Y. Algamal, Classification of breast cancer histopathology images based on adaptive sparse support vector machine, Journal of Applied Mathematics & Bioinformatics 7(1) (2017) 49-69.
[27] P. McCullagh, J.A. Nelder, Generalized Linear Models, 2nd ed., Chapman and Hall Inc., London, 1989. https://doi.org/10.1007/978-1-4899-3242-6.
[28] R.H. Myers, D.C. Montgomery, G.G. Vining, Generalized Linear Models with Applications in Engineering and the Sciences, John Wiley & Sons, Inc., New York, 2002.
[29] P. Vidoni, Prediction and Calibration in Generalized Linear Models, the Annals of Institute of Statistical Mathematics 55(1) (2003) 169-185.
[30] C. Zhukovskaya, Use of the Generalized Linear Model in Forecasting the Air Passengers Conveyances from EU Countries, Journal of Computer Modeling and New Technologies 11(1) (2007) 62-72.
[31] M.A. Kahya, W. Al-Hayani, Z.Y. Algamal, Classification of breast cancer histopathology images based on adaptive sparse support vector machine, Journal of Applied Mathematics & Bioinformatics 7(1) (2017) 49-69.
[32] Z.Y. Algamal, M.K. Qasim, H.T.M. Ali, A QSAR classification model for neuraminidase inhibitors of influenza A viruses (H1N1) based on weighted penalized support vector machine, SAR and QSAR in Environmental Research (2017) 1-12.