IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 13, Issue 3 Ver. III (May - June 2017), PP 17-23
www.iosrjournals.org

Gamma Regression Model Estimation Using Bootstrapping Procedure

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*

Abstract: *Gamma regression is a member of generalized liner models and often used when the phenomenon under study is skewed and the mean is proportional to the standard deviation. It can find applications in several areas such as life-testing problems, forecasting cancer incidences, weather extremes and quality control. Also it is a natural candidate when modeling the variance and it has been increasingly used over the past decade. paper attempts to introduce readers with the concept of the gamma regression model, in which the dependent variable has the gamma distribution, and the use of the paired bootstrapping resampling associated with the "boot" package in R program. Three confidence intervals were computed.*
Keywords*: Gamma distribution, gamma regression, paired bootstrapping, confidence intervals.*

I. Introduction

Inference procedures for regression models assume that the response variable follows the normal distribution. There are, however, many situations in social sciences where this assumption fails to hold. Common examples are count data models, qualitative response models, and duration data models (Sapra, 2005). The utility of the uses of gamma regression model arises in two different ways. Certainly, if we believe that the response variable to have a gamma distribution, the model is clearly applicable. However, the model can also be useful in other situations where we may be willing to think about the relationship between the mean and the variance of the response variable (Faraway, 2006). In the normal linear regression model, the variance of the response variable is constant as a function of the mean response. This is a fundamental assumption necessary for the optimality of least squares method (Faraway, 2006).

The bootstrap by pairs, proposed in Freedman (1981) consists of resampling the regression and regressors together from the original data. Bootstrapping pairs is less sensitive to assumptions than bootstrapping residuals (Efron & Tibshirani, 1993). In this paper we introduce the gamma regression model and use the paired bootstrap, all the implementation were done using R program.

The rest of this paper is organized as follows. Section 2 discusses the gamma regression model. Section 3 presents the concept of bootstrap resampling and section 4 shows the bootstrap packages that in R program. Sections 5 and 6 show the data and the final results. Finally, section 7 concludes the paper short conclusion.

II. Gamma Regression Model

In classical models of regression the following relationship is adopted

$$y_i = \beta_0 + \beta_j\, x_j + e_i \quad , \quad i = 1,2,..,n \;\; ; \;\; j = 1,2,..,k \qquad \text{…………(1)}$$

Where the random variables e_i are independent and have a normal distribution with mean zero and variance equal to σ_e^2 model (1) assume that the variance of the response is constant as a function of the mean response (Faraway ,2006). A model of gamma regression is consider when the dependant variable y_i has a gamma distribution with p.d.f.

$$f(y;v,\lambda) = \frac{\lambda^v}{\Gamma v}\, y^{v-1}\, e^{-\lambda y} \;,\;\; y > 0 \qquad \text{…………….(2)}$$

Where $\lambda > 0$ is the scale parameter, $v > 0$ is the shape parameter, and $\Gamma\bullet$ is the gamma function. The expected value of y is $\frac{v}{\lambda}$ and the variance is $\frac{v}{\lambda^2}$ (Krishnamoorthy, 2006). In gamma regression, we have the variance of the response variable y_i is not constant but rather is proportional to square of the mean (i.e.

$Var(y) = \sigma^2 (E(y))^2$). For gamma regression, the coefficient of variation ($C.V$), defined to be the ratio of $\sqrt{\frac{Var(y)}{E(y)}}$
is constant (Faraway ,2006).

Using generalized linear model (GLM) framework the equation (2) can reparamrterize by putting $\mu = \frac{v}{\lambda}$

$$f(y) = Exp\{\frac{y(-\frac{1}{\mu}) - \log(\mu)}{\frac{1}{v}} + v\log(v) + (v-1)\log(y) - \log(\rceil v)\}$$
…………..(3)

The canonical parameter is $(-\frac{1}{\mu})$, so , the canonical link function (the reciprocal link) is (Uusipaikka, 2009)

$$g(\mu_i) = -\frac{1}{\mu} = \sum_{j=1}^{k} x_{ij}\beta_j$$ ………….(4)

The equation (4) has the drawback that it does not guarantee $\mu > 0$ which could cause problems and might require restrictions on β on the range of possible predictor values (Faraway, 2006). As well as to the link function in (4) , there are other two commonly used link functions, they are :
The log link function, which is used when the effect of the predictors is suspected to be multiplicative on the mean

$$g(\mu_i) = \log(\mu_i)$$ …………..(5)

and the identity link function

$$g(\mu_i) = \mu_i$$ …………..(6)

The Gamma regression equations for the reciprocal and log link function respectively, are

$$E(y_i) = \frac{1}{\hat{\beta}_0 + \sum_{j=1}^{k} x_{ij}\beta_j}$$ ………….(7)

$$E(y_i) = Exp\{\hat{\beta}_0 + \sum_{j=1}^{k} x_{ij}\beta_j\}$$ .. ……….(8)

III. Bootstrap Resampling

The term bootstrap which is due to the Efron (1979) is an illusion to the expression "pulling on self up by one's bootstraps" meaning doing the impossible (Efron & Tibshirani, 1993). The bootstrap is a method to derive properties link standard error, confidence intervals, of the sampling distribution of estimators. The bootstrap resampling consists of n elements that are drawn randomly from the n original data points with replacement (Friedl & Stampfer, 2001).

In the term of regression analysis, we have two kind of bootstrapping, residual bootstrapping and paired bootstrapping. Consider a sample with n independent observations of the response variable y and $k+1$ explanatory variables x . A paired bootstrap sample is obtained by independently drawing rows with replacement from the pairs (y_i , x_i).

The bootstrap sample has the same number of observations, however some observations appear several time and others never. The bootstrap involves drawing a large number B of bootstrap samples. An individual bootstrap sample is denoted (y_b^*, x_b^*) (Carroll & et al., 2006).

The estimated $Se(\hat{\beta})$ and the bias are:

$$Se(\hat{\beta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\beta}_B - \hat{\beta})^2} \qquad \ldots\ldots\ldots\ldots(9)$$

and

$$bias = \hat{\beta}_b - \hat{\beta} \qquad \ldots\ldots\ldots\ldots\ldots(10)$$

where $\hat{\beta}_B = \frac{1}{B}\sum_{b=1}^{B}\beta_b$ is the estimated bootstrap parameter .

Three widely used bootstrap confidence intervals are: Normal theory interval, percentile interval, and bias – corrected accelerated (BCa) percentile interval.

To construct a $100(1-\alpha)\%$ confidence interval for B_b based on the bootstrap estimator $\hat{\beta}_b$ (Efron & Tibshirani, 1993)

$$B_b = \hat{\beta}_b \mp t_{(n-k,\frac{\alpha}{2})}\, Se(\hat{\beta}_b) \qquad \ldots\ldots\ldots\ldots\ldots(11)$$

To produce a $100(1-\alpha)\%$ percentile interval

$$\hat{\beta}_{b(lower)} < \beta_b < \hat{\beta}_{b(upper)}$$

where is $(\frac{\alpha}{2})B$ and upper is $(1-\frac{\alpha}{2})B$. For more details on bias – corrected accelerated percentile interval see (Efron & Tibshirani, 1993).

IV. Data

In this section we present two data sets to illustrate our study. The first data set is the **coalition data**, which is a part of Zelig package (Venables & Ripley, 2002). This data set contains survival data on government in parliamentary democracies from the period 1945-1987. The coalition data frame has 814 observations. The second data set is **wafer data** which is a part of faraway package (Faraway, 2006). The response variable is the resistivity of the test wafer.

V. Results

A gamma linear regression model is fitted to the two data sets. For **coalition data**, we examine the influence of selected two covariate *fract* and *numst2* on *duration* in R by using the following command: glm(duration ~ fract+numst2,family =Gamma ("inverse"),data =coalition)
The results of gamma regression model are given in table (1).

Table (1): Gamma regression coefficients

Coefficient	Value	$Se(\hat{\beta})$	t- value
Intercept	-0.01296	0.0133	-0.98
fract	0.000115	0.000017	6.67**
numst2	-0.01738	0.0058	-2.96**

** Significant at $\alpha = 0.01$

Dispersion parameter for gamma regression is (0.6291), the null deviance is (300.71) on (313) degrees of freedom, and the residual deviance is (272.19) on (311) degrees of freedom. Table (1) shows that all two covariates are statistically significant. The paired bootstrap step of the gamma regression model for the coalition data is

```
coal.boot<-function(data,indices){
```

```
+ data<-data[indices,]                    # select observation
+gam<-glm(duration~fract+ umst2,family=Gamma("inverse"),data=data)
+coefficients(gam)                        # return coefficient vector
+ }
coalboot<-boot(coalition, coal.boot, R=10000)
```

Table (2): Shows the results of the bootstrapped gamma regression for coalition data :

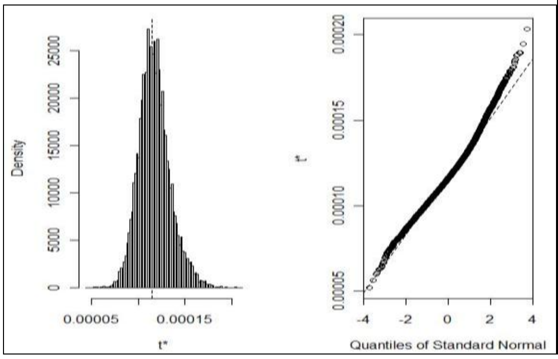
Coefficient	$Se(\hat{\beta})$	bias
Intercept	0.0135	-0.00132
fract	0.0000174	0.000002
numst2	0.0062	-0.00031

Based on $B = 10000$ bootstrap replication the confidence intervals showed in table (3).
boot.ci (coalboot, type=c ("norm","prec","bca"), index=1) is the confidence interval for the intercept, by changing the index into index=2 and index=3 we can get confidence interval for *fract* and *numst2* covariates.

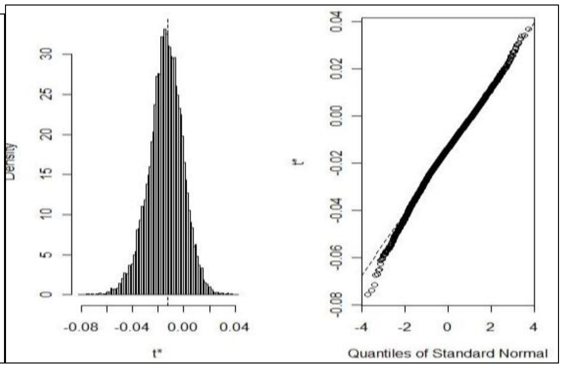
Table (3): 95% confidence intervals for parameters

Coefficient	Normal	Percentile	BCa
Intercept	(-0.0382,0.0149)	(-0.043,0.0109)	(-0.0408,0.0131)
fract	(-0.000,0.0001)	(0.0001,0.0002)	(0.0001,0.0002)
numst2	(-0.0293,-0.0048)	(-0.0307,-0.0062)	(-0.0304,-0.006)

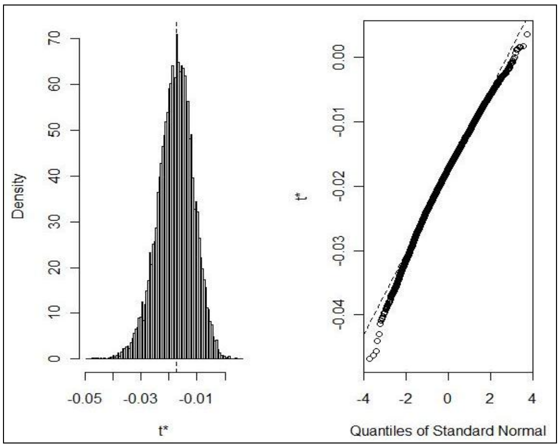
Figure 1-3 show the histograms and the normal quantile plot for bootstrap replication of the *intercept* (Figure 1), *fract* (Figure 2), and *numst2* (Figure 3).



Figure(1): The histogram and the normal quantile plot for the *intercept*



Figure(2): The histogram and the normal quantile plot for the *fract*



Figure(3): The histogram and the normal quantile plot for the *numst2*

Now, the results of fitted gamma regression model for the *wafer data* set, table (4) shows the results. glm (resist $\sim X_1 + X_2 + X_3 + X_4$ family = Gamma ("log") , data =wafer)

Table (4): Fitted Gamma regression model:

Coefficient	Value	$Se(\hat{\beta})$	t- value
Intercept	5.502	0.1593	34.54**
X_1	0.1211	0.0523	2.313*
X_2	-0.3004	0.0523	-5.736**
X_3	0.1797	0.0523	3.432**
X_4	-0.057	0.0523	-1.099

(**) significant at $\alpha = 0.01$, (*) significant at $\alpha = 0.05$

The dispersion parameter taken to be (0.0109), the null deviance (0.6978) on d.f =15, and the residual deviance (0.1241) on d.f=11. Table (5) shows the paired bootstrap gamma regression for the *wafer data* in R.

```
wafer.boot<-function(data,indices){
+ data<-data[indices,]
+ gam<-glm(resist ~ x1+ x2+ x3+ x4),family=Gamma("log"),data=data)
+ coefficients(gam)
+ }
waferboot <- boot(data=wafer, wafer.boot, R=10000)
```

Table (5): bootstrapped standard error and Bias

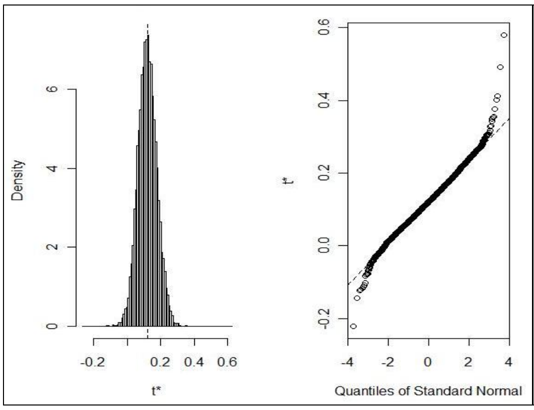
Coefficient	$Se(\hat{\beta})$	bias
Intercept	0.2168	0.000764
X_1	0.057	-0.0000103
X_2	0.0569	-0.000469
X_3	0.0614	-0.000971
X_4	0.0583	0.0001616

The bootstrapped confidence interval is showed in table (6) and it implemented in R by boot.ci(waferboot, type=c("norm","prec","bca"),index=1)

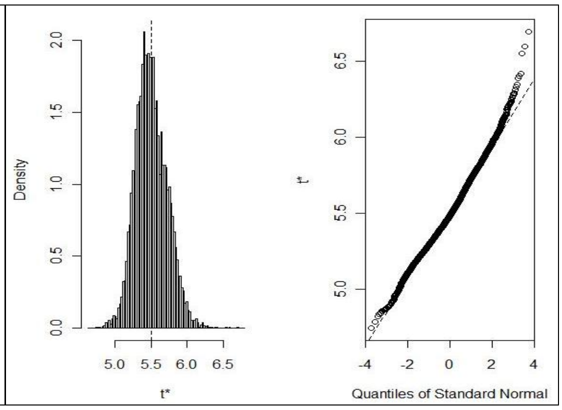
Table (6): 95% confidence intervals for parameters

Coefficient	Normal	Percentile	Bca
Intercept	(5.077,5.927)	(5.117,5.59)	(5.06,5.912)
X_1	(0.0093,0.233)	(0.0145,0.235)	(0.0184,0.2408)
X_2	(-0.4117,-0.1883)	(-0.414,-0.188)	(-0.4126,-0.1866)
X_3	(0.0607,0.3003)	(0.055,0.3018)	(0.0627,0.3087)
X_4	(-0.172,0.054)	(-0.168,0.056)	(-0.1686,0.0559)

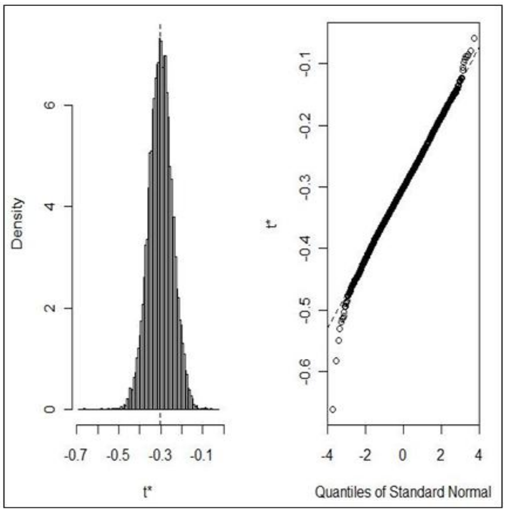
Figure (4-8) show the histograms and the normal quantile plots for bootstrap replication.



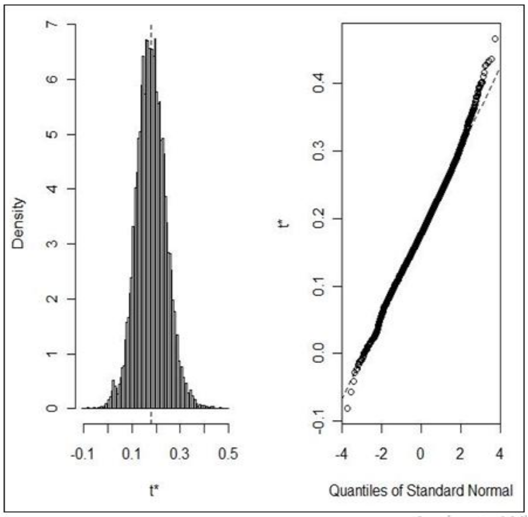
Figure(4): The histogram and the normal quantile plot for the *intercept*



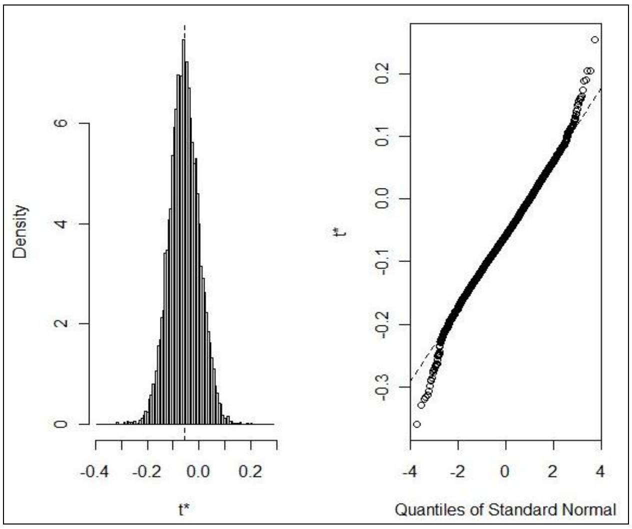
Figure(5): The histogram and the normal quantile plot for the X_1



Figure(6): The histogram and the normal quantile plot for the X_2



Figure(7): The histogram and the normal quantile plot for the $X3$



Figure(8): The histogram and the normal quantile plot for the X_4

VI. Conclusion

In this article we have used the gamma regression model to fit the **coalition** and **wafer data**. All figures with the histogram and normal quantile plot show asymptotic normal theory. So, it may be conclude that the bootstrap by pairs could potentially be applied.

References

[1] Carroll, R., J., Ruppert, D., Stefanski, L., A., and Crainiceanu, C., M.,(2006),"Measurement Error in Nonlinear Models, A Modern Prespective", 2nd ed., Chapman & Hall/CRC, Florida.
[2] Davison, A., D. and Kuonen, D.,(2002), " An Introduction to Bootstrap with Applications in R", Statistical computing & Statistical Graphics Newsletter, Vol.13, No.1, pp.6-11.
[3] Efron, B. (1979) "Bootstrap Methods: Another look at Jackknife", Annals of Statistics,Vol.7, pp.1-26.
[4] Efron, B. and Tibshirani, R., (1993), "An introduction to the bootstrap", Chapman and Hall , New York.
[5] Everitt, B., S. and Hothorm, T., (2010), " A Handbook of Statistical Analysis Using R", 2nd ed., Chapman & Hall/CRC, Florida.
[6] Faraway, J., J., (2006),"Extending the Linear Model with R, Generalized Linear, Mixed Effects and Nonparametric Regression Models", Chapman & Hall/CRC, Florida.
[7] Freedman, D.,A.,(1981) "Bootstrapping Regression Models", Annals of Statistics,Vol.9, No.6, pp.1218-1228.
[8] Friedl, H. and Stampfer, E.,(2002), "Jackknife Resampling", Encyclopedia of Environmetrics, 2, pp.1089-1098.
[9] Krishnamoorthy, K., (2006), "Handbook of Statistical Distributions with Applications", Chapman & Hall/CRC, Florida.
[10] Sapra, S., (2005),"A Regression Error Specification Test (RESET) for Generalized Linear Models", Economics Bulletin, Vol.3, No.1, pp.1-6.
[11] Uusipaikka, E.,(2009),"Confidence Intervals in Generalized Regression Models", Chapman & Hall/CRC, Florida.
[12] Venables. S. and Ripley, B., D., (2002),"Modern Applied Statistics with S", 4th ed., Springer-Verlage.

[13] Al-Fakih, A. M., Algamal, Z. Y., Lee, M. H., Abdallah, H. H., Maarof, H., & Aziz, M. (2016). Quantitative structure-activity relationship model for prediction study of corrosion inhibition efficiency using two-stage sparse multiple linear regression. Journal of Chemometrics, 30(7), 361-368.
[14] Al-Fakih, A. M., Aziz, M., Abdallah, H. H., Algamal, Z. Y., Lee, M. H., & Maarof, H. (2015). High dimensional QSAR study of mild steel corrosion inhibition in acidic medium by furan derivatives. International Journal of Electrochemical Science, 10, 3568-3583.
[15] Algamal, Z. Y. (2008). Exponentiated exponential distribution as a failure time distribution. IRAQI Journal of Statistical science, 14, 63-75.
[16] Algamal, Z. Y. (2011). Paired Bootstrapping procedure in Gamma Regression Model using R. Journal of Basrah Researches, 37(4), 201-211.
[17] Algamal, Z. Y. (2012). Diagnostic in poisson regression models. Electronic Journal of Applied Statistical Analysis, 5(2), 178-186.
[18] Algamal, Z. Y. (2017). Using maximum likelihood ratio test to discriminate between the inverse Gaussian and gamma distributions. International Journal of Statistical Distributions, 1(1), 27-32.
[19] Algamal, Z. Y., & Ali, H. T. M. (2017). An efficient gene selection method for high-dimensional microarray data based on sparse logistic regression. Electronic Journal of Applied Statistical Analysis, 10(1), 242-256.
[20] Algamal, Z. Y., & Ali, H. T. M. (2017). Bootstrapping pseudo - R2 measures for binary response variable model. Biomedical Statistics and Informatics, 2(3), 107-110.
[21] Algamal, Z. Y., & Lee, M. H. (2015). Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. Expert Systems with Applications, 42(23), 9326-9332.
[22] Algamal, Z. Y., & Lee, M. H. (2015). Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. Computers in Biology and Medicine, 67, 136-145.
[23] Algamal, Z. Y., & Lee, M. H. (2015). Penalized Poisson regression model using adaptive modified elastic net penalty. Electronic Journal of Applied Statistical Analysis, 8(2), 236-245.
[24] Algamal, Z. Y., & Lee, M. H. (2015). High dimensional logistic regression model using adjusted elastic net penalty. Pakistan Journal of Statistics and Operation Research, 11(4), 667-676.
[25] Algamal, Z. Y., & Lee, M. H. (2015). Adjusted adaptive lasso in high-dimensional Poisson regression model. Modern Applied Science, 9(4), 170-176.
[26] Algamal, Z. Y., & Lee, M. H. (2015). Applying penalized binary logistic regression with correlation based elastic net for variables selection. Journal of Modern Applied Statistical Methods, 14(1), 168-179.
[27] Algamal, Z. Y., & Lee, M. H. (2017). 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), 75-90.
[28] Algamal, Z. Y., Lee, M. H., & Al-Fakih, A. M. (2016). 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), 50-57.
[29] Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M., & Aziz, M. (2015). High-dimensional QSAR prediction of anticancer potency of imidazo[4,5-b]pyridine derivatives using adjusted adaptive LASSO. Journal of Chemometrics, 29(10), 547-556.
[30] Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M., & Aziz, M. (2016). High-dimensional QSAR modelling using penalized linear regression model with L1/2-norm. SAR and QSAR in Environmental Research, 27(9), 703-719.
[31] Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M., & Aziz, M. (2017). 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 (doi:10.1002/cem.2889).
[32] Algamal, Z. Y., Qasim, M. K., & Ali, H. T. M. (2017). 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, 1-12.
[33] Kahya, M. A., Al-Hayani, W., & Algamal, Z. Y. (2017). Classification of breast cancer histopathology images based on adaptive sparse support vector machine. Journal of Applied Mathematics & Bioinformatics, 7(1), 49-69.