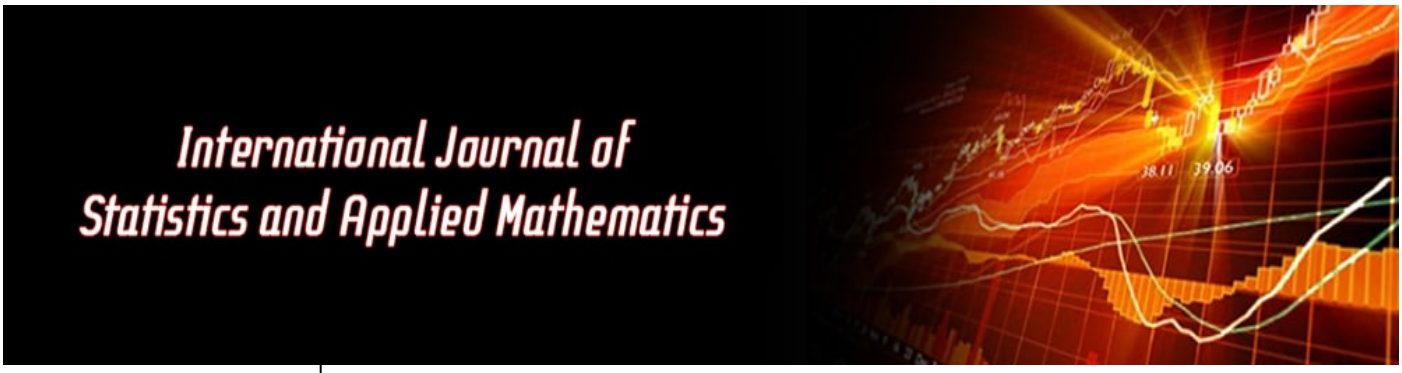
International Journal of Statistics and Applied Mathematics 2021; 6(6): 83-90



ISSN: 2456-1452
Maths 2021; 6(6): 83-90

www.mathsjournal.com
Received: 15-09-2021
Accepted: 18-10-2021

Obaji Ifeoma
Department of Mathematics and Statistics, Ignatius Ajuru University of Education Rivers State P.M.B. 5047, Port Harcourt, Rivers Nigeria

Nwagor Peters
Department of Mathematics and Statistics, Ignatius Ajuru University of Education Rivers State P.M.B. 5047, Port Harcourt, Rivers Nigeria

Corresponding Author:
Obaji Ifeoma
Department of Mathematics and Statistics, Ignatius Ajuru University of Education Rivers State P.M.B. 5047, Port Harcourt, Rivers Nigeria

Multiple regression model selection via birth weight, mother age and gestation variables

Obaji Ifeoma and Nwagor Peters

Abstract
This study is on multiple regression model selection. The source of data is from records of 2019 to 2020 live deliveries in Federal Medical Center (FMC) Umuahia Abia State, Nigeria. The hospital keeps records of deliveries, ages of mothers and their birth weights. The mothers' records that do not have the exact record of last menstrual period were not used for this study. Miscarriages i.e., pregnancies that did not go beyond 28 weeks gestation were not included in the study. It is a retrospective study of 100 pregnancies with outcomes of live births that exceeded 28 weeks gestation. The dependent variable is birth weight, while the independent variables are mother's age and gestation. Four regression models; Lin-Log, Linear, Inverse, and Polynomial were examined in this study. The E-views software was used in this study. Four model selection techniques known as; coefficient of determination, Akaike Information Criterion, Schwarz Information Criterion, and Hannan-Quinn Information Criterion were used to select the best model. From the analysis in the overall goodness of fit assessment, the study concluded that the Lin-Log regression model performs slightly better than the other three regression models used in this study. Therefore, future researchers should look at a similar work by incorporating other nonlinear regression models like Double-Log and Log-Lin Regression models to compare results.

Keywords: akaike information criterion, schwarz information criterion, hannan-quinn information criterion, coefficient of determination, regression models

Introduction
Fitting of straight and non-linear models to information generated is ordinarily utilized inside every area of science and medical related test evaluation, in spite of the fact that fitting a straight model to collected information does not often occur since most information tend to take after non-linear models. Non-linear models are in existence, but the option of choosing the proper model for the information could be a blend of involvement, information around the basic process and measurable translation of the fitting result. It is of vital in measuring the legitimacy of a fit by a few degrees which segregates a 'good' from a 'bad' fit. Numerous analysts ordinarily utilize a common degree known as the coefficient of assurance R^2 utilized in regression of linear form when evaluating calibration tests for tests to be measured (Montgomery *et al*, 2006) [8]. Subsequently, within the direct viewpoint, this degree is exceptionally instinctive as values between 0 and 1 deliver a straightforward translation of how much of the fluctuation within the information is clarified by the fit. Indeed in spite of the fact that for a few time, it has been set up that R^2 is not an adequate measure for non-linear regression, numerous researchers and analysts still make utilize of it in ponders managing with non-linear information analysis (Nagelkerke, 1991; Magee, 1990) [9, 7]. Concurring to Juliano & Williams (1987) [6], a few introductory and more seasoned portrayals for R^2 being of no profit in non-linear fitting had pointed out this issue but have probably fallen into blankness. This perception may well be due to contrasts within the numerical foundation of prepared analysts and analysts who regularly utilize measurable strategies but lack detailed statistical knowledge according to Spiess and Neumeyer (2010) [12].
Having expressed that analysts unpredictably utilize R^2 as a means of surveying the legitimacy of a specific model when managing with nonlinear information fit, it is expressed that R^2 is not an ideal option in a non-linear assessment as the total sum of squares (SS_T) isn't break even with the regression sum of squares (SS_R) together with the error (residual) sum of squares

(SS_R), as in the case of linear regression, and consequently it lacks the adequate explanation. The method of reasoning behind a high occurrence in exclusively utilizing R^2 values within the legitimacy of non-linear models can be as a result of analysts not being mindful of this misguided judgment.

Indeed in spite of the fact that the implementation of R^2 to get to the execution of nonlinear data investigation has been disheartened, this study will utilize it, in line with other three model determination strategies known as; Akaike Information Criterion, Schwarz Information Criterion, and Hannan-Quinn Information Criterion for appropriate elucidation and conclusion.

Statement of Problem

Medically speaking, it has been established that there is a relationship between birth weight versus age and gestation of a mother. However, many researchers especially those in other areas who probably do not have sufficient knowledge of statistics usually employ the multiple linear regression technique to establish such relationships, without looking at the nonlinear form of models. It is as a result of the situation that this study intends to look at various multiple non-linear models versus linear model to establish the best model for the variables aforementioned.

Review of Related Literature

Hamidian *et al* (2008) inquired on comparison of direct and nonlinear models for assessing brain deformation utilizing limited component strategy. The study displayed limited component computation for brain distortion amid craniotomy. The study was utilized to demonstrate the comparison between two mechanical models: direct solid-mechanic demonstrates, and non direct limited component demonstrate. To this conclusion, the study utilized a test circle as a show of the brain, tetrahedral limited component work; two models that portrayed the fabric property of the brain tissue, and work optimization that optimized the model's parameters by minimizing remove between the coming about distortion and the accepted distortion. Straight and nonlinear model expected limited and expansive deformation of the brain after opening the cranium individually. By utilizing the precision of the optimization handle, the think about concluded that the precision of nonlinear demonstrate was higher but its execution time was six time of the direct show.

Aristizábal-Giraldo *et al* (2016) carried out a work on a comparison of straight and nonlinear show execution of shia_landslide: a determining show for rainfall-induced avalanches. The work clarified that avalanches are one of the most causes of worldwide human and financial misfortunes. The study compared the estimating execution of straight and nonlinear SHIA_Landslide show. The results gotten for the La Arenosa Catchment amid the September 21, 1990 rainstorm appeared that the nonlinear SHIA_Landslide duplicated more precisely avalanches activated by precipitation highlights.

Hunt and Maurer (2016) carried out a study on comparison of direct and nonlinear input control of heart rate for treadmill running. The reason of the study was to compare straight (L) and non-linear (NL) controllers utilizing quantitative execution measures. Sixteen sound male subjects taken part within the test L vs. NL comparison. The linear controller was calculated employing a coordinate expository plan that utilized an existing inexact plant demonstrate. The nonlinear controller had the same straight component, but it was increased utilizing inactive plant-nonlinearity emolument. At moderate-to-vigorous power, no noteworthy contrasts were found between the direct and nonlinear controllers in cruel RMS following blunder and normal control flag control, but scattering of the last mentioned was significantly higher for NL. At moo speed, RMS following mistakes was comparative, but normal control flag control was considerably and altogether higher for NL. The execution results for direct and non-linear control were not essentially diverse for moderate-to-vigorous force, but NL control was excessively delicate at moo running speed. Precise, steady and vigorous in general execution was accomplished for all 16 subjects with the direct controller.

Methodology

Regression Models

Four Regression models are considered in this study, which are Lin-Log, Linear, Inverse and Polynomial as written in Equations (1), (2), (3) and (4) respectively

$$Y = \lambda_0 + \lambda_1 \ln Z_1 + \lambda_2 Z_2 \quad \ldots (1)$$

$$Y = \lambda_0 + \lambda_1 Z_1 + \lambda_2 Z_2 \quad \ldots (2)$$

$$Y = \lambda_0 + \lambda_1 (1/Z_1) + \lambda_2 Z_2 \quad \ldots (3)$$

$$Y = \lambda_0 + \lambda_1 Z_1 + \lambda_2 Z_1^2 + \lambda_3 Z_2 \quad \ldots (4)$$

A factual method that communicates numerically the affiliation between two or more quantitative variables in such a way that the response variable can be anticipated from the informative factors is known as Regression analysis. It can be utilized to see at the impacts that some variables apply on others. It may be direct or non-linear.

Considering a linear regression model (multiple) as in Equation (5);

$$Y_i = \lambda_0 + \lambda_1 Z_1 + \cdots + \lambda_p Z_{pi} + e_i \quad \ldots (5)$$

For the case of this study, where there are two explanatory variables in the data collected for this study, Equation (5) is stated as [See Equation (6)].

$$Y_i = \lambda_0 + \lambda_1 Z_{1i} + \lambda_2 Z_{2i} + e_i \quad \ldots (6)$$

$Y \Rightarrow$ Dependent variable,
Z_1 & $Z_2 \Rightarrow$ Independent variables,
$e_i \Rightarrow$ Error term, and
$i \Rightarrow$ *ith* observation

If Y, Z_1, and Z_2 are in deviation forms, then Equation (6) is now written as shown in Equation (7)

$$y_i = \lambda_1 z_{1i} + \lambda_2 z_{2i} + e_i \quad \ldots (7)$$

Where
$y_i = Y_i - \overline{Y}$, $z_{1i} = Z_{1i} - \overline{Z}_{1i}$ and $z_{2i} = Z_{2i} - \overline{Z}_{2i}$

$$e_i = y_i - \hat{y}_i = y_i - \hat{\lambda}_1 z_{1i} - \hat{\lambda}_2 z_{2i} \quad \ldots (8)$$

$$\Sigma e_i^2 = \Sigma (y_i - \hat{\lambda}_1 z_{1i} - \hat{\lambda}_2 z_{2i})^2 \quad \ldots (9)$$

The estimate of the parameters using OLS technique is shown in Equations (10), (11) & (12)

$$\hat{\lambda}_1 = \frac{\Sigma z_1 y \Sigma z_2^2 - \Sigma z_1 z_2 \Sigma z_2 y}{\Sigma z_1^2 \Sigma z_2^2 - (\Sigma z_1 z_2)^2} \quad \ldots (10)$$

$$\hat{\lambda}_2 = \frac{\Sigma z_1^2 \Sigma z_2 y - \Sigma z_1 z_2 \Sigma z_1 y}{\Sigma z_1^2 \Sigma z_2^2 - (\Sigma z_1 z_2)^2} \quad \ldots (11)$$

and

$$\hat{\lambda}_0 = \overline{Y} - \hat{\lambda}_1 \overline{Z}_1 - \hat{\lambda}_2 \overline{Z}_2 \quad \ldots (12)$$

The following formulas are used to obtain the sum of squares and cross products in deviation forms.

$$\Sigma z_1^2 = \Sigma (Z_1 - \overline{Z}_1)^2 = \Sigma Z_1^2 - \frac{(\Sigma Z_1)^2}{n} \quad \ldots (13)$$

$$\Sigma y^2 = \Sigma (Y - \overline{Y})^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} \quad \ldots (14)$$

$$\Sigma z_2^2 = \Sigma(Z_2 - \overline{Z}_2)^2 = \Sigma Z_2^2 - \frac{(\Sigma Z_2)^2}{n} \quad \ldots (15)$$

$$\Sigma z_1 z_2 = \Sigma(Z_1 - \overline{Z})(Z_2 - \overline{Z}) = \Sigma Z_1 Z_2 - \frac{\Sigma Z_1 \Sigma Z_2}{n} \quad \ldots (16)$$

$$\Sigma z_1 y = \Sigma(Z_1 - \overline{Z}_1)(Y - \overline{Y}) = \Sigma Z_1 Y - \frac{\Sigma Z_1 \Sigma Y}{n} \quad \ldots (17)$$

$$\Sigma z_2 y = \Sigma(Z_2 - \overline{Z}_2)(Y - \overline{Y}) = \Sigma Z_2 Y - \frac{\Sigma Z_2 \Sigma Y}{n} \quad \ldots (18)$$

Coefficient of Determination

$$R^2 = \frac{\hat{\lambda}_1 \Sigma z_1 y + \hat{\lambda}_2 \Sigma z_2 y}{\Sigma y^2} \quad \ldots (19)$$

The adjusted R^2 is stated in Equation (20)

$$\overline{R}^2 = t\frac{n-1}{n-p} \quad \ldots (20)$$

Where $t = 1 - (1 - R^2)$

Table 1: Analysis of Variance Table

SV	DF	SS	MS
Regression	2	SS_R	MS_R
Error	n – 3	SS_E	MS_E
Total	n – 1	SS_T	

$$F_{cal} = \frac{MS_R}{MS_E} \quad \ldots (21)$$

$$SS_T = \Sigma y_i^2 \quad \ldots (22)$$

$$SS_R = \hat{\lambda}_1 \Sigma z_1 y + \hat{\lambda}_2 \Sigma z_2 y \quad \ldots (23)$$

$$SS_E = SS_T - SS_R \quad \ldots (24)$$

The same steps employed in linear model are applicable to the nonlinear models as used in this study.

Akaike Information Criterion (AIC)

The degree of the goodness of fit of an assessed measurable model is known as AIC (Akaike, 1974) and it can be employed for model choice. It is scientifically characterized as;

$$AIC = \exp^{\frac{2p}{n}} \frac{\sum \hat{e}_i^2}{n} = \exp^{\frac{2p}{n}} \frac{SS_E}{n} \quad \ldots (25)$$

where p is the number of parameters with the inclusion of the intercept. Equation (25) is stated mathematically for convenience sake as;

$$\ln(AIC) = \left(\frac{2p}{n}\right) + \ln\left(\frac{SS_E}{n}\right) \quad \ldots (26)$$

Schwarz Information Criterion (SIC)

The degree of the goodness of fit of an evaluated measurable model is known as Schwarz Information Criterion SIC (Schwarz, 1978) and it can be employed for model choice. It is mathematically characterized as;

$$SIC = n^{\frac{p}{n}} \frac{\sum \hat{e}_i^2}{n} = n^{\frac{p}{n}} \frac{SS_E}{n} \quad \ldots (27)$$

Taking the natural logarithm of both sides of Equation (27) to give Equation (28);

$$\log_e(SIC) = \frac{p}{n}\log_e(n) + \log_e\left(\frac{SS_E}{n}\right) \quad \ldots (28)$$

Hannan-Quinn Information Criterion (HQIC)

The degree of the goodness of fit of an evaluated measurable model is known as HQIC (Hannan and Quinn, 1979) and it can be utilized for model choice. It is mathematically characterized as;

$$HQIC = n\ln\frac{SS_E}{n} + 2p\ln(\ln n) \quad \ldots (29)$$

The model with the least AIC, SIC or HQIC score is chosen for model comparison.

Analysis of Data

The source of data is from records of (2019 to 2020) live deliveries in Federal Medical Center (FMC) Umuahia Abia State, Nigeria. The hospital keeps records of deliveries, ages of mothers and their birth weights. The mothers' records that do not have the exact record of last menstrual period were not used for this study. Miscarriages i.e., pregnancies that did not go beyond 28 weeks gestation were not included in the study. It is a retrospective study of 100 pregnancies with outcomes of live births that exceeded 28 weeks gestation. The four regression models employed in this study were tackled. The set of data is presented (See Table 2).

Table 2: Data on New Born Babies According to Maternal Age, Gestation and Birth Weight

S/N	Birth Weight(kg)	Mother Age	Gestation (weeks)	S/N	Birth weight(kg)	Mother Age	Gestation (weeks)
1	2.7	30	37	51	4.1	33	41
2	2.5	27	31	52	3.5	26	38
3	3.3	31	36	53	3.3	23	36
4	3.0	25	30	54	3.5	33	38
5	2.9	23	29	55	3.0	32	36
6	3.5	27	37	56	3.2	34	38
7	3.7	35	39	57	3.5	37	39
8	3.3	30	38	58	3.9	38	40
9	3.4	32	40	59	3.3	35	36
10	2.7	24	36	60	3.1	30	32
11	2.4	21	32	61	3.0	27	30
12	2.9	25	36	62	3.6	30	38
13	3.5	32	39	63	3.2	28	29
14	3.7	34	41	64	2.4	22	32
15	2.9	29	35	65	3.5	35	39
16	3.3	31	39	66	2.6	26	31
17	4.1	33	41	67	4.1	36	39
18	3.5	26	38	68	2.9	31	35
19	3.3	23	36	69	3.5	38	40
20	3.5	33	38	70	2.7	30	37
21	3.0	32	36	71	2.5	27	29
22	3.2	34	38	72	3.3	31	36

23	3.9	38	40	73	3.0	25	30
24	3.3	35	36	74	2.9	23	29
25	3.1	30	32	75	3.5	27	37
26	3.0	27	30	76	3.7	35	39
27	3.6	30	38	77	3.3	30	38
28	3.2	28	29	78	3.4	32	40
29	2.4	22	31	79	2.7	24	36
30	3.5	35	39	80	2.4	21	32
31	2.6	26	31	81	2.9	25	36
32	4.1	36	39	82	3.5	32	39
33	3.0	32	37	83	3.7	34	41
34	3.5	38	40	84	2.9	29	35
35	2.7	30	37	85	3.3	31	39
36	2.5	27	29	86	4.1	33	41
37	3.3	31	36	87	3.5	26	38
38	3.0	25	30	88	3.3	23	36
39	2.9	23	30	89	3.5	33	38
40	3.5	27	37	90	3.0	32	36
41	3.7	35	39	91	3.2	34	38
42	3.3	30	38	92	3.5	37	39
43	3.4	32	40	93	3.9	38	40
44	2.7	24	36	94	3.3	35	36
45	2.4	21	32	95	3.1	30	32
46	2.9	25	36	96	3.0	27	30
47	3.5	32	39	97	3.6	30	38
48	3.7	34	41	98	3.7	35	40
49	2.9	29	35	99	3.4	32	40
50	3.3	31	39	100	2.7	24	36

Y = Birth Weight (BWT)
Z_1 = Mother Age (MA)
Z_2 = Gestation (GS)

Table 3: Lin-Log Model Output

Dep. Var: BWT				
Meth: LS				
Date: 06/05/21 Time: 20:51				
Samp: 1 100				
Used Obsn: 100				
Var.	Coeffs	Std Error	t-Stat	Prob
C	-2.355331	0.602379	-3.910051	0.0002
LOG(MAG)	1.047864	0.232408	4.508717	0.0000
GS	0.056324	0.010174	5.535819	0.0000
R^2	0.617813	Mean dep. var		3.229000
Adj. R^2	0.609933	S.D. dep. var		0.427181
Std Error	0.266798	AIC		0.224887
Sum squared resid	6.904549	SIC		0.303042
Log likelihood	-8.244365	HQIC		0.256518
F-stat	78.40129	D-W stat		1.201148
Prob(F-stat)	0.000000			

Table 4: Linear Model Output

Dep. Var: BWT				
Meth: LS				
Date: 06/05/21 Time: 20:55				
Samp: 1 100				
Used Obsn: 100				
Var.	Coeffs	Std Error	t-Stat	Prob
C	0.119709	0.272202	0.439778	0.6611
MAG	0.036331	0.008073	4.500044	0.0000
GS	0.055956	0.010241	5.463957	0.0000
R^2	0.617558	Mean dep. var.		3.229000
Adj. R^2	0.609673	S.D. dep. var.		0.427181
Std Error	0.266886	AIC		0.225553

Sum squared resid.	6.909150	SIC	0.303709
Log likelihood	-8.277674	HQIC	0.257184
F-stat	78.31678	D-W stat.	1.166709
Prob.(F-stat)	0.000000		

Table 5: Inverse Model Output

Dep. Var: BWT				
Meth: LS				
Date: 06/05/21 Time: 20:57				
Samp: 1 100				
Used Obsn: 100				
Var.	Coeffs	Std Error	t-Stat	Prob
C	2.155639	0.539073	3.998787	0.0001
1/MAG	-29.14957	6.498550	-4.485549	0.0000
GS	0.057318	0.010054	5.700791	0.0000
R^2	0.617133	Mean dep. var.		3.229000
Adj. R^2	0.609239	S.D. dep. var.		0.427181
Std Error	0.267035	AIC		0.226665
Sum squared resid.	6.916833	SIC		0.304820
Log likelihood	-8.333245	HQIC		0.258296
F-stat	78.17591	D-W stat.		1.242146
Prob.(F-stat)	0.000000			

Table 6: Polynomial Model Output

Dep. Var: BWT				
Meth: LS				
Date: 06/05/21 Time: 20:59				
Samp: 1 100				
Used Obsn: 100				
Var.	Coeffs	Std Error	t-Stat	Prob
C	0.019105	1.112751	0.017169	0.9863
MAG	0.043257	0.074697	0.579098	0.5639
MAG^2	-0.000117	0.001258	-0.093274	0.9259
GS	0.055979	0.010297	5.436636	0.0000
R^2	0.617593	Mean dep. var.		3.229000
Adj. R^2	0.605643	S.D. dep. var.		0.427181
Std Error	0.268261	AIC		0.245463
Sum squared resid.	6.908524	SIC		0.349670
Log likelihood	-8.273143	HQIC		0.287637
F-stat	51.68051	D-W stat.		1.172868
Prob.(F-stat)	0.000000			

Discussion of Result

Evaluation on the linear with different non-linear models has been done and the results based on their performance are presented in Table 7.

Table 7: Linear and Non-linear Models Summary Result

Model	AIC	SIC	HQIC	R^2
Lin-Log	0.22489	0.30304	0.25652	0.61781
Linear	0.22555	0.30371	0.25718	0.61756
Inverse	0.22667	0.30482	0.25830	0.61713
Polynomial	0.24546	0.34967	0.28764	0.61759

The results in Table 7 revealed that Lin-Log model has slightly the largest R^2 (0.61781) with the smallest AIC (0.22489), SIC (0.30304), and HQIC (0.25652), which makes it the leading model with regard to the information utilized in this study.

Conclusion

From the investigation, the study concluded that one of the nonlinear models performed FAR better than the linear model in line with the information utilized in this study. In any case, within the generally goodness of fit appraisal, the study concluded that the Lin-Log model performs somewhat superior than the other three regression models utilized in this study. Subsequently, future analysts ought to see at a comparable work by consolidating other nonlinear models like Double-Log and Log-Lin models to compare outcomes.

References

1. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control 1974;19(6):716-723.

2. Aristizábal-Giraldo EV, Vélez-Upegui JI, Martínez-Carvaja HE. A comparison of linear and nonlinear model performance of shia_landslide: a forecasting model for rainfall-induced landslides. Revista Facultad de Ingeniería 2016;80:74-88.
3. Hamidian H, Soltanian-Zadeh H, Akhondi-Asl A, Faraji-Dana R. Comparison of Linear and Nonlinear Models for Estimating Brain Deformation Using Finite Element Method. Advances in Computer Science and Engineering. CSICC 2008. Communications in Computer and Information Science 2008;6:340-347.
4. Hannan EJ, Quinn BG. The Determination of the order of an autoregression. Journal of the Royal Statistical Society, Series, 1979;41:190-195.
5. Hunt KJ, Maurer RR. Comparison of linear and nonlinear feedback control of heart rate for treadmill running. Systems Science & Control Engineering 2016;4(1):87- 98.
6. Juliano SA, Williams FM. A comparison of methods for estimating the functional response parameters of the random predator equation. Journal of Animal Ecology 1987;56:641- 653.
7. Magee LR. measures based on Wald and likelihood ratio joint significance tests. American Statistics 1990;44:250-253.
8. Montgomery DC, Peck EA, Vining GGIntroduction to Linear Regression Analysis. Wiley & Sons, Hoboken 2006.
9. Nagelkerke NJD. A note on a general definition of the coefficient of determination. Biometrika 1991;78:691-692.
10. Scarneciu CC, Sangeorzan L, Rus H, Scarneciu VD, Varciu MS, Andreescu O, Scarneciu I. Comparison of Linear and Non-linear Regression Analysis to Determine Pulmonary Pressure in Hyperthyroidism. Pakistan Journal of Medical Sci*ence*, 2017;*33*(1):111-120.
11. Schwarz G. Estimating the dimension of a model. The Annals of Statistics, 1978;6:461– 464.
12. Spiess A, Neumeyer N. An evaluation of R^2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. BMC Pharmacol 2010;10:2010.