Forecasting Parameter Estimates: A Modeling Approach Using Exponential and Linear Regression

W. M. A. W. Ahmad
School of Dental Sciences
Universiti Sains Malaysia, Malaysia
wmamir@usm.my

R. A. A. Rohim
Universiti Sains Malaysia
Malaysia
adawiyah5350@yahoo.com

N. H. Ismail
Universiti Sains Malaysia
Malaysia
noorhuda@usm.my

***Abstract*-This paper supplies a calculation method for the parameter estimates of an exponential equation through SAS algorithm. The aim of this paper is to investigate the efficiency of the gained parameter estimates through the forecasting performance. The proposed calculation method can provide a very useful technique to develop an exponential equation with better accuracy performance. This research paper illustrates a sample of the data obtained from the established study, which characterize the proliferative capacity of mesenchymal stem cells. This paper also provides the specific algorithm for the parameter estimates.**

Keywords-exponential; SAS algorithm; parameter estimates

I. INTRODUCTION

Regression analysis is a statistical methodology that uses the relationship between two or more quantitative variables in a way that one variable can be predicted from the other, or others. This methodology is widely used in business, social, behavioral and biological sciences, including agriculture and fishery research [1]. For example, fish weight at harvest can be predicted by utilizing the relationship between fish weights and other growth affecting factors like water temperature, dissolved oxygen, and free carbon dioxide. There are other situations in a fishery where relationships among variables can be exploited through regression analysis [1]. Regression analysis serves three major purposes: (1) description, (2) control and (3) prediction. We frequently use equations to summarize or describe data. Regression analysis is helpful in developing such equations. For example, we may collect a considerable amount of fish growth data and a data on a number of biotic and abiotic factors and a regression model would probably be a much more convenient and useful summary of those data than a table or a graph. Besides prediction, regression models may be used for control purposes. A cause and effect relationship may not be necessary if the equation is to be used only for prediction [2]. A functional relationship between two variables is expressed by a mathematical formula. If x denotes the independent variable and y the dependent variables, a functional relationship is of the form $y = f(\mathrm{x})$.

Given a particular value of x, the function indicates the corresponding value of y. A statistical relation, unlike a function, is not a perfect one. In general, the observations for a statistical relation do not fall directly on the relationship's curve. Depending on the nature of the relationship between x and y, regression approach may be classified into two categories, linear regression and nonlinear regression models. The models that are linear in these parameters are known as linear models, whereas in nonlinear models parameters show nonlinearity. Linear models are generally satisfactory approximations for most regression applications. There are occasions, however, when an empirically indicated or a theoretically justified nonlinear model is more appropriate [3].

A. Linear Regression

Linear regression is used to study the linear relationship between a dependent variable Y and one or more independent variables X. The dependent variable Y must be continuous, while the independent variables may be either continuous, binary, or categorical. The initial judgment of a possible relationship between two continuous variables should always be made on the basis of a scatter plot (scatter graph). This type of plot will show whether the relationship is linear or nonlinear. Performing a linear regression makes sense only if the relationship is linear. Other methods must be employed to study nonlinear relationships [4]. A model with more than predictor variables is a straightforward one. The model can be stated as follows:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_{\mathrm{i}} \quad (1)$$

where y_{i} is the value of the response variable in the i^{th} trial, β_0 and β_1 are parameters x_i is a known constant, namely the i^{th} value of the predictor variable and ε_i is a random error term with mean zero and variance σ^2 and their covariance is zero [5].

B. History of the Exponential Function

The exponential is one of the most significant and widely occurring functions. In biology, it may depict the growth of bacteria or animal populations, the reduction of the number of bacteria in response to a sterilization process, the development of a tumor or the absorption or elimination of a drug. Exponential growth cannot go on forever because of limitations of nutrients, etc. Knowledge of the exponential function makes it more comfortable to understand birth and death rates, even when they are not perpetual. In physics, the exponential function describes the disintegration of radioactive nuclei, the

emission of light by atoms, the assimilation of light as it passes through matter, the change of voltage or current in some electrical circuits, the variance of temperature with time as a warm object cools, and the rate of some chemical reactions [1]. Although the exponential distribution provides a simple, elegant and closed form solution to many problems, it does not offer a reasonable parametric fit for some practical applications where the underlying failure rates are nonconstant, presenting monotone shapes. Recently, in the procedure of overcoming such problems, new categories of examples were introduced based on adjustments of the exponential distribution. Authors in [6] offered a generalized exponential distribution, which can hold data with increasing and decreasing failure rate function. Authors in [7] ushered in the exponential geometric distribution with decreasing failure rate, authors in [8] proposed a two-parameter distribution known as exponential-Poisson distribution, which takes in a decreasing failure rate and authors in [3] proposed another modification of the exponential distribution with decreasing failure rate function This model is inferred in a complementary risk scenario [9] where the lifetime associated with particular danger is not evident, rather we observe just the maximum lifetime value among all risks .

C. Exponential Growth

Exponential growth is often used to model the growth of organism populations in a resource-rich environment. Here "resource-rich" means that there is abudance of food and other resources necessary for the population to grow. For example, the initial growth of a cell bacteria in a mouth is often modeled as exponential. The justification for this model is that the rate at which a population of organisms grows should be proportional to their number, assuming that the organisms reproduce at a constant rate. For example, if you double the size of a population, then this should precisely double the rate at which the population bears an offspring, and should, therefore, double the rate at which the size of the population increases. What this means is that the population A of a given organism in a resource-rich environment should satisfy the differential equation

$$\frac{dA}{dx} = Ax,$$

where x is some constant that depends on the rate of reproduction. Thus the population grows exponentially

$$A = A_0 e^{bx}$$

This model predicts that the population A will grow indefinitely, which cannot be true in any real situation. Eventually, any population will run out of resources such as food or space to grow. However, the exponential model often gives fairly accurate results in cases where the short-term growth of a population is not inhibited by limited resources[10].

D. Interpreting R^2

R^2 is frequently defined as the proportion of variance of the response that is predictable (or explained) from the regressor variables, that is the variability explained by the model. A low value of R^2 can suggest that the assumptions of linear regression are not satisfied. Plots and diagnostics will substantiate this suspicion.

II. Materials and Methods

We used the data which characterize the proliferative capacity of mesenchymal stem cells. The data are composed of two variables which are the days of the culture (X) and population doubling level ($\ln Y$). First, we bootstrap the data in order to increase the sample size and also to optimize the parameter estimates. Then, we estimate the parameters through the exponential curve fitting and transform the nonlinear model into a linear form. This would bring a linear equation form. From the equation, we estimate the value of the independent variable (x) and fit the data with robust weighted regression by Cauchy, robust Fair weighted regression and robust weighted regression by Huber. Then a covariate-dependent variable is used to examine the differences in performance of the model suitability.

A. The Algorithm of Exponential Calculation

The algorithm showed below is the way of inserting data in SAS algorithm and the way of calculating the bootstrapping method.

- Data in SAS format. The name of the dataset is given as *cell_growth*. The data consist of two variables x and $\ln y$

```
Data cell  growth;
input x y lny;
cards;
0.00   38.00   3.64
5.00   39.31   3.67
8.00   39.74   3.68
10.00  40.98   3.71
13.00  43.10   3.76
17.00  45.78   3.82
20.00  49.15   3.89
22.00  49.90   3.91
24.00  53.98   3.99
28.00  57.46   4.05
31.00  61.03   4.11
34.00  63.80   4.16
37.00  65.52   4.18
40.00  68.54   4.23
44.00  72.62   4.29
47.00  75.42   4.32
50.00  79.38   4.37
53.00  83.31   4.42
;
run;
```

- Adding bootstrapping algorithm to the methodology *building. cell growth* data were bootstrapped two times with resampling. The following procedure was given in SAS syntax as follows. The new data which are generated by the SAS procedure will be named as booted. The produce data in the study will be print through the print procedure.

- We also add the syntax of 'ods rtf file='abc.rtf' style=journal' in the SAS language in order to get the output in Microsoft Word format.

```
%MACRO bootstrap(data=_last_, booted=booted, boots=2,
```

```
seed=1234);
DATA &booted;
pickobs = INT(RANUNI(&seed)*n)+1;
ET &data POINT = pickobs NOBS = n;
REPLICATE=int(i/n)+1;
i+1;
IF i> n*&boots THEN STOP;
RUN;
%MEND bootstrap;
ods rtf file='abc.rtf' style=journal;
%bootstrap(data= cell_growth, boots=2);
run;
proc print data=booted;
run;
```

- PROC SQL is a procedure developed in SQL. We can use this procedure to adjust, retrieve and report data in tables and views. SAS will create the output in the form of tables.
- The syntax below shows the calculation of the corrected sum of squares and the exponential parameter estimates. Through the syntax provided, we are able to measure the fitness of the parameter through the R-Square and AIC value.

```
Title "CORRECTED SUM OF SQUARES";
proc sql;
select css(y) into :CORRECTED_SUM_OF_SQUARESy from
cell_growth;
quit;

Title "EXPONENTIAL PARAMETER ESTIMATES";
ods graphics/imagename="ExponentialFit";
proc nlin data= booted plots=fit;
parameters A=1 b=0;
model y = A * exp(b*x);
ods output EstSummary=summExp;
run;

proc sql;
select  N.nValue1 as n,
  SSE.nValue1 as SSE,
1 - SSE/&CORRECTED_SUM_OF_SQUARESy as RSquare,
n * log(SSE/n) + 2*2 as AIC from summExp as SSE,
 summExp as N
 where N.Label1="Observations Used" and
SSE.Label1="Objective";
quit;
```

B. The Syntax of Regression Modeling Based on Four Types of Calculation

- Calculation regression based on robust regression. The full syntax is given as follows.

```
Title "REGRESSION";
/* ROBUST REGRESSION */
procrobustreg method=mm data=booted;
model lny = x / diagnostics leverage;
output out=robout r=resid sr=stdres;
run;
```

- Calculation regression based on robust (weighted) Huber. The full syntax is given as follows.

```
/* ROBUST (WEIGHTED) HUBER */
Title "ROBUST (WEIGHTED) HUBER";
procrobustreg method=m(wf=huber(c=1.345)) data=booted;
model lny = x / diagnostics leverage;
output out=robout r=resid sr=stdres;
run;
```

- Calculation regression based on robust (weighted) Cauchy. The full syntax is given as follows.

```
/* ROBUST (WEIGHTED) CAUCHY */
Title "ROBUST (WEIGHTED) CAUCHY";
procrobustreg method=m(wf=cauchy(c=2.385)) data=booted;
model lny = x / diagnostics leverage;
output out=robout r=resid sr=stdres;
run;
```

- Calculation regression based on robust (weighted) Fair. The full syntax is given as follows.

```
/* ROBUST (WEIGHTED) FAIR */
Title "ROBUST (WEIGHTED) FAIR";
procrobustreg method=m(wf=fair(c=1.4)) data=booted;
model lny = x / diagnostics leverage;
output out=robout r=resid sr=stdres;
run;
ods rtf close;
```

The syntax of "ods rtf close" gives an order to close the file in Microsoft Word. This means that the output will be generated in the Microsoft Word format.

III. RESULTS

The results for the first model without involving weighted procedures is given in Tables I and II. Table I shows that the model predicts the dependent variable well. The p-value is less than 0.05, and this indicates that, overall, the model statistically significantly predicts the outcome variable and is a good fit for the data. Figure 1, indicates the fit plot for lny.

A. Result for Exponential Fit

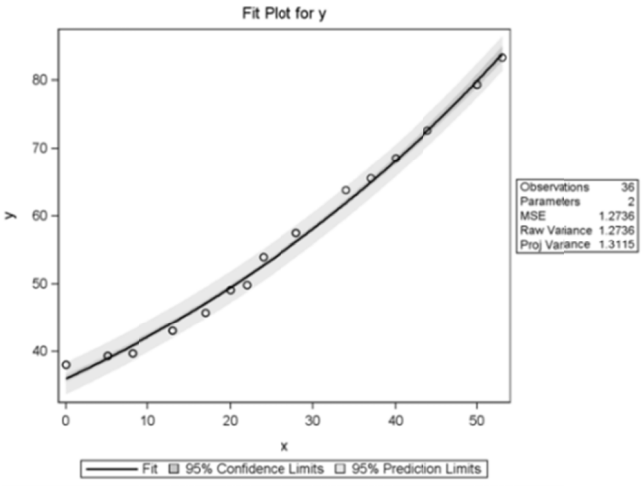

Fig. 1. The fit plot for y vs x

TABLE I. ANOVA

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr>F
Model	2	100489	50244	39450.4	<.0001
Error	34	43.3029	1.273		
Total	36	100533			

TABLE II. PARAMETER ESTIMATES

Parameter	Estimate	Approx Std. Error	Approximate 95% Confidence Limits	
A	35.9735	0.2564	35.4525	36.4945
b	0.0160	0.000219	0.0155	0.0164

From Table II, we can write an exponential model as

$$y = 35.9735\, e^{0.0160x} \quad (2)$$

Model (2) can be transformed into a linear form by taking, it can be written as follows

$$ln\,(y) = ln\,\,(35.9735) + 0.0160x$$

$$ln(y) = 3.5827 + 0.0160x$$

Table III gives the information of exponential fit. From Table III, the R-square value indicates how much of the total variation in the dependent variable or variability of the data is explained by the regression model. In this case, 97.98 can be explained, which is very large. The AIC value is about 10.649, which is the smallest value among all proposed methods. A good model is the one that has minimum AIC value.

TABLE III. EXPONENTIAL FIT

n	SSE	R-Square	AIC
36	43.302936	0.988339	10.64925

B. Result for Robust Regression

Figure 2 shows the fit plot for lny vs x. Previously we used the original data, while in this section we are using the transformed data. Table IV shows the parameter estimates. We can write an exponential model as:

$$\ln y = 3.5729 + 0.0162x \quad (3)$$

The R-Square value indicates the total variation in the dependent variable. In this case, 78.94 can be explained and the value of AIC is given as 28.24. Detailed information is given in Table V.

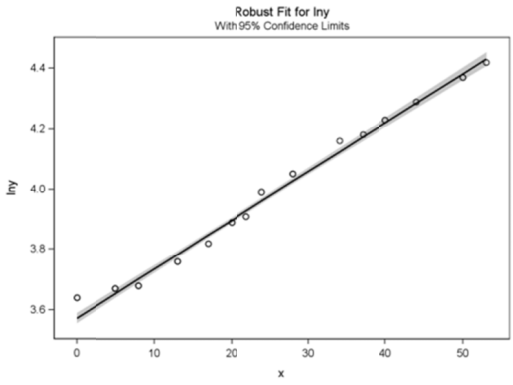

Fig. 2. Fit plot for lny

C. Result for Robust (Weighted) Huber

Figure 3 gives the plot for lny vs x for Huber robust regression. From the Table VI, we can write a robust regression which is weighted by Huber as:

$$\ln y = 3.5799 + 0.0160x \quad (4)$$

From the robust regression which is weighted by Huber, the value of R-Square is 0.9577 and AIC value is given as 24.3544.

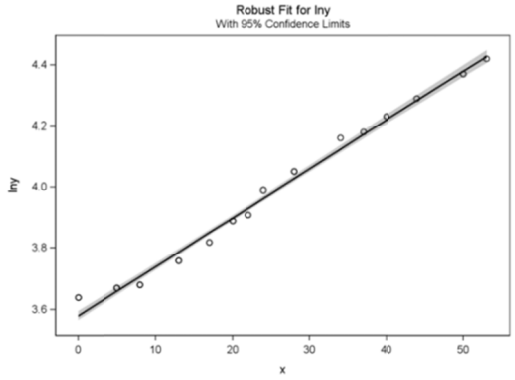

Fig. 3. Fit plot for lny

TABLE IV. PARAMETER ESTIMATES

Parameter Estimates							
Parameter	DF	Estimate	Std. Error	95% Confidence Limits		Chi-Square	Pr > Chi.Sq.
Intercept	1	3.5729	0.0087	3.554	3.590	168673	<0001
x	1	0.0162	0.0003	0.016	0.017	2383.2	<0001
Scale	0	0.0282					

TABLE V. PARAMETER ESTIMATION

Goodness-of-Fit	
Statistic	Value
R-Square	0.7894
AICR	28.2413
BICR	32.9392
Deviance	0.0205

D. Result for Robust (Weighted) Cauchy

In Figure 4 the result of robust regression which is weighted by Cauchy is plotted. From Table VIII, we can write a robust regression which is weighted by Huber as follows

$$\ln y = 3.5798 + 0.0160x \quad (5)$$

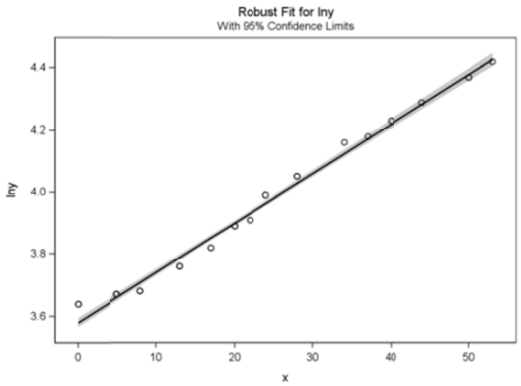

Fig. 4. Fit plot for lny

The value of R-Square is 0.4797 and AIC value is given as 377.98. A robust regression which is weighted by Cauchy seems not to be a good procedure for the forecasting.

TABLE VI. PARAMETER ESTIMATES FOR HUBER REGRESSION

Parameter Estimates							
Parameter	**DF**	**Estimate**	**Std. Error**	**95% Confidence Limits**		**Chi-Square**	**Pr > ChiSq.**
Intercept	1	3.5799	0.0076	3.5651	3.594	222668	<.0001
x	1	0.0160	0.0003	0.0154	0.016	2807.09	<.0001
Scale	1	0.0335					

TABLE VII. PARAMETER ESTIMATES FOR HUBER REGRESSION

Goodness-of-Fit	
Statistic	Value
R-Square	0.9577
AICR	24.3544
BICR	29.2774
Deviance	0.0249

TABLE VIII. PARAMETER ESTIMATES FOR CAUCHY REGRESSION

Parameter Estimates							
Parameter	**DF**	**Estimate**	**Std. Error**	**95% Confidence Limits**		**Chi-Square**	**Pr > ChiSq.**
Intercept	1	3.5798	0.0080	3.5642	3.595	201478	<0.0001
x	1	0.0160	0.0003	0.0154	0.016	2540.99	<0.0001
Scale	1	0.0336					

TABLE IX. PARAMETER ESTIMATES FOR CAUCHY REGRESSION

Goodness-of-Fit	
Statistic	Value
R-Square	0.4797
AICR	377.9854
BICR	383.0731
Deviance	0.4252

E. Result for Robust (Weighted) Fair

Below is the result of Fair robust regression. From Table X, we can write a fair robust regression is given in (6). The value of R-Square is 0.9616 and AIC value is given 12.309.

$$\ln y = 3.5803 + 0.0160x \quad (6)$$

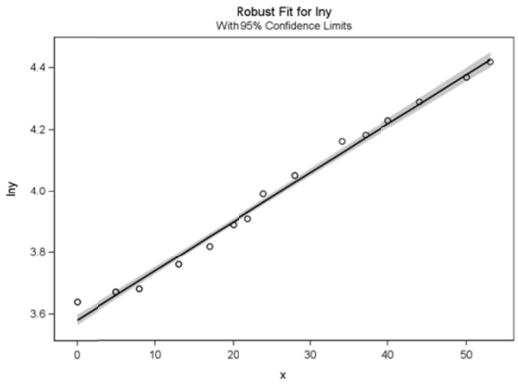

Fig. 5. Fit plot for lny

TABLE X. PARAMETER ESTIMATES FOR FAIR REGRESSION

Parameter Estimates							
Parameter	**DF**	**Estimate**	**Std. Error**	**95% Confidence Limits**		**Chi-Square**	**Pr > ChiSq.**
Intercept	1	3.5803	0.0086	3.563	3.597	171908	<0.0001
x	1	0.0160	0.0003	0.015	0.016	2162.9	<0.0001
Scale	1	0.0336					

TABLE XI. PARAMETER ESTIMATES FOR FAIR REGRESSION

Goodness-of-Fit	
Statistic	Value
R-Square	0.9616
AICR	12.3093
BICR	17.8695
Deviance	0.0121

F. Method Comparison

Table XII shows the comparison of the five different methods. We summarized the output of the gained model with their parameter estimates.

TABLE XII. PARAMETER ESTIMATES COMPARISON

Parameter Estimates	**Model**	**R-Square**	**AIC**
Exponential Fit	$y = 35.9735e^{0.0160x}$ or in a linear form $\ln(y) = 3.5827 + 0.0160x$	0.988339	1.064.925
Robust Regression	$\ln y = 3.5729 + 0.0162x$	0.7894	282.413
Robust Huber	$\ln y = 3.5799 + 0.0160x$	0.9577	243.544
Robust Cauchy	$\ln y = 3.5798 + 0.0160x$	0.4797	3.779.854
Robust Fair	$\ln y = 3.5803 + 0.0160x$	0.9616	123.093

IV. DISCUSSION AND CONCLUSION

The main objective of this research is to compare the parameter estimate between several proposed calculations and also to find the best calculation which can represent the data through modeling techniques. We have given an example of cell doubling data by using PROC NLIN and PROC ROBUSTREG. In this paper, five different methods were used (Table XII): (i) Exponential fit (ii) Robust regression (iii) Robust (weighted) Huber (iv) Robust (weighted) Cauchy (v) Robust (weighted) fair. This paper provides only a preliminary overview of the above mentioned different techniques that can be employed. From the results, we can see that the exponent fit shows a very good fitting result, followed by robust regression weighted by fair techniques. Both methods produced the highest R^2 and the lowest AIC. This indicates that exponential fit is the best method, followed by robust fair regression. The other methods fit the model poorly. This may be due to some outliers through the output. The best way to handle outliers is to delete them from the set of data and then rerun to the analysis once again. To keep the efficiency and accuracy of the proposed model, it is necessary to have a good way of calculation with some improvements of the proposed strategy. The exponential fit reveals the findings more explicitly compared to the other proposed methods.

ACKNOWLEDGMENT

Authors would like to express their gratitude to Universiti Sains Malaysia for providing the research funding (Grant no.1001/PPSG/8012278, School of Dental Sciences, Universiti Sains Malaysia).

REFERENCES

[1] N. R. Draper, H. Smith, Applied Regression Analysis, Wiley Eastern, 1998

[2] R. J. Tallarida, R. B. Murray. Exponential growth,and decay, Manual of Pharmacologic Calculations, Springer, 1987

[3] R. Tahmasbi, S. Rezaei, "A two-parameter lifetime distribution with decreasing failure rate", Computational Statistics & Data Analysis,Vol. 52, No. 8, pp. 3889-390, 2008

[4] A. Schneider, G. Hommel, M. Blettner, "Linear Regression Analysis", Deutsches Arzteblatt International, Vol. 107, No. 44, pp. 776-82, 2010

[5] D. C. Montgomery, E. Peck, G. Vining, Introduction to linear regression analysis, 3rd Edition, John Wiley and Sons, 2003

[6] R. Gupta, D. Kundu, "Generalized Exponential Distributions", Australian and New Zealand Journal of Statistics, Vol. 41, No. 2, pp. 173-188, 1999

[7] K. Adamidis, S. Loukas, "A lifetime distribution with decreasing failure rate", Statistics & Probability Letters, Vol. 39, No. 1, pp. 35-42,1998

[8] D. Kus, "A new lifetime distribution distributions", Computational Statistics and Data Analysis, Vol. 11, No. 9, pp. 4497-4509, 2007

[9] F. Louzada-Neto, "Poly hazard regression models for lifetime data", Biometric,Vol. 55, No. 4, pp. 1281-1285, 1999

[10] J. U. Kreft, G. Booth, J. W. T. Wimpenny, "BacSim, a simulator for individual-based modeling of bacterial colony growth", Microbiology, Vol. 144, No. 12, pp. 3275-3287, 1998