



Evaluating the Efficiency of an Aaziz Regression Model

Abdulaziz Usman^{1*}, and Tasi'u Musa¹

¹Department of Statistics, Ahmadu Bello University, Zaria, Nigeria

*Correspondence: rabdulazizusman@gmail.com

Abstract

Regression involves a set of procedures statistically designed to study relationships among variables such that one variable is defined as the dependent variable (or response variable) and the other is defined as the independent variable (or predictor variable). With the aid of regression, the cause-and-effect relationship between the dependent and independent variables can be evaluated. The aim of this paper is to propose a model called Aaziz's regression model (also called Aaziz indicators rule) in order to eliminate/reduce the under- or overestimation of a response variable and compare its forecast performance with the linear regression (standard regression) model. Least squares estimation and maximum likelihood estimation were used to obtain parameters of the model. Statistical techniques such as Root Mean Square Error (RMSE) and coefficient of determination were considered. In all the empirical analyses, the results show that the RMSE in Aaziz's model are less than that of the linear regression model, while the coefficient of determination in Aaziz's model is higher > 93% (averagely) than the standard regression model (averagely < 54%) in all cases. Furthermore, the sum of squared errors in Aaziz's regression model is far less than that of the standard regression model, indicating that estimated values of the response variable are closer to the true value in Aaziz's regression model than in the standard regression model. In conclusion, it turns out that Aaziz's regression model outperformed the standard regression model.

Keywords: Aaziz's regression model, standard regression model, predictor and response variables, overestimation

1. Introduction

Without a doubt, it is often of interest to determine how change(s) in the value(s) of some variables influence the change in value(s) of other variable(s). For instance, how feed intake influences hen weight, or how changes in air temperature affect feed intake. In both examples, the association (relationship) between variables can be determined with a function: a function of feed intake to describe hen weight, or a function of temperature to describe feed intake. A function that describes such an association (relationship) is called a regression function, and analysis of such problems and

estimation of the regression function is called regression analysis. Regression involves a set of procedures statistically designed to study relationships among variables such that one variable is defined as the dependent variable (or response variable) and the other defined as the independent variable (or predictor variable). With the aid of regression, the cause-and-effect relationship between the dependent and independent variables can be evaluated. If the change of the dependent variable is determined with just one independent variable and the association (relationship) between them is linear, the appropriate procedure is called simple linear regression. When the change in a dependent variable is explained by two or more independent variables, the procedure is called multiple regression. Between these procedures, only simple linear regression will be considered.

Regression analysis makes several assumptions, which are quite akin to those for correlation analysis (Aggarwal & Ranganathan, 2017). First, the relationship between the dependent and independent variables should be linear. Second, all the observations in a sample must be independent of each other; thus, the method should not be used if the data includes more than one observation on any individual. Furthermore, the data must not include one or a few extreme values since these may create a false sense of relationship in the data even when none exists. If these assumptions are not met, the results of linear regression analysis may be misleading (Aggarwal & Ranganathan, 2017).

Iqbal (2020) researched the application of regression techniques with their advantages and disadvantages. Some of the disadvantages of linear regression are: independence of the variables, linearity of the relationships, which is not always the case, outliers, etc. Linear regression may suffer from underfitting, a situation where a machine learning model fails to encapsulate the data of interest properly (Iqbal 2020). This usually happens when the hypothesis function cannot fit the data well (Iqbal 2020).

Flom (2022) highlighted that linear regression is limited to linearity in the relationships between the variables, which is not always the case. For example, the relationship between income and age is curved (Flom, 2022). Income tends to rise in the early parts of adulthood, flatten out in later adulthood and decline after people retire (Flom, 2022). Linear regression is sensitive to outliers, only looks at the mean of the response variable, requires that data be independent, and so on, which are not always the case in reality (Flom, 2022).

John (2025) performed research on the uses and abuses of regression models. John (2025) explained that statistics is widely understood to provide a body of techniques for “modelling data”, underpinned by what is called “true model myth”. The role of a statistician/analyst is to build a model that closely approximates the true data-generating process. John (2025) further explained that, based on historical examples and a few reviews of mainstream clinical research journals, this perspective has led to a range of problems in the uses of regression methods, including misguided “adjustment” for covariates, misinterpretation of regression coefficients and the widespread fitting of regression models without a clear purpose.

Reid C. (2020) researched the limitations of regression models and explained that regression models are limited across two important dimensions. The first is known as “unobservable” or “omitted” variable bias, and the second is known as “endogeneity”. Endogeneity refers to the causal direction of the two variables included in the model. Reid (2020) realised that, for example, elevators are associated with higher development costs, but it could be that projects that are more expensive require elevators.

Jihye (2015) conducted research on the strengths and limitations of multiple regression and other statistical modelling. The results revealed that while a regression model helps in predicting the

response variable, it also has the limitation of predicting response variables when there are outliers in the predictor variable.

It is generally known that simple linear regression has limitations such as normality of the response variable, independence, under- or overestimation of response variables and many more, which are known to every statistician worldwide. So, delving more into the literature review may not be necessary.

This paper is concerned with the fact that while simple linear regression methods dominate much of biostatistics (and other fields) practice, numerous applications of regression analysis in the medical and health research literature either overestimate or underestimate the desired variables. The core aim of this study is to propose and evaluate the efficiency of a statistical linear model called the Aaziz regression model. The objectives of this study are to eliminate or reduce under- or overestimation of response variables and to examine the forecast performance of the Aaziz regression model and that of the standard model.

This study is of interest to answer the following research questions such as: Does the Aaziz regression model predict better than the standard regression model? Is the RMSE in the Aaziz regression model smaller than that of the standard model? Is the coefficient of determination in the Aaziz regression model higher than that of the standard model?

The following aforementioned research questions do not really need hypothesis testing, but rather use forecast measures, RMSE and coefficient of determination, R^2 .

The results of this study are expected to contribute to the world on the newly developed model (Aaziz regression model) through the methodology used and empirical evidence.

The study will be useful to the world (especially data analysts or statisticians) by understanding the efficiency of the Aaziz regression model as compared to the standard regression model when examining a linear relationship between two variables and predicting the response variable.

The findings will aid policymakers and non-governmental organisations in choosing the best linear model when analysing the relationship between two variables.

From the perspective of an academican, the study will contribute to the increase of literature on the generalised linear regression model by conducting numerical statistical analysis such as standard linear regression analysis and Aaziz regression analysis, to examine the relationships between response and predictor variables. The findings could also serve as an important reference for future researchers who are interested in developing a regression model better than the existing ones.

2. Materials and Methods

2.1. The Simple Linear Regression Function

A regression that describes linear change of a response variable with respect to (w.r.t) the changes of one predictor variable is called a simple linear regression. For example, the weight of hens can be predicted by the amount of feed intake. The aim will be to determine a linear function that will explain changes in weight as feed intake changes. Weight is the response variable and feed intake is the predictor variable.

A regression function uses pairs of measurements $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. According to the function, each observation y_i can be described as follows:

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad i = 1, 2, \dots, n \quad (1)$$

where α represents the intercept, a value of y_i when $x_i = 0$; β (called the slope) explains the rate of change of the response variable per unit change of the predictor variable. When β is $+ve$, it implies y increases as x increases. If β is $-ve$, y decreases as x increases. A regression with $\beta = 0$, indicates no linear relationship between the variables.

2.1.1. Model Assumptions

- i. $E(\varepsilon_i) = 0$
- ii. $Var(\varepsilon_i) = \sigma^2$
- iii. $Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$
- iv. $\varepsilon_i \sim N(0, \sigma^2)$

2.1.2. Model Properties

- i. $E(y_i | x_i) = \alpha + \beta x_i$
- ii. $Var(y_i) = \sigma^2$
- iii. $Cov(y_i, y_j) = 0, i \neq j$

2.1.3. Parameter Estimation of Regression

Population parameters of the regression function are always unknown; as such, a sample of data is usually collected to estimate the population parameters. Parameter estimators will be denoted by $\hat{\alpha}$ and $\hat{\beta}$. The line of regression $E(y|x)$ is not known and can be estimated by using a sample with:

$$y_i = \hat{\alpha} + \hat{\beta} x_i \quad (2)$$

Equation (2) is called the estimated regression line or estimated model. This best-fitted line will be as close to the dependent variables as possible. The least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ for a given set of observations from a sample are such that:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \text{ is minimized}$$

It follows that the estimators $\hat{\alpha}$ and $\hat{\beta}$ are obtained using either Least squares estimation or maximum likelihood estimation as:

$$\hat{\beta} = \frac{SS_{xy}}{SS_{xx}} = \frac{Cov(x_i, y_i)}{Var(x)} \quad (3)$$

where:

$$SS_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n} \quad (4)$$

$$SS_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n} \quad (5)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad (6)$$

2.1. Aaziz's Regression Model

Consider the model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 D_i + \varepsilon_i \quad (7)$$

where $D_i = (x_i - y_i)$, $\hat{\beta}_0$ represent the intercept, a value of y_i when $D_i = 0$; $\hat{\beta}_1$ (called the slope) explains the rate of change of the response variable per unit change in the differences between the variables. When $\hat{\beta}_1$ is +ve, it implies y increases as D_i increases. If $\hat{\beta}_1$ is -ve, y decreases as D_i increases. A regression with $\hat{\beta}_1 = 0$, indicates no linear relationship between the variables.

2.2.1. Model Assumptions and Properties

The model assumptions and properties are the same as highlighted in section 2.1.1 and section 2.1.2, respectively.

2.2.2. Parameter Estimation of Aaziz's Regression Model

The parameter estimation of Aaziz's regression model can be obtained using Least squares procedures and maximum likelihood procedures.

2.2.2.1. Least Squares Estimation

The estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained as follows:

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n \left[y_i - \hat{\beta}_0 - \hat{\beta}_1 (x_i - y_i) \right]^2 \quad (8)$$

Differentiating equation (8) w.r.t the estimators, we have:

$$\frac{\partial \sum_{i=1}^n \varepsilon_i^2}{\partial \sum_{i=1}^n \hat{\beta}_0} = 2 \sum_{i=1}^n \left[y_i - \hat{\beta}_0 - \hat{\beta}_1 (x_i - y_i) \right] (-1) \quad (9)$$

$$\frac{\partial \sum_{i=1}^n \varepsilon_i^2}{\partial \sum_{i=1}^n \hat{\beta}_1} = -2 \sum_{i=1}^n \left[y_i - \hat{\beta}_o - \hat{\beta}_1(x_i - y_i) \right] (x_i - y_i) \quad (10)$$

where equations (9) and (10) are called normal equations. Solving equations (9) and (10) simultaneously;

$$\sum_{i=1}^n \left[y_i - \hat{\beta}_o - \hat{\beta}_1(x_i - y_i) \right] = 0 \quad (11)$$

$$\sum_{i=1}^n (x_i - y_i) \left[y_i - \hat{\beta}_o - \hat{\beta}_1(x_i - y_i) \right] = 0 \quad (12)$$

Expanding equations (11) and (12) and solving simultaneously, we have:

$$\sum_{i=1}^n y_i - n\hat{\beta}_o - \hat{\beta}_1 \sum_{i=1}^n (x_i - y_i) = 0 \quad (13)$$

$$\sum_{i=1}^n (x_i - y_i) y_i - \hat{\beta}_o \sum_{i=1}^n (x_i - y_i) - \hat{\beta}_1 \sum_{i=1}^n (x_i - y_i)^2 = 0 \quad (14)$$

From equation (13), we have:

$$\hat{\beta}_o = \frac{\sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n (x_i - y_i)}{n} = \bar{y} - \frac{\hat{\beta}_1 \sum_{i=1}^n (x_i - y_i)}{n} = \bar{y} - \hat{\beta}_1 \bar{D} \quad (15)$$

Multiplying equation (12) by $\sum_{i=1}^n (x_i - y_i)$ and equation (14) by n to eliminate $\hat{\beta}_o$, we have:

$$\sum_{i=1}^n (x_i - y_i) \sum_{i=1}^n y_i - n\hat{\beta}_o \sum_{i=1}^n (x_i - y_i) - \hat{\beta}_1 \sum_{i=1}^n (x_i - y_i)^2 = 0 \quad (16)$$

$$n \sum_{i=1}^n (x_i - y_i) y_i - n\hat{\beta}_o \sum_{i=1}^n (x_i - y_i) - n\hat{\beta}_1 \sum_{i=1}^n (x_i - y_i)^2 = 0 \quad (17)$$

Subtracting equation (17) from (16), we have:

$$\sum_{i=1}^n (x_i - y_i) \sum_{i=1}^n y_i - n \sum_{i=1}^n (x_i - y_i) y_i - \hat{\beta}_1 \left[\sum_{i=1}^n (x_i - y_i) \right]^2 + n\hat{\beta}_1 \sum_{i=1}^n (x_i - y_i)^2 = 0 \quad (18)$$

$$n\hat{\beta}_1 \sum_{i=1}^n (x_i - y_i)^2 - \hat{\beta}_1 \left[\sum_{i=1}^n (x_i - y_i) \right]^2 = n \sum_{i=1}^n (x_i - y_i) y_i - \sum_{i=1}^n (x_i - y_i) \sum_{i=1}^n y_i \quad (19)$$

$$\hat{\beta}_1 \left\{ n \sum_{i=1}^n (x_i - y_i)^2 - \left[\sum_{i=1}^n (x_i - y_i) \right]^2 \right\} = n \sum_{i=1}^n (x_i - y_i) y_i - \sum_{i=1}^n (x_i - y_i) \sum_{i=1}^n y_i \quad (20)$$

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^n (x_i - y_i) y_i - \sum_{i=1}^n (x_i - y_i) \sum_{i=1}^n y_i}{n \sum_{i=1}^n (x_i - y_i)^2 - \left[\sum_{i=1}^n (x_i - y_i) \right]^2} = \frac{n \sum_{i=1}^n D_i y_i - \sum_{i=1}^n D_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n D_i^2 - \left[\sum_{i=1}^n D_i \right]^2} \quad (21)$$

2.2.2.2. Maximum Likelihood Estimation

Maximum likelihood estimation can also be used to determine the parameter estimates of Aaziz's regression model. Under the normality assumption of the dependent variable, the likelihood function is a function of the parameters $(\hat{\beta}_0, \hat{\beta}_1, \sigma^2)$ for a given set of n values of dependent and independent variables.

$$L(\hat{\beta}_0, \hat{\beta}_1, \sigma^2 | y) = \frac{1}{\left(\sqrt{2\pi\sigma^2} \right)^n} e^{-\frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i)^2}{2\sigma^2}} \quad (22)$$

The log-likelihood function is:

$$\log L(\hat{\beta}_0, \hat{\beta}_1, \sigma^2 | y) = -\frac{n}{2} \log(\sigma^2) - \frac{n}{2} \log(2\pi) - \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i)^2}{2\sigma^2} \quad (23)$$

The maximum likelihood estimation is obtained by taking the partial derivatives of the log-likelihood function w.r.t the parameters:

$$\frac{\partial \log L(\hat{\beta}_0, \hat{\beta}_1, \sigma^2 | y)}{\partial \hat{\beta}_0} = \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i)}{\sigma^2} \quad (24)$$

$$\frac{\partial \log L(\hat{\beta}_0, \hat{\beta}_1, \sigma^2 | y)}{\partial \hat{\beta}_1} = \frac{\sum_{i=1}^n D_i \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i)}{\sigma^2} \quad (25)$$

$$\frac{\partial \log L(\hat{\beta}_0, \hat{\beta}_1, \sigma^2 | y)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i)^2}{2\sigma^4} \quad (26)$$

To find the estimators, we equate equations (24), (25) and (26) to zero, which gives:

$$\frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i)}{\sigma^2} = 0 \quad (27)$$

$$\frac{\sum_{i=1}^n D_i \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i)}{\sigma^2} = 0 \quad (28)$$

$$-\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i)^2}{2\sigma^4} = 0 \quad (29)$$

Simplifying the equations, we have:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n D_i = \sum_{i=1}^n y_i \quad (30)$$

$$\hat{\beta}_0 \sum_{i=1}^n D_i + \hat{\beta}_1 \sum_{i=1}^n D_i^2 = \sum_{i=1}^n D_i y_i \quad (31)$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i)^2 \quad (32)$$

Solving equations (30) and (31) results in identical estimators as with least squares estimation.

2.3. Model Accuracy

The accuracy of a model can be determined using various techniques such as Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), etc. In this paper, RMSE and the coefficient of determination will be used to determine whether or not Aaziz's regression model is better than the standard linear regression model. The techniques are defined as follows:

2.3.1. Root Mean Square Error

The RMSE is among the best statistical techniques used to determine the best model among the models. The lower the RMSE, the better the model. It is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (33)$$

where y_i and $\hat{y}_i$ are the respective actual and predicted values of the response variable, n is the sample size

2.3.2. Coefficient of Determination

The coefficient of determination is used to measure the correctness of a model. It describes the proportion of the total variability explained by the model. The higher the coefficient of determination, the better the model. It is given by:

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}} \quad (34)$$

where:

$$SS_{tot} = \sum_{i=1}^n (y_i - \bar{y}_i)^2 \quad (35)$$

$$SS_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i)^2 \quad (36)$$

$$SS_{reg} = SS_{tot} - SS_{res} = \sum_{i=1}^n (y_i - \bar{y}_i)^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y}_i)^2 \quad (37)$$

The coefficient of determination lies within $0 \leq R^2 \leq 1$. The closer the value of R^2 to 1, the better the model.

3. Results and Discussions

This section depicts the applications of Aaziz's and the standard linear regression model.

Table 1. Empirical Analysis of Weight(Y) and Heart girth(X) of Six Cows, n = 6

Estimates	Standard Model				Aaziz's Model		
	y_i	$\hat{y}_i$	ε_i	ε_i^2	$\hat{y}_i$	ε_i	ε_i^2
Mean	659.8333	659.7450	-0.0883	77.2562	659.8519	-0.0186	1.3729
RMSE		8.7896				1.1717	
R^2 (%)		0.81(80.612)				0.99(99.655)	

From Table 1, the RMSE of Aaziz's regression model (1.1717) is considerably less than that of standard linear regression (8.7896); while for the coefficient of determination, that of Aaziz's regression model (0.9966) is higher than that of the standard model (0.8061). Indicating Aaziz's regression model is better than that of the standard regression model.

Table 2. Empirical Analysis of the Influence of Hen Weights(X) on the Feed Intake(Y) in a Year, n=10

Estimates	Standard Model				Aaziz's Model		
	y_i	$\hat{y}_i$	ε_i	ε_i^2	$\hat{y}_i$	ε_i	ε_i^2
Mean	46.6000	46.5990	0.0010	1.4260	46.6002	-0.0002	0.0136
RMSE		1.1942				0.1167	
R^2 (%)		0.61(60.824)				0.99(99.625)	

According to Table 2, there is a 99.625% coefficient of determination in Aaziz's model, implying the model explains 99.625% of the total variability, while in the standard model, the total variability that is explained by the model is 60.824%. Furthermore, the RMSE of Aaziz's model (0.1167) is less than that of the standard regression model (1.1942), implying that Aaziz's model outperforms the standard model.

Table 3. The Marks in Economics by the variable X and the Marks in Statistics by the variable Y

Estimates	Standard Model				Aaziz's Model		
	y_i	$\hat{y}_i$	ε_i	ε_i^2	$\hat{y}_i$	ε_i	ε_i^2
Mean	38.0000	37.9999	0.0001	33.6221	38.0000	0.0000	6.5015
RMSE		5.7986				2.5498	
R^2 (%)		0.16(15.52)				0.84(83.67)	

Looking at Table 3, we can clearly deduce that Aaziz's regression model is by far better than that of the standard regression model.

4. Conclusion

Statistically speaking, backed by the empirical results, Aaziz's regression model accounts for more variability in the regression model than the standard regression model. Indicating that using Aaziz's model to predict future response variables is better than the standard regression model. This improvement was further confirmed by the Root Mean Square Error (RMSE). In addition, the sum of squared errors in Aaziz's regression model is far less than that of the standard regression model,

indicating that estimated values of the response variable are closer to the true value in Aaziz's regression model than in the standard regression model.

From the perspective of the linear regression model properties and assumptions, Aaziz's regression model complies more than the standard regression model; the expected mean of the response variable in Aaziz's regression model is closer to the true mean of the response variable than that of the standard model. In addition, the mean of the error terms in the Aaziz regression model is closer to zero than that of the standard regression model.

Despite the results revealing valuable insight, it is limited by reliance on self-reported results, which may be subject to bias. Future researchers should consider more data from different fields.

References

- Aggarwal, R., & Ranganathan, P. (2017). Common pitfalls in statistical analysis: Linear Regression Analysis. *National Library of Medicine*. April-June; 8(2): 100-102. DOI: 10.4103/2229-3485.203040
- Flom, P. (2022). The disadvantage of linear regression. <https://www.science.com>
- Iqbal, M. H. (2020). Application of linear regression with its advantages and disadvantages. L....eid, C. (2020). The limitations of regression models: <https://turnercenter.berkeley.edu>
- Jihye, J. (2015). Strengths and limitations of regression analysis, *International Scholarly and Scientific Research and Innovation*, 9(5), 2015.
- John, B. (2025). limitations of regression methods, 44(13-14).